SANDIA REPORT
SAND2008-2619
Unlimited Release
Printed April 2008

Issues in Benchmarking Human Reliability Analysis Methods: A Literature Review

Ronald L. Boring, Stacey M.L. Hendrickson, John A. Forester, Tuan Q. Tran, Erasmia Lois

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia is a multiprogram laboratory operated by Sandia Corporation,
a Lockheed Martin Company, for the United States Department of Energy's
National Nuclear Security Administration under Contract DE-AC04-94AL85000.

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd.
Springfield, VA 22161

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.fedworld.gov
Online order: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2008-2619
Unlimited Release
Printed April 2008

Issues in Benchmarking Human Reliability Analysis Methods: A Literature Review

Ronald L. Boring and Tuan Q. Tran
Human Factors, Instrumentation and Control Systems Department
Idaho National Laboratory
P.O. Box 1625
Idaho Falls, Idaho 83415-3605

Stacey M. L. Hendrickson and John A. Forester
Risk and Reliability Analysis Department
Sandia National Laboratories
P.O. Box 5800
Albuquerque, New Mexico 87185-0748

Erasmia Lois
Probabilistic Risk Analysis Branch
US Nuclear Regulatory Commission
Mail Stop 10-E50
Washington, DC 20555-0001

ABSTRACT

There is a diversity of human reliability analysis (HRA) methods available for use in assessing human performance within probabilistic risk assessment (PRA). Due to the significant differences in the methods, including the scope, approach, and underlying models, there is a need for an empirical comparison investigating the validity and reliability of the methods. To accomplish this empirical comparison, a benchmarking study is currently underway that compares HRA methods with each other and against operator performance in simulator studies. In order to account for as many effects as possible in the construction of this benchmarking study, a literature review was conducted, reviewing past benchmarking studies in the areas of psychology and risk assessment. A number of lessons learned through these studies are presented in order to aid in the design of future HRA benchmarking endeavors.

Acknowledgments

This work was funded by the U.S. Nuclear Regulatory Commission (USNRC) and performed at Sandia National Laboratories and at Idaho National Laboratory (under contract from Sandia). The opinions expressed in this paper are those of the authors and not of USNRC or of the other organizations. We would like to acknowledge the contributions of members of the steering committee for the ongoing HRA benchmarking study being conducted at the Halden Man-Machine Laboratory (HAMMLAB) simulators at the Halden Reactor Project in Norway. The identification of various HRA benchmarking experimental design issues in this literature review benefited from discussions with these individuals and others at various times. They include: Andreas Bye, Helena Broberg, Vinh Dang, Jeff Julius, Alan Kolaczkowski, Bruce Hallbert, and Gareth Parry.

Table of Contents

1. INTRODUCTION 7
2. GENERAL BENCHMARKING EXAMPLES 8
   2.1. Introduction 8
   2.2. Example of Benchmarking in Educational Psychology 8
      2.2.1. Introduction 8
      2.2.2. Example: Benchmark of Essay Grading Methods 9
      2.2.3. Lessons Learned 10
      2.2.4. Implications for HRA Benchmarking 11
   2.3. Example of Benchmarking in Usability Evaluation 11
      2.3.1. Introduction 11
      2.3.2. Example: Method-to-Method Usability Benchmarking 11
      2.3.3 Example: Product-to-Product Competitive Usability Benchmarking 12
      2.3.4. Lessons Learned 13
      2.3.5. Implications for HRA Benchmarking 14
   2.4. Example of Benchmarking in Cognitive Modeling 14
      2.4.1. Introduction 14
      2.4.2. Example: Validation of a Specific Cognitive Model to Human Performance 15
      2.4.3. Example: Comparative Validation of Cognitive Models to Human Performance 15
      2.4.4. Lessons Learned 17
      2.4.5. Implications for HRA Benchmarking 17
   2.5. Example of Benchmarking in Risk Assessment of Chemical Plants 18
      2.5.1. Introduction 18
      2.5.2. Example: The ASSURANCE project: Assessment of Uncertainties in Risk Analysis of Chemical Establishments 18
      2.5.3. Example: Uncertainties in Chemical Risk Assessment: Results of a European Benchmark Exercise 19
      2.5.4. Example: BEQUAR (Benchmark Exercise in Quantitative Area Risk Assessment) in Central and Eastern European Countries 20
      2.5.5. Lessons Learned 21
      2.5.6. Implications for HRA Benchmarking 21
3. BENCHMARKING STUDIES IN HUMAN RELIABILITY ANALYSIS 22
   3.1. Introduction 22
   3.2. Qualitative Comparisons 22
   3.3. European Benchmark Exercise in Human Reliability Analysis 23
      3.3.1. Introduction 23
      3.3.2. Lessons Learned 25
   3.4. Zimolong's Empirical Benchmark 27
      3.4.1. Introduction 27
      3.4.2. Lessons Learned 28
   3.5. Kirwan's Quantitative Benchmarking Exercise 29
      3.5.1. Introduction 29
      3.5.2. Lessons Learned 29

3.6. Maguire's Validation Benchmark 30
3.6.1. Introduction 30
3.6.2. Lessons Learned 31
4. CONCLUSION 32
5. REFERENCES 35

1. INTRODUCTION

A diversity of human reliability analysis (HRA) methods is currently available to treat human performance in probabilistic risk assessment (PRA) and other applications. Given the significant differences in the scope, approach, and underlying models of these methods, there has been a growing interest on the part of HRA method developers and users (including PRA analysts that typically rely on HRA results) to empirically validate and test the consistency and reliability of the various methods. To this end, there is an ongoing international effort to begin this benchmarking process by testing the application of HRA methods to nuclear power plant operating crew performance in the Halden Man-Machine Laboratory (HAMMLAB) simulator at the Halden Reactor Project in Norway. Another goal of this effort is to develop an empirically based understanding of the performance, strength and weaknesses of the methods, in order to provide the technical basis for the development of improved HRA guidance, and if necessary, improved HRA methods.

The purpose of this report is to review literature relevant to performing benchmarking of HRA methods, with the goal of identifying experimental design and other issues that need to be addressed in order to adequately benchmark HRA methods. A benchmark in conventional language usage refers to a standard to which something can be compared. Benchmarking in the present context refers to comparing HRA methods both to each other and to actual operating crew performance in a nuclear power plant simulator. Thus, HRA benchmarking in this study involves both the validation of method predictions against actual performance (i.e., do the methods make accurate predictions?) and against each other in terms of their ability to accurately predict and explain the basis for the predictions (i.e., what factors influence the expected performance?). In addition, benchmarking should strive to evaluate methods in terms of their ability to produce consistent results. In other words, to be valid, a method must also be able to produce consistently accurate results across the range of situations to which it is assumed to be applicable and across competent HRA teams.

In order to facilitate an effective comparison among HRA methods and between HRA methods and human performance, it is necessary to define the measures upon which HRA methods will be compared and to provide adequate experimental controls. This report provides an overview of general benchmarking approaches, a review of examples of benchmarking as found in the psychological and risk assessment literatures (see Section 2), a review of previous HRA benchmarking studies (see Section 3), and a summary of lessons learned from the previous research relevant to appropriate experimental design (see Section 4). These results will serve as the basis for designing studies to validate HRA methods against actual human performance and to empirically and definitively contrast the findings of different HRA methods with each other.

2. GENERAL BENCHMARKING EXAMPLES

2.1. Introduction

The purpose of this section is to provide general examples of benchmarking and identify lessons learned from the studies that could be relevant to the methods used to benchmark HRA methods. These examples are meant to be illustrative, not to provide an exhaustive list of all types of benchmarking that are found in the research literature. We discuss four types of benchmarking. The first three types are derived from psychological research in education, usability evaluation, and cognitive modeling, where benchmarking is a well-established research undertaking. To demonstrate the methodological overlap between psychological and other benchmarks, the final type of benchmarking is taken from risk assessment. At the conclusion of each example of benchmarking, we provide a brief discussion of the lessons learned and best practices that can be gleaned from these benchmarking studies that could be relevant to the planned HRA benchmarking effort.

2.2. Example of Benchmarking in Educational Psychology

2.2.1. Introduction

The field of educational psychology is engaged in the development of tools, methods, and policies for improving learning. One hallmark of the field is the need to determine the efficacy of particular educational approaches as well as the need to evaluate novel methods that purport to improve on existing approaches. Such validation of existing approaches and comparison between different approaches are classic examples of benchmarking. Without the evaluative component of benchmarking, educational psychology could still be engaged in improving tools, methods, and policies, but there would be no way to verify the quality of those improvements.

One of the greatest challenges to benchmarking in educational psychology is the determination of the evaluative measure. The diversity of learners, instructors, learning environments, and instructional tools necessitates the use of a broad range of measures. Consider, for example, the diversity of learners. Learners vary in age, background, aptitude, ability, and learning style, among other factors. An educational benchmark must either carefully control for the type of learner participating in a study or establish a nuanced range of measures that fully reflect the learners' level of success with a particular approach. Each approach represents a trade-off. A study focused on a narrowly defined type of learner avoids confounds and has a greater ability to demonstrate clear results for that learner population. On the other hand, the narrowly defined sample of learners may not generalize to a broader population of learners. As a consequence, the narrowly defined learner population may omit facets of learner performance that have significant implications for the efficacy of an educational approach.

2.2.2. Example: Benchmark of Essay Grading Methods

To illustrate issues related to benchmarking in educational psychology, consider work to compare the efficacy of two types of essay grading methods [1]. Holistic essay grading considers the essay according to the grader's overall impression of the essay. Analytic essay grading considers individual elements of the quality of the essay that combine to make an overall grade. Whereas typically holistic grading provides an overall grade followed by explanatory comments related to the quality, analytic grading provides a rubric of grading factors that are individually tallied.

