RIL 2024-04
PNNL-35181

United States Nuclear Regulatory Commission
Protecting People and the Environment

Human Reliability Analysis Methods for Calculating Effects of Performance Influencing Factors: A Review of Historical Approaches

November 2023

Prepared by:
J. Baweja
G. Coles
B. Jefferson
R. Prasad

Pacific Northwest National Laboratory
P.O. Box 999
Richland, WA 99352

Y. James Chang, NRC Project Manager

Research Information Letter
Office of Nuclear Regulatory Research

Disclaimer

TABLE OF CONTENTS

EXECUTIVE SUMMARY

The U.S. Nuclear Regulatory Commission funded Pacific Northwest National Laboratory (PNNL) to evaluate the technical basis for the form used in the Integrated Human Event Analysis System for Event and Condition Assessment (IDHEAS-ECA) presented in NUREG-2256, *Integrated Human Event Analysis System for Event and Condition Assessment* to account for the effects of performance influencing factors (PIFs) on human error probabilities (HEPs). This report describes the work accomplished for Task 3 of this project and contains:

- An overview of the literature review and the insights learned.
- An overview of the interviews performed with developers of human reliability analysis (HRA) methods, including discussion of selection of the experts, design of interview questions, and documentation of the interview results.
- A discussion of insights from the results of the HRA developer interviews that may play a role in informing the HRA guidance provided in IDHEAS-ECA, which is based on the concepts published in NUREG-2198, *General Methodology of An Integrated Human Event Analysis System*.

The purpose of this task was to evaluate the technical basis of (1) distinguishing the PIF effects on HEP into base HEPs and HEP modifiers, and (2) using the probabilistic sum and linear addition to calculate the combined effects of the base and modifier PIFs, respectively. To accomplish these objectives, PNNL staff conducted a literature review of existing HRA methods to assess how each quantifies HEPs with respect to the impact of PIFs. PNNL staff also reviewed literature on existing research that has explored the impact of multiple PIFs in combination on human performance.

In addition, during interviews conducted with the developers of selected HRA methods, PNNL focused on first- and second-generation approaches for obtaining input regarding the justification for their quantification approaches. In the continuum of first-, second-, and third-generation HRA methodologies, second-generation methodologies generally focus more on the concepts of cognition and context than first-generation methodologies. Both are used in support of nuclear power plant probabilistic risk assessments. Third-generation methods include dynamic simulations and are not in mainstream use for supporting nuclear power plant probabilistic risk assessments. The primary goal of the semi-structured interviews with HRA method developers was to better understand the quantification approaches used in the methods that they developed or helped develop, particularly as it pertains to factors that may influence human performance. Participants were developers or researchers of HRA methods and were selected for their expertise and knowledge in the creation or application of HRA methods and the field in general. Nine interviewees were contacted and eight agreed to participate.

The results of this effort found that none of the methods reviewed with the HRA developers included a distinction between base and modifier PIFs comparable to the one proposed in IDHEAS. The experts were asked about the differences in importance between PIFs and whether there are PIFs that should be quantified differently in the method. The experts acknowledged that certain PIFs may generally be more important, but did not report that they made a distinction between base and modifier PIFs or quantified the impact of certain PIFs differently than other PIFs.

Although some reports in the literature have explored the additive versus multiplicative combination of PIFs and their impact on HEP, developers who participated in this study did not express strong beliefs or opinions on this topic. In general, they reported that the functional form in their model (often multiplicative) was a matter of historical convention. Additional data and analyses are necessary to provide a more definitive answer to this question. The results of the interviews suggest that in past and current HRA approaches, a multiplicative form is more common; however, the papers that explicitly explore this question seem to indicate an additive form is more appropriate. Most of the HRA methods treat PIFs as independent, with the exception of the Cognitive Reliability and Error Analysis Method; however, most developers (with a single exception) expressed the belief that PIFs are likely interdependent. As mentioned, however, additional data collection and analysis on this question would be informative.

The following recommendations are provided based on the results of this research:

- Additional data collection is required to provide a technical basis or justification for the distinction between base and modifier PIFs. The results presented in this study do not support this proposition but do not necessarily exclude it as a possibility. Because the existing data have not examined this question in depth, additional research is required.
- There is a clear link between the independence or interdependence of PIFs and the functional form of their impact on HEP. The existing literature suggests that PIFs combine additively. However, the results of interviews conducted in this study suggested that PIFs do interrelate; thus, a multiplicative form might be appropriate. A systematic experimental exploration of this question would provide additional information that cannot be obtained solely through a review of the literature. A potential family of experiments exploring triads of PIFs might be useful to explore their interactions.

ABBREVIATIONS AND ACRONYMS

ASEP	Accident Sequence Evaluation Program
CBDT	cause-based decision tree
CFM	cognitive failure mode
CPC	common performance condition
CREAM	cognitive reliability and error analysis method
EPC	error producing condition
EPRI	Electric Power Research Institute
GTT	generic task type
HCR	human cognitive reliability
HEART	human error assessment and reduction technique
HEP	human error probability
HRA	human reliability analysis
IDHEAS-ECA	Integrated Human Event Analysis System for Event and Condition Assessment
IDHEAS-DATA	Integrated Human Event Analysis System for Human Reliability Data
IDHEAS-G	General Methodology of An Integrated Human Event Analysis System
NARA	Nuclear Action Reliability Assessment
NRC	U.S. Nuclear Regulatory Commission
PIF	performance influencing factor
PNNL	Pacific Northwest National Laboratory
PRA	probabilistic risk assessment
PSF	performance shaping factor
SLI	success likelihood index
SLIM	success likelihood index methodology
SPAR-H	Standardized Plant Analysis Risk HRA
THERP	technique for human error rate prediction
TRC	Time-Reliability Correlation

1 INTRODUCTION

The U.S. Nuclear Regulatory Commission (NRC) funded Pacific Northwest National Laboratory (PNNL) to evaluate the technical basis for the form used in the Integrated Human Event Analysis System for Event and Condition Assessment (IDHEAS-ECA) presented in NUREG-2256, *Integrated Human Event Analysis System for Event and Condition Assessment* [1] to account for the effects of performance influencing factors (PIF) on human error probabilities (HEP). The IDHEAS-ECA approach is based on guidance in IDHEAS-G published in NUREG-2198, *General Methodology of An Integrated Human Event Analysis System (IDHEAS-G)* [2]. The purpose of this PNNL report is to examine the technical basis for impact of PIFs on HEP, to evaluate the technical basis of the functional form of the combined effects of different PIFs on a single HEP, and to specify the applicable scope of that equation. This report presents the results of an effort to address these research objectives.

1.1 Background

Human reliability analysis (HRA) methods strive to estimate the HEPs on the basis of the conditions under which the actions are performed. The factors that influence performance are called either PIFs or, historically, performance shaping factors (PSFs). The purpose of these PIFs is to operationalize or define the conditions under which human errors might occur. Many HRA methods have been developed, and these methods have defined different sets of PIFs. PIFs are used to modify the likelihood of error using a variety of functional forms. Different HRA methods make assumptions regarding the relationships among the PIFs, including whether the effects of PIFs are dependent or independent (e.g., [3, 4]). This report explores the question of how PIF impacts are aggregated and whether the treatment of some PIFs should be different from the treatment of others.

1.2 Integrated Human Event Analysis System

Despite their differences, many of the existing HRA methods approach the quantification of human error in a similar manner. They assume a nominal rate of human error based on the nature of the task being completed (generally determined using a task analysis). The nominal error rate is then modified based on different levels of PIFs that affect the task. Developed by NRC staff, the IDHEAS-G framework makes several modifications to this traditional HRA methodology. The nominal error rates are used when the task performance condition in the aspects of three base PIFs is optimal. The three base PIFs are information availability and reliability, task complexity, and scenario familiarity. When the task performance condition is not optimal in the three base PIFs (i.e., there is an attribute of these three base PIFs in the task performance conditions that negatively affects the task reliability), the nominal error rates are replaced by the base HEPs of these base PIFs (NUREG-2198 [2]). In addition to the three base PIFs, IDHEAS-G identifies 17 "modifier" PIFs to represent the effects other than the three base PIFs. These modifier PIFs adjust the task reliability by multiplying the base HEP.

In summary, the task reliability (i.e., HEP) is generated using the base HEP, which is calculated by the base PIFs and nominal error rate and then adjusted by the modifier PIFs by multiplying the base HEP with modifiers (or weights). Each PIF has several attributes, which are different observations that may apply to a given event. Each PIF attribute has a base HEP (for base PIFs) or modifier (for modifier PIFs) to represent their effects on task reliability.

The method of calculating the base HEP used in IDHEAS-G is a departure from the traditional approach, which generally assumes that the base error rate is a function of the type of task and does not explore the underlying cognitive causes of error. In IDHEAS-G, base HEPs are constructed using data on human reliability documented in the *Draft Integrated Human Event Analysis System for Human Reliability Data* (*IDHEAS-DATA* [5]). The base HEPs then are modified by any relevant PIFs. In this effort, the NRC sought input from PNNL to evaluate the technical basis of the quantification of base versus modifier PIFs in the IDHEAS-G model.

Although many HRA methods traditionally consider the impact of multiple PIFs in a multiplicative fashion, the IDHEAS-G quantification is somewhat more complex. In IDHEAS-G, multiple base PIFs are treated probabilistically assuming independence among base PIFs (probabilistic addition). For the modifier PIFs, however, the impact of multiple PIFs is realized by summing weights assigned to each PIF, with the cumulative weight used to *multiply* the base HEP. In this way, HEPs are computed for each cognitive failure mode (CFM) and then combined using a probabilistic sum. Specifically, IDHEAS-ECA [1], which is NRC's HRA tool and provides guidance based on IDHEAS-G, formulates the HEP as shown in Eq. (1):

$$HEP_{CFM} = \left[1 - \prod_{i=1}^{N}(1 - Base\ HEP_i)\right]\left[1 + \sum_{j=1}^{M}(W_j - 1)\right] \cdot \frac{1}{Re} \qquad (1)$$

where N = the number of base PIF attributes identified by analysts for the human-failure event (≤3)
M = the number of modifier PIF attributes identified by analysts for the human-failure event (≤17)
W = the weights of the modifier PIF attributes
Re = the recovery factor.

