INL/CON-05-00108
PREPRINT

Advancing Usability Evaluation Through Human Reliability Analysis

Human Computer Interaction International 2005

Ronald L. Boring
David I. Gertman

July 2005

The INL is a
U.S. Department of Energy
National Laboratory
operated by
Battelle Energy Alliance

Advancing Usability Evaluation through Human Reliability Analysis

Ronald L. Boring and David I. Gertman

Idaho National Laboratory
Idaho Falls, Idaho 83415, USA
{ronald.boring,david.gertman}@inl.gov

Abstract

This paper introduces a novel augmentation to the Mohlich and Nielsen's heuristic usability evaluation methodology. The SPAR-H human reliability analysis method was developed for categorizing human performance in nuclear power plants. Despite the specialized use of SPAR-H for safety critical scenarios, the method also holds promise for use in commercial off-the-shelf software usability evaluations. The SPAR-H method shares task analysis underpinnings with usability approaches to human-computer interaction, and it can be easily adapted to incorporate usability heuristics as performance shaping factors. By assigning probabilistic modifiers to heuristics, it is possible to arrive at the usability error probability (UEP). This UEP is not a literal probability of error but nonetheless provides a quantitative basis to heuristic evaluation. Because HRA provides estimates of human error, it offers a seamless method for prioritizing usability issues, as high error rates typically require more immediate fixes.

1 Introduction

Human-computer interaction (HCI) centers on the iterative design and usability testing of software and hardware devices. Designers and usability testers are routinely employed or contracted by corporations and organizations to use a myriad of techniques to improve software and hardware. For commercial off-the-shelf (COTS) software and hardware, the goal of these efforts is to make usable, appealing, and useful systems (Nielsen, 1993). For specialized software and hardware, such as in nuclear power plant control rooms or human-robot interfaces for urban search and rescue, the goal of HCI is to make safe, usable, and standards-compliant systems.

An emerging issue within human-computer interaction (HCI) is the need for simplified or "discount" methods. The current economic slowdown has necessitated innovative methods that are both results driven and cost effective (Lindgaard, 2004). The myriad methods of design and usability are currently being cost-justified, and new techniques are actively being explored that meet current budgets and needs. In this paper, we present an approach to usability evaluation that combines existing cost-justified heuristic evaluation techniques with a novel method of human reliability analysis (HRA).

Recent efforts in HRA are highlighted by the ten-year development of the Standardized Plant Analysis Risk HRA (SPAR-H) method for the US Nuclear Regulatory Commission (Gertman et al., in press). The SPAR-H method provides a taxonomy of common human errors and quantifies them in terms of human error probabilities. A SPAR-H analysis incorporates a thorough task analysis of events in order to model the sequence of events, error precursors, and consequences of errors. The SPAR-H method has been used primarily for determining human-centered risk at nuclear power plants (Boring et al., 2004). However, the SPAR-H method, like other HRA methods, shares task analysis underpinnings with HCI. Despite this methodological overlap, there is currently no HRA approach deployed in usability evaluation (Gertman et al., 2004). This paper presents an extension of the existing SPAR-H method to be used as part of usability evaluation in HCI.

2 Usability Heuristics

Heuristic evaluation is one of the key methods available in the list of streamlined methods for assessing the usability of interfaces. Heuristics, as defined by Molich and Nielsen (1990), are short lists of key factors that comprise a usable interface. More specifically, it is the absence of these factors that contributes to user errors and dissatisfaction with interfaces. Typically, a list of relevant usability characteristics is used as a checklist by a

usability evaluator or design expert. In reviewing the interface, the usability evaluator or design expert identifies specific areas in which the interface violates these usability characteristics.

Heuristic evaluation is not without shortcomings. While it is estimated that the probability of heuristic evaluation detecting any given usability problem is around 32% (Nielsen & Landauer, 1993), the likelihood that separate evaluators will detect the same problems is considerably lower (Kessner, Wood, Dillon, & West, 2001). However, various approaches have been employed to improve the problem hit rate as well as the interrater reliability of the method (Chattratichart & Brodie, 2004; Law & Hvannberg, 2004).

