INL/CON-06-11187
PREPRINT

A Human Reliability Based Usability Evaluation Method for Safety-Critical Software

Human Factors in Computing Systems: CHI 2006

Ronald L. Boring
Tuan Q. Tran
David L. Gertman
Austin Ragsdale

April 2006

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

This is a preprint of a paper intended for publication in a journal or proceedings. Since changes may be made before publication, this preprint should not be cited or reproduced without permission of the author. This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, or any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility for any third party's use, or the results of such use, of any information, apparatus, product or process disclosed in this report, or represents that its use by such third party would not infringe privately owned rights. The views expressed in this paper are not necessarily those of the United States Government or the sponsoring agency.

A Human Reliability Based Usability Evaluation Method for Safety-Critical Software

Ronald L. Boring, Tuan Q. Tran, & David I. Gertman
Human Factors, Instrumentation and Control Systems
Idaho National Laboratory
Idaho Falls, ID 83415, USA
{ronald.boring,tuan.tran,david.gertman}@inl.gov

Austin Ragsdale
Department of Psychology
Brigham Young, University-Idaho
Rexburg, ID 83460, USA
ragsdale22@gmail.com

Abstract – *Boring and Gertman (2005) introduced a novel method that augments heuristic usability evaluation methods with that of the human reliability analysis method of SPAR-H. By assigning probabilistic modifiers to individual heuristics, it is possible to arrive at the usability error probability (UEP). Although this UEP is not a literal probability of error, it nonetheless provides a quantitative basis to heuristic evaluation. This method allows one to seamlessly prioritize and identify usability issues (i.e., a higher UEP requires more immediate fixes). However, the original version of this method required the usability evaluator to assign priority weights to the final UEP, thus allowing the priority of a usability issue to differ among usability evaluators. The purpose of this paper is to explore an alternative approach to standardize the priority weighting of the UEP in an effort to improve the method's reliability.*

I. INTRODUCTION

Heuristics are short lists of key factors that comprise a usable interface [1]. More specifically, it is the absence of these factors that contributes to user errors and dissatisfaction with interfaces. Typically, in heuristic usability evaluation, a list of relevant usability characteristics is used as a checklist by an evaluator or design expert. In reviewing the interface, the usability evaluator identifies specific areas in which the interface violates these usability characteristics. Importantly, 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. To mitigate this concern, Boring and Gertman [2] borrowed quantification procedures from the field of human reliability analysis (HRA), specifically, the Standardized Plant Analysis Risk HRA (SPAR-H) method [3]. Boring and Gertman's [2] HRA-based Usability (HRA-U) method provides a clear and easy way to prioritize usability issues.

However, a limitation to the original HRA-U method was its reliance on usability evaluators to assign priority weights to the final usability error probability (UEP). Thus, the initial method allowed two usability evaluators to differ in their prioritization of usability issues that need to be fixed, even though both evaluators may have identified identical usability concerns. The subjective activity of assigning priority weights to the UEP can seriously hamper the method's reliability. This paper presents an extension of Boring and Gertman's method by exploring an alternative protocol to standardizing the UEP priority weights. Before addressing the priority weighting, a brief description of Boring and Gertman's method is provided (see [2] for a more detailed description).

II. HRA-BASED USABILITY METHOD

The HRA-U procedures heavily draw upon the SPAR-H method, which was developed to assess human error probabilities (HEPs) in nuclear power plants [3]. 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 (PSF) 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 PSF 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. By replacing the SPAR-H list of PSFs with a list of usability heuristics, an evaluator can use a method akin to SPAR-H to prioritize usability concerns.

An important aspect of the SPAR-H method is that human activity is assigned 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. Based upon operational research, base-rate or nominal human error probabilities (NHEP) for diagnosis tasks are assumed to be 0.01 (or 1E-2) while action tasks are assumed to be 0.001 (or 1E-3), excluding any adjustment for PSFs or dependencies between a chain of events.