Assuming the purpose of the essay is to articulate the knowledge of the essay writer on a particular topic, a quality essay is considered one that reflects a high level of topic mastery. Thus, a benchmark of these two essay grading methods should compare the degree to which the grades awarded for essays reflect the quality of knowledge mastery represented in the essays. Boring [1] identified several challenges to this type of benchmarking research, including:

- *Difficulty in establishing an objective baseline upon which the two grading methods can be compared*. This measure corresponds to the validity of the assessment—while it is relatively simple to compare the grades generated by the two methods, it is no simple matter to gauge which of those grades is the valid reflection of essay quality.
- *Control of extraneous factors known from the research literature to influence grades.* Factors that may or may not be reflective of the essay's quality are known to impact the grade awarded for an essay. These factors include the quality of spelling in the essay, the correctness of the essay, the relative quality of other essays, the particular topic chosen in open-ended essay assignments, the quality of the introduction of the essay, the pragmatic form of the essay, the essay writer's handwriting, the writer's name, the writer's race, the writer's physical attractiveness, and the writer's level of native English proficiency. In addition, factors that are in no way related to the essay or essay writer are known to influence the grade awarded for the essay. These factors include the grader's interaction style, the grader's personal prejudices, the grader's mood, the grader's training, the grader's expertise, the grader's severity, and the grader's psychological personality type.
- *Reliability of the measure and grader*. Even when a seemingly solid measure for benchmarking is selected, there are issues related to reliability that need to be carefully considered. Two or more graders using the same grading method may award different grades. This potentially low inter-rater reliability could be a byproduct of the method not adequately controlling for extraneous factors that influence grading. It could also be the result of the particular scaling used by the grading method. Graders who perceive essays in a similar manner may not translate their perception of the quality of the essay into a grade in the same manner. So, even if the grading method succeeds at eliciting a common quality perception among different graders, it may fail to constrain the process of assigning grades for that common quality perception, resulting in variability in grades between grading. A poor mapping of quality perception to grades may likewise result in inconsistency within the grades awarded for comparable essays by the same grader, causing poor intra-rater reliability.
- *Measures of efficiency*. Even if two grading measures prove equally valid and reliable, there may still be significant differences between the methods in terms of the time required to grade essays. This measure of efficiency is confounded by several

considerations. For example, if a method requires extensive comments to be written by the grader, there may be considerable variability in the amount of time different graders spend on comments.

2.2.3. Lessons Learned

It should be noted that many of the potential issues with essay grading are due to individual differences between graders. A typical way to control for such differences is to draw a large enough sample of data to capture the range of performance while neutralizing strong outliers when arriving at the expected average performance. Well established techniques in statistical power analysis [2, 3] help the designer of an experiment determine the acceptable sample size to arrive at statistically significant differences between study groups. A sample size of sufficient graders will help diffuse the effects of a single extraneous factor on grades. It will also allow the computation of inter-rater reliability across graders per method, an averaged inter-rater reliability per method, and average time to grade essays per method, all of which are particularly susceptible to outlier graders with small sample sizes.

Using an optimal number of graders and essays may not be practicable. It may not be possible, for example, to use a large number of professional graders for a series of essays due to budget constraints. Thus, most benchmarking studies in educational psychology represent a trade-off between the ideal study with an ample sample size of graders and essays coupled with careful control of all confounds, and the realistically implementable study with a limited sample size coupled with reasonable efforts to control for confounds. This trade-off means that benchmarking findings may not always be conclusive statistically speaking. A small study comparing methods may not provide definitive enough findings to change educational policy. However, it may generate enough interest in the results to warrant an eventual larger benchmarking effort. In this manner, educational benchmarking tends to follow a tiered approach. Smaller studies with strong results may serve as pilot studies for eventual larger studies that have the potential for changing educational policy.

The issue of the validity of the findings remains elusive. To the extent objective measures of the quality of an essay can be established, these measures are necessary to calibrate the performance of competing methods. The inter-rater and intra-rater reliabilities and efficiency of competing methods do not tell which method strikes closer to reality. In the case of essay grading methods, reliability and efficiency methods do not tell which method actually best maps the grade awarded by the grader to the quality of the essay writer's topical mastery. This relationship may be established through other forms of testing, e.g., determining the correlation between the grade for the essay and the grade on a multiple choice test on the same subject matter (to the extent the test can be considered a valid measure of knowledge). Other techniques, such as assessing the underlying semantic knowledge representation of the topic [4], have been explored and hold promise. In the case of Boring [1], the essay grades were validated against exam scores and the relationship between the graders' and the essay writers' structural knowledge representations. Both served as external empirical measures with which the measures under investigation could be compared.

2.2.4. Implications for HRA Benchmarking

Educational benchmarking has clear implications for HRA benchmarking. Graders may be seen as analogous to analysts, while grading methods may be seen as analogous to HRA methods. The challenge of controlling for grader variability is mirrored in any study that uses multiple human reliability analysts; striking a balance between the optimal number of graders and methods is reflected in the need to balance qualified analysts with particular methods. The ideal may not be practicable, either for educational psychology or for HRA. The challenge therefore remains how best to control variability while constraining the study design to something that is realistically implementable. Finally, there is the issue of validity. Just as there is a presumed objective quality of an essay, there is the objective performance against which HRA methods need to be compared. Human performance may, however, not prove as elusive to categorize as the quality of an essay—when appropriately designed, it is possible for a study to measure an HRA method's estimate against empirically observed human performance, thereby validating that estimate.

2.3. Example of Benchmarking in Usability Evaluation

2.3.1. Introduction

The field of usability is the subfield of human-machine interaction concerned with designing systems that are easy, enjoyable, and safe to use. Usability consists of the systematic interplay of design and evaluation. A cornerstone of usability is the user-centered design cycle, in which interface designs are tested on representative users to identify usability issues, which are corrected in the next design iteration.

Usability evaluation avails itself of two types of benchmarking: method-to-method comparison and product-to-product comparison. Method-to-method comparison is similar to the techniques employed in educational psychology—the efficacy of usability evaluation methods is compared. A product-to-product comparison is also known as a competitive usability benchmark. The goal of such evaluation is generally to determine the market viability in terms of usability of one product vs. another product.

2.3.2. Example: Method-to-Method Usability Benchmarking

The field of usability evaluation has emerged with a number of different techniques to help the usability engineer to test products. Many of these techniques have emerged from practice, and there is the desire within the usability community to compare the results of new techniques with the results of established industry standard techniques. Thus, there are ongoing method-to-method benchmarks within usability, usually comparing a novel technique with an established technique.

Consider, for example, research by Faulkner [5] on usability heuristics. Usability heuristics consist of a list of factors thought to influence the usability of a product. In a heuristic evaluation, an experienced usability engineer or a group of experienced usability engineers

reviews the quality of a particular product interface according to the heuristics. The goal of the heuristic evaluation is to identify key issues that would hamper the usability of the product. Typically, such an evaluation is done as a formative evaluation—an evaluation conducted on an early or prototype product, in which feedback from the evaluation is used to improve the design of the product prior to product release. Heuristic evaluation may also be summative—occurring after product completion and used to refine the design of the product interface. A problem with heuristic evaluation is seen in the reliability of results. Different usability engineers tend to identify different usability issues when given the same interface to review [6]. Faulkner's Usability Professional's Evaluation Checklist attempts to provide a structured, systematic approach to heuristic evaluation in order to increase the inter-rater reliability of heuristic evaluation. Faulkner compared usability engineers with traditional unstructured heuristics (in which the usability heuristics were provided but no explicit procedure was provided on systematically conducting the analysis) vs. the updated approach (in which usability heuristics were paired with a procedure). This method-to-method comparison afforded a direct comparison between an established technique and a novel technique. By demonstrating that the novel technique improved on the existing technique (i.e., usability engineers were more consistent with the novel technique, resulting in significantly improved inter-rater reliability), the novel technique is able to establish itself as a viable alternative method.

2.3.3 Example: Product-to-Product Competitive Usability Benchmarking

As an example of product-to-product competitive usability benchmarking consider a comparison of three travel Web sites—Expedia.com, Travelocity.com, and orbitz.com [7]. The benchmark is summative, because all three Web sites are finished, commercially available products. These three travel Web sites lead the online travel market in terms of full-service travel options including plane tickets, hotel bookings, and car rental reservations. The travel sites essentially provide a suite of search tools that allow online users to retrieve travel information from multiple vendors, thereby effecting comparison shopping with respect to features, cost, or routes. Because of the complexity of searching multiple sources of information, each of these travel Web sites takes a slightly different approach to retrieving the relevant travel information. The approach may have a significant impact on the ultimate search returned to the user.

Consider, for example, a search for a flight between two cities. Each travel Web site must formulate thousands and sometimes millions of flight segment combinations to arrive at reasonable and cost effective flights. Because this process is time intensive for the computers the travel Web site uses, the travel Web site may adopt strategies (e.g., only searching particular airlines or only searching pre-determined flight segment combinations). This approach may result in a faster search time or simpler flight combinations, but it may neglect flight combinations that are significantly less expensive. Such a trade-off has negligible adverse effect in terms of user perception unless a competing Web site consistently returns less expensive flight combinations. A Web site that consistently returns better deals on flights may take consistently longer to produce search results. However, as Meyer and colleagues [8, 9] point out, perception of search times may be influenced by aesthetic considerations in the presentation of interstitial displays—the screens typically displayed while the user is waiting for the host system to complete its search. A travel Web site that features a more interactive or engaging interstitial

display may compensate for the slow search retrieval time by shortening the perceived interstitial display time. Thus, a travel Web site may take measures to mitigate apparent trade-offs.