A shortcoming of heuristic evaluation that is seldom discussed is the need to prioritize those usability issues that have been identified (McInerney, Pantel, & Melder, 2001). Heuristic evaluation provides a concise checklist of usability issues, but it does not provide the usability evaluator with a clear means to prioritize the list of issues that are identified. Without a method to prioritize usability issues, the evaluator must use his or her subjective best judgment to highlight the severity of those issues that he or she believes will have the greatest overall impact on the product's usability. A particularly lamentable consequence of such a systematic lack of prioritizing issues is that the impact of some issues is underestimated by evaluators. Moreover, because prioritization is not always tractable by objective metrics, development effort allocations may not correlate to the actual severity of software dysfunction that is attributable to a usability issue. The result is that much needed usability ameliorations may sometimes be omitted in the translation from a heuristic evaluation to software refinement.

3 Human Reliability Analysis (HRA)

The SPAR-H method was developed to assess the probability of human error in nuclear power plants (Gertman et al., in press). Human error probabilities (HEPs) are incorporated into overall probabilistic fault and event trees. Combined with system and component error probabilities, HEPs allow probabilistic risk analysts to identify end states in which the safety of the power plant could be compromised. Because these end states have associated probabilities, the analysts are able to determine which areas need to be addressed to increase plant safety. The analysts operate within certain acceptable bounds of error, such that any end state with a probability over a specific set point is flagged for evaluation and immediate system and system-operator redesign.

The SPAR-H method is based on eight performance shaping factors that encapsulate the majority of the contributors to human error. These eight performance shaping factors are as follows: *available time to complete task*, *stress and stressors*, *experience and training*, *task complexity*, *ergonomics*, *the quality of any procedures in use*, *fitness for duty*, and *work processes*. Each performance shaping factor features a list of levels and associated multipliers. For example, the presence of extremely high stress would receive a higher multiplier than moderate stress. A higher multiplier results in a higher decrement in human performance and a corresponding increase in the likelihood of human error (see Figure 1).

It is important to note that these performance shaping factors are not truly orthogonal. Certain performance shaping factors tend to co-occur, and, as illustrated in Figure 1, some performance shaping factors exert influence on other performance shaping factors. Since the exact interrelationship between performance shaping factors is not known, the SPAR-H method does not currently attempt to partition co-variance between performance shaping factors.

The SPAR-H method assigns human activity to one of two general task categories: *action* or *diagnosis*. Examples of action tasks include operating equipment, conducting calibration or testing, and other activities performed during the course of system operations. Diagnosis tasks consist of planning and prioritizing activities, determining appropriate courses of action, and using knowledge and experience to understand existing conditions. Operational research suggests that for cognitively engaging tasks such as diagnosis, people tend to exhibit a base human error rate equal to 0.01 (or 1E-2). This means that people have about a 1 in a 100 chance of making a diagnosis error, excluding any adjustment for performance shaping factors or dependencies between a chain of events. This base or default error rate is called the nominal human error probability (NHEP). For tasks that are more action oriented, the base human error rate is equal to about 0.001 (or 1E-3), suggesting about a 1 in a 1000 chance of making an error. If a control room operator is working in the area of developed skills, the analyst uses the action NHEP. However, if considerable domain extrapolation is required on the part of the control room operator, the diagnosis nominal HEP would be used for quantification. Base error rates for the two task types associated with the SPAR-H method have been calibrated against other human reliability analysis methods. This calibration reveals that the SPAR-H human error rates fall

within the range of rates predicted by other methods and accords with human performance data found the behavioral and cognitive sciences literature.

