Comparing Generic vs. Specific Heuristics: Illustrating a New UEM Comparison Technique

Jacob Somervell[1] and D. Scott McCrickard
Department of Computer Science and Center for Human-Computer Interaction
Virginia Polytechnic Institute and State University
Blacksburg, VA 24061
{jsomerve, mccricks}@cs.vt.edu

Heuristic evaluation method comparison is important for developing new heuristic sets, to ensure effectiveness and utility. However, comparing different sets of heuristics requires a common baseline upon which a comparison can be made, usually some set of usability problems from a particular interface. This is often accomplished by having evaluators perform system evaluation to produce a set of usability problems for each method in question. A problem arises in that different methods produce different sets of problems, thus introducing validity concerns and ambiguity in resolution of disparate problem sets. We address this problem by illustrating a new comparison technique in which predetermined usability issues are presented to the evaluators up front, followed by assessment of thoroughness, reliability, and cost for the target methods. Comparison of method effectiveness is simplified, and validity concerns are ameliorated.

INTRODUCTION

Usability evaluation is one of the most important and costly steps in developing human computer interfaces. The goal is to identify and fix problems in the interface, so as to improve the user experience with the interface. Usability evaluation methods (UEM) are techniques for extracting usability feedback from investigations and can be classified as either *analytical* or *empirical*. Analytic methods rely on reasoning and common sense inspection of the system, usually with expert usability professionals. Empirical methods focus on observing and/or testing real users with the system, typically performing set tasks, with the goal of seeing where users run into problems with the interface.

Each type of UEM has advantages and disadvantages related to cost, implementation, finding participants, and applicability. It is mostly accepted that given unlimited resources, empirical testing is the preferred technique. However, real design projects are faced with both time and money budgets, and designers need to optimize the feedback they receive from whatever evaluation they perform; hence, many designers and developers seek alternative approaches.

Indeed, there is significant interest in developing, testing, and comparing analytic evaluation methods (specifically heuristics). Researchers are actively working to create new sets of heuristics for different system classes (Somervell et al., 2003, Baker et al., 2002, Mankoff et al., 2003), compare heuristics to user testing (Karat et al., 1992, Lavery et al., 1997), or to simply further understand heuristics (Cockton & Woolrych, 2002). These efforts stem from the fact that empirical methods (user testing) are often difficult to setup and execute, especially for new, emerging systems like ubiquitous interfaces, notification systems (McCrickard et al. 2003), handheld devices, etc. Why? Because observing users in these situations or setting up experimental models is quite difficult. These technologies are new, and are often imbedded in the environment, making controlled lab studies difficult or impossible. Hence, designers and usability professionals must rely on analytic techniques to gain feedback early in the design process.

What this means is that researchers are tasked with developing effective evaluation tools for many different system classes. A key part of this development work involves testing the new methods and comparing them to existing alternatives. Traditionally UEM comparison has been difficult and wrought with problems that lead to debate over validity and utility of previous comparison studies (Gray & Saltzman, 1998). A large part of the problem with these comparison studies involves validity in problem sets and results. Hartson et al. suggest a set of metrics to use in UEM comparison studies that are designed to help with accurately comparing different methods (2001). However, calculating these measures relies upon knowledge of the *real problem sets* for the target system. This can be problematic for traditional comparison tests because it is not clear what constitutes the real problem set (Hartson et al., 2001).

Our work addresses this problem in a novel way. Instead of having evaluators uncover usability problems in a traditional evaluation; we provide a list of problems for the target system and ask the evaluators to rate the applicability of specific heuristics to that problem. In other words, we ask the usability professionals to assess the heuristics in terms of how much the heuristic would help in identifying the issue in a traditional evaluation. This relies on the evaluators' experience with usability problems and their ability to reason about the heuristics in an abstract manner. The strength of this

[1] Authors current address: University of Virginia's College at Wise, Wise, VA 24293

approach comes from the fact that we have a specific set of problems to serve as the real problem set and we can easily calculate the Hartson et al. measures (reliability, effectiveness, thoroughness, validity, cost) from this setup.

METHOD

One goal of this study is to illustrate the utility of the new experimental method for UEM comparison. We conducted an experiment with this new testing procedure to compare three sets of heuristics, each representing a different level of specificity for the *large screen information exhibit* (LSIE) system class.

LSIEs are notification systems (McCrickard et al., 2003) that run on large displays. These displays leverage the natural breakpoints in a person's task to allow them to decide when they need to look at the display (self-defined interruption) and facilitate long term storage of the information (high comprehension).