Thus, the underlying quantification is both multiplicative (at the base level) and additive (among modifiers). The question of multiplicative versus additive functional form has been the subject of debate (e.g., [6, 7]). However, questions remain regarding the empirical support for either model and also regarding the scope of Eq. (1). One aspect of this project was to assess the basis for the quantification of the impact of multiple PIFs.

1.3 Purpose and Scope of Work

The purpose of the project task reported here was to evaluate the technical basis of distinguishing the PIF effects on HEP into base and modifiers and of using the probabilistic sum and linear addition to calculate the combined effects of the base and the modifier PIFs, respectively. To accomplish these objectives, PNNL staff conducted a literature review of existing HRA methods to assess how each method quantifies HEPs with respect to the impact of PIFs. In addition, PNNL conducted interviews with developers of important HRA methods focusing on first- and second-generation approaches[1] to obtain input regarding the basis for their quantification approaches. The outline of the report is as follows:

1 In the continuum of first-, second-, and third-generation HRA methodologies, second-generation HRA methodologies in general focus more than first-generation on the concepts of cognition and context and both are used to support nuclear power plant probabilistic risk assessments (PRAs). Third-generation methods contain dynamic simulation and are not used to support nuclear power plant PRAs.

- Chapter 2 provides an overview of the literature reviewed and the insights learned.
- Chapter 3 provides an overview of the interviews performed with developers of HRA methods. It includes discussion of the selection of experts and design of interview questions.
- Chapter 4 documents the results of the HRA developer interviews.
- Chapter 5 presents a discussion of the conclusions of the literature review and interviews; Appendix A provides more detail on these results.

2 LITERATURE REVIEW OF BASIS FOR CALCULATING PIF EFFECTS

This chapter provides an overview of the literature reviewed and the insights learned. The literature review focused on addressing the questions specified in the scope of work, beginning with how PIFs were selected for inclusion in different HRA methods and noting any differences in their quantification or importance within the method. Second, the literature review explored how the impact of multiple PIFs are quantified to assess their effect on the HEP. Specifically, the review explored the functional form used to quantify the combined effects of multiple PIFs on the probability of error and whether that functional form was additive, multiplicative, or something else. Finally, the literature was reviewed to gain an understanding of how the different HRA methods addressed the possible interrelationships among PIFs; that is, whether the PIFs are treated as independent or, if treated as interdependent, how that interdependence was quantified.

2.1 Human Reliability Analysis Approaches

HRA methods have approached the modeling of multiple PIFs in different ways. This section provides a summary and overview of the ways HRA methods have modeled and quantified PIFs.

2.1.1 Standardized Plant Analysis Risk HRA

After some evolution, the guidance for performing the Standardized Plant Analysis Risk HRA (SPAR-H) was published August 2005 in NUREG/CR-6883, “The SPAR-Human Reliability Method” [3]. The SPAR-H method is built on an explicit information-processing model of human performance derived from the behavioral sciences literature that was then interpreted in light of activities at nuclear power plants. A major component of the SPAR-H method is the worksheet presented in NUREG/CR-6883. The SPAR-H method assigns human activity to one of two general task categories: action or diagnosis. Base error rates for the two task types associated with the SPAR-H method were calibrated against the Technique for Human Error-rate Prediction (THERP) [8] and Accident Sequence Evaluation Program (ASEP) [9] HRA methods. These methods employ eight summary operational factors that impact performance and the likelihood of human error. The operational factors are broad categories of PSFs that are relevant to nuclear power plants and include:

1. Available time
2. Stress and stressors
3. Experience and training
4. Complexity
5. Ergonomics (including the human-machine interface)
6. Procedures
7. Fitness for duty
8. Work processes

The impact of these PSFs on the probability of error is quantified in SPAR-H using multipliers that were selected based on an extensive literature review. Other than the values that each multiplier can be, the PSFs are treated equivalently in the model. The impact of multiple PSFs is quantified multiplicatively as shown in Eq. (2):

$$PSF_{composite} = PSF_1 \times PSF_2 \times \ldots \times PSF_n \quad (2)$$

The base HEP then is multiplied by the $PSF_{composite}$ to estimate the final HEP. The SPAR-H guidance acknowledges that the true independence (or not) of the eight PSFs is not known but treats the PSFs as independent within the model. The authors note, "Historically, in quantifying HEPs, HRA practitioners have treated these influencing factors as independent. In reality, dependence is unknown when simultaneously considering such a large group of factors" (NUREG/CR-6883 [3]). That is, the treatment of PSFs as independent significantly simplifies the (likely complex) interrelationships among PSFs and is the traditional convention. Notably, because of the multiplicative combination of the impact of multiple PSFs, the resulting HEP can, in some cases, exceed 1.0. However, an adjustment factor provided in SPAR-H worksheets allows analysts to reduce the effect of this impact if desired.

In summary, relevant to this effort, the approach used in SPAR-H models all PSFs in an equivalent fashion quantitatively (i.e., does not distinguish between base and modifier PSFs), combines the impact of those PSFs on performance in a multiplicative fashion, and treats PSFs independently.

2.1.2 Cognitive Reliability and Error Analysis Method

The cognitive reliability and error analysis method (CREAM) was developed in the late 1990s and, like SPAR-H, used a model of human cognition as its basis [10]. The CREAM method is based on a simplified set of cognitive functions that provide a profile of the cognitive demands of the task, as well as 13 CFMs that describe the ways the task might fail. These CFMs are modified by common performance conditions (CPCs), which are functional modifiers that determine the crew context. Although not entirely identical to the concept of PSFs, CPCs can be considered analogous to them, as they are meant to describe the contexts under which performance might vary. There are nine CPCs in total:

1. Adequacy of organization
2. Working conditions
3. Adequacy of machine-man interface and operational support
4. Availability of procedures/plans
5. Number of simultaneous goals
6. Available time
7. Time of day (Circadian rhythm)
8. Adequacy of training and experience
9. Crew collaboration quality

CREAM has two HEP quantification methods: basic and extended. This section describes the extended CREAM quantification method.

In the first step of applying CREAM, the analyst creates a cognitive demands profile of the task to determine the functions required for task completion. In the second step, the likely cognitive function failures are identified. Once the failures have been identified, the analyst can assess the probability of failure for each cognitive failure type (i.e., the cognitive failure probability). CREAM provides a set of nominal values for each cognitive function failure. Next, the analyst must account for the dependencies between the CPCs. This is done by identifying the probably control mode, for which there are four: strategic; tactical; opportunistic; and scrambled. The analyst identifies the control mode by determining the number of CPCs that reduce, improve, or have no effect on performance. Using the values in this triplet, CREAM provides a table that identifies the control mode. Once the control mode is identified, the cognitive failure probability is adjusted according to a specified weight. If desired, the analyst can take a more complex approach which accounts for the differences between the CPCs in the strength of their impact on the cognitive failure probability.

Unlike SPAR-H, which assumes independence among PSFs, CREAM proposes that there are relationships between CPCs. Hollnagel [10] outlines the relationships between CPCs and the conditions under which CPCs might be adjusted based on the effect of another CPC. For example, CPC working conditions depend on the CPCs adequacy of organization, adequacy of man-machine interface and operational support, available time, time of day, and adequacy of training experience. If four out of five of those CPCs affect performance in the same manner (e.g., reduce), the effect of working conditions also should be adjusted in the same manner.

In the basis for CREAM, Hollnagel [10] acknowledges that CPCs cannot conceptually be independent of each other, although that is the assumption made for most prior, first-generation HRAs. Using this approach, the total effect of the CPCs could be derived simply by summing the effects of the separate CPCs using appropriate weights. However, CREAM treats this issue using a more comprehensive description of human performance by defining how CPCs are coupled. Assessment of the effects of CPCs on cognitive function failures are made and assigned a weight from strong to weak. For cognitive function failures associated with more than one CPC, only the strong and medium impacts are considered. The total effect of the influence from the CPCs for each cognitive function failure is found by multiplying them. The method provides nominal values and uncertainty bounds for a range of 13 cognitive function failures based on examination of past HRA methods. The HEP is determined by summing applicable adjusted cognitive functions failures. The guidance warns that the resulting failure probability might still be greater than 1.0 but should be truncated at 1.0.

CREAM can be used qualitatively and or quantitatively, but limitations were highlighted for using it as a quantitative tool. The CREAM approach models all PSFs in an equivalent fashion quantitatively (i.e., does not distinguish between base and modifier PSFs) though PSFs can be weighted differently through a coupling process. The approach combines the impact of those PSFs on performance in a multiplicative fashion and treats PSFs independently.