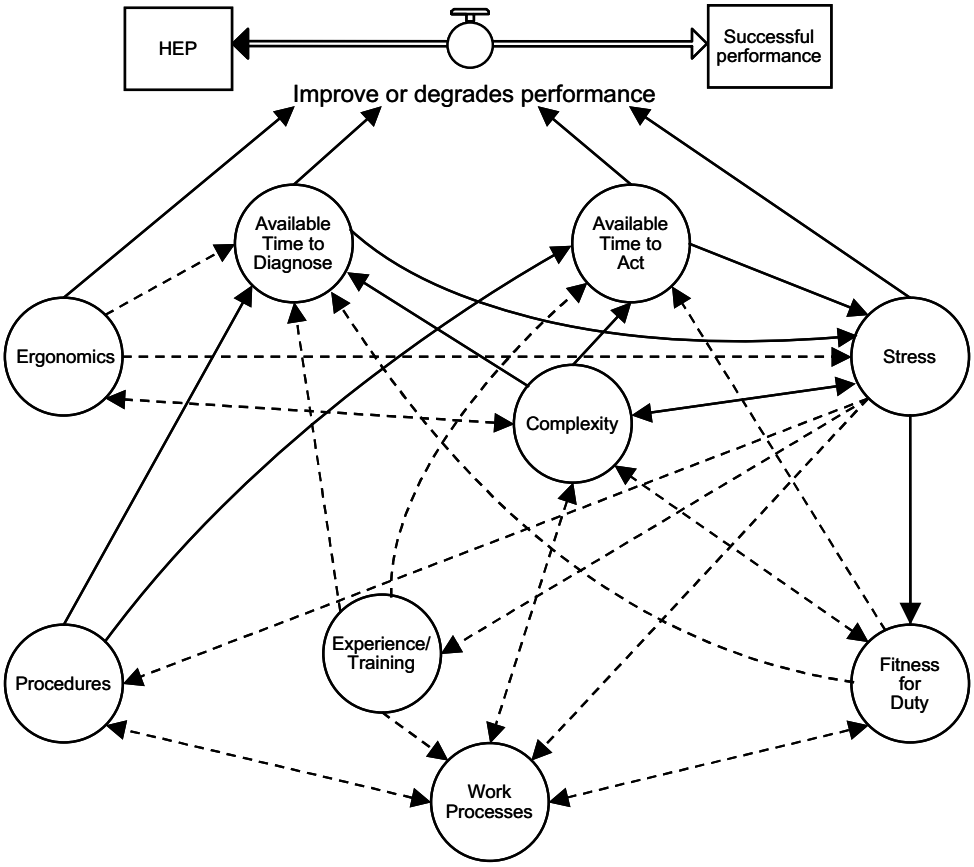

Figure 1: Relationships between performance shaping factors in SPAR-H, where solid lines indicate a strong relationship and dotted lines indicate a moderate relationship (from Gertman et al., in press)

4 Heuristic Evaluation and HRA

4.1 Basic Process

While the SPAR-H performance shaping factors have been employed for analyzing human performance among highly trained staff at nuclear power plants, the method has not yet been validated in other domains. Nonetheless, the applicability of the SPAR-H performance shaping factors to the domain of HCI remains a promising possibility. Until such time as the existing performance shaping factors can be calibrated for use in HCI, it is fruitful to borrow other identified contributors to human error. In terms of usability, heuristics provide a readily available list of surrogate performance shaping factors. Moreover, by assigning quantitative multipliers to the heuristics akin to the performance shaping factors used in SPAR-H, the methodology provides an easy way to prioritize usability issues.

Table 1 illustrates a rubric of usability heuristics (Nielsen, 1993) that have been quantified using generic performance shaping factor multipliers from SPAR-H. As a proof of concept, the multipliers have been held constant across each heuristic, whereas in SPAR-H, the precise multipliers have been validated for individual performance shaping factors. Using Table 1, the usability evaluator performing the heuristic evaluation simply identifies the correct level of usability violation for each heuristic. Associated with each level is a multiplier. To tally the total usability error probability (UEP), the evaluator multiplies the product of the individual heuristic

multipliers by the diagnosis or action NHEP. A higher number suggests that the usability issue has a higher likelihood of occurrence and, therefore, a higher need or priority to be addressed.