Usability of a product is not unidimensional. Chaparro and Gibson [7], for example, considered three overall measures of usability when comparing the usability of the three travel Web sites—user satisfaction, navigational efficiency, and general preference. User satisfaction utilizes a series of questions that together produce a subjective scale to reflect an overall negative, neutral, or positive impression of each Web site. Navigational efficiency is measured in terms of the ratio of the number of pages to complete a task to the actual number of pages visited during a particular task. General preference simply uses a rank ordering of the three travel Web sites. In all cases, it is possible to evaluate the three measures of usability across different representative travel tasks, thus producing a usability comparison of the three Web sites for particular tasks as well as a composite or overall impression.

2.3.4. Lessons Learned

In comparing method-to-method and product-to-product usability benchmarking, a number of trends are revealed, which are informative in terms of lessons learned:

- *Emphasis on reliability and efficiency measures.* For method-to-method usability comparisons, similar considerations prevail as in educational psychology. To date, there has been a strong emphasis on the reliability of methods (as seen in the Faulkner study [5]) and on efficiency (e.g., earlier research [10] comparing heuristic evaluation to traditional laboratory usability studies involving sample users). In contrast to educational psychology, method-to-method usability comparisons have minimally addressed the validity of findings (for an exception involving validating the usability of mobile phone designs, see [11]). It is expected that as appropriate measures of usability validity emerge, method-to-method usability benchmarking will follow the course of educational psychology and review the full spectrum of benchmarking measures, including reliability, efficiency, and validity.
- *Formative vs. summative evaluation.* Usability evaluation draws a distinction between evaluations conducted early vs. late in the product development cycle. Early (formative) evaluations tend to be qualitative and diagnostic in nature, designed to identify significant usability problems and fix them in the design. Later (summative) evaluations tend to focus on quantitative measurements to arrive at an index of product performance. Usability benchmarking is almost always summative.
- *Use of experimental test participants.* As in other disciplines of psychology, usability evaluation tends to draw a sample of participants (e.g., prospective users) and average the results of those participants. Sample test participants are carefully screened to meet the demographics and characteristics expected of the product end users. With the exception of the experts utilized in heuristic evaluations, usability studies have usability test participants conduct and complete a series of representative tasks or scenarios. Unlike many other disciplines of psychology, usability evaluation tends to use a relatively small number of participants per study. The goal of most usability studies is not to develop a test design capable of producing significant inferential statistics. Rather, the goal is to identify problems with the product interface and to baseline performance. The

descriptive nature of these data warrants a small sample size. Even in comparative benchmarking studies, the number of test participants tends to remain small, with an emphasis on qualitative, descriptive comparisons vs. quantitative, inferential comparisons.

- *Emphasis on psychometric measures to augment actual performance measures.* Usability is construed as a combination of actual performance and perceived performance. User perceptions may not always prove to be accurate or valid reflections of product performance. It is therefore necessary to collect objective and subjective measures and, where possible, to explain disparities between those measures.

2.3.5. Implications for HRA Benchmarking

As with educational psychology, there is a clear analogue between usability and HRA benchmarking. Method-to-method comparisons are sensible in both usability and HRA benchmarking. Perhaps less obvious but equally powerful is the applicability of product-to-product comparisons for HRA. The "product" in HRA can be seen as the task or scenario that is analyzed. It is important, thus, to focus not only on an overall comparison of different HRA methods but also to consider the spectrum of applications for which those HRA methods are designed. Different HRA methods may have different "product" areas—tasks, scenarios, or even types of analyses for which they are optimized. A benchmark of HRA methods should therefore consider a range of tasks and scenarios in order to have a complete representation of the capabilities of different HRA methods.

2.4. Example of Benchmarking in Cognitive Modeling

2.4.1. Introduction

Cognitive or human performance modeling is a field focused on developing simulations that mimic human decision making and behavior. Cognitive modeling differs from traditional artificial intelligence (AI) work in that AI has focused on creating software programs that match or exceed human performance, often without regard to replicating the processes used by humans to make decisions or act. Cognitive modeling, in contrast, typically utilizes documented human cognitive processes (e.g., goal setting or focus of attention) as the basis for its simulated performance. The performance of the cognitive model on a particular task is closely baselined or validated to empirically observed human performance to determine the match of the cognitive model to actual performance. As expected, much of the emphasis in cognitive benchmarking is on validating the model's performance against human performance. This validation may take the form of comparing the results of a particular cognitive model to human performance. Alternatively, when there is a comparative or competitive benchmark, different or competing cognitive models are validated against each other and human behavior. These two types of benchmarking are elucidated in the following examples.

2.4.2. Example: Validation of a Specific Cognitive Model to Human Performance

Best, Lebiere, Schunk, Johnson, and Archer [12] compared an Adaptive Control of Thought-Rational (ACT-R) model of flight approach during landing to the performance of human pilots. For this simulation, a procedural task model based on human operators was embedded in an ACT-R goal and decision making simulation model. The ACT-R model served as a virtual pilot with perceptual inputs, cognitive processing, and psychomotor outputs, while an external flight simulation model provided plane and environmental states. For comparison purposes, human performance of actual pilots was obtained in a flight simulator.

The authors [12] discuss three levels of validation, depending on the level of validation granularity sought. Coarser levels of granularity correspond to qualitative validation, while finer levels correspond to quantitative validation. The first level, *successful task completion*, reflects validation at the level that if a human completes a task successfully, the cognitive model will likewise complete the task successfully. Conversely, a human's failure to complete a task should be reflected by the model's failure to complete the same task. This level of validation is centered on the overall task performance, not performance at the subtask level. Subtask performance is captured at the second level of validation—*subtask correspondence*. While the first level may reflect overall performance similarity between the model and the human, it fails to ensure that the steps taken to achieve the task map between the model and the human. This second level provides the granularity suitable for determining that the model is making decisions and acting in a manner compatible with human decisions and actions. The third level of validation represents the finest granularity. This level attempts to capture not only a mapping between the model and human in terms of similar decisions and actions; it also captures the *quantitative performance* of the model and human. If, for example, a human consistently performs an action in a given time, validation at this level would suggest that the cognitive model would perform the action in a similar relative time. Note that not all cognitive modeling architectures perform actions in real time; it is therefore necessary to use relative time to complete a task when examining cognitive models.

Validation inclusive of all three levels is rare. In practice, validation benchmarks are planned in stages, with early success marked by successful task completion and latter success of mature cognitive models marked by close approximation of the model to human performance. The so-called Turing test [13], in which an AI or cognitive model is indistinguishable from a human, remains elusive. Validation criteria are therefore typically set to focus on qualitative levels. After a successful map is found at the task and subtask level, performance metrics are introduced. It is at this point that the close approximation of human cognition adopted by cognitive models translates better into quantitative validation than do AI models.

2.4.3. Example: Comparative Validation of Cognitive Models to Human Performance

Comparative validation of cognitive models builds on the individual model validation approach but must accommodate a reasonable range of features included in each model. Thus, while particular models may excel in a particular task domain, there may be other domains in which they perform poorly. Given the complexity of human cognition, no single cognitive

modeling architecture claims universal coverage of all human cognitive functions. A comparative validation must therefore be designed to consider those facets of human cognition that are modeled. The comparison must strike a balance between a realistic breadth of cognitive functions that would be encountered in real human performance and those cognitive functions that are reasonably addressed by particular cognitive modeling architectures. A comparative benchmark may choose to focus on those functions covered by all cognitive models under consideration, or it may focus on functions that only select cognitive models can cover. When a cognitive model does not encompass a particular human cognitive function, model performance is expected to diverge considerably from actual human performance. In some cases, the benchmark may focus as much on the completeness of model coverage as on the detailed validation of particular cognitive functions or tasks.

A representative example of a comparative validation between cognitive models is found in Leiden and Best [14]. The authors compared five cognitive modeling architectures across a variety of aviation tasks, including the landing approach task described in Best et al. [12]. The cognitive modeling architectures under review included two variants of ACT-R, the Air Man-machine Integration Design and Analysis System (Air MIDAS), a Bayesian Belief Network system entitled the Distributed Operator Model Architecture (D-OMAR), and the Attention-Situation Awareness (A-SA) model. The benchmark considered ten sets of modeling capabilities, including modeling of: external environment, memory, scheduling and multi-tasking, crew interactions, visual attention, workload, situation awareness, error prediction, learning, and emergent behaviors. As expected, not every cognitive modeling architecture successfully addressed each of these cognitive functions. For example, only Air MIDAS and D-OMAR included explicit modeling of crew interactions. ACT-R and A-SA were designed to model individual cognitive performance and were only able to accomplish crew modeling using coding workarounds. In both ACT-R and A-SA, co-pilot behavior was pre-scripted rather than dynamically generated by the cognitive model.

Leiden and Best [14] distinguish two levels of validation that are useful for comparative benchmarking of different cognitive models. The first level corresponds to the verification phase, which is akin to the debugging phase in software development. Verification of a model entails a process by which model functions are tested to ensure they perform as desired. For example, if a cognitive model must communicate with an external flight simulation system, verification would ensure the proper transmittal of information between the cognitive model and the flight simulation system. Verification ensures elimination of spurious model performance due to faulty configuration of the model. In essence, verification ensures that all models are tested on equal terms and are calibrated properly to the performance task and environment to which they will be benchmarked. The second level of benchmarking involves validation. Leiden and Best [14] refer to this phase as predictive validation, which is the "ability of the model to provide sound predictions within certain bounds" (p. 45). The authors acknowledge the subjectivity of the terms "sound predictions" and "certain bounds." The point remains that the team conducting the benchmark must establish criteria for success or failure when comparing model and human performance. Validation, as noted earlier, can occur at the task, subtask, or performance level—the appropriate level being closely linked to the particular goals of the benchmark study.