Previous research describes a set of heuristics tailored to the LSIE system class (Somervell et al., 2003). We use those heuristics as well as heuristics designed for notification systems (Berry, 2003), and Nielsen's general interface heuristics (Nielsen & Mack, 1994). We created a new testing platform designed to alleviate the difficulty associated with calculating typical UEM comparison measures. We used structured presentation of identified problems in the target systems and asked the evaluators to indicate the level of applicability each heuristic held for the problem. This rating was provided through a 7-point Likert scale where the evaluator indicated their level of agreement with the heuristic when asked if the heuristic applied to the problem. An answer of seven would indicate strong agreement (highly applicable), while a one indicates strong disagreement (not applicable at all).

We used three example large screen information exhibits for the test and used pre-identified problem sets to serve as the "real" problem sets:

1. The Notification Collage provides a communication mechanism for lab members upon which they can post various types of information, from personal communication to documents and even video clips (Greenberg & Rounding, 2001). Users are often busy with work at their desks but can choose to look up at the NC to check on postings and keep track of lab information.
2. The Plasma Poster performs similar tasks, but is placed in common areas like break rooms, kitchens, or atriums (Churchill et al., 2003). Users can find information on local events, user postings, and automatically generated content.
3. The Source Viewer is an LSIE system found in a local television station. Program control managers must ensure proper source switching between commercials and standard program content. This system allows the manager to see all of the upcoming sources simultaneously and thus facilitate source switching.

Each evaluator was randomly assigned to one of the three heuristic sets given the constraint of keeping equal numbers for each set. The evaluators then proceeded to rank the heuristics according to the problems for each of the three systems. System presentation order was completely balanced using a Latin Square ordering.

Determining Real Problem Sets

One problem identified in other UEM comparison studies involves the calculation of specific metrics that rely upon something referred to as the "real" problem set (Hartson et al., 2001). In most cases, this problem set is the union of the problems found by each of the methods in the comparison study. In other words, each UEM is applied in a standard usability evaluation of a system, and the "real" problem set is simply the union of the problems found by each of the methods. There are issues with this approach because there is no guarantee that the problems found by the UEMs are the problems that would be experienced by real users in normal day to day activity with the system in question.

This comparison study also faced the same challenge. Instead of relying on evaluators to produce sets of problems from each method, then using the union of those problem sets as the "real" problem set, analysis and testing was performed on the target systems *beforehand* and the problem reports from those efforts were used to come up with a standard set of real problems for each system. Coupled with a new testing approach, this eliminated much of the variability inherent in most UEM comparison studies that arises from having to read through problem reports and deduce (perhaps erroneously) the intention of the evaluator.

Source Viewer problem set. To determine the real problem set for the Source Viewer, two types of field study were employed. First, observations were made of control managers as they proceeded with their everyday work tasks. This observation provided insight into the everyday usage of the system and what problems were encountered. In addition, interviews were held with one of the control managers, probing Source Viewer usage and usability concerns. These efforts produced a list of 11 usability concerns for the Source Viewer. An example involved the use of green color for source labels. The manager stated that it was often difficult to read the source label when the background was dark, due to poor contrast between the fore and background colors.

Plasma Poster and Notification Collage problem sets. A different approach to real problem set identification was required for both the Plasma Poster and the Notification Collage. In these instances, we relied upon existing published literature to verify problems identified through claims analysis. In addition, discussions with the developers of the systems were also used to confirm the claims analysis.

Claims analysis is an analytic technique in which design decisions are analyzed according to potential upside and downside psychological impacts on users (Carroll & Rosson, 1992). For example, consider the following claim about using blinking text:

Using blinking text in stock tickers can,

+ attract attention to important information

BUT – can distract users from primary work.

Designers can easily see the tradeoff that would arise from choosing to incorporate the design element (blinking text) in a system.

Claims analysis techniques were used to uncover usability issues with the systems through system inspections. Literature reviews served as mechanisms for supporting or refuting the claims analysis. If a claim was determined to have been observed through published usability studies of the systems, it was included in the experiment, claims that were not refuted by published studies were also kept, and refuted claims were discarded.

In addition, system developers were contacted and asked to further verify the non-refuted claims. The developer responded by providing indication that a particular claim was supported or refuted, based on first-hand knowledge of the system and its use. This process allows us to capture a subset of the real problems in a system.