Table 1: The SPAR-H based heuristic evaluation matrix for calculating usability error probabilities

➲ ***Circle the appropriate multiplier for each heuristic.***

Heuristic	Multipliers				
Simple and natural dialog	10 Poor	5 Available	1 Nominal	0.2 Good	0.1 Excellent
Speak the users' language	10 Poor	5 Available	1 Nominal	0.2 Good	0.1 Excellent
Minimize users' memory load	10 Poor	5 Available	1 Nominal	0.2 Good	0.1 Excellent
Consistency	10 Poor	5 Available	1 Nominal	0.2 Good	0.1 Excellent
Clearly marked exits	10 Poor	5 Available	1 Nominal	0.2 Good	0.1 Excellent
Shortcuts	10 Poor	5 Available	1 Nominal	0.2 Good	0.1 Excellent
Good error messages	10 Poor	5 Available	1 Nominal	0.2 Good	0.1 Excellent
Prevent errors	10 Poor	5 Available	1 Nominal	0.2 Good	0.1 Excellent
Help and documentation	10 Poor	5 Available	1 Nominal	0.2 Good	0.1 Excellent

➲ ***Multiply the product of the heuristic multipliers by the nominal human error probability to arrive at the usability error probability.***

☐ Diagnosis: 1.0E-2 x ___ x ___ x ___ x ___ x ___ x ___ x ___ x ___ x ___ = _____

☐ Action: 1.0E-3 x ___ x ___ x ___ x ___ x ___ x ___ x ___ x ___ x ___ = _____

Note that the performance shaping factor heuristics in Table 1 may have either positive or negative effects, implying good or poor usability, respectively. Negatively weighted performance shaping factors serve to increase the UEP. These multipliers have a value greater than 1. If a performance shaping factor is notably positive, the performance shaping factor may serve to decrease the UEP. These multipliers have a value less than 1.

4.2 Special Calculations

Using the multipliers, it is sometimes possible to arrive at a UEP that is greater than 1.0. A raw UEP that is greater than 1.0 suggests that the probability of a significant usability error is near 100%. The number must be truncated at 1.0, but the uncertainty surrounding the estimate considerably diminishes as the raw value exceeds 1.0. To compensate for UEPs that are greater than 1.0, a correction factor is applied to standardize the number over a range from 0.0 to 1.0:

$$UEP = \frac{NHEP \cdot PSF_{composite}}{NHEP \cdot \left(PSF_{composite} - 1\right) + 1} \quad (1)$$

UEP signifies the corrected usability error probability, *NHEP* signifies the nominal HEP value for diagnosis or action usability error types, and $PSF_{composite}$ signifies the product of the multipliers for the performance shaping factors.

In some cases, the usability evaluator may find that it is not possible to parse a task into solely a cognitively engaging diagnosis or a routine action task. In such a case, the evaluator should treat the task as a joint diagnosis and action task. The joint UEP is calculated by taking the sum of the corrected diagnosis and action UEPs. If the joint UEP should exceed 1.0, it is truncated at 1.0.

4.3 Task Granularity

A particular challenge in all usability evaluation, but especially in heuristic evaluation, is a determination of the ideal level of detail for the task decomposition. In most cases, a GOMS level task analysis (Card, Moran, & Newell, 1983) provides too fine a level of detail for consumer software usability evaluations, although it is well suited for the level of precision required in safety critical HCI. The usability evaluator employing the method outlined in this paper will generally find the appropriate level of task decomposition in *use cases*, or typical usage scenarios (Brinck, Gergle, & Wood, 2002). For simplicity and expedience, it is convenient to evaluate the entire usage scenario according to one set of heuristics. In other words, the entire usage case scenario produces a single UEP, which highlights usability strengths and weaknesses across the user's experience.

However, each use case may also be broken down into a series of steps that are required by the user to complete a desired goal. The usability evaluator may present an individual UEP and analysis for each subtask. Alternately, the evaluator may wish to present a composite UEP of all subtask UEPs. In this case, the evaluator must determine the logical relationship between subtasks. A logical AND relationship implies that two subtasks are interdependent, such that both must concurrently have significant usability issues to prevent successful completion of the overall task. In contrast, a logical OR relationship implies that a significant usability issue in any subtask could prevent successful completion of the overall task. Figure 2 provides a fault tree diagram of the logical AND and OR relationships between subtasks. A logical AND relationship is treated multiplicatively; the composite UEP is the product of the subtask UEPs. A logical OR relationship is treated cumulatively; the composite UEP is the sum of the subtask UEPs, truncated at 1.0.