2.4.4. Lessons Learned

There are two important lessons that can be learned from the validation benchmarking illustrated in the cognitive modeling examples:

- *Importance of considering different levels of validation.* The granularity of validation varies from a coarse-grained qualitative level of task and subtask mapping to the fine-grained quantitative level of performance matches between the cognitive model and humans. It is not appropriate or necessary for all validation to occur at a quantitative level. Insight into cognitive processes is readily demonstrated when the cognitive model completes tasks in a similar manner to humans. For many tasks, validation at the task and subtask level is a true litmus test for successful cognitive modeling of a particular facet of human cognition.
- *Importance of comparing equivalent cognitive functions.* It is crucial that comparative benchmarking studies take adequate steps at verification of the experimental apparatus. A model-to-model (or, by analog, a method-to-method) comparison must ensure that the model can function in the domain to be investigated. A failure of the model to utilize required information or communicate with other simulation systems to which it is tied can serve as a significant limitation on the performance of the cognitive model. Further, once verified, the models should be compared in the functional domains for which they were designed. Cognitive models have emerged for specific purposes and capture vastly different aspects of human cognition. A comparison of cognitive functions should attempt to control for these differences and ensure that comparisons are only made in truly overlapping functions.

2.4.5. Implications for HRA Benchmarking

The lessons learned from benchmarking in cognitive modeling also generalize to HRA benchmarking. Cognitive modeling typically uses a rich data collection of both qualitative and quantitative information. An HRA benchmark should likewise carefully consider what data are available to it and how to collect rich data. Cognitive modeling especially demonstrates the value of capturing qualitative information. A comparison of HRA methods needs not focus solely on the end result—a comparison of human error probabilities generated by different HRA methods. Instead, it should also review the processes that led to that end state. Qualitative measures—including the performance shaping factors considered in the HRA models, assumptions made, and task decomposition—may prove equally informative to a benchmark comparison as purely quantitative measures. Consideration of different levels of validation, both quantitative and qualitative, ensures the completeness of the comparison.

2.5. Example of Benchmarking in Risk Assessment of Chemical Plants

2.5.1. Introduction

Although the current review focuses specifically on the evaluation *of human* reliability analysis methods, there is a history of benchmarking studies done in the broader context of evaluating hardware risk assessment procedures. The chemical industry employs quantitative risk assessment (QRA), a process similar to probabilistic risk assessment (PRA), to evaluate the safety of chemical storage and processing plants. QRA typically progresses through the four steps of identifying hazards, assessing dose-response, evaluating exposure, and characterizing risks. The variety of methods possible in completing each of the steps within the QRA allows the researcher flexibility in fitting numerous situations.

Although the purpose of the QRA is to model the uncertainties associated with a major accident within the plant, the estimates themselves are prone to high variability. Modeling the systems of the chemical plant is made all the more complex by the great number of components and control equipment to be included and the interactions between these items. Uncertainty is increased by the use of imperfect knowledge and expert judgment in the identification of hazards and in estimating the dose-response relationships. Finally, uncertainties are introduced by the modeling tools based on the assumptions made regarding the impact by the weather and environmental conditions (e.g., release and dispersion phenomena). One of the purposes of benchmarking for QRA studies is to outline the methods available for each of the steps and propose more reliable methods. Furthermore, a goal is to standardize the presentation of results so that comparisons can be more easily made between QRA studies.

2.5.2. Example: The ASSURANCE project: Assessment of Uncertainties in Risk Analysis of Chemical Establishments

The Joint Research Centre (JRC) of the European Commission along with Risø National Laboratory conducted a benchmarking project to better understand the uncertainties that arise in QRA [15]. Seven teams from countries within the European Union participated in the exercise involving the risk assessment of an ammonia storage plant. The study progressed in five stages of which the first was a documentation phase that included a plant tour and time for questioning of plant personnel. Efforts were made by the coordinators to ensure that all teams were provided with the same information resulting from the tour and questioning period. Following this first stage, there were three working phases involving the qualitative analysis by the teams (e.g., hazard identification and qualitative risk rankings), the quantitative analysis by the teams (i.e., perform the QRA), and the use of case studies to attempt to distinguish uncertainties due to methods and those due to the current particular exercise. The final phase was evaluation and dissemination of the results.

The participants within the study engaged in one of three different forms for identifying hazards in the qualitative analysis phase. The three approaches were: 1) top-down analysis, 2) bottom-up analysis, and 3) use of checklists or national standards. The top-down approach is

most similar to fault trees and progresses by first characterizing some top event and then working down to combinations of basic events capable of provoking accidents. The bottom-up analysis (e.g., Hazard and Operability [HAZOP], Structured What-If Technique [SWIFT], and Hazardous Scenario Analysis [HAZSCAN]) investigates basic elements first to determine if a major accident could be caused by the failure of one of these individual parts or through deviations within the process variables. Finally, the use of checklists or national standards was employed to perform the systematic identification of possible release events within the plant. The use of the three methods produced different estimates for hazards. Teams also raised inconsistencies within their identification of hazards based on differing definitions of key terms such as "catastrophic event" or "likely event". Finally, the participants differed both in the scenarios selected for the hazard identification and in the ranking of the hazardous scenarios based on the use of probabilistic or deterministic approaches. However, the participants did all agree on the most important hazardous events.

The coordinators of the study were particularly interested in three types of uncertainty (parameter, modeling, and completeness) evident in the estimates proposed by the study participants. Parameter uncertainty deals with the variability resulting from the usage of different numerical input data for quantifying the same events and characteristics. Modeling uncertainty involves questioning the appropriateness of the model. This type of uncertainty results in estimates produced from the use of different models. Finally, completeness uncertainty is associated with how well the model of choice addresses the phenomena of the system being studied. A large portion of completeness uncertainty is due to differences in the number of basic, intermediate, or initiating events included in the model. By understanding these types of uncertainty, and perhaps controlling for them, an effort can be made to distinguish between the variability in answers due to differences in input data versus the models being used.

2.5.3. Example: Uncertainties in Chemical Risk Assessment: Results of a European Benchmark Exercise

In an earlier but similar effort, the Joint Research Centre (JRC) of the European Commission recruited 11 teams to perform a chemical risk analysis on an ammonia storage plant [16, 17]. The goal of the exercise was to investigate differences in methods for performing a chemical risk analysis and better understand and estimate the uncertainties associated with the results. The study was conducted in two phases. In the first stage, each of the 11 teams was instructed to perform a full risk analysis including hazard identification and the calculation of the risk estimates. At the end of this phase, the results were compared between the teams. A follow-up phase was then conducted in which the coordinators attempted to discern the sources of the variability evident in the first phase results. In this second phase, the teams were presented with a set of partial exercises and asked to conduct a risk analysis of each exercise.

Two methods were employed by the teams for conducting the risk analysis. These methods differed in the selection of hazards and failure events identified to be included in the quantification of risks. Most of the teams identified hazards through system analysis techniques (e.g. HAZOP, Failure Mode and Effects Analysis [FMEA], or Master Logic Diagram) followed by the selection of relevant release scenarios. This method involves a systematic review of the full description of the process and determining where in that process failures, either through errors or design defects, may occur. For those failures considered significant enough, the team

analyzes their consequences. The second method, used by four of the teams, performs hazard identification through a review of every plausible break of components and pipes within the plant that might lead to a release of flammable or toxic material. This list of failures may be further supplemented with the use of an event tree technique applied to investigate responses to these initial failures. The failure frequency is calculated for each of the significant contributors by using historical data and engineering judgment for the failure events.

The teams produced quite different end results as a consequence of using these two different methods for identifying failure cases. Even a within group comparison of the teams applying the same method showed team-to-team variability due to the wide amount of flexibility offered in using each of the methods. Many of the differences found between the teams' results were due to differences in input data. For instance, failure frequencies calculated through the use of fault tree analysis differed considerably from the frequencies calculated by teams using historical data banks. Differences were also evident between the teams' responses due to differing assumptions under which the models were performed.

2.5.4. Example: BEQUAR (Benchmark Exercise in Quantitative Area Risk Assessment) in Central and Eastern European Countries

A benchmarking exercise was undertaken by the Joint Research Centre (JRC) of the European Commission to gain a better understanding of the current risk analysis practices and methods in use by the European Union (EU) Candidate Countries [18]. Unlike the previous benchmarking studies [15, 16] in which participants independently conducted a risk analysis, this study requested participants to review an already existing risk analysis. The main intent of the study was to investigate the differences in findings and conclusions reached by independent evaluations conducted by individual experts of the same risk analysis and to determine how these findings might impact the calculation of risk estimates. The initial risk analysis was completed on a lower tier Seveso II chemical plant. Experts from nine EU countries actively participated in the study.

Participants in the study were provided with data on the surrounding territory of the plant (e.g., topographic map, population density and vulnerability data), on the establishment of the plant (e.g., layout diagrams, and process and instrumentation diagrams), and on the risk analysis (e.g., hazard identification analysis, and frequencies of and damage profiles for each accident scenario). The participants were instructed to use this information and focus their review on the following three items: 1) an analysis of the postulated accident scenarios, 2) frequencies of accident scenarios, and 3) accident consequences. These variables then became input for the ARIPAR 4.0 software, a QRA tool used to determine local, individual, and societal risks as a consequence of major accidents in industrial areas where hazardous substances are stored, processed, and transported.

The JRC team reviewed the analyses presented by the benchmark members and compared the different approaches pursued by the benchmark members for assessing the risk analysis. Based on this review, they hoped to glean insight into how some parameters (e.g., specific scenarios, frequencies, and consequences) can affect the risk assessment or sensitivity analysis. Furthermore, they hoped to be able to identify areas of improvement and make recommendations to new EU Member States in methods for conducting risk analyses.