Data Collection

Data were collected through a pen-and-paper setup in which the evaluators ranked each heuristics from a given set in terms of whether it applied to the problem. A single problem was provided to the evaluator, and he/she indicated the level of agreement for each heuristic by marking the appropriate spot on a Likert-scale. Figure 1 provides an example of the test layout.

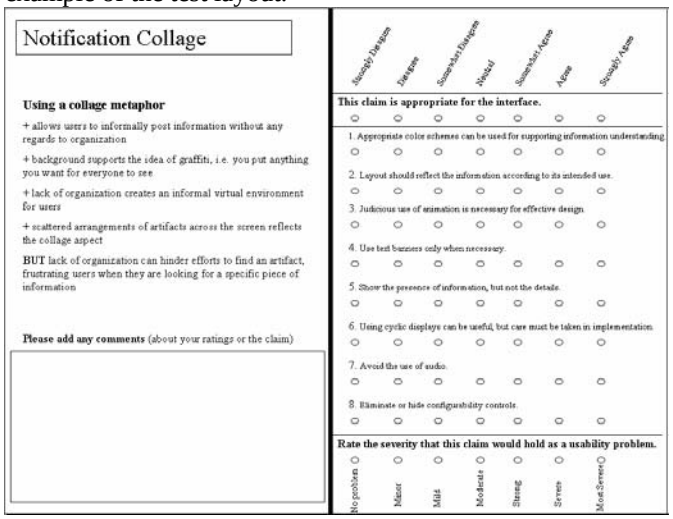

Figure 1. Layout of data collection form.

RESULTS

Information on the evaluators' experience levels with respect to usability engineering, heuristic evaluation, and large screen information exhibits was gathered before the test, and afterwards, each evaluator reported his/her completion time.

Calculation of thoroughness, validity, effectiveness, and reliability is straightforward when using this new testing platform. Applicability scores indicate that a problem is "found" by a heuristic if the average rating for the heuristic is greater than 5. The cutoff value of 5 was used because averages above this value indicate "agreement", which suggests that the heuristic applies to the problem.

Thoroughness is found by dividing the number of problems found by a single method (set of heuristics) by the total number found by all methods. In other words, determine the total number of real problems found by all of the heuristic sets, then divide the number for each individual set by this total. Validity is found through a similar division but relies on the cardinality of the real problem set as the denominator instead of the union of problems found. Effectiveness is the product of thoroughness and validity. Reliability is found by calculating the average difference in the evaluators for each of the problems. This represents an accurate measure of the total difference in answers among the evaluators. Alternatively, one could measure the number of agreements among the evaluators, yielding a separate measure of reliability. To simplify calculations, all of the problems across all three systems were grouped together.

Results indicate that the more specific heuristics held the best scores for the aforementioned metrics. There was also a general trend that the more specific heuristics were better suited to the large screen information exhibit system class, evident through the resulting ordering of the methods based on comparison metrics.

Thoroughness. Somervell's heuristics had the highest thoroughness rating of the three heuristic sets with 96% (27 of 28 claims). Berry's heuristics came next with a thoroughness score of 86% (24 of 28) and Nielsen's heuristics had a score of 61% (17 of 28). Somervell's heuristics had significantly higher thoroughness scores than Nielsen's heuristics, according to test of proportions ($z=3.26$, $p<0.05$). Berry's heuristics also held significantly higher thoroughness over Nielsen's ($z=2.11$, $p=0.04$). No significant differences were found between Somervell's and Berry's sets ($z=1.41$, $p=0.16$). Figure 2 provides a graphical representation of the thoroughness scores.

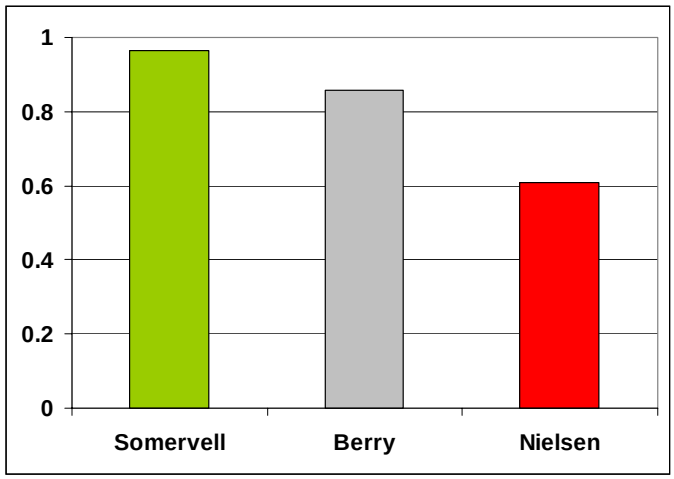

Figure 2. Thoroughness scores for each heuristic type.