2.1.3 Human Error Assessment and Reduction Technique and Nuclear Action Reliability Assessment

As the name implies, human error assessment and reduction technique (HEART) was developed to assess the likelihood of error, but much of the intent in developing the method was for reduction of errors [11] as opposed to quantitative predications. HEART begins with the assumption that human reliability is dependent on the generic nature of the task to be performed, and that given perfect conditions, reliability consistently will be within certain probabilistic limits of a nominal likelihood of error. However, because conditions generally are not perfect, reliability may decrease as a function of the conditions. A HEART assessment begins with a set of nine generic task types (GTTs), each associated with a nominal HEP. These basic probabilities of failure are modified by a set of error producing conditions (EPCs) for which there are 38 in total. Each EPC is assigned a multiplier ranging from 1.02 to 17. Then, the assessor determines the assessed portion of affect, which is essentially the likelihood that a given EPC might exist (e.g., a 30% chance that it applies to a given scenario). Using the EPC multiplier and assessed portion of affect, the assessed effect of the EPC is calculated using Eq. (3).

$$Assessed\ Effect\ (AE) = \left[\left(EPC_{multiplier} - 1\right) \times APOA\right] + 1 \tag{3}$$

The EPCs are treated independently, with the overall probability of error calculated using Eq. (4):

$$HEP = GTT_1 \times AE_1 \times AE_2 \times AE_3 \ldots \times AE_n \tag{4}$$

where GTT = the nominal HEP associated with the GTT.

HEART was developed to assess or estimate the reliability of human performance, but not to specifically assess activities in a nuclear power plant context. As a result, there are some nuclear power plant tasks that were not captured well using HEART. In response to this weakness, Kirwan et al. [12] developed the Nuclear Action Reliability Assessment (NARA). NARA was designed to modify HEART by incorporating more recent data, providing additional guidance in the development of GTTs and EPCs, and focusing specifically on the nuclear power plant context. However, the quantification of HEP within NARA is identical to HEART, although it is more concisely written as Eq. (5):

$$HEP = GTT \times \left\{ \begin{matrix} [(EPC_1 - 1) \times APOA_1 + 1] \times [(EPC_2 - 1) \times APOA_2 + 1] \times \ldots \\ \times [(EPC_n - 1) \times APOA_n + 1] \end{matrix} \right\} \tag{5}$$

Again, GTT here refers to the nominal HEP associated with the task type.

The HEART and NARA methods quantify the impact of all conditions or contexts in the same fashion, treating all EPCs in the same manner (i.e., through the use of weights) and combining the results multiplicatively. Notably, HEART lists the EPCs in order of importance—meaning that those EPCs listed first had the highest multiplier. In HEART, then, the most impactful EPCs were scenario familiarity, time available, a low signal-to-noise ratio, ease of information suppression, and ease of information assimilation. However, these EPCs are treated equivalently to any others with the exception of the varying multipliers associated with them.

2.1.4 Cause-Based Decision Tree

The cause-based decision tree (CBDT) method was developed as a complement to the human cognitive reliability (HCR [13]) method and is presented in Electric Power Research Institute (EPRI) TR-100259, *An Approach to the Analysis of Operator Actions in Probabilistic Risk Assessment* [13]. HCR was a time-based method that assessed situations in which time was a critical factor, whereas the CBDT method was designed to address situations where time was not a limiting factor. The HCR method determines HEPs directly using time reliability correlations, while the CBDT method determines HEPs from the likelihood of errors. The CBDT method uses a structured causal-based approach and addresses the diagnostic as well as execution portion of an operator action. The CBDT method uses two situation-specific failure modes each with four error mechanisms. The two failure modes and their associated error mechanisms are described below.

Failure mode 1: Failure of the plant information-operator interface

- Error mechanism a: Availability of information
- Error mechanism b: Failure of attention
- Error mechanism c: Misread/miscommunicate data
- Error mechanism d: Information misleading

Failure mode 2: Failure of the plant procedure-operator interface

- Error mechanism e: Skip a step in the procedure
- Error mechanism f: Misinterpret instruction
- Error mechanism g: Misinterpret decision logic
- Error mechanism h: Deliberate violation

Each error mechanism has an associated decision tree that asks questions about the failure mechanism primarily regarding PSFs that modify the likelihood of the failure mechanism. Starting on the left side of the tree, each decision adjusts the probability of error based on the choice made (i.e., a selected branch of the tree). Each branch results in a different end-state on the right side of the tree, and each end-state is given a nominal failure probability based on multiplication of the branch point probabilities leading to that end-state. The overall probability of failure (i.e., HEP) is the sum of all the applicable end-states for the eight error mechanisms. A sample decision tree is presented in Figure 1 [13].

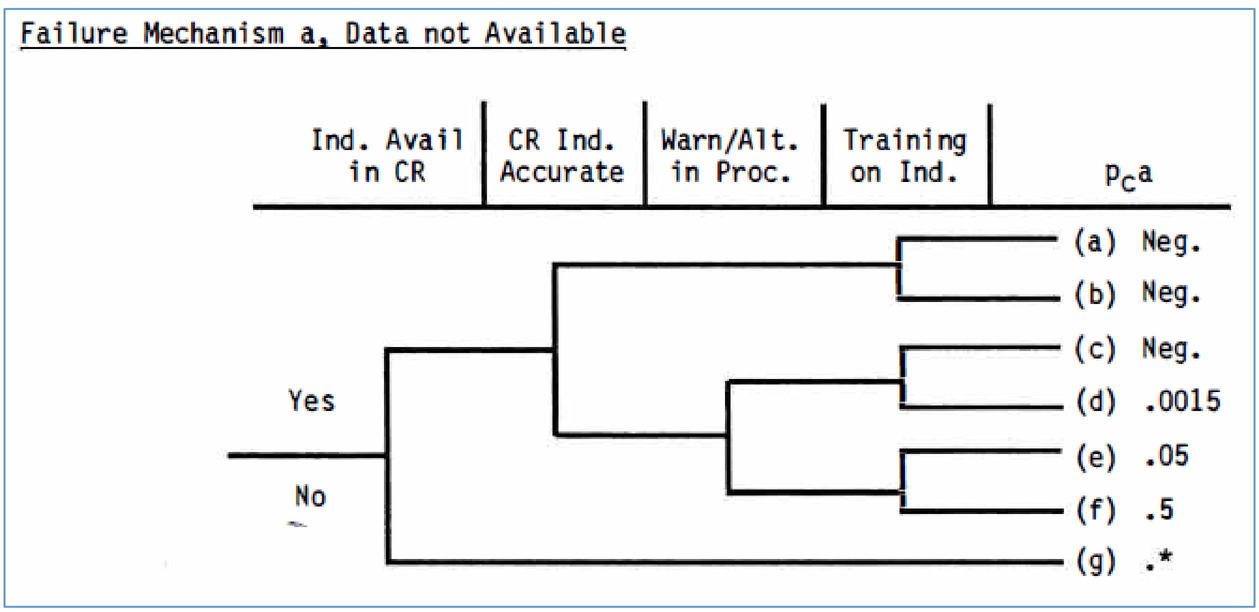

Figure 1. Decision Tree for Failure Mechanism a, Data Not Available (Parry et al. 1992)

Note: "Neg." in Figure 1 indicates that the impact of the error mechanism on the HEP is negligible, and thus the HEP is not adjusted.

If relevant, the resulting HEP can be adjusted for recovery from failure of the action. The probability of failure of the recovery action is evaluated in much the same way as the failure of the original action is evaluated.

Thus, the CBDT method differs from the other reviewed methods. Generally, it is similar to the technique for human error-rate prediction presented in THERP (NUREG/CR-1278 [8]) in its use of decision trees to quantify HEPs. In the CBDT method, the probabilities of the error mechanism decision tree branch points are treated independently, and the HEP is a simple arithmetic sum of the applicable nominal end-state probability results of the decision trees. No distinctions are made between the different error mechanisms in terms of their impact on the final HEP (i.e., all error mechanisms and associated decision tree branch points are treated equivalently and independently).

2.1.5 Success Likelihood Index Methodology

The success likelihood index methodology (SLIM) was developed in the mid-1980s to be implemented as part of a computer system called Multi-Attribute Utility Decomposition (together, SLIM-MAUD [14, 15]). SLIM relies on experts (e.g., HRA analysts) to evaluate the relative importance (i.e., weight) of different PSFs on a given HEP. The methodology involves a panel of experts—the assessors—who are presented with a detailed description of the tasks to be completed and the factors that might influence success within the scenario. The tasks may be classified in groups or subsets based on subsequent discussion. Then, the experts are presented with a random set of three tasks out of those that are to be assessed to encourage them to consider how the tasks differ from each other. Each expert is asked to rate the tasks on each PSF. There were six PSFs in the original formulation of SLIM: procedures; stress levels; time pressures; consequences (i.e., risk level); task complexity; and teamwork required. Once all the tasks have received a PSF rating, using Eq. (6), the ratings are multiplied by their respective weights, and the products are summed to give an overall success likelihood index (SLI):

$$SLI_j = \sum W_i \times R_{ij} \tag{6}$$

where SLI_j = the SLI for task j (j = 1…number of tasks),
W_i = normalized importance weight for the i^{th} PSF (the weights for all the PSFs sum to 1), and
R_{ij} = rating of task j corresponding to the i^{th} PSF.

The SLI represents, in numerical form, the experts' overall assessment of the likelihood of task success. It is a measure of the relative likelihood of success of the tasks presented in a given session. Using Eq. (7), the SLI can be subsequently transformed to a corresponding HEP estimate:

$$\log HEP = aSLI + b \tag{7}$$

where a and b are constants.

To determine a and b, at least two tasks for which the HEPs are known must be presented in each session, which are used to calibrate the SLI in order to transform it into a HEP. Embrey [14] notes that for this method of calculation to be valid, the PSFs must be independent. This assumption is therefore evaluated in the estimation process. If the scales for two PSFs are found to be too highly correlated, the assessor is asked to create a new, combined scale that captures both ratings.

Unlike the other methods, SLIM treats the combination of multiple PSFs additively (weighted sum) rather than multiplicatively. The weights multiplied with relative importance ratings are added together rather than multiplied. As part of this additive combination, the model enforces independence on the PSFs.