2.5.5. Lessons Learned

The original goal of these benchmarking efforts on the use of QRA methods was to gain a better understanding of the sources of variability in the risk estimates. They offer many lessons learned on what these sources of variability are:

- *Variability in definitions and use of terms.* The teams participating within the QRA benchmarking studies did not share the same definitions of key terms (e.g., catastrophic event, likely event) in identifying hazards. Due to the lack of standardization, the comparison of the teams' results and the inter-rater reliability is unclear. The differences between the teams could be due to method differences or to terminological differences. This variability calls for a standardization in the definitions of terms commonly used within a QRA context.
- *Variability in historical data and expert judgment.* When the participating teams were performing their respective QRAs, there were uncertainties evident in the assessment of frequencies and consequences. The major causes behind the variability in frequency assessment were that the teams considered different failure causes and used different generic data sets for the assessment of the failure frequencies. Furthermore, differences were evident based on the judgments offered by the subject matter experts. These differences can again interfere with the inter-rater reliability between teams. Therefore, it is important for teams in the process of completing risk assessments to carefully document the source of their information.
- *Variability in presentation of results.* In evaluating results and comparing computed risk estimates across teams, the coordinators of the studies remarked on the great difficulty in comparing the teams' estimates due to the divergent presentation of results. These differences present the need for a unified or standardized method and format for presenting results.
- *Lack of validation.* Each of the benchmarking studies was undertaken with the goal of comparing results to each other to determine how the estimates differed. However, no effort was made to compare the results to real data to determine which method came closer to the true state of affairs. Therefore, the studies lacked validation and where unable to attest to a method's ability to accurately measure or predict true events.

2.5.6. Implications for HRA Benchmarking

The lessons learned through benchmarking QRA methods have direct applicability to the benchmarking of HRA methods. The QRA methods struggle with controlling and explaining variability in the same manner that any benchmarking endeavor of HRA methods will. QRA and HRA studies should work to control the variability between the applications of the methods by attempting to control extraneous variables within the study. For instance, a first step is to ensure approximately equal expertise between the teams in applying methods. Second, an effort should be made to ensure multiple teams are using the same method to check for intra-method reliability. To assist in checking intra- and inter-method reliability, teams should be directed to evaluate specific failure events, or in the case of HRA, human failure events (HFEs). Finally, in order to evaluate the method's accuracy, the method's results should be compared to real data.

3. BENCHMARKING STUDIES IN HUMAN RELIABILITY ANALYSIS

3.1. Introduction

The previous section reviewed benchmarking in a broad sense as it has been applied in psychology and risk assessment. Additionally, a number of studies have specifically looked at various aspects of benchmarking for HRA methods. Of primary concern to this review are those benchmarking efforts that have empirically compared the results of human reliability analyses performed by different analysts using different HRA methods. These benchmarks, similar to the method-to-method benchmarks in psychology, allow direct comparison of modeling and quantification as well as, in some cases, reliability measures. Variations on this model-to-model benchmarking have also enlisted validation techniques, whereby the estimated human error probabilities (HEPs) have been compared to actual human performance data. Before looking at these quantitative benchmarks in detail, we briefly review previous qualitative comparisons in HRA—those attempts to compare HRA methods on the basis of subjective criteria.

3.2. Qualitative Comparisons

Some benchmarking efforts, briefly noted here, have consisted of a qualitative comparison among methods along certain criteria deemed important. These reviews, although inherently subjective, have provided considerable insight into the strengths and weaknesses of various HRA methods. For example, Swain [19] conducted a study, funded by the Gesellschaft für Reaktorsicherheit (German Commission for Reactor Safety), with the goal of using expert opinion to evaluate 14 HRA methods for their effectiveness in the context of PRA. The study utilized a then recently developed questionnaire, "Evaluation Criteria for Selecting HRA Methods" (HRACRIT), as the major performance measure. The HRACRIT consists of three broad categories of criteria—*usefulness*, *acceptability*, and *practicality*—that a good HRA method should possess. Furthermore, within each of the three broad criteria, there is a list of detailed sub-criteria that specify the scope of each overarching criterion. For example, under *usefulness* are detailed sub-criteria such as validity, consistency, and quantitative and qualitative performance. While the criteria for evaluating HRA methods are useful and the report is interesting from a historical point of view, the results from this evaluation are not directly related to benchmarking against actual empirical data.

In another qualitative review, Kirwan [20-21] provided a list of eight factors to consider in making a qualitative assessment of the most appropriate HRA method to implement for a particular use. In addition to proving a useful cross-sectional comparison of HRA methods, Kirwan's eight qualitative HRA selection criteria (comprehensiveness, accuracy, consistency, theoretical validity, usefulness, resource usage, auditablity, and acceptability) serve as a comprehensive list of benchmarking criteria for qualitative comparisons and validation. In

addition, several of the criteria, when metricized, provide a solid basis for quantitative benchmarking.

Finally, several HRA methods used in regulatory applications are reviewed by Forester et al. in NUREG-1842 [22] and compared against HRA good practices. *Good practices* are defined by NUREG-1792 [23] as "those processes and individual analytical tasks and judgments that would be expected in an HRA in order for the HRA results to sufficiently represent the anticipated operator performance as a basis for risk-informed decisions" (p. 1-2). In other words, good practices are a form of HRA method verification, ensuring that the method specifications and processes are conducted as intended by the developer. The comparative study examined the strengths and limitations of each method so that practitioners may make a more informed choice when choosing the method dependent on the type of application.

3.3. European Benchmark Exercise in Human Reliability Analysis

3.3.1. Introduction

The Joint Research Centre of the European Commission conducted a quantitative benchmark exercise in HRA [24-26]. This study, referred to as the Human Factors Reliability Benchmark Exercise (HF-RBE) (also colloquially known in the HRA community as the "Ispra Study" after the town in which the Joint Research Centre is located), reviewed human performance in two nuclear power plant scenarios—following routine test and maintenance procedures and correcting an operational transient. The goal of the study was to compare procedures, modeling techniques, and quantification methods used in HRA; to obtain insights into sources of variability; to identify preferred HRA approaches; and to identify limitations in terms of the then current state of the art in HRA.

The scenarios were centered around human activities related to the steam generator feed systems of a German Pressurized Water Reactor plant, consisting of a two train start-up and shut-down system and a four train emergency feedwater system. Two tasks were analyzed. The first task consisted of analyzing human errors during routine execution of written procedures. Three functional "Test and Maintenance" procedures were tested. Each of these routine procedure tasks was analyzed by HRA evaluation teams in three phases:

- *Phase 1:* Each HRA evaluation team performed an analysis individually according to its own procedures, models, methods, and data. Each evaluation team was provided with common documentation, consisting of: a description of the affected plant system, copies of the procedures, a description of the control panels, a description of electronic modules used in the control interface, a description of available shift and control room personnel and their level of training, an overview video of the control panels and actual performance of the tasks, and answers to specific questions asked by other evaluation teams. No information was provided on the psychological state of the hypothetical crew performing the task, in accord with a prospective HRA such as used in the design phase of a plant.
- *Phase 2:* Subsequent to the individual analyses, the coordination team eliminated differences in underlying assumptions and boundary conditions that might have

contributed to variability in results. Analysis was limited to two types of human failure events (HFEs): non-detection of failures of the free flow check valve, and venting valves left open after completion of the test. Each HRA evaluation team, in turn, repeated the analysis using the constraints provided by the coordination team. This approach allowed a direct comparison between the quantitative findings of the various methods and teams.

- *Phase 3:* Finally, a precisely defined HFE was provided in which all steps of the analysis were provided except quantification. This phase investigated quantification of a pre-modeled condition to eliminate modeling as a source of variability between teams. Each HRA evaluation team recalculated their HEPs based on the provided HFE.

The second task was aimed at analyzing human errors in a diagnosis activity and in the corresponding selection of a response strategy. Specifically, the HRA evaluation teams reviewed an operational transient caused by loss of offsite power, leading to a reactor scram and turbine trip. The scenario presents a complex case in which the operator must decide the appropriate course of action within a significant time constraint. Without appropriate mitigative actions between control room and balance of plant personnel, the steam generator level would drop to a critical level in 40 minutes. The documentation provided to the evaluation teams was comparable to that provided for the test and maintenance procedure tasks.

Fifteen teams representing industry, utilities, regulators, and research institutions participated in various phases of the benchmark study by completing an HRA on the scenarios. The teams utilized the following methods:

- Technique for Human Error Rate Prediction (THERP),
- Success Likelihood Index Methodology (SLIM),
- Human Cognitive Reliability (HCR) Model,
- Human Error Assessment and Reduction Technique (HEART),
- Technica Empirica Stima Errori Operatori (TESEO),
- Absolute Probability Judgement (APJ), and
- Maintenance Personnel Performance Simulation Model (MAPPS).

The majority of evaluation teams utilized THERP and SLIM, with one or two teams applying each of the other methods. While teams were allowed to use whichever method they preferred for the first phase, all teams were instructed to use THERP and SLIM in the second and third phases. MAPPS was not primarily considered an HRA method and produced lower bound HEP estimates of 1E-2. Consequently, it was not benchmarked against the other methods. MAPPS analyses were conducted parallel to the benchmark activities in an exploratory investigation of its uses for HRA.

For the initial task involving potential errors on routine procedural tasks, comparisons between teams and methods were rendered virtually impossible, because the teams decomposed the scenarios differently or focused their analysis on different subtasks. As a consequence, qualitative and quantitative results could not be readily aligned for direct method-to-method or team-to-team comparisons. The second phase significantly constrained the analysis by providing the same HFEs to all teams. This phase also constrained the HRA method by specifying that all evaluation teams should primarily use THERP and SLIM for their quantification.

The comparison of quantification results suggested a wide spread of HEP values, as can be seen in Figure 17 of the original report [26]. Note that some strong overlap between THERP

and SLIM was attributable to teams using THERP to calibrate the Success Likelihood Index (SLI) in SLIM, a practice that strongly coupled SLIM and THERP HEPs.