Procedurally similar to SPAR-H, to conduct an HRA-U analysis, an evaluator would complete the following steps:

Step 1: After identifying the appropriate level of task decomposition, the evaluator (using Table 1) performs a heuristic evaluation by simply identifying the correct level of usability for each heuristic. Associated with each usability level is a multiplier that will be used later in Step 2.

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

Step 2: To tally the total UEP, the evaluator multiplies the product of the individual heuristic multipliers (that was identified in Step 1) by the diagnosis or action NHEP. A higher number suggests that the usability issue has a higher likelihood of occurrence. Note that heuristics in Table 1 may have either positive or negative effects, implying good or poor usability, respectively. Poor designs serve to increase the UEP by having multipliers valued greater than 1. On the other hand, if the design is notably positive, the usability level may serve to decrease the UEP by having multipliers valued less than 1.

Step 3: 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:

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

where *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 heuristics.

In some cases, the evaluator may find it is not possible to parse a task into a solely cognitive 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.

Finally, ***Step 4***: To prioritize the weights among heuristics (i.e., to construct a list of heuristics in magnitude of what needs to be fixed immediately), the initial version of HRA-U method asks the evaluator to determine a consequence level using the usability consequence matrix (UCM) in Table 2. Each of the three consequence levels has a corresponding consequence multiplier that is then multiplied by the overall UEP to

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

obtain a usability consequence coefficient (UCC). Based upon the UCC rubic in Table 3, evaluators are informed to whether a usability concern warrants immediate fix or not.

Table 3: Usability consequence coefficient range

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*

To summarize the steps required for the HRA-U method, 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.

As depicted in Step 4, the HRA-U method requires evaluators to *determine* the appropriate consequence level. This subjective judgment may lead to inconsistency between evaluators and thus, seriously deflate the method's reliability. The current paper proposes an alternative protocol to prioritize usability issues, by standardizing the consequence matrix for each heuristic and placing it within the UEP calculation process. This revision will shorten the HRA-U procedures and, more importantly, strengthen their reliability.

III. STANDARDIZING PRIORITY WEIGHTS

III.A. Goal

The goal of the current study is to standardize priority weighting procedures by eliminating the process of having evaluators assess the appropriate consequence level (Step 4 of the HRA-U procedure). To achieve this aim, we developed a protocol that consists of the following steps:

1. Obtain standardized rankings of heuristics.
2. Incorporate the mean rankings into the UEP calculation while adjusting the heuristic multipliers to reflect the standardized ranking.

In standardizing the rankings of heuristics, we conducted a small study in asking usability students and a human factors expert to evaluate and rank-order a list of heuristics. To adjust the heuristic multipliers, we use values derived from Boring and Gertman [2] and SPAR-H method [3].

III.B. Small Usability Heuristic Study

III.B.I. Participants

Participants were five human factors students and one human factors expert from a small college in the Northwest region of the United States. Participants were all males and completed the study as part of their human factor class project.

III.B.II. Materials and Procedures

The initial heuristics proposed in [1] have been updated by Nielsen [4]. Participants were given a list of the ten updated heuristics: *aesthetic and minimalist design; match between system and real world; consistency and standards; recognize rather than recall; error prevention; user control and freedom; flexibility and efficiency of use; help users recognize, diagnose, and recover from errors; visibility of system status;* and *help and documentation.*

Upon receiving the list, participants were instructed to rank-order the heuristics based upon the seriousness of the heuristics, starting with "1" being the most serious heuristic, "2" being the second most serious heuristic, and continuing to "10" being the least serious heuristic. Participants rank-ordered the heuristics independently and were given no time limits to complete the task.

III.B.III. Results

The ranking means (*M*) and standard deviations (*SD*) are presented in Table 4.