2.1.5 Summary

The literature review provides an overview of the way other HRA methods have approached the quantification of PIF effects. First, many of the methods selected the factors that influence performance in the model (e.g., PSFs, CPCs, EPCs) based on behavioral science literature and experience in the study of human error. The literature review found neither a standard set of PIFs among the models nor a clear ranking of hierarchy of importance. HEART rank orders the influencing factors based on the strength of their impact on human performance, but that ranking was not reflected in other models. Of course, given the variety of terms used across models and, in some cases, the lack of clear and precise definitions, it is not always obvious whether factors in different models refer to the same or similar concepts. Nonetheless, the review of the various HRA models neither reveals that there is a clearly identified set of underlying PIFs nor provides clear indications of relative importance among PIFs. Relevant to this effort, there was little indication in the literature that there are any PIFs that might be considered more fundamental than others in terms of their impact on human error. Furthermore, none of the models distinguish between the PIFs quantitatively by treating some as probabilities and others as modifiers. The impact of influencing factors is treated consistently within the method for all approaches as only the weight of the impacts differs.

Regarding review of the functional form used in combining the impact of multiple PIFs, the impacts are typically assessed in a multiplicative fashion. This appears to be largely due to historical convention, as THERP [8] modeled multiple PIFs in this way. SPAR-H [3], CREAM [10], HEART [4, 11], and NARA [12] all treat the combination of PSFs multiplicatively. In CBDT [13], the overall HEP is calculated by summing the individual HEP values from each decision tree, but the impact of factors associated with PIFs are factored in as branch point probabilities that are multiplied to determine the failure mechanism end-state probabilities. SLIM [14] is an exception because it combines multiple PIFs in an additive fashion. However, none of the models explore the functional form of the PIF combination. Recent work (e.g., [6, 7]) has examined existing literature to assess the degree to which findings support an additive or multiplicative model; however, there is not a consensus regarding the combination of PIFs in past HRA models. Another consideration impacting this question is whether the PIFs are represented as probabilities directly or if there is a weight representation for their effects.

Finally, all models except for CREAM [10] either assume or (in the case of SLIM [14]) enforce independence among the PIFs. Although SPAR-H [3] acknowledges that PIFs likely influence each other, the SPAR-H developers also note that those interrelationships are likely to be highly complex. In at least one case, independence is assumed for the sake of parsimony of the model. CREAM does consider the influence of CPCs on each other in a somewhat qualitative fashion using the notion of coupling strength of the CPCs to cognitive function failures. SLIM addresses the question of independence by requiring analysts to consider relationships amongst PIFs holistically, combining their influences when necessary.

2.2 Literature Exploring Multiple PIFs

There is a variety of existing research that has explored the impact of multiple PIFs in combination on human performance. Xing et al. [7] and Liu and Liu [6] reviewed a number of these papers; in addition, several papers are summarized in the IDHEAS-G report (NUREG-2198 [2]). For instance, using personnel from the U.S. Army Management Training Activity Department, Mount et al. [16] explored whether ability and conscientiousness relate to job performance independently or whether they interact. They found across two studies that there was no evidence for an interaction, suggesting that (in PIF terms), the influence of these factors should be considered independently and additively.

Other papers present meta-analyses of existing research to assess if specific combinations of PIFs have an additive or multiplicative effect on performance (e.g., Grether [17]; Van Iddekinge et al. [18]). Murray and McCally [19], Grether [17], Broadbent [20], and Hancock and Pierce [21] all explored the combined effects of different PIFs on performance. For example, Hancock and Pierce [21] specifically looked at the possible combinations of heat and noise on human performance. Although the results of these studies were somewhat mixed, they generally suggested that there was an additive effect when multiple PIFs were present. There was very rarely any indication of a greater-than-additive effect or subtractive effect when multiple PIFs were present. This provides some support for the notion that PIFs should be combined additively within HRA models.

However, many of the studies focused on only one PIF, or two PIFs in combination, and were not designed to address the broader question of how effects of PIFs in general should be combined. Here, we hope to understand through semi-structured interviews how and why developers of different HRA methods selected a specific approach for PIF effect quantification to help assess the technical basis of combining the effects of PIFs on human performance in an additive versus multiplicative manner in IDHEAS-G.

3 INTERVIEWS WITH HRA MODEL DEVELOPERS

The primary goal of the semi-structured interviews with HRA method developers was to better understand the quantifications used in the models they developed or researched, particularly as it pertains to factors that may influence human performance.

3.1 Participants

Participants were developers of HRA methods. They were selected due to their expertise and knowledge of the creation and use of HRA methods and the field in general. Nine interviewees were contacted, and eight agreed to participate. Table 1 shows the models discussed during the interviews. All interviews focused on a single model, with two exceptions: the second interview (with a HRA researcher) focused on HRA methods generally (noted as "General" in Table 1) and the fifth interview discussed both the CBDT method and the IDHEAS-At Power model.

Table 1. HRA Methods Discussed During the Interviews

Interview #	Method
1	SPAR-H
2	General
3	CREAM
4	SPAR-H
5	CBDT/IDHEAS-At Power
6	HEART
7	NARA
8	SLIM

3.2 Procedure

Interviews were conducted virtually via Microsoft Teams™ and were scheduled for no more than 90 minutes. Three members of the PNNL research team attended the interviews, during which the PNNL researchers took notes to capture participant responses. Prior to the interview, participants were given a general description of the research effort and the questions to be discussed. Participants were provided with an informed consent document and assured that participation in the discussion was voluntary.

The interviews began with introductions followed by an overview of the topic of the discussion. The participant was reminded of the length of the interview and given an opportunity to ask clarifying questions about the study purpose. Then, the interviewers proceeded with the questions. The questions varied slightly for each HRA method by using terms and language consistent with the method being discussed. For example, some methods call the variables that influence the likelihood of successful performance "performance influencing factors," while other methods use the terms "performance shaping factors" or "error producing conditions." Throughout, the interviewers asked additional clarifying questions to better understand participants' responses.

The basic questions used during interviews are listed below.

- In [method name], how were the PIFs selected? Are there any PIFs that are more important than others? Are there any differences in the ways that different PIFs are used to modify the probability of human error?
- Can you explain the reasoning or justification for how [method name] combines multiple PIFs together for a single human action (e.g., additive, multiplicative)?
- [Method name] treats multiple PIFs as [independent/non-independent]. What was the reason for that decision?
- In [method name], only the negative effect of PIFs was considered. Can you explain the reasoning behind that decision? Should positive PIFs be quantified differently than negative PIFs?

3.3 Analysis

After completion of each interview, notes from the PNNL researchers were combined and shared with the participant to ensure completeness and accuracy. These comprehensive notes of the interviews are archived in the project records. Upon completion of the interviews, the notes were consolidated into a summary document and reviewed to identify patterns and synthesize the results of the conversations. This consolidated summary document is presented in Appendix A.

4 RESULTS

Results of the interviews are summarized in this section by the topics of the questions: selection of PIFs and the relative importance of PIFs; combination of PIF effects (mathematically); interdependence of PIFs; and consideration of positive versus negative impact of PIFs. For simplicity, the factors that influence performance generally are referred to as PIFs; however, within the context of the interview or a method, a different term (e.g., EPCs, PSFs) may have been used.

4.1 Selection of PIFs

When discussing the selection of PIFs within the context of a method, developers described this as either a theory-driven (i.e., top-down) or data-driven (i.e., bottom-up) process. In some cases, there was a model of human cognition used to drive the selection of factors that might influence performance. In other cases, a review of the literature revealed patterns that helped to identify the factors that were empirically shown as impacting human performance (i.e., the likelihood of failure). In at least one case, the initial selection of PIFs was verified through data from probabilistic risk assessment (PRA) and from HRA analysts' expert knowledge. Many of the other methods later verified the PIFs with analysts or with data in some fashion; however, only a single developer referred to this as driving the selection of PIFs from the outset.

Generally, the developers suggested that PIFs serve to describe or operationalize the context or conditions that have an impact on human performance. PIFs were included in the HRA methods depending on the perspective of the developer and the research that was available at the time the method was developed. For example, one method (i.e., SPAR-H) relied on information-processing theory to drive the factors that influence performance. Others relied heavily on learning, memory, and response time literature (i.e., cognitive psychology research) to identify patterns in the factors that contribute to the likelihood of error. In all cases, the intent of the PIFs was to identify the underlying factors that the developers believed had an important impact on performance. PIFs were identified using theory (e.g., cognitive psychology) or the patterns identified in empirical research (e.g., human performance research), or some combination of the two.

HRA methods employ different PIFs and number of PIFs, and certain methods employ a large number of PIFs compared to other methods. However, developers generally agreed for any given operator action there are likely only a few PIFs that have a significant impact. The importance of addressing further PIFs diminishes as more are considered. Developers seemed to acknowledge that some degree of variability exists in how the PIFs are categorized and described. There is some emerging consensus about the PIFs that are important and that should be considered in HRA; however, no strong attachment (in most cases) to specific PIFs used in any particular method seemed to exist. Instead, the developers emphasized that use of PIFs is a way to understand and characterize the conditions under which an action is performed. They suggested that data and theory can be used to identify general factors for consideration when attempting to assess the likelihood of operator failure.

4.2 Relative Importance of PIFs and Quantification Structure

The next questions explored the relative importance of PIFs and whether there are any PIFs that should be quantified differently within the context of an HRA method. Participants were asked whether there were any PIFs that were more important than others or that should always be evaluated, and if so, which PIFs those were. Overall, the participants did not think that any PIFs should always be evaluated, but in general, they thought some PIFs may be more important than others. Almost all participants emphasized that all PIFs in a method need to be considered and evaluated, because depending on the conditions, any PIF may be important to the likelihood of failure. Developers also were asked whether there were any PIFs that should be quantified differently in the method, and their general response was no. One developer stated that the impacts of some PIFs are addressed qualitatively as compared to quantitatively due to the nature of the factor. However, no developers suggested any PIFs should be handled differently in a mathematical sense in terms of their impact on the probability of error. That is, although the PIFs varied in the strength of their impact on the HEP, they all modified the HEP in the same manner.