The spread in HEP values for comparable tasks is, according to Poucet [26], attributable to differences in the interpretation of the event, the decomposition level, the necessity of certain actions, and recovery and dependency modeling. To control for these influences, the third and final phase of this task further constrained the HFE. The HRA evaluation teams were instructed to quantify only the following HFE: *Non-detection of free flow check valve side port failure to close.* The results showed less spread in the HEPs than for phase 2, but there were still considerable differences between teams. A further analysis revealed that the teams did not agree on the appropriate THERP values to use in quantification—the HRA evaluation teams sometimes used different THERP lookup tables and frequently used different items even when utilizing the same lookup table.

For the task involving the operational transient, analysis focused on the qualitative modeling of the task as well as the quantification of the task. No effort was made to constrain the modeling or type of quantification used. Because the hypothetical task contained a significant diagnostic or cognitive element, many teams incorporated HCR in their analysis. The results showed considerable differences in the decomposition of the task and in the calculated HEPs (see, for example, Figure 21 in [26]). The author notes that the use of HCR calculations had greater variability than the other dominant method used, namely THERP.

3.3.2. Lessons Learned

The HF-RBE provided an early and comprehensive empirical benchmark study on HRA. It remains, in the present authors' views, an invaluable reference document in that several of the methods discussed are still in use today. It also affords considerable insights for a large-scale comparison using different methods and different evaluation teams, which can guide future benchmarking efforts. It is also, as a first of its kind study, not without certain shortcomings.

The findings of the study were not favorable to HRA methods of the time. These findings showed a high degree of variability between methods and analysis teams. Poucet [26] attributed this variability to the following factors:

- *Differences in decomposition and modeling.* Poucet suggested that analysts made different assumptions about task characteristics, thus identifying different tasks as starting points in the analyses. Tasks were also often decomposed at a different level. Some analyses, performed at a fine grain of analysis, tended to elaborate on details that could not properly be inferred from the scenario source descriptions. Counterintuitively, these fine grained analyses tended to overlook the most important human errors in a given task, perhaps because the level of detail failed to highlight significant actions. HRA methods tended to treat recovery actions differently, but analysts using the same methods also tended to model recovery differently from each other. Finally, analysts and methods were not consistent in their modeling of dependency.
- *Differences in quantification.* Differences in HEPs were partially attributable to differences in the underlying modeling produced by the teams of analysts. However, even when these differences are eliminated, the calculated HEPs were in many cases several orders of magnitude different. THERP and SLIM showed generally strong

concurrence between methods, due in part to SLIM's calibration to THERP data. HCR produced significantly different HEPs than other methods. More troubling than the low inter-method reliability was the low inter-rater reliability. Analysts using the same method showed considerable variability, although THERP achieved the highest inter-rater reliability.

Attempts were made to control for differences in decomposition and modeling, especially those inherent in the first task of the study. By constraining the modeling differences, the author was able to affect a comparison between methods and teams. Although Poucet concludes that HRA modeling and decomposition processes are mature, the study findings made it clear that these processes accounted for some of the greatest sources of variability in the study. The study did not, however, identify ways to improve such modeling and decomposition in practice.

The study constrained the methods used for quantification in the second and third phases of the study. The decision to have evaluation teams only use THERP and SLIM in the latter phases of the first task was not fully explained in the final report [26]. Because expertise in the application of a particular HRA method may vary considerably and because the expertise of the evaluation teams with SLIM was not documented, it is difficult to know if this forced method use introduced an unaccounted-for source of variability within and between HRA evaluation teams.

Beyond the concluding discussion in [26], there are additional lessons to be learned from the design and conduct of the study. These lessons learned are not provided as a critique of the study but serve as a review of factors that the present authors believe should be addressed in future benchmark studies in HRA.

- *The study did not consider uncertainties.* The quantitative comparison provided in the study focused on single-point estimates without providing uncertainty bounds for the analysis. This point-to-point comparison overlooks the potential that many of the divergent findings might be covered within the uncertainty bounds of the analysis. Future benchmark studies of quantitative risk estimates should, whenever practicable, include uncertainty information to allow comparison of the range of estimated performance.
- *The study did not provide a standardized framework for comparing qualitative aspects of the analysis.* While some aspects of the analysis such as task decomposition were classified and compared across methods, other qualitative aspects of the analysis such as performance shaping factors, contextual information, and analysis assumptions were not articulated or compared. It would be possible, for example, to translate qualitative information provided by the evaluation teams into a standard template for analysis. Future benchmark studies should consider quantitative and qualitative comparisons and provide frameworks for comparing qualitative aspects of the analysis.
- *The study did not validate findings.* Although the study revealed considerable variability in the quantitative estimates provided by the HRA evaluation teams, the study failed to provide an empirical basis against which these estimates could be evaluated. Consequently, the main finding of the study is a lack of reliability in HRA methods, but the study fails to provide insights into the validity of particular findings, nor does it provide clues as to which HRA method or which HRA evaluation team's analysis most closely approximates actual human performance in the scenarios. Future benchmark

studies should provide both reliability and validation results to allow not only a method-to-method or team-to-team comparison but also a comparison of findings to empirical data.

3.4. Zimolong's Empirical Benchmark

3.4.1. Introduction

Zimolong [27] performed an analysis to validate and compare THERP, SLIM, and an expert estimation procedure. A simulated batch manufacturing scenario was performed under twelve task conditions by 48 experimental test participants. Three performance shaping factors (PSFs) were manipulated within the scenario—motivation (2 levels) x task load (2 levels) x information load (3 levels), creating the 12 conditions. Motivation was varied by offering different levels of incentive for successful task completion on the manufacturing scenarios. A low motivation group was not offered monetary remuneration for good performance, while a high motivation group was offered monetary remuneration for good performance. Task load was varied on two levels, whereby participants in the low task load group simply had to monitor tool wear, while the high task load group had to monitor tool wear in the face of an additional time-critical task. Finally, information load (three levels) was manipulated by the interval of required maintenance of the simulated machines. The task was said to be cognitively demanding, requiring diagnosis of machine state, comparison of graphical indicators, and decision making regarding whether or not to perform maintenance. The study defined the failure to shut down the machine in time for maintenance as a human error. However, there was some incentive for delaying the machine shutdown to increase the number of parts completed. The point at which the machine failed was a randomized.

The human error rates across the 12 scenarios were derived from the participants' task performance. These values were compared to results generated by THERP, SLIM, and the expert estimation procedure. SLIM estimates were derived by asking judges to rank the importance of the three PSFs in influencing participants' ease and ability in accurately completing the task. The expert estimation procedure was accomplished by asking these same judges to rank order the 12 scenarios as to the difficulty they felt the participants' would have in accurately completing each one. The author of the study, based on a task analysis, used THERP in deriving an overall HEP for the study, but noted that THERP was not able to be used to generate HEPs for each of the individual scenarios because it did not appropriately address the relevant factors.

The average weightings of the three PSFs made by the judges using SLIM were compared to the empirical estimates achieved by the participants. Although, overall, the judges felt that task load would account for the greatest amount of variance in the participants' performance, the actual data showed that motivation and information load accounted for the greatest proportion. A Spearman rank correlation between the ratings given by the judges using the expert estimation technique and the actual rankings of the scenarios based on errors made by the participants showed low agreement between the judges and the real data (*rho* = 0.38). THERP was unable to be evaluated in its ability to estimate the difficulty of the PSFs or the individual scenarios as THERP was only applied in determining an overall HEP. In comparing

the overall HEP computed by SLIM and THERP, both came relatively close to the actual HEP (0.029). The HEP was overestimated by SLIM (0.037) and underestimated by THERP (0.016 using a modifier of 2, and 0.024 using a modifier of 3).

3.4.2. Lessons Learned

The study presents a useful example of how psychological research can inform HRA, even when not using expert performers such as control room crews as participants. Nonetheless, the study reveals some obstacles to overcome in an effective benchmark study design.

- *Definition of human error.* The validity of the study may be questioned somewhat due to the definition of human error. Human error is defined within the study as the breakdown of the machine from not being serviced in time. The classification of such a failure as a human error is debatable as it does not fit the traditional definition of error as a mistake made by the operator. The action of the operator in servicing the machine at the very moment when one of the machine tools reaches 80% of its life is a deliberate decision by the operator and not necessarily a mistake, lapse or slip. And as noted above, there was some incentive for the participants to try and complete parts currently being machined when the 80% mark was reached.
- *Use of non-experts.* The judges employed within this study to use the SLIM and expert estimation techniques were not experts of either method, of HRA in general, or of the expected effects of the manipulated factors. The lack of experience by these judges in performing such HRA tasks calls into question the reliability of their performance and the derived estimates. The validity of comparing the results from SLIM and the expert elicitation technique to the results of THERP, which was employed by an expert, may be suspect.
- *Significant impact of PSFs on performance.* The three PSFs were found to impact performance in the direction as expected. High motivation resulted in lower error rates than did low motivation. High task load and information load resulted in higher error rates than did low task load and information load. While the effects of all PSFs were statistically significant, motivation and information load exhibited a greater effect on performance than did task load.
- *Difference between calculated HEPs and actual performance.* The application of THERP in this context does not allow the calculation of individual HEPs for each of the 12 scenarios investigated. Therefore, no comparison can be made between the other two methods and THERP in regards to the individual HEPs. It is unclear a priori how the three PSFs were predicted to influence the participants' performance within the HRA methods, and it would require a fairly complex analytical method to derive estimates of performance. However, non-expert judges were asked to forego these questions and apply SLIM and the expert estimation technique to obtain estimates or rankings according to the manipulations. Therefore, the inexperienced use of SLIM and the expert estimation technique in deriving the individual scenario HEPs without further guidance as to how the PSFs would affect performance is an unreliable portrayal of the methods.