Validity. Somervell's heuristics had the highest validity, with 27 of 33 claims yielding applicability scores greater than five, for a validity score of 82%. Berry's heuristics had the next highest validity with 24 of 33 claims, for a validity score of 73%. Nielsen's heuristics had the lowest validity score,

with 17 of 33 claims for a score of 52%. Test of proportions reveals significant differences between Somervell's heuristics and Nielsen's heuristics ($z = 2.61$, $p = 0.01$). No significant differences were found between Berry's heuristics and Nielsen's heuristics ($z = 1.78$, $p = 0.08$), nor between Somervell's heuristics and Berry's heuristics ($z = 0.88$, $p = 0.38$). Figure 3 provides a graphical representation of the validity scores.

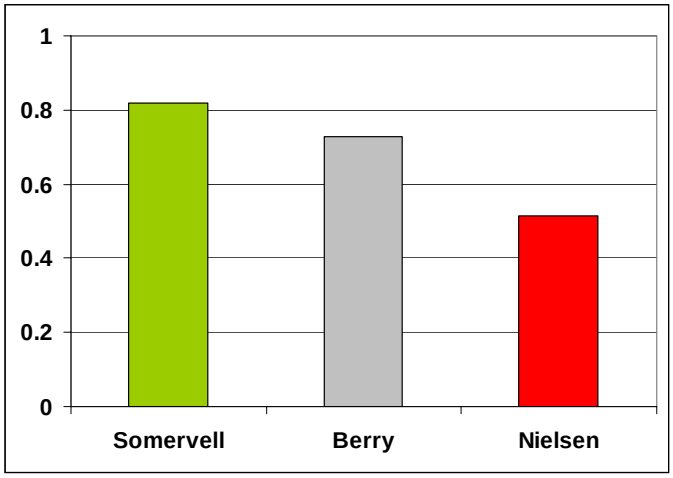

Figure 3. Validity scores for each heuristic set.

Effectiveness. Recall that effectiveness is the product of thoroughness and validity. This calculation is straightforward, and reflects the general trends observed in both the thoroughness and validity measures. Considering the effectiveness scores across all three systems reveals that Somervell's heuristics had the highest effectiveness with a score of 0.79. Berry's heuristics came next with a score of 0.62. Nielsen's heuristics had the lowest overall effectiveness with a score of 0.31. Figure 4 shows the effectiveness scores.

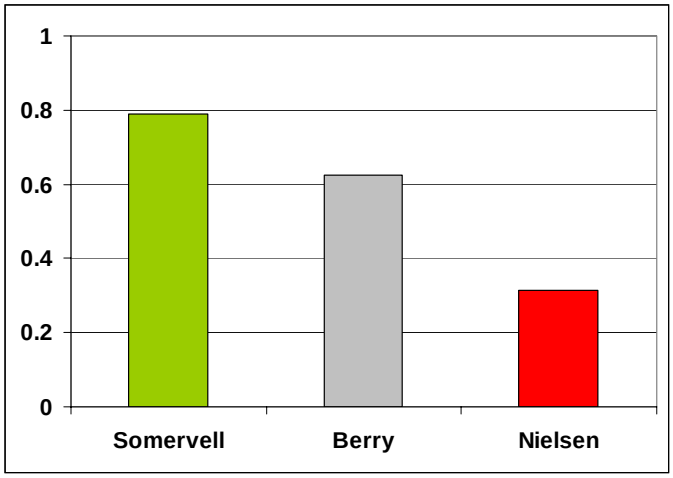

Figure 4. Effectiveness scores for each heuristic set.