Some methods (HEART and NARA, specifically) contain a relative ranking of importance for PIFs, and the developers of those methods referred to this rating as reflecting the general importance of these factors. Unfamiliarity with the situation and time availability were highlighted as highly influential. In addition, these developers mentioned the importance of procedures and the plant controls (i.e., the interface). In contrast, another participant explicitly stated that PIFs are not hierarchical, although some may have a greater impact than others. In addition, some PIFs may have a more proximate or distal impact on performance than others. Although there was some acknowledgement that there are factors that tend to have a greater impact on performance, overall there was no indication from the developers that any PIFs should be treated functionally (i.e., additive or multiplicative) differently due to the nature of their impact.

Several interviewees discussed the impact of time as potentially overwhelming; that is, if time was not available to complete a task, then this factor overcomes the influence of any other PIF as it essentially guarantees operator failure. This observation still was not considered justification for treating any specific PIFs differently in an HRA. Instead, participants emphasized the importance of considering the situation and determining the influential factors, and noted that depending on the context, any given PIF might have a large impact on error.

4.3 Combinations of PIFs

The next questions to the participants were about the methods by which effects of multiple PIFs are combined within a method. Most methods examined in this study use a multiplicative approach versus an additive approach, where the impact of multiple PIFs is combined in a product. The participants generally acknowledged that there was no strong mathematical justification for the way that effects of PIFs were combined in a method; instead, the decision was largely historical, as this is how effects of multiple factors have traditionally been combined. In addition, some participants preferred to avoid strongly focusing on quantification altogether, emphasizing instead a qualitative assessment that relied on quantification only for the purposes of communication and integration into a PRA.

In addition, several of the participants acknowledged that a multiplicative combination of PIF effects can occasionally result in HEPs that exceed 1.0; in some methods, this situation is addressed with a correction factor when there are many PIFs (i.e., more than three) that are considered relevant to performance in a given scenario. Although this situation is problematic from a mathematical standpoint, the participants expressed that this was simply an indication that, given the PIFs, there is a high probability of error. The quantification of the HEP was not considered to be a primary concern. Instead, the high HEP indicates a need for intervention in the scenario to reduce the likelihood of error (i.e., a practitioner focus).

Overall, the interviews suggested that there was little empirical evidence to support any specific method of mathematically combining effects of multiple PIFs. Participants described that the multiplicative method was selected for historical convention in the field of HRA and general conservatism of the resulting numbers (i.e., the tendency to produce relatively higher estimates of HEPs than an additive approach). Moreover, the question of how effects of PIFs are combined was not a central concern. Many participants expressed an interest in deemphasizing the quantitative application of methods and focusing instead on using the methods qualitatively, with the quantification necessary only for consideration within a PRA. Absolute precision is not always needed from an HRA for evaluation of operator errors modeled in a PRA, because not all errors are risk significant. Also, the PRA propagates the parametric uncertainty of failures that are modeled including operator errors; these approaches help lessen the impact of the imprecision in HRA.

4.4 Independence versus Interdependence of PIFs

Almost all participants acknowledged that PIFs are interdependent rather than independent or orthogonal. A single developer argued for independence insofar as there is little empirical evidence suggesting that multiple PIFs combined create a greater-than-additive effect; that is, there was no indication that PIFs produced a stronger effect in combination than when considered independently. All but one of the participants acknowledged that PIFs are interrelated and that they can either exacerbate or (in some cases) ameliorate the effects of other PIFs. However, all methods explored here treat PIFs as independent. Participants generally described that this decision was made because the interrelationships among PIFs are highly complex, and that there is no clear way to model those interrelationships mathematically. Attempts to do so did not result in clear evidence for any specific interrelationships for inclusion in an HRA method.

CREAM was the only method that explicitly considered the interaction of PIFs, but did so primarily in a qualitative manner using the notion of coupling strength of the CPCs to cognitive function failures described in its documentation [10]. In addition, one participant observed that the interrelationships among PIFs are currently being modeled using Bayesian belief networks in so-called third-generation HRA methods. Another developer suggested that additional research might support a theoretical model for which the factors are mostly, if not completely, orthogonal (e.g., through factor analysis).

Overall, the participants acknowledged the limitation that most HRA methods do not consider interactions among PIFs, but that the interactions do exist. The decision to consider them independently was based on the pragmatic need to keep the methods themselves simple given the complexity of the interrelationships amongst PIFs. Generally, the results of the interviews suggest that more research is needed to understand and accurately model the relationships among the PIFs. One developer in particular suggested that experiments using triads of PIFs might be beneficial to understanding how multiple PIFs are interrelated without the exploration becoming too complex.

4.5 Positive versus Negative PIFs

The interviews with participants also explored the question of both positive and negative impacts of PIFs on performance. Overall, the participants expressed there can be conditions that improve performance or compensate for other negative factors; however, they also noted that defining a reference or a baseline level of performance was a significant challenge. That is, understanding what a better-than-baseline level of performance might be is difficult even though some factors can clearly improve or degrade performance. One developer mentioned that application of a positive effect is a popular feature among nuclear power HRA practitioners. However, to accurately model positive impacts, it is necessary to first develop a point of reference (e.g., a baseline performance level), which was considered difficult.

Certain discussion about the positive effects of PIFs focused on ways that factors might help an operator recover from a potential error, rather than improving performance above some level. However, some participants noted that there was little empirical evidence to support how and when positive influences should be considered and that there was little consistent data that could be applied to addressing those questions. That is, positive influences may exist, but they have not reliably been found in the literature or in past events, and thus there is no evidence about how they should be quantified. Regardless, participants noted that if positive influences are included, they should be handled quantitatively in the same way as negative influences.

5 CONCLUSIONS AND RECOMMENDATIONS

In this study, PNNL interviewed HRA method developers and reviewed the literature to evaluate the technical basis of the IDHEAS method with respect to three elements:

1. The division of PIFs into base and modifier PIFs
2. The functional form of the impact of multiple PIFs (i.e., additive or multiplicative)
3. The independence (or non-independence) of PIFs.

Below we review the implications of the findings.

5.1 Base versus Modifier PIFs

None of the HRA methods reviewed here include a distinction between base and modifier PIFs comparable to the one proposed in IDHEAS. However, the participants were asked about the differences between relative importance and whether there are PIFs that should be quantified differently in the methods. In general, the participants did not distinguish between the PIFs. As already mentioned, HEART (and also NARA) does rank the PIFs in order of importance based on the weight of their perceived impact on an HEP. However, there was no method that proposed a separate quantification for any of the PIFs in a way similar to IDHEAS.

The traditional method for developing a base HEP is to use a task analysis to understand how often, on average, a task results in success or failure (e.g., GTT in HEART). IDHEAS-G, in contrast, proposes these three base PIFs as being major causes of error based on signal detection theory (NUREG-2198 [2]). However, there was no clear opinion from the participants or in the literature that there is a distinction between base versus modifier PIFs, or that any PIFs might be so fundamental as to require evaluation in every scenario. Instead, the task context was emphasized as being critical for identifying relevant PIFs, and the importance of the PIFs was informed by that context. Again, however, the IDHEAS-G framework is a novel approach to the calculation of HEPs through the base PIFs; the fact that prior methods do not take this approach is not a surprise. Regardless, there was no strong indication from the interviewed participants that any PIFs should differ in their functional form.

The current functional form in IDHEAS-G does affect the resulting HEPs because of distinctions made between the base and modifier PIFs. In the current formulation, the impacts of base PIFs are treated as probabilities, whereas impacts of modifier PIFs are expressed as multiplicative effects or ratios of error rates. A quantitative simulation[2] was conducted to explore the potential for the HEPs to exceed a probability of 1.0 depending upon the weights assigned to modifier PIFs. Results showed that, depending on the relevant CFM, the percentage of scenarios in which the HEP exceeded 1.0 ranged from approximately 1% to nearly 50%. This suggests that the quantification currently used in IDHEAS does result in HEPs that exceed a probability of 1.0 in severe cases. Thus, some modification to the HEP formulation to address this issue might be warranted.

[2] Additional details on this exploration are presented in Appendix B.

5.2 Additive versus Multiplicative

Although some past literature does explore the additive versus multiplicative combination of PIFs and their impact on HEP (e.g., [6]), the participants who participated in this study did not express strong beliefs or opinions on this topic. In general, they reported that the functional form in their methods (often multiplicative) was a matter of historical convention. Also, most reported that PIFs likely interact, although the nature of that interaction is largely unknown. The literature on this topic was similarly limited. Although some studies and reviews explored one or two PIFs in a single scenario or experiment, there is no definitive answer to the question of how *all* PIFs interact or how they should be quantified in combination. In fact, outside of the two studies previously mentioned (i.e., [6, 7]), few papers have directly discussed or explored this topic.

There is clearly a link between the quantification of PIFs and their independence or interaction. If PIFs are independent, an additive form for combining their effects is appropriate. If they are not independent, then a multiplicative form may be appropriate. Thus, to address the question of the functional form of the impact of multiple PIFs on HEP, it is important to understand how the PIFs are interrelated. One of the participants suggested that a family of experiments to test different combinations of three PIFs could help to elucidate relationships among different PIFs. In the current literature, such a systematic exploration has not been reported. Additional data are necessary to provide a clearer answer to this question. The results here suggest that in past HRA approaches, a multiplicative form is more common; however, papers that explicitly explore this question seem to indicate that an additive form is more appropriate.