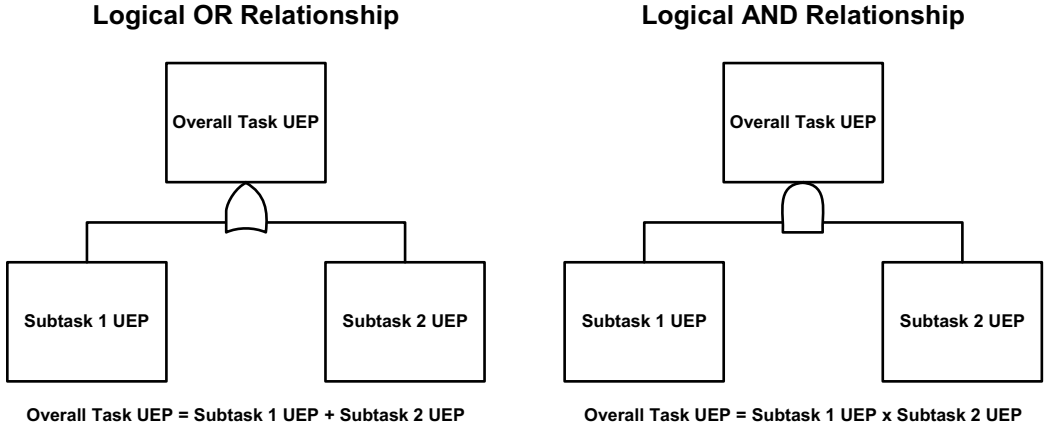

Figure 2: Fault tree diagram depicting logical OR and AND relationships for task decomposition

5 Consequence Determination

Just because a usability issue results in a high UEP value, this does not automatically mean that it is a high priority item. In much of probabilistic risk assessment, there is a further step in evaluating the importance of a condition. The classic conception of probabilistic risk holds risk is the product of a likelihood and a consequence (Garrick et al, 2004). In our discussion thus far, we have treated all usability issues as having the same consequence. In fact, some issues may have greater consequence. For example, a software usability issue that leads to loss of data is of greater consequence than a usability issue in which the user misunderstands a harmless command from which there is easy recovery. A separate consequence multiplier aids the usability evaluator in fine tuning the prioritization of usability issues extracted through heuristic evaluations.

Table 2 presents a consequence matrix consisting of four levels of usability consequence and three resulting priority levels for corrective action. Each of the four usability consequences has a corresponding consequence multiplier. Multiplying the overall UEP by the consequence multiplier produces the usability consequence coefficient (UCC). The UCC maps directly to three levels of prioritization, ranging from low (fix is not required) to high (fix is required). The usability evaluator should fine tune the mapping from the UCC to the prioritization levels as appropriate to meet the application-specific acceptable levels of usability.

Table 2: Usability consequence matrix

Usability Consequence	Consequence Multiplier	Usability Consequence Coefficient (UCC)
High *Serious usability problem that may cause loss of data, system malfunction, or user attrition*	5	UCC = UEP x 5 =
Medium *Moderate usability problem that inconveniences user but affords sufficient recovery that most users can carry out task*	2	UCC = UEP x 2 =
Low *Usability inconvenience that does not impede overall system usage or inconvenience user*	1	UCC = UEP =
None *No usability consequence*	0	UCC = 0

UCC Range	Priority
UCC > 0.09	**High** *Serious usability problem that requires immediate fix*
0.02 < UCC < 0.09	**Medium** *Usability problem that should be fixed for optimal usability*
UCC ≤ 0.02	**Low** *Usability has minimal impact on product and does not require fix*

6 Illustration of Concept

A summary of the steps required for HRA informed heuristic usability evaluation is depicted in Figure 3. The usability evaluator must first determine the appropriate level of task decomposition. Then, he or she performs the heuristic evaluation and calculates the UEP, including consideration of joint diagnosis and action tasks as well as the correction factor in Equation 1 for raw UEP values that exceed 1.0. Finally, the evaluator determines the consequence of the usability issues and calculates the usability priority.