Reliability. This difference is found by subtracting the ratings of each evaluator from every other evaluator and summing up each of the differences, then dividing by the number of differences (or the average difference). Suppose that an evaluator rated the first heuristic with a 6 (agree) and another rated it as a 4 (neutral) and a third rated it as a 5 (somewhat agree). The difference in this particular instance would be: (6 - 4) + (6 - 5) + (5 - 4)/3= 1.33

Considering all 33 claims across the three systems gives an overall indication of the average differences for the heuristic sets. One-way ANOVA suggests significant differences among the three heuristic sets ($F(2, 23) = 23.02$,$MSE = 0.84$, $p < 0.05$). Pair-wise t-tests show that Somervell's heuristics had significantly lower average differences than both Berry's heuristics ($df = 14$, $t = 4.3$, $p < 0.05$) and Nielsen's heuristics ($df = 16$, $t = 6.8$, $p < 0.05$). No significant differences were found between Berry's heuristics and Nielsen's heuristics ($df = 16$, $t = 1.43$, $p = 0.17$), but Berry's set had a slightly lower average difference ($M_B = 2.02$, $SD_B = 0.21$; $M_N = 2.14$, $SD_N = 0.13$). Figure 5 provides a graphical depiction of the difference scores. Note that a lower difference indicates better reliability.

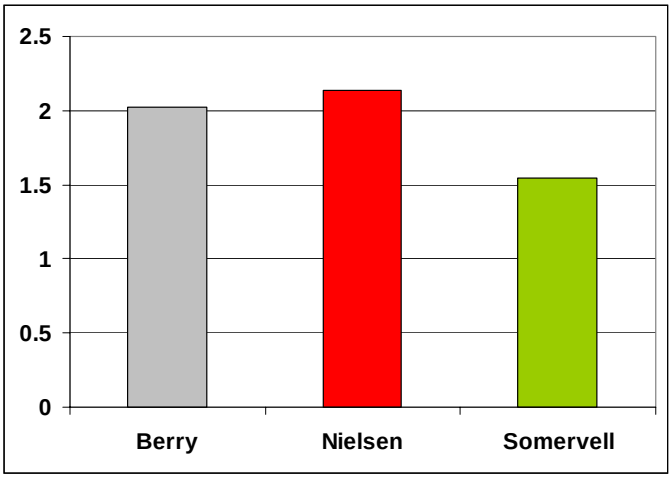

Figure 5. Average differences (reliability) for each heuristic set.

Cost. In this comparison study, method cost was estimated as the time required (in minutes) for an evaluator to complete the ratings for each problem across all three systems. This time was self-reported by the individual evaluators. Averaging reported times across evaluators for each method suggests that Somervell's set required the least amount of time ($M = 103.17$, $SD = 27.07$), but one-way ANOVA reveals no significant differences ($F(2, 17) = 0.26$, $p = 0.77$). Berry's set required the most time ($M = 119.14$, $SD = 60.69$) while Nielsen's set ($M = 104.29$, $SD = 38.56$) required slightly more than Somervell's. Figure 6 provides a graphical representation of the average times.

Thus, Somervell's heuristics seemed to fare better for all of the measures calculated in this comparison experiment. It should be noted that each of the calculations required simple spreadsheet operations, and no interpretation of evaluator intent were necessary.

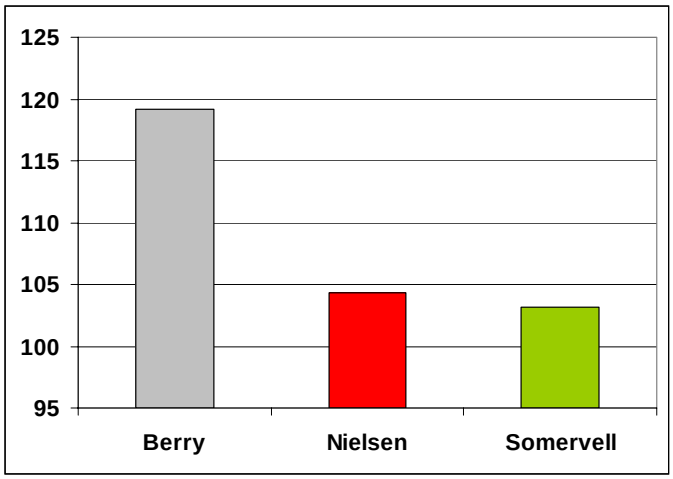

Figure 6. Average time spent for completing test, by heuristic set.

DISCUSSION

Based on these findings, one can argue that system class level heuristics are the best choice for formative heuristic evaluation of large screen information exhibits. In this case they are tailored to the unique user goals associated with that system class and hence provide better usability feedback. This finding falls in line with other heuristic comparison work (Mankoff et al., 2003, Baker et al., 2002) that suggests the appropriate level of UEM specificity should be at the system-class level.

By providing a set of problems to the evaluators, one can more accurately determine the applicability of a set of heuristics to that problem set. This allows quick, accurate calculation of several measures of the method and compare different methods on the same basis. Other comparison studies usually must deal with validity issues that arise from evaluator differences, investigation of lengthy, wordy problem reports, and then mapping multiple descriptions of problems to an accepted set of problems.