5.3 Independence versus Non-independence

Most HRA methods treat PIFs as independent; however, study participants (with a single exception) expressed the belief that PIFs are likely interdependent. As previously mentioned, additional data collection would be informative. Again, little research has explored this question for different combinations of PIFs. The paucity of evidence of greater-than-additive (or subtractive) effects of PIFs in the current literature is not necessarily an indication that such effects do not exist. The intuition of the participants who participated in this effort, albeit with little empirical data, was generally that PIFs are likely to interact in complex ways, and that additional research on this topic would be of value to the HRA community.

5.4 Recommendations

The following recommendations are provided based on the results of this research:

- Additional data collection is required to provide a technical basis or justification for the distinction between base and modifier PIFs. The results presented in this study do not support this proposition but do not necessarily exclude it as a possibility. Because the existing data have not examined this question in depth, additional research is required.
- There is a clear link between the independence or interdependence of PIFs and the functional form of their impact on HEP. The existing literature suggests that PIFs combine additively. However, the results of interviews conducted in this study suggested that PIFs do interrelate; thus, a multiplicative form might be appropriate. A systematic experimental exploration of this question would provide additional information that cannot be obtained solely through a review of the literature. A potential family of experiments exploring triads of PIFs might be useful to explore their interactions.

6 REFERENCES[3]

1. Xing, J., Y.J. Chang, and J. DeJesus, *Integrated Human Event Analysis System for Event and Condition Assessment (IDHEAS-ECA)*, 2020, NUREG-2256, U.S. Nuclear Regulatory Commission, ADAMS Accession No. ML22300A117: https://www.nrc.gov/docs/ML2230/ML22300A117.pdf.
2. Xing, J., Y.J. Chang, and J. DeJesus, *The General Methodology of an Integrated Human Event Analysis System (IDHEAS-G)*, 2021, NUREG-2198 U.S. Nuclear Regulatory Commission, ADAMS Accession No. ML21127A272.
3. Gertman, D., H. Blackman, J. Marble, J. Byers, and C. Smith, *The SPAR-H Human Reliability Analysis Method*, 2005, NUREG/CR-6883 (ADAMS Accession No. ML051950061), U.S. Nuclear Regulatory Commission: https://www.nrc.gov/docs/ML0519/ML051950061.pdf.
4. Williams, J. *HEART: A Proposed Method for Assessing and Reducing Human Error*. in *The 9th Advances in Reliability Technology Symposium*. 1986. University of Bradford.
5. Xing, J., Y.J. Chang, and J. DeJesus, *Draft - Integrated Human Event Analysis System for Human Reliability Data (IDHEAS-DATA)*, 2020, RIL-2021-XX U.S. Nuclear Regulatory Commission, ADAMS Accession No. ML20238B982.
6. Liu, P. and J. Liu, *Combined Effect of Multiple Performance Shaping Factors on Human Reliability: Multiplicative or Additive?* International Journal of Human–Computer Interaction, 2020. **36**(9): p. 828-838: https://www.tandfonline.com/doi/full/10.1080/10447318.2019.1695461.
7. Xing, J., Y.J. Chang, and N. Siu. *Insights on human error probability from cognitive experiment literature*. in *2015 International Topical Meeting on Probabilistic Safety Assessment and Analysis (PSA 2015)*. 2015.2015/04/26/30.
8. Swain, A.D. and H.E. Guttmann, *Handbook of Human Reliability Analysis with Emphasis on Nuclear Power Plant Applications*, 1983, NUREG/CR-1278 (ADAMS Accession No. ML071210299), U.S. Nuclear Regulatory Commission: https://www.nrc.gov/docs/ML0712/ML071210299.pdf.
9. Swain, A.D., *Accident Sequence Evaluation Program Human Reliability Analysis Procedure*, 1987, NUREG/CR-4772, U.S. Nuclear Regulatory Commission, ADAMS Accession No. ML071630060: https://inis.iaea.org/collection/NCLCollectionStore/_Public/19/005/19005695.pdf.
10. Hollnagel, E., *Cognitive Reliability and Error Analysis Method (CREAM)*. 1998, Elsevier.0-08-0428487.
11. Williams, J. *A data-based method for assessing and reducing human error to improve operational performance*. in *IEEE Conference on Human Factors in Power Plants*. 1988. Monterey California.June 5-9, 1988.
12. Kirwan, B., H. Gibson, R. Kennedy, J. Edmunds, G. Cooksley, and I. Umbers. *Nuclear Action Reliability Assessment (NARA): A Data-Based HRA Tool*. in *The 7th Probabilistic Safety Assessment and Management*. 2004. Berlin, Germany: Springer.June 14-18, 2004.

[3] Publicly available NRC published documents are available electronically through the NRC Library on the NRC's public web site at http://www.nrc.gov/reading-rm/doc-collections/. The documents can also be viewed online or printed for a fee in the NRC's Public Document Room (PDR) at 11555 Rockville Pike, Rockville, MD; the mailing address is USNRC PDR, Washington, DC 20555; telephone (301) 415-4737 or (800) 397-4209; fax (301) 415-3548; and e-mail pdr.resource@nrc.gov.

13. Parry, G.W., B.O.Y. Lydell, A.J. Spurgin, P. Moieni, and A.N. Beare, *An approach to analysis of operator actions in probabilistic risk assessment*, 1992, TR-100259, Electric Power Research Institute.
14. Embrey, D.E., P. Humphreys, E.A. Rosa, B. Kirwan, and K. Rea, *SLIM-MAUD: An Approach to Assessing Human Error Probabilities Using Structured Expert Judgment. Volume I: Overview of SLIM-MAUD*, 1984, NUREG/CR-3518, Vol. I, U.S. Nuclear Regulatory Commission, ADAMS Accession No. ML071220404.
15. Embrey, D.E., P. Humphreys, E.A. Rosa, B. Kirwan, and K. Rea, *SLIM-MAUD: An Approach to Assessing Human Error Probabilities Using Structured Expert Judgment. Volume II: Detailed Analysis of the Technical Issues*, 1984, NUREG/CR-3518, Vol. II, U.S. Nuclear Regulatory Commission, ADAMS Accession No. ML071220406.
16. Mount, M.K., M.R. Barrick, and J.P. Strauss, *The Joint Relationship of Conscientiousness and Ability with Performance: Test of the Interaction Hypothesis.* Journal of Management, 1999. **25**(5): p. 707-721: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.455.967&rep=rep1&type=pdf#:~:text=Overall%2C%20the%20preponderance%20of%20evidence,than%20those%20who%20are%20not.
17. Grether, W.F., *Effects on Human Performance of Combined Environmental Stress*, 1970, AMRL-TR-70-68, Aerospace Medical Research Laboratory: https://apps.dtic.mil/dtic/tr/fulltext/u2/727041.pdf.
18. Van Iddekinge, C.H., H. Aguinis, J.D. Mackey, and P.S. DeOrtentiis, *A Meta-Analysis of the Interactive, Additive, and Relative Effects of Cognitive Ability and Motivation on Performance.* Journal of Management, 2018. **44**(1): p. 249-279: http://journals.sagepub.com/doi/10.1177/0149206317702220.
19. Murray, R. and M. McCally, *Combined environmental stresses.* 1973: https://www.researchgate.net/publication/23622718_Combined_environmental_stresses.
20. Broadbent, D.E., *Differences and interactions between stresses.* Quarterly Journal of Experimental Psychology, 1963. **15**(3): p. 205-211: http://www.tandfonline.com/doi/abs/10.1080/17470216308416325.
21. Hancock, P.A. and J.O. Pierce, *Combined Effects of Heat and Noise on Human Performance: A Review.* American Industrial Hygiene Association Journal, 1985. **46**(10): p. 555-566: https://www.researchgate.net/profile/Peter_Hancock2/publication/19101231_Combined_Effects_of_Heat_and_Noise_on_Human_Performance_A_Review/links/556db0e408aefcb861d9ad94/Combined-Effects-of-Heat-and-Noise-on-Human-Performance-A-Review.pdf.
22. Xing, J., G. Parry, M. Presley, J. Forester, S. Hendrickson, and V. Dang, *An Integrated Human Event Analysis System (IDHEAS) for Nuclear Power Plant Internal Events At-Power Application, Volume 1*, 2017, NUREG-2199, U.S. Nuclear Regulatory Commission; Electric Power Research Institute, ADAMS Accession No. ML17073A041.
23. Hall, R.E., J. Fragola, and J. Wreathall, *Post Event Human Decision Errors: Operator Action Tree/Time Reliability Correlation*, 1982, NUREG/CR-3010; BNL-NUREG-51601, U.S. Nuclear Regulatory Commission.
24. Ekanem, N.J., A. Mosleh, and S.-H. Shen, *Phoenix - A model-based human reliability analysis.* Reliability Engineering and System Safety, 2016. **145**: p. 301-315.

APPENDIX A – INTERVIEW RESULTS

As discussed in Chapter 3 of this report, interviews were conducted with seven participants of HRA methods. Results of the interviews are summarized by the HRA method that was discussed. These results are summaries, not detailed accountings of the interviews. The intent is to capture the nature of the participants' responses relative to the primary research questions of interest.

