INL/CON-05-00423
PREPRINT

Proof of Concept for a Human Reliability Analysis Method for Heuristic Usability Evaluation of Software

Human Factors and Ergonomics Society 49th Annual Meeting – 2005

Ronald L. Boring
David I. Gertman
Jeffrey C. Joe
Julie L. Marble

September 2005

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

PROOF OF CONCEPT FOR A HUMAN RELIABILITY ANALYSIS METHOD FOR HEURISTIC USABILITY EVALUATION OF SOFTWARE

Ronald L. Boring, David I. Gertman, Jeffrey C. Joe, and Julie L. Marble
Idaho National Laboratory, Idaho Falls, Idaho 83415, USA

An ongoing issue within human-computer interaction (HCI) is the need for simplified or "discount" methods. The current economic slowdown has necessitated innovative methods that are results driven and cost effective. 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. Recent efforts in human reliability analysis (HRA) are highlighted by the ten-year development of the Standardized Plant Analysis Risk HRA (SPAR-H) method. The SPAR-H method has been used primarily for determining human-centered risk at nuclear power plants. The SPAR-H method, however, shares task analysis underpinnings with HCI. Despite this methodological overlap, there is currently no HRA approach deployed in heuristic usability evaluation. This paper presents an extension of the existing SPAR-H method to be used as part of heuristic usability evaluation in HCI.

INTRODUCTION

Human-computer interaction (HCI) centers on the iterative design and usability testing of software and hardware devices. Designers and usability evaluators are routinely employed or contracted by corporations and organizations to use a variety 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.

Recent efforts in human reliability analysis (HRA) are highlighted by the ten-year development of the Standardized Plant Analysis Risk HRA (SPAR-H) method for the US Nuclear Regulatory Commission (Boring et al., 2004; Gertman et al., in press). The SPAR-H method provides a taxonomy of common human error contributors 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). 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 heuristic usability evaluation in HCI.

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 or design expert. In reviewing the interface, the usability 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 and Landauer, 1993), the likelihood that separate experts will detect the same problems is considerably lower (Kessner, Wood, Dillon, and West, 2001). However, various approaches have been employed to improve the problem hit rate as well as the interrater reliability of the method (Chattratichart and Brodie, 2004; Law and Hvannberg, 2004).

An issue regarding heuristic evaluation that is seldom discussed in the literature is the need to prioritize those usability issues that have been identified. Heuristic evaluation provides a possible checklist of usability issues, but it does not provide the usability expert with a clear means to prioritize the list of issues that are identified. Without a method to prioritize usability issues, the expert must use his or her subjective best judgment to highlight those issues that he or she believes will have the greatest overall impact on the product's usability.

HUMAN RELIABILITY ANALYSIS

The SPAR-H method was developed to assess the probability of human error in nuclear power plants. 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 possible system and system-operator redesign.

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 ___ = _____*

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

***NOTE:** Any value greater than 1 is treated as 1.

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 HEP.

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 control 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 the system user is working in the area of developed skills, for purposes of risk estimation, the analyst uses the action NHEP. However, if considerable domain extrapolation is required on the part of the user, the diagnosis NHEP should be used for quantification. The multipliers may be updated throughout each phase of operations to enable dynamic human error modeling. In such cases, a correction factor is incorporated in the modeling to account for dependencies between events (see Gertman et al., in press).

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 (Gertman et al., in press).

HEURISTIC EVALUATION AND HRA

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 for other domains. The applicability of the SPAR-H performance shaping factors to the domain of HCI remains an interesting possibility. However, 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 performance shaping factors for the determination of usability. Moreover, by assigning quantitative multipliers to the heuristics akin to those 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 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 each individual performance shaping factor. The usability expert 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 expert 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.

For example, consider a software interface that has cumbersome dialogue and no discernible exits but that has good shortcuts. In this case, the user is confused and goes down a path from which he or she has difficulty returning. 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. The *dialogue* heuristic would be marked as "poor" and receive a corresponding multiplier of 10. For the *clear exit* heuristic, the usability expert would denote that it was "poor" with the corresponding multiplier of 10. Both the poor dialogue and the unavailable 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 *dialogue* and *exit* heuristics. For the *shortcuts* heuristic, the expert 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. The overall likelihood that this series of issues will result in a significant disruption to the usability of the software is 1 in 10. By comparison, if a different portion of the interface featured a nominal value for the shortcuts heuristic, the composite UEP would by 10. Any UEP greater than 1 is treated as 1, suggesting that there is nearly a 100% chance that this series of issues will prove a significant impediment to software usability.

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 that 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.

DISCUSSION

This augmentation of heuristic evaluation does not purport to offer a validated 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

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*

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 and 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 citable metrics of the overall usability of a product.

Despite these limitations, HRA influenced heuristic evaluation affords distinct advantages over current usability 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 engineer's subjective judgment to determine which issues have the most pressing need for correction. Because HRA provides human error probabilities, it offers a seamless method for prioritizing usability issues, as high error rates typically require more immediate fixes.

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.

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 views expressed in this paper do not necessarily represent the views of Idaho National Laboratory, the US Department of Energy, or Battelle Energy Alliance. Portions of this paper first appeared in a paper entitled "Advancing Usability Evaluation through Human Reliability Analysis" presented at the 2005 HCI International Conference.

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.

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

Dworman, G., and Rosenbaum, S. (2004). Helping users to use help: Improving interaction with help systems. *Proceedings of CHI*, 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., & 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*, 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*, 241-250.

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 the CHI*, 206-213.