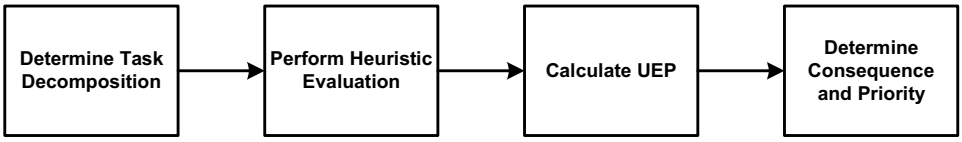

Figure 3: Required steps in HRA informed heuristic usability evaluation

To illustrate this method in practice, consider a software interface that has cumbersome dialog and no discernible exits but that has good shortcuts. The user is confused and goes down a path from which he or she has difficulty backtracking. However, the user is aware of a keyboard shortcut, which allows him or her to backtrack in the software to a more comprehensible area of the interface.

In considering this example, the usability evaluator would first determine the appropriate level of task decomposition. For purposes of parsimony, the evaluator elects for a one-task heuristic evaluation. Next, the evaluator performs the heuristic evaluation. The dialog heuristic would be marked as "poor" and receive a corresponding multiplier of 10. For the clear exit heuristic, the usability evaluator would similarly denote that it was "poor" with the corresponding multiplier of 10. Both the poor dialog and the poor exit in the interface serve to decrease the UEP. However, the excellent availability of shortcuts—in this case a readily known keystroke combination to backtrack—would also be noted and would counteract the negative influence of the dialog and exit heuristics. For the shortcuts heuristic, the evaluator would circle "excellent" with the corresponding multiplier of 0.1. All other heuristics would be treated as nominal, with a null-effect multiplier of 1. Taking the product of the three non-nominal heuristic multipliers, 10 x 10 x 0.1, yields a value of 10. This value is in turn multiplied by the diagnosis NHEP of 0.01 (to signify a cognitively engaging task) to produce a composite UEP equal to 0.1. Since this value does not exceed a UEP value of 1.0, it is not necessary to apply the correction factor in Equation 1. Thus, the overall likelihood that this series of issues will result in a significant disruption to the usability of the software is 1 in 10. The consequence of this combination of usability heuristics is determined to be "medium," implying that it inconveniences the user but the user is generally able to recover from this inconvenience. A "medium" usability consequence has a multiplier of 2. Thus, the UCC equals the UEP (0.1) multiplied by the consequence (2), or 0.2. In Table 2, this UCC value maps to a "high priority" usability item that requires a fix. While the user may be able to backtrack in the interface, the combination of negative heuristics may significantly impede software usability, warranting a fix to one or more of the problem areas identified in the heuristics. This simple example helps illustrate how it is possible for a heuristic usability evaluator to produce a systematic, quantitative, and tractable metric for prioritizing usability issues in an interface.

7 Discussion

7.1 Shortcomings of the Method

This augmentation of heuristic evaluation does not purport to offer a literal metric for calculating the usability error likelihood. The multipliers are provided merely as examples of how heuristic multipliers may be used to provide a quantification based prioritization of usability issues. The values provided as proof of concept are also not weighted according to their overall contribution to the usability of the system. There is evidence, for example, that help and documentation are seldom used in software (Dworman & Rosenbaum, 2004). It is therefore likely that this heuristic would not receive the same weighting as other heuristics that are more likely to impinge on a product's usability.

Moreover, the exact selection of heuristics is a matter open for debate. It is therefore crucial that the prospective user of this method keep in mind the restricted generalizability of the quantities that this method provides. The quantities are aids toward prioritizing usability issues; they are not literal, citable probabilistic metrics of the overall usability of a product.

7.2 Advantages of the Method

Despite these limitations, HRA influenced heuristic evaluation affords distinct advantages over current heuristic evaluation practices. Current heuristic evaluation techniques provide little guidance on prioritizing usability issues. If current heuristic evaluation reveals a usability issue, it is typically a matter of the usability evaluator's subjective judgment to determine which issues have the most pressing need for correction. Because HRA provides estimates of human error, it offers a seamless method for prioritizing usability issues, as high error rates typically require more immediate fixes. The prioritization is further simplified by the incorporation of the consequence matrix.

Further, current usability evaluation techniques are minimally cost justified. An organization that invests in HCI receives design guidance that is often not clearly tied to a product's return on investment (ROI). While HRA driven HCI cannot answer all ROI questions, it provides an end state mapping of user interaction with the product. By providing quantitative error and potential consequence estimates, HRA influenced heuristic evaluation better informs investment decisions pertaining to product design and refinements.