Table 4: Ranking means and standard deviations

Overall Ranking	Heuristics	M	SD
1	Match between system and real world	2.83	1.17
2	Consistency and standards	3.67	2.73
3	Flexibility and efficiency of use	4.17	3.37
4	Recognize rather than recall	4.83	1.14
5	User control and freedom	5.33	1.17
6	Error prevention	5.50	2.43
7	Visibility of system status	6.33	1.15
8	Help users recognize, diagnose, and recover	6.50	1.89
9	Aesthetic and minimalist design	6.50	0.96
10	Help and documentation	9.33	0.71

We rank-ordered the heuristics based upon their mean rankings—the lower the mean ranking, the higher the overall ranking. For example, *Match between system and* *real world* was overall ranked the most serious heuristic because on average, our participants rank-ordered it in the top three positions as indicated by its ranking mean of 2.83 and a small standard deviation of 1.17. *Help and documentation* was overall ranked the least serious heuristic because, on average, our participants rank-ordered it in the bottom 2 or 3 positions with a ranking mean of 9.33 and a very low standard deviation of 0.71. Thus, by examining the means and the standard deviations, one can assess, on average, the ranking of each heuristic as well as its variability in ranks.

Having obtained standardized rankings of heuristics, we next mapped our standardized ranks into priority weights for use in UEP calculations.

III.C. Mapping Standardized Heuristic Rankings to Heuristic Multipliers in Usability Error Probabilities Calculations

Recall, the purpose of this paper is to standardize heuristic priority weights, not to revise Boring and Gertman's [2] framework. Thus, our mapping protocol is constrained to Boring and Gertman's UCM table (referred to in Table 2), with heuristics updated according to [4]. As a consequence, our mapping protocol consists of the following three steps:

1. Classifying our list of standardized heuristic rankings into meaningful groups by identifying clusters in the list (i.e., heuristics that tend to be ranked similarity as high, medium, low, etc).
2. Ensuring that our identified clusters of standardized heuristics can correspond to the four levels of the UCM.
3. Adjusting the heuristic multipliers in the UEP calculations (i.e., Table 1) to reflect our standardized heuristic priority weights.

Ideally, we would make use of advanced statistical techniques (e.g., cluster analysis, factor analysis) in identifying clusters in our standardized heuristic list. However, such analyses cannot be performed at this time due to our small sample size. To compensate for this limitation, we relied on the face value of mean rankings. As can be seen in Table 4, the first three heuristic ranks tend to cluster around a mean value of 3, the middle three heuristic ranks tend to cluster around a mean value of 5, the next three heuristic ranks tend to cluster around a mean value of 6, and finally, the lowest heuristic rank is isolated by itself with a mean value of 9.33. Thus, we can argue that four clusters emerged from our list of standardized heuristic rankings, as follow:

Cluster 1: *match between system and real world, consistency and standards,* and *flexibility and efficiency of use*
Cluster 2: *recognize rather than recall, user control and freedom,* and *error prevention*
Cluster 3: *visibility of system status; help user recognize, diagnose, and recover;* and *aesthetic and minimalist design*
Cluster 4: *help and documentation*

The next step in our revised HRA-U protocol is to map the identified clusters to the four levels of UCM. Thus, we make the following categorizations: Cluster 1 as "High" usability consequence with a multiplier of 5, Cluster 2 as "Medium" usability consequence with a multiplier of 2, Cluster 3 as "Low" usability consequence with a multiplier of 1, and, finally, Cluster 4 as "None" usability consequence with a multiplier of 0.

The final step in our mapping protocol is to adjust the heuristic multipliers table to reflect our standardized heuristic priority weights. To do this, we took the consequence multiplier value and multiply it by 10 under the label "poor". We then use the SPAR-H's PSF table as a guide to adjust the values for the remaining labels (see Table 5).