These comparison studies are often also plagued with having questionable or weak "real" problem sets. For example, a common technique is to use the union of problems found by all the methods in a comparison study as the real problem set. One problem with this approach arises from the fact that this set of problems may not be the ones that real users would experience during typical system use. In this approach, actual problems encountered by the users of the systems, as found through system inspection and feedback with developers, or through direct user studies are used as the real problem sets.

Our implementation of this new comparison technique suggests a better approach to UEM assessment. Instead of relying on highly variable problem sets from traditional evaluation approaches, one can establish a common base set to use in the calculation of comparison metrics. This reduces the variability in the calculations, ensuring that the comparison is fair and balanced.

Furthermore, this approach can be somewhat automated. By relying upon existing design knowledge, one can create a new testing setup by importing usability problems and heuristic sets to dynamically create new tests, either for evaluating the problems or for comparison tests. In fact, this particular effort is underway as part of the LINK-UP system for evaluating notification systems (Chewar et al., 2004). The testing platform used in this work can be automated to retrieve specific claims from a database, which can then be used in analytic evaluations.

REFERENCES

Baker, K., Greenberg, S., Gutwin, C. (2002) Empirical development of a heuristic evaluation methodology for shared workspace groupware. In *ACM Conference on Computer Supported Cooperative Work*, pages 96-105, New Orleans, LA.

Berry, B. (2003) Adapting heuristics for notification systems. In *41st Annual ACM Southeast Conference*, pages 144–149, Savannah, GA.

John M. Carroll and Mary Beth Rosson (1992). Getting around the task-artifact cycle: how to make claims and design by scenario. *ACM Transactions on Information Systems*, 10(2), pages 181–212.

Chewar, C. M., Bachetti, E., McCrickard, D. S., Booker, J. (2004) Automating a Design Reuse Facility with Critical Parameters: Lessons Learned in Developing the LINK-UP System. In *Proceedings of the 2004 International Conference on Computer-Aided Design of User Interfaces*, pages 236-247, Island of Madeira, Portugal.

Churchill, E., Nelson, L., Denoue, L., Girgensohn, A. (2003). The plasma poster network: Posting multimedia content in public places. In *Ninth IFIP TC13 International Conference on Human-Computer Interaction*, pages 599–606, Zurich, Switzerland.

Cockton, G, Woolrych, A. (2002) Sale Must End: Should Discount Methods be Cleared off HCI's Shelves? *interactions.* September + October, pages 13-18.

Gray, W., Salzman, M. (1998) Damaged merchandise? a review of experiments that compare usability evaluation methods. *Human-Computer Interaction, 13(4)* pages 203–261.

Greenberg, S., Rounding, M. (2001) The notification collage: Posting information to public and personal displays. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 515–521, Seattle, WA, April 2001.

Hartson, H. R., Andre, T. S., Williges, R. C. (2001) Criteria for evaluating usability evaluation methods. *International Journal of Human-Computer Interaction, 13(4)* pages 373–410.

Karat, CM, Campbell, R., Fiegel, T. (1992) Comparison of empirical testing and walkthrough methods in user interface evaluation. In *Conference on Human Factors and Computing Systems*, pages 397–404, Monterrey, California.

Lavery, D., Cockton, G., Atkinson, M. (1997) Comparison of evaluation methods using structured usability problem reports. *Behaviour and Information Technology, 16(4/5)* pages 246–266.

Mankoff, J., Dey, A., Hsieh, G., Kientz, J., Lederer, S., Ames, M. (2003) Heuristic evaluation of ambient displays. In *Proceedings of the ACM Conference on Human Factors and Computing Systems,pages169-176* , Ft. Lauderdale, FL.

McCrickard, D. S., Chewar, C. M., Somervell, J., Ndiwalana, A. (2003) A Model for Notification Systems Evaluation--Assessing User Goals for Multitasking Activity. *ACM Transactions on Computer-Human Interaction*, 10(4), pages 312-338.

Nielsen, J., Mack, R. L. (1994) *Usability Inspection Methods* . John Wiley and Sons, New York, NY.

Somervell, J., Wahid, S., McCrickard, D. S. (2003) Usability heuristics for large screen information exhibits. In *Proceedings of the Ninth IFIP TC13 International Conference on Human Computer Interaction*, pages 904-907, Zurich, Switzerland.