Finally, current usability evaluation techniques are not standardized for safety critical applications. Because HRA is grounded in the safety arena, its implementation in HCI allows the method to scale from consumer grade COTS software and hardware to safety critical systems. Where HCI standards guidance is available for safety critical systems, HRA driven HCI is better able to ensure standards compliance and resolve standards issues than current usability evaluation techniques by potentially incorporating standards as performance shaping factors.

Much research still remains in developing HRA driven usability evaluation. Future efforts will focus on refining and validating the heuristics that have been identified as performance shaping factors. Initial research will aim to determine the appropriate weightings of individual heuristics as well as the appropriate multipliers for probabilistic estimation. Additionally, further examples will be developed to illustrate the utility of this method across a wide range of usability domains. Ultimately, the authors trust that this method will prove a useful and robust addition to current usability evaluation methods.

8 Acknowledgements

This research was funded by a Laboratory Directed Research and Development grant to Dr. Ronald L. Boring at Idaho National Laboratory, a US Department of Energy laboratory operated by Battelle Energy Alliance. The authors gratefully acknowledge the contributions of Harold S. Blackman, Bruce P. Hallbert, Jeffrey C. Joe, and Julie L. Marble in developing the ideas presented in this paper.

References

Boring, R.L., Gertman, D.I., & Marble, J.L. (2004). Temporal factors of human error in SPAR-H human reliability analysis modeling. *Proceedings of the 48th Annual Meeting of the Human Factors and Ergonomics Society*, 1165-1169.

Brinck, T., Gergle, D., and Wood, S.D. (2002). *Usability for the web: Designing web sites that work* San Francisco: Morgan Kaufmann.

Card, S.K., Moran, T.P., and Newell, A. (1983). *The psychology of human-computer interaction.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Chattratichart, J., and Brodie, J. (2004). Applying user testing data to UEM performance metrics. *Proceedings of CHI 2004*, 1119-1122.

Dworman, G., and Rosenbaum, S. (2004). Helping users to use help: Improving interaction with help systems. *Proceedings of CHI 2004*, 1717-1718.

Garrick, B.J., Hall, J.E., Kilger, M., McDonald, J.C., O'Toole, T., Probst, P.S., Parker, E.R., Rosenthal, R., Trivelpiece, A.W., Van Arsdale, L.A., and Zebroski, E.L. (2004). Confronting the risk of terrorism: Making the right decisions. *Reliability Engineering & System Safety*, *86*, 129-176.

Gertman, D., Blackman, H., Marble, J., Byers, J., Haney, L., and Smith, C. (In press). *The SPAR-H human reliability analysis method, NUREG/CR-in press*. Washington, DC: US Nuclear Regulatory Commission.

Gertman, D.I., Boring, R.L., Marble, J.L., and Blackman, H.S. (2004). Mixed model usability evaluation of the SPAR-H human reliability analysis method. *Proceedings of the Fourth American Nuclear Society International Topical Meeting on Nuclear Power Plant Instrumentation, Controls, and Human-Machine Interface Technologies,* 59-67.

Kessner, M., Wood, J., Dillon, R.F., and West, R.L. (2001). On the reliability of usability testing. *Proceedings of CHI 2001*, 97-98.

Law, E.L.-C., and Hvannberg, E.T. (2004). Analysis of strategies for improving and estimating the effectiveness of heuristic evaluation. *Proceedings of NordiCHI 2004*, 241-250.

Lindgaard, G. (2004). Making the business our business: One path to value-added HCI. *Interactions*, *11*(*3*), 12-17.

McInerney, P., Pantel, C., and Melder, K. (2001). Managing usability defects from identification to closure. *CHI 2001 Extended Abstracts*, 497-498.

Mohlich, R., and Nielsen, J. (1990). Improving a human-computer dialogue. *Communications of the ACM*, *33*, 338-348.

Nielsen, J. (1993). *Usability engineering*. Boston: AP Professional.

Nielsen, J., and Landauer, T.K. (1993). A mathematical model of the finding of usability problems. *Proceedings of InterCHI 1993*, 206-213.