3.5. Kirwan's Quantitative Benchmarking Exercise

3.5.1. Introduction

Kirwan [28-30] expands upon several of the criteria identified in his qualitative assessment of HRA [20-21] in a series of three papers to validate the HRA methods THERP, HEART, and Justification of Human Error Data Information (JHEDI). The first paper introduces the three HRA methods, which were widely used in the British nuclear industry. The second paper reviews the results of an empirical benchmark to compare the three methods. The final paper reviews practical aspects of using the techniques, with a lengthy discussion on ways in which to improve the consistency of HEP generation in the three methods.

Of particular interest to the present discussion is the empirical validation presented in the second paper [29]. Kirwan selected 30 human error scenarios from the Computerized Operator Reliability and Error Database (CORE-DATA) of human performance [31]. With the exception of one scenario based on expert judgment, the scenarios featured human performance and error data obtained from empirical observation during plant operations or simulator studies. These scenarios featured skill and rule based behavior and excluded knowledge based operator actions. In addition, the errors were representative of executing maintenance related tasks rather than the diagnostic related tasks involved in operating crews responding to accident conditions. A between-subjects design was employed in which 10 analysts were engaged for each of the three methods (i.e., a total of 30 analysts across the three methods). The analysts were screened to ensure they had proper experience in using each technique. The analysts were given scenario descriptions and then asked to use each technique to model and quantify the errors found in the scenarios within a two-day time limit.

The analysts' estimates using the three HRA methods were compared with the actual HEPs derived from the CORE-DATA. The results showed that there was a significant correlation between the estimates produced by the analysts and the values found in the CORE-DATA. Nearly 77% of analysts achieved a significant correlation between the computed HEP and the actual error rate contained in the CORE-DATA database. Moreover, no individual assessor had lower than 60% precision between their analysis and the CORE-DATA values. JHEDI showed the most conservative results, followed by HEART and THERP. Moreover, an average of 72% of HEP estimates fell within a factor of 10 of the actual values contained in the CORE-DATA, suggesting a high degree of estimate precision. However, the degree of overlap between the actual accuracy of the analysts and the analysts' perceived accuracy was low, with only a 23% overlap. Analysts were equally under and over-confident of the accuracy of their results.

3.5.2. Lessons Learned

Kirwan's quantitative benchmark offers some of the most tractable data available in HRA in terms of providing an external, empirically derived validation and of systematically comparing methods on relevant criteria. Apart from the data to support the general validity of the three HRA methods, Kirwan's study offers insights into the reasons for inconsistencies (or poor

reliability) between and within methods. Lessons learned with respect to validity and consistency include:

- *Sources of inconsistency.* Although consistency between methods and across analysts was generally strong, there were several areas where improvement was possible. In particular, there was variability in analysts' selection of the Generic Error Probability in HEART and in modeling and decomposition in THERP and JHEDI. Also, there appeared to be difficulty in consistently modeling errors of commission in HEART and JHEDI, slips in HEART, cognitive or diagnostic tasks in THERP, low-probability administrative tasks in THERP and JHEDI, human-machine interface tasks in THERP, and rule violations in HEART. These shortcomings point to the fact that no HRA method is comprehensive in its coverage of human errors and that each method represents strengths and weaknesses in terms of its coverage and quantification.
- *Identification of several important benchmarking design issues.* Although the generalizability of the results to the validity of HRA methods in quantifying post-accident operating crew actions is limited by the nature of the errors addressed, several important benchmarking design issues were identified. They included the importance of having: (1) a range of tasks/scenarios against which to test the methods, (2) enough HRA teams assigned to each method to be able to control for potential outlier effects in the results, and (3) actual performance data against which to compare the HRA method predictions.

3.6. Maguire's Validation Benchmark

3.6.1. Introduction

HRA methods harken to empirical data as sources for their HEP estimations. While the nominal HEP values intrinsic to a method may have strong ties to human performance data, it is rare that HEPs generated by the method in application are validated to human performance. Maguire [32] provides an informative and straightforward validation exercise using HEART. HEART, like most HRA methods, was created based on and for use in the nuclear industry. Maguire worked in the domain of aviation and sought to validate HEPs generated by HEART to aviation performance data. Across a set of tasks typical for aviation (i.e., landing, hover, and transit), Maguire and his team generated HEP estimates using the HEART method and subsequently polled aviation operating databases to arrive at actual human error rates in aviation.

Maguire's investigation into the use of HEART for aircrew tasks began with an identification of what errors might occur in a flight landing task using Hierarchical Task Analysis (HTA) and through the construction of critical task lists. Following the identification of errors, HEART was used to quantify the human error rates. A secondary research analysis was undertaken to validate the findings produced from the utilization of HEART in producing human error probabilities. Raw data were available from the US Army Safety Center database of aircrew performance, from which human error rates could be calculated in terms of the frequency of errors per task. These rates were adjusted because errors that do not lead to incidents are typically underreported. Maguire used the number of sorties per year to extrapolate

the per task frequency. Maguire found that the HEART generated estimates were within a factor of ten of the referent data.

3.6.2. Lessons Learned

Although the comparison is of particular interest to users of HEART, the study is included here because of its illustrative use of operational data to validate HEP estimates produced by an HRA method. As well, Maguire's study raises a key point regarding HRA validation:

- *Validation of human performance across domains*. To date, there have been minimal attempts to validate nuclear-derived HEPs to other domains such as aviation. It might be argued that when HRA methods are applied to non-nuclear domains, there is an over-reliance on the underlying human performance data in a nuclear power plant setting. It is important that existing data sources be validated to new domains, scenarios, and tasks prior to tacit acceptance of the generalizability of these data sources. While it is assumed that human performance is stable across domains, it is necessary to prove this assumption before HRA methods are applied to areas for which they were not necessarily designed.

4. CONCLUSION

This paper presented results from several relevant HRA benchmarking studies. The lessons learned from these studies were supplemented with knowledge gained by examining other benchmarking studies within the psychology and risk assessment literatures. Although the lessons learned and associated experimental design issues in these domains are often discussed in somewhat different terminology, in most cases they point out parallel issues relevant to HRA benchmarking. The implications of these lessons learned for HRA benchmarking are discussed at the end of each of the domain reviews in Section 2. The most important points derived from these studies, in conjunction with those from existing HRA benchmarking studies and some logical inferences, are summarized below in terms directly relevant to HRA benchmarking.

A benchmarking study should begin with a clear explanation of the scenarios or tasks to be evaluated within the study. Unless HRA approaches to identifying and structuring the scenarios and tasks are part of what is being evaluated, as little as possible should be left to interpretation by the HRA teams. Variations in the teams' understanding of the events being analyzed have been shown to lead to undesired and intractable variations in their results (e.g., [26]). In fact, Poucet [26] attempted to control for this effect by providing teams with a precisely defined HFE within the third phase of his benchmarking study. Since an important aspect of benchmarking studies will be to understand why different results are obtained between methods and to identify ways to improve them, it is important to provide adequate controls that allow inferences about the causes of variation in results. In some cases this may require testing baseline conditions in initial experiments and then allowing important factors to vary in later experiments in order to track their effects.

The assortment of identified tasks should be relatively large and varied to encompass the full spectrum of activities possible and to test the strengths and weaknesses of methods along multiple dimensions, as revealed in the discussion of method-to-method and product-to-product usability comparisons [5-7]. Within the domain of HRA, Kirwan [28–30] demonstrated this requirement well by selecting 30 tasks and scenarios to test each method. Further, an attempt should be made to include a scenario that would assume a detectable failure rate. Kirwan's scenarios were derived from actual, previously observed erroneous human activities.

Of particular importance following the identification of scenarios and tasks is collecting empirical data to compare to the HRA methods' estimates. A comparison between the HRA methods' results and empirical data is necessary to validate the methods. For instance, Zimolong [27], Kirwan [28-30], and Maguire [32] each compared team results to real data to validate the HRA methods on their ability to match empirically derived data. Such validation is key to building confidence in the use of the method. Validation is not strictly a quantitative comparison of HEPs. As the cognitive modeling benchmark examples [12, 14] revealed, there are multiple levels of granularity to consider in validation, ranging from qualitative to quantitative comparisons. Each level of validation may provide different and valuable insights into the methods under comparison—insights on processes, PSFs, and analysis assumptions that might not be gleaned from a strict quantitative comparison.

In general, human error performance tends to be infrequent and stochastic. This is particularly the case in nuclear power accident scenarios (Maguire [32] was able to avoid this problem by using data outside of nuclear power production and used instead aviation data to test HRA methods.) An issue stemming from infrequent human error occurrence is being able to

examine an error of interest or error of significant importance. A technique to circumvent this issue is called *seeding*, in which the researcher includes a situation that would strongly induce an error to occur in the scenario. For example, if the researcher induces a series of co-occurring PSFs that would not necessarily occur during normal event response, the researcher has seeded a particular scenario of highly loaded PSFs. While the seeding method allows the researcher to examine the type of error of interest, the external validity of the error may be questionable. That is, if the error of interest does not occur in an unintrusive simulated study, then the importance of that error in the real world becomes questionable. Thus, it is important for researchers to identify plausible scenarios with the potential for some detectable human errors, without seeding in such a way as to virtually guarantee the errors.

An alternative approach is to allow the operators' performance in the simulator to dictate the type of error to emerge. Again, because of the infrequent occurrence of error, this approach may be limited in providing adequate numbers of errors to validate HRA methods. For example, if most HRA methods specify that a particular error will only occur 1 in a 1,000 times a task is performed, it is questionable whether such a prediction can reasonably be validated in 1,000 trials in a simulator study, nor is it typically practical or cost-effective to run crews through this many simulator trials of a given scenario.

Another way to address this issue is to identify potential measures other than errors per se in order to validate HRA predictions. For example, most HRA methods attempt to identify the important PSFs that will influence the potential for success or failure, and this information is used in deriving HEPs. Debriefs of operating crews and the observations of experimenters on the factors and scenario conditions influencing the crews' behavior in the test scenarios can be compared against the predictions of the different methods in order to provide another measure for assessing validity. Poucet [26] attempted to elaborate on differences between teams' estimates by comparing how the teams decomposed tasks. However, he may have been able to further explain team differences by continuing his exploration into examining the PSFs identified by each method.