Table 5: Revised heuristic evaluation matrix

Heuristic	Multipliers				
Match between system and real world	50 Poor	10 Available	1 Nominal	0.2 Good	0.1 Excellent
Consistency and standards	50 Poor	10 Available	1 Nominal	0.2 Good	0.1 Excellent
Flexibility and efficiency of use	50 Poor	10 Available	1 Nominal	0.2 Good	0.1 Excellent
Recognize rather than recall	20 Poor	5 Available	1 Nominal	0.5 Good	0.1 Excellent
User control and freedom	20 Poor	5 Available	1 Nominal	0.5 Good	0.1 Excellent
Error prevention	20 Poor	5 Available	1 Nominal	0.5 Good	0.1 Excellent
Visibility of system status	10 Poor	3 Available	1 Nominal	0.7 Good	0.1 Excellent
Help user recognize, diagnose, and recover	10 Poor	3 Available	1 Nominal	0.7 Good	0.1 Excellent
Aesthetic and minimalist design	10 Poor	3 Available	1 Nominal	0.7 Good	0.1 Excellent
Help and documentation	5 Poor	2 Available	1 Nominal	0.9 Good	0.1 Excellent

III.D. An Example in Using the Newly Revised Priority Weighted Table

To illustrate the revised method in practice as well as provide a basis for comparison between the original and revised methods, we use an example taken from Boring and Gertman [2]. 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.

III.D.I. Method A: Boring and Gertman [2]

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 based on Table 1. 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. 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. The consequence of this combination of heuristics is determined to be "medium" by the evaluator, 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 3, this UCC value maps to a *high priority* usability item that requires a fix.

III.D.I. Method B: Standardized Priority Weights

The procedures for this method are exactly the same as Boring and Gertman [2] with an exception of using the weighted multipliers in Table 5 and the exclusion of evaluator's consequence level judgment. Thus, the *aesthetic and minimalist design* (a.k.a., dialog) heuristic would be marked as "poor" and receive a corresponding multiplier of 10. Similarly, the *user control and freedom* (a.k.a., clear exit) heuristic would be judged as "poor" with the corresponding multiplier of 20. Finally, the *flexibility and efficiency of use* (a.k.a. shortcuts) heuristic would be judged as "excellent" with the corresponding multiplier of 0.1. Again, 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 20 x 0.1, yields a value of 20. This value is in turn multiplied by the diagnosis NHEP of 0.01 to produce a composite UEP equal to 0.2. Comparing the composite UEP of 0.2 to Table 3, this UEP maps to a *high priority* usability issue that needs to be fixed.

Accordingly, in this example, the standardized priority weights method produces the identical outcome as Boring and Gertman's original method [2]. More importantly, the standardized priority weights method eliminates the subjective judgment of assigning priority weights by different evaluators (i.e., evaluator A may choose "medium" level while evaluator B may choose "high" level). As a result, the revised HRA-U approach increases the method's reliability.

IV. DISCUSSION/CONCLUSION

This paper was successful in demonstrating an approach to standardize the priority weighting of the UEP based upon rank-order data. As evident in the above scenario, using the newly developed heuristic priority weights embedded in the UEP calculation, we obtained similar results of that of Boring and Gertman [2] without reliance on evaluator's subjective consequence judgment.

Despite this success, there are several limitations to our paper that should to be addressed with further research. First, the study's results are based upon a small sample size of six participants. Second, because of the small sample size, the study utilizes a face value technique of mean ratings instead of statistical techniques in classifying and mapping heuristics onto the four levels of the UCM. Nevertheless, we should emphasize that the purpose of this paper was to explore a protocol procedure to standardizing heuristic priority weights. Given that this approach proved successful, we now venture a further study to validate this protocol procedure with a larger human factors population. Such a study would allow us to utilize more advanced statistical techniques (e.g., cluster analysis or factor analysis) in our heuristic classification scheme.

REFERENCES

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

[2] R. L. Boring and D. I. Gertman, "Advancing Usability Evaluation through Human Reliability Analysis," *Proceedings of HCII 2005*, (2005).

[3] D. Gertman, H. Blackman, J. Marble, J. Byers, L. Haney, and C. Smith, "*The SPAR-H human reliability analysis method, NUREG/CR-6883,"* Washington, DC: US Nuclear Regulatory Commission, (2005).

[4] J. Nielsen, "Heuristic evaluation" in J. Nielsen and R.L. Mack (eds.), "*Usability inspection methods,"* Wiley, New York City, (1994).