A.1 Standardized Plant Analysis Risk Human Reliability Analysis (SPAR-H)

- How were PIF selected?
  - Selection of PIFs came from a cognitive model; developed a basic information processing model to describe the nature of human knowledge, memory, performance, etc. Then, the cognitive model was applied to a nuclear power plant context, examining the actions that occur in that context, and the potential actors that impinge on performance. This allows you to identify the PIFs and what is important. Information processing informs the factors that influence performance.
  - Operational factors are the intervening factors between the PIF and performance—they describe how the PIFs influence human performance.
- Why are time available, stress, and complexity always evaluated, and other factors evaluated only when relevant? Are there differences between the PIFs in their importance?
  - This decision was made following initial development, and the developer did not necessarily agree with it. This appears to be an artifact of the SPAR models, but neither the developer nor the researcher agreed with it.
  - Generally, all factors should be evaluated each time (i.e., considered in an analysis); analysts should evaluate the factors that may have an impact.
  - There are some factors, such as time available, that can overwhelm the others regardless of their level, but they all need to be considered as they may all have an impact.
  - PIFs are a way of operationalizing context; as such, depending on the context, any one of them might be important.
- What is the reasoning between how PIFs are combined in the model, and were other avenues explored?
  - There is mathematically no clear justification. The original source of the error rates was the technique for human error rate prediction (THERP [8]); they continued to combine PIFs in that way because that is the way it has traditionally been done, and the data supported the decision.
  - The PIFs should be a product because it is a good way of accounting for the counterbalancing and the ways that the PIFs may interact. There is not a more complex way that is currently available and accounts for the relationships amongst PIFs.
  - There is a corrective factor when there are many PIFs because the error rate cannot exceed a probability of 1.0; that came from statisticians.
- Why were positive impacts of PIFs considered in the model?

- This came from discussions with the regulator; it made sense to include positive effects as well as negative ones. It is important to consider how things might improve. It should be done in the same way for positive and negative PIFs (i.e., quantitatively equivalent).
- It is important to give credit to the operator but determining the baseline performance or the reference point is challenging. What is the baseline, and at what point should estimates of performance be improved?

- Why are PIFs treated as independent in SPAR-H?
  - The PIFs are interdependent (interact), but there was not a clear way to account for the interaction in the model, and the criticism of the model based on that is valid. The PIFs were selected to be as independent as possible, but with human behavior, they are not completely independent. Additional measurement development (e.g., through factor analysis) might help to account for the interdependencies to create more orthogonal factors.

A.2 Cognitive Reliability and Error Analysis Method (CREAM)

- How were PIFs selected?
  - The need for HRA arises out of PRA. It uses a system analogy to describe human behavior—just as machines or systems sometimes fail so do humans. PIFs have not been selected or created in any systematic way; they were created as a means to explain performance. Rather than doing that, in the second generation of HRA methods, developers explored the PIFs as the signal, not the noise. This removes human error from the equation and focuses on the context. That is why in CREAM they are called CPCs rather than PIFs. The CPCs were intended to be a manageable number of conditions that affect performance. They were designed to capture what it is that influences what people do.
- Are there differences in the effects that different conditions might have on the probability of human error?
  - It is not sensible to talk about error probability. If you look at a specific activity in a specific situation or context, and if it is a regular activity that happens many times, you might say there is a probability of doing things in a certain way. This is different than defining it as an error. However, time pressure is always a factor because there is an efficiency tradeoff.
  - It is important consider the factors that make work difficult rather than focusing on human error. The factors that make work difficult are going to depend heavily on the type of job the person is completing.
- What is the reasoning between how PIFs are combined in the model, and were other avenues explored?
  - Human performance is variable, and the variability is systematic, not random. There is a need to describe functions rather than task steps; a function is something that is being done, and that can be variable in the way that it is being carried out. Rather than focusing on human error, examine the functions involved in doing something and describe how the function is carried out. This allows quantification.
  - The current proposed approach is to use the proposed functional resonance analysis method to describe how a function is carried out; quantification may result from that analysis but should only be done once the functions are described.

- Quantification is important primarily for communication but is not central, especially because there is a lot of variability in the accuracy of quantifications.

- Why were positive impacts of PIFs considered in the model?
  - It was a way to account for the fact that sometimes things go well. Many approaches handle positive influences differently than negative ones; if a model contains both, it should handle them equivalently.
- The CPCs operationalize context. Is that right? How were they related?
  - When CREAM was developed, there needed to be a manageable number of CPCs because we needed to consider their interdependence—because they are not independent. The HEP is just the summation of PIFs. We need to account for their independence, but we need a practical number of PIFs.
  - You need to think about what has an impact on what people do—what is it that influences what people do. It is things like goals, workload, adequacy of the organizational procedures, crew collaboration quality, variability, training, adequacy of the human-machine interface. They all influence each other. If you really want to use the idea of conditions that affect performance, you need some kind of model or theory that explains how these conditions are interrelated. Take any specific condition such as available time, which is something that everyone would say or concede is important. Available time is just an expression of the conditions. It is the time that the person who is doing the work think they have available. It is not a metric of time; it is psychological time. If I think I have little time available, I think I have to do it very quickly.
  - There were around 12 CPCs in CREAM, and there are some diagrams about how they interconnect.

A.3 <u>Human Error Assessment and Reduction Technique (HEART)</u>

- How were EPCs selected?
  - Essentially, they emerged from a literature review of experiments exploring response time, learning, and memory and examining it from the perspective of human error. After reviewing and categorizing the papers, patterns emerged, these patterns showed the mechanisms of human failure, and there were common factors across experiments. This later resulted in HEART.
- Are there any differences in the importance of EPCs?
  - Some factors are extremely important, like unfamiliarity. There can be variability depending on the context; procedures can be important, and controls can be important. The factors should be applied in the same way; the factors that are important need to be considered.
- What is the reasoning between how EPCs are combined in the model, and were other avenues explored?
  - It was simple and intuitive. The possibility of using percentages was explored, but people had difficulty understanding it. Occasionally it results in a probability higher than one, but that generally indicates a serious problem that needs to be considered. The precision of the numbers is less critical than identifying the most influential factors.
- Why were positive impacts of factors not considered in the model?

- It's very difficult to find evidence of positive impacts on performance. It matters, but it's not clear how much or when, and so it was difficult to quantify.

- Why are EPCs treated as independent in HEART?
  - There is no evidence of super-additivity—that is, of factors combining to be larger than the sum of their parts. There is some evidence for a subtractive effects (e.g., noise canceling the effects of sleep deprivation) but there is no consistent evidence of super-additivity.

A.4 Cause-Based Decision Tree / Integrated Human Event Analysis System–At Power (CBDT/IDHEAS-At Power)

- How were PIFs selected (for IDHEAS-At Power [22])?
  - PIFs were primarily chosen out of the cognitive psychology literature. Cognitive failure modes (CFM) were chosen to represent the ways to determine that yes, there had been a failure. The PIFs that related to those came out of the cognitive science literature.
  - IDHEAS has this notion of base versus modifier PIFs. What do you think about that idea? (I should have questioned this because I do not recall the concept of base versus modifier PIFs. Maybe this comes from IDHEAS-G?)
  - The PIFs are categories. They're names. The way we dealt with that in IDHEAS-At Power because they could affect things differently depending on the failure mode, depending on which aspects they would affect. We tried to explain which aspects of the PIFs would be most relevant to a particular CFM by the guidance questions in the description of the trees – workload might be a good example.
- Are there any error mechanisms that should be captured or represented differently in quantitative models?
  - When we first developed IDHEAS, we were not trying to quantify it (at least not to my recollection, although that might have been the project managers' goals). We were trying to build a structure to help people decide which combinations of PIFs and which aspects of those could result in a failure. Obviously, with different combinations of PIFs for a particular CFM you get an ordering of the likelihood of failure. We were structuring it to identify what the factors that would increase or decrease the likelihood of a failure. To some extent, take the precise numbers with a grain of salt. It was more qualitative than quantitative because it would not have been data based. My preference, although it does not fit into the PRA paradigm well, is to have more of a relative scale of things.
  - CBDT combines multiple PIFs in the decision trees. Is that right? Models have different ways of combining PIFs. Can you talk about how you combined them?
  - There was not a real basis for the quantification of CBDT; the THERP probabilities were used as an example.
  - The IDHEAS approach looked at the whole sequence and considered it holistically. It is not additive or multiplicative. They are combined effects. The combined effects were what the expert panels tried to look at when they came up with the probabilities of the endpoints. I do not think the impacts are strictly additive or multiplicative; it is more complex with than that.
- CBDT/IDHEAS considers only the negative impact of error mechanisms, rather than any potential positive impact of context. Did you consider including a positive impact of context on performance?

- Positive influences in IDHEAS come in through the recovery factors. The baseline assumption is that the operators do what the procedures tell them to do correctly. You can't improve on that. What you can do is if there are complications in the context in which they are working, can you find compensating factors? That is what we focused on in IDHEAS. We were not trying to look at improving performance because I do not know if that can be accomplished. If they press the button within the requisite number of seconds, does it matter if they do it a second earlier? Probably not.
  - I suppose that if you were starting with some base probability for an average, but not necessarily optimal context, you could include a positive influence, but we did not structure IDHEAS-At Power that way.
- In these environments, you studied and looked at these causes of errors. Were there some errors that stuck out to you as always being there? Or not necessarily always there, but really important to consider?
  - Not really. We were looking at the cognitive literature and worked from that basis.
- What about relationships among the PIFs?
  - PIFs are definitely related in the sense that they make the impact of others worse. If you're working in a high workload, high stress environment—that is going to make it worse. How much worse is probably dependent on context.

A.5 Nuclear Action Reliability Assessment (NARA)

- How were PIFs selected?
  - Started with THERP—in practice, people overloaded stress as PIF. Starting with HEART, there were 38 EPCs. NARA was an attempt to identify the factors that mattered in the nuclear industry, and we used data from PRAs from existing plants. There were 18 PSF that were identified as useful, and those were verified with about 30 assessors (HRA analysts who are considered experts in the field) in the United Kingdom who agreed that the other factors were of academic interest only.
- Are there any EPCs that are more important than others in the model?
  - Yes and no. Both in NARA and SPAR-H, if there is an unfamiliar situation—something they have never encountered before, that is a real challenge. In HEART and NARA, it is called unfamiliarity; it's a novel situation that they've never considered before. The interface—either digital or analog—if it is going to confuse the operator and send them the wrong message, then that is a big problem. Of course, time pressure is a big factor. They are three of the big ones. Sometimes you can consider training, but then training and novelty are interlinked.
- Are there, or should there be, any differences in the way that EPCs are quantitatively represented?
  - Maybe. We have never felt the need to do that. We might change the risk model structure instead.
- Can you explain why you chose the multiplication method of combining EPCs instead of another approach, for example, a simple addition of the assessed effect?
  - That was the traditional method. There have been other attempts to use Bayesian approaches or fuzzy approaches. In certain cases, looking at data with multiple PSFs together, they had a greater impact than just adding them.

  - In addition, there is a preference conservative numbers, which come from a multiplicative approach.
- How are PSFs interrelated? They seem to be treated as independent in NARA. Can you elaborate on the reasoning behind this decision, and how they might be related to each other, if they are?
  - The interaction of PIFs depends on the scenario. There is not a clear pattern. We explored this with a limited amount of data and also had some expert judgment sessions. Robust results did not emerge from data or experts about the interrelationships.
- NARA considers only the negative impact of EPCs, rather than any potential positive impact of context. Can you explain the reasoning behind this decision? Should positive context be quantified differently than negative context?
  - We tried very hard to put this into NARA. It never really worked. We tried it for a couple of years, using it prospectively. There was no consistent effect. We literally call it the miracle effect or the John Wayne effect. But you cannot rely on it in a risk assessment, because maybe John Wayne would not be on shift that day.
- If there are EPCs that are not independent, what might they be? How might this play out quantitatively?
  - Examining EPCs, they all seem to interact. One way to test this might be through triads of PIFs and examine their associations. As some possible interrelationships, teamwork can compensate for time pressure and stress. Ergonomics of the interface and time pressure and stress might interact. Stress and time pressure both reduce your cognitive resources.

A.6 General

- How are PIFs selected?
  - PIFs are surrogates for a causal model of human behavior. They are not precise, but they help to define the factors that conceptually, theoretically, or experimentally influence behavior. People began to list PIFs for the purpose of adjusting some reference probabilities; they serve as adjustment factors for base error probabilities. SLIM-MAUD [14, 15] was the first to transform these factors into scales or weights, and, using some anchor points, give a relative probability of error for different situations based on the corresponding level of the PIFs. The PIFs were specified by the domain experts.
  - Through the years, models have used several different cognitive models and developed some different classifications of PIFs. Recent work has explored Bayesian models as an alternative.
- Are there PIFs that are more important than others, and if so, why?
  - If PIFs are intended to provide pictures of context, they can be divided into dynamic and static PIFs. In addition, there can be personal/individual, organizational, and contextual/situational PIFs. Depending on how you view those different PIFs, they may need to be quantified differently. For example, available time—in some models (e.g., Time-Reliability Correlation (TRC)[23], HCR), time was the central factor. These are all hypotheses or assumptions, but there is no empirical validation for this different treatment in any conventional way. There is no way to test and validate the differences

among PIFs. In addition, causal factors are not hierarchical in a mathematical sense; they are interdependent. There are layers and levels.

- Sometimes, we use the notion of proximate or root cause. In the HRA and PIF domain, there is a causal chain that goes from deeper or broader factors or categories to more specific. One way of exploring PIFs is to consider the depth of causality.
- Many years ago, developed a “hybrid method” sponsored by a utility, and came up with two to three PIFs that had what we called veto power. If those are present, the others do not matter. This also is done in common-cause failure modeling. We have the separation between lethal shocks that affect the system and have a global impact and non-lethal shocks that have a distributed impact. That is a characterization between the dominant things in terms of the value and the way they affect the numbers.
- One has to look at the list of PIFs in a model to decide which ones are the dominant factors. The level of training—if you don’t have a procedure for the situation you are facing, you are set up for failure. That may not be a good example, but it is an extreme case. If you say, the procedures are there, these are the operators, but they did not have adequate training in terms of the class of an accident. That is probably secondary in terms of level of influence compared to something else—for example, time constraint.

- Are there PIFs that should be treated differently quantitatively?
  - Some are qualitative, whereas other are more quantitative; some are more directly observable than others. This is one of the appeals of Bayesian models, which allow for soft and fuzzy relationships; they do not have to be all quantitative or all qualitative. There are different reasons and requirements for different PIFs; not all of them should be treated in the same way.
- Can you discuss the methods of combining multiple PIFs/PSFs? Some take a multiplicative approach; others take an additive approach. Can you discuss your thoughts on these different approaches?
  - There is the question of interdependencies among the PIFs. Assume that they are not interdependent.
  - THERP had the base numbers (HEPs), and then you multiply them. SLIM-MAUD used a loglinear relationship. The log of the error probability is a linear combination of PIFs times the corresponding weight. There you introduce the weight, and you create a composite index. Things are still nominally independent. The SLIM-MAUD approach is consistent with the way that we do it in many other sciences. We take a linear or loglinear model, and it’s a good approximation. That ignores the interdependencies. Considering the interdependencies, that’s when you move into the Bayesian belief network. SPAR-H is multiplicative. You take each of the PIF weights and you multiply them.
  - If you’re not concerned about the interdependencies, if it’s additive or multiplicative, there’s no theoretical reason for one or the other. There’s no fundamental advantage to additive or multiplicative.
  - The multiplicative approach gives you a way of looking at a wide range of numbers; it may capture the range of variability and uncertainty that is as vague and uncertain as this. It’s almost loglinear. That is why lognormal distribution is so popular in the PRA discipline. It gives the ability to cover a few orders of magnitude in terms of uncertainty. The additive approach is a bit more restrictive. Both need to ensure that there is normalization.

- Would you generally agree that PIFs are interdependent and the assumption was for mathematical convenience?
  - HRA estimations are made with huge amounts of uncertainty. PRAs can absorb that and still be useful, even with huge uncertainties. From a practical point of view, if we simplify the method and assume independence, we will probably still produce reasonable numbers if the reference probabilities are reasonable.
  - There's also the question of dependencies between human-failure events—the probabilities of error #2 given error #1. Correct modeling of interdependencies of PIFs and interdependencies that they create for errors has the best return on investment, because currently, if you look at methods accounting for HEP interdependence, they are totally ad hoc. They have no basis. All of them except those that have started looking at Bayesian belief networks.
  - Phoenix [24] has eight PSFs because of practical considerations. IDHEAS has 20. Others have 50. In HRA, we function under two constraints:
    - Practical considerations in terms of assessment of these PSFs. When you have many factors, there are many things that contribute to the noise. Having many factors may not be advantageous even though they may be more descriptive.
    - There have challenges in implementing heavy computation using computers among practitioners. The flexibility, the more sophisticated mathematical formulation, such as bringing realism in terms of interdependency. The PIFs are interdependent because of the causality of the relationship. The combined impact of the PIFs would not be the same as if they were working independently.
- Can you talk a bit about the inclusion of positive as well as negative PIFs in some models. Should positive PIFs be included? Should positive context be quantified differently than negative context?
  - That is the way it should have been from day 1. As recognized by SPAR-H and a couple of other methods, PIFs have a positive or a negative impact. Most scaling done for HEP calculations, if people are aware of how these are or quantified or what the middle point is, many of the methods can account for the positive impact.
- Do you think that the effect of positive PIFs and the effect of negative PIFs should be symmetrical? Do they have the same weight? Are they equally important?
  - That is not really the way to think about the concept. PIFs are not positive or negative, fundamentally. They influence behavior one way or the other. They have traditionally been used to alter the probability of error. There isn't a study that shows there are five positive, five negative. Some quantifications go from 0–5, and 1 is the norm. Others use the midpoint as the norm. Understanding that means that mathematically treating numbers differently. If my scale is between 0.1 and 10. 1 now is in the middle multiplicatively. You increase or decrease the probability based on the effect of that PIF. To highlight the two possible roles of PIFs in this given situation, the quantitative method needs to be appreciative of these possible roads and say the scales go from 0.1 to 10 instead of 0 to 10. Your mathematics needs to be friendlier with respect to thinking about quantifying these things. The usual methods we have used in the past. There is no a priori way of defining the influence or degree by labeling something as positive or negative.

A.7 Success Likelihood Index Methodology (SLIM)

- How are PIFs selected?
  - Primarily through review of human factors risk analyses over 40 years of consultancy work. Custom PIFs may be created for specific projects where not available in a database. We also conducted workshop sessions with experienced subject matter experts (e.g., nuclear power plant operators).
- How are multiple PIFs combined?
  - Most of the time, default to the simplest model possible. Simplest combination rules—the effects of PIFs are essentially linear additive combinations. That is what is used to derive numeric input that is converted to a probability based on a calibration.
- Why additive?
  - Partially based on decision analysis. We also did experiments on a range of models. Non-linear models have not been shown to be better predictors than simple linear models. Start off as simple as we can and see where more complexity is necessary.
  - PIFs are quantified by the weight multiplied by the rating; rather than having a discrete value, it possible to use a distribution. Can have different factors. Put those into a Monte Carlo simulation package. Shows you the distribution and shows how the quality index (probability) changes over time. This is important because if the probability is high, it's of interest.
- Are some PIFs are non-independent? If so, how is that reconciled?
  - The work focused on non-interactivity; simple way to handle it is a matrix with x and y axis, look at relationships. Where they are both present simultaneously, they might become more important in combination. However, the analysis generally assumes that they are independent.

APPENDIX B – EXPLORATORY ANALYSIS OF CALCULATING PERFORMANCE INFLUENCING FACTOR EFFECTS

This appendix presents the conclusion including the technical bases recommended for distinguishing the effects of PIF attributes on HEP into base HEP and HEP modifier and calculating their combined effects on HEP CFMs. The IDHEAS-G formulation separates factors into base PIFs and modifier PIFs. In the mathematical formulation, base PIFs are treated as probabilities while modifier PIFs are expressed as multiplicative effects using