A further requirement of any benchmarking study is an explanation of the participants or teams of participants who will be applying the methods. A common concern in past benchmarking studies (e.g., the QRA benchmarks [15-18]) has been large performance variability between and within teams or analysts. Such considerations have been successfully controlled in previous HRA benchmark activities. For example, in order to fully examine intra-method reliability, Kirwan [28-30] employed 10 teams to use each of the HRA methods. Performance variability within an HRA team leads to inadequate method reliability, while high performance variability between HRA teams leads to measurement insensitivity for traditional inferential statistical methods and precludes inferences about the accuracy of methods as opposed to the team applying the method. Variability may reflect unexpected confounding influences, artificial task characteristics of the work situation(s) in the experiment, or the valid reflection of the HRA method performance.

An approach to reduce performance variability is to utilize subject-matter experts (SME) to identify key individual difference variables within and between HRA teams and set these variables as covariates in the experimental design. An SME can also be used to review experimental task demand characteristics to ensure that the task is highly similar to the real world. Furthermore, every attempt should be made to equalize the teams in amount of experience and knowledge they have in applying their chosen method. For instance, it was unclear in the results from Poucet's benchmarking study [26] if differences seen in estimates

produced by the teams were due to differing levels of expertise in applying the HRA methods. Standardizing teams in experience levels lends confidence to findings that differences between method results are dependent upon the method being applied and not on the team applying it.

In order to better compare results across teams, participants should be encouraged to include uncertainty bounds around their HEP estimates. Although a point-to-point comparison between teams may show divergence, it is possible there is overlap within the uncertainty bounds that could demonstrate greater agreement between the teams. The failure to include uncertainty bounds particularly plagued the results of Poucet [26], who may have been able to show more consistency between results by comparing uncertainty estimates.

Past HRA benchmarking studies using a large number of methods coupled with a small number of teams for each of the HRA methods have found large performance variability in the benchmarking results, which has limited the confidence of their results. While there is no clear solution to this concern, it is recommended that careful consideration is given to the number of methods to be analyzed as well as the number of teams within a method. For example, given that there is a small number of participants in a benchmarking study, it may be beneficial to decrease the number of methods (between-subjects variable) to be analyzed while increasing the number of teams within each of the methods (within-subject variable). Allowing more data points within each method can increase power and facilitate inferential statistical analyses, allowing more conclusive findings to emerge from the benchmark.

Note that while these lessons learned provide an important foundation for future benchmarking, perhaps the most important lesson learned is that benchmarking represents a compromise. As discussed in the section on benchmarking in educational psychology [1], there are necessary trade-offs that typically need to be made in order to strike a balance between the experimental ideal and the experimental practicable. A study that meets all of the lessons learned remains an ideal. A study that acknowledges the ideal criteria and addresses the most important ones remains a real possibility. Nevertheless, given the wide range of variables that can influence the results of benchmarking studies, it is unlikely that a single experiment will be definitive. Rather a series of experiments, systematically varying important variables while ensuring adequate controls, will be necessary to demonstrate the validity and reliability of HRA methods.

5. REFERENCES

[1] Boring R. Human and computerized essay assessment: A comparative analysis of holistic, analytic, and latent semantic methods. Unpublished Master's Thesis, New Mexico State University, Las Cruces, 2002.

[2] Cohen J. Statistical power analyses for the behavioral sciences, 2nd edition. Hillsdale: Lawrence Erlbaum Associates, 1988.

[3] Murphy K, Myors B. Statistical power analysis. A simple and general model for traditional and modern hypothesis tests, 2nd edition. Mahwah: Lawrence Erlbaum Associates, 2004.

[4] Jonassen D, Beissner K, Yacci M. Structural knowledge: Techniques for representing, conveying, and acquiring structural knowledge. Hillsdale: Erlbaum, 1993.

[5] Faulkner L. Reducing variability: Research into structured approaches to usability testing and evaluation. In: Proceedings of the Annual Conference of the Usability Professionals Association. Florida, 2002.

[6] Kessner M. On the reliability of usability testing. Unpublished Master's Thesis, Carleton University, Ottawa, 2000.

[7] Chaparro B, Gibson A. Planning your next vacation: Orbitz, Expedia, or Travelocity? Usability News, 2002; 4.1: 1-5.

[8] Meyer J, Bitan Y, Shinar D. Displaying a boundary in graphic and symbolic "wait" displays: Duration estimates and users' preferences. International Journal of Human-Computer Interaction, 1995; 7: 273-290.

[9] Meyer J, Shinar D, Bitan Y, Leiser D. (1996). Duration estimates and users' preferences in human-computer interaction. Ergonomics, 1996: 39; 46-60.

[10] Nielsen J. Usability engineering. Boston: AP Professional, 1993.

[11] Ryu Y, Smith-Jackson T. Reliability and validity of the Mobile Phone Usability Questionnaire (MPUQ). Journal of Usability Studies, 2006; 2: 39-53.

[12] Best B, Lebiere C, Schunk D, Johnson I, Archer R. Validating a cognitive model of approach based on the ACT-R architecture. Boulder: Micro Analysis & Design, Inc., 2005.

[13] Turing A. Computing machinery and intelligence. Mind, 1950; 59: 433-460.

[14] Leiden K, Best B. A cross-model comparison of human performance modeling tools applied to aviation safety. Boulder: Micro Analysis & Design, Inc., 2005.

[15] Lauridsen K, Kozine I, Markert F, Amendola A, Christou M, Fiori M. Assessment of uncertainties in risk analysis of chemical establishments: The ASSURANCE project, Final Report, Risø-R-1344(EN), Roskilde: Risø National Laboratory, 2002.

[16] Contini S, Amendola A, Ziomas I. Benchmark exercise on major hazard analysis. Vol. 1 Description of the project, discussion of the results and conclusions, Final Report, EUR 13386, Ispra: CEC-JRC, 1991.

[17] Amendola A, Contini S, Ziomas I. Uncertainties in chemical risk assessment: Results of a European benchmark exercise. Journal of Hazardous Materials 1992; 29: 347-363.

[18] Fabbri L, Jirsa P, Contini S. Benchmark exercise in quantitative area risk assessment in central and eastern European countries (BEQUAR), Final Report, EUR 22619, Ispra, CEC-JRC, 2007.

[19] Swain A. Comparative Evaluation of Methods for Human Reliability Analysis, GRS-71. Cologne: Gesellschaft für Reaktorsicherheit, 1989.

[20] Kirwan B. Human error identification in human reliability assessment. Part 1: Overview of approaches. Applied Ergonomics, 1992; 23: 299-318.

[21] Kirwan B. Human error identification in human reliability assessment. Part 2: Detailed comparison of techniques. Applied Ergonomics, 1992; 23: 371-381.
[22] Forester J, Kolaczkowski A, Lois E, Kelly D. Evaluation of human reliability analysis methods against good practices, Final report, NUREG-1842. Washington, DC: US Nuclear Regulatory Commission, 2006.
[23] Kolaczkowski A, Forester J, Lois E, Cooper S. Good practices for implementing human reliability analysis: Final report, NUREG-1792. Washington, DC: US Nuclear Regulatory Commission, 2005.
[24] Poucet A. Survey of methods used to assess human reliability in the human factors reliability benchmark exercise. Reliability Engineering and System Safety, 1988; 22: 257-268.
[25] Poucet A. The European benchmark exercise on human reliability analysis. In: Proceedings of the American Nuclear Society International Topical Meeting on Probability, Reliability, and Safety Assessment. Pittsburgh, 1989. p. 103-110.
[26] Poucet A. Human factors reliability benchmark exercise, Final Report, EUR 12222, Ispra: CEC-JRC, 1989.
[27] Zimolong B. Empirical evaluation of THERP, SLIM, and ranking to estimate HEPs. Reliability Engineering and System Safety, 1992; 35: 1-11.
[28] Kirwan B. The validation of three human reliability quantification techniques—THERP, HEART, and JHEDI: Part I—Technique descriptions and validation issues. Applied Ergonomics, 1996; 27: 359-373.
[29] Kirwan B. The validation of three human reliability quantification techniques—THERP, HEART, and JHEDI: Part II—Results of validation exercise. Applied Ergonomics, 1997; 28: 17-25.
[30] Kirwan B. The validation of three human reliability quantification techniques—THERP, HEART, and JHEDI: Part III—Practical aspects of the usage of the techniques. Applied Ergonomics, 1997; 28: 27-39.
[31] Kirwan B, Basra G, Taylor-Adams, SE. CORE-DATA: A computerised human error database for human reliability support. In: Proceedings of the 1997 IEEE Sixth Conference on Human Factors and Power Plants. New York: Institute of Electrical and Electronic Engineers, Inc., 1997, p. 9.7–9.12.
[32] Maguire R. Validating a process for understanding human error probabilities in complex human computer interfaces. In: Proceedings of the Second Workshop on Complexity in Design. Glasgow: Glasgow University Press, 2005. p. 81-89.

Distribution List

2	MS 3605	Idaho National Laboratory Attn: Ronald L. Boring (1 electronic copy) Tuan Q. Tran (1 electronic copy) P.O. Box 1625 Idaho Falls, Idaho 83415-3605
1	MS 10E50	US Nuclear Regulatory Commission Attn: Erasmia Lois (1 electronic copy) Washington, DC 20555-0001
1	MS0748	J.A. Forester, 06761 (1 electronic copy)
1	MS 0748	S. L. Hendrickson, 06761 (1 electronic copy)
1	MS 0899	Technical Library, 9536 (1 electronic copy)


Sandia National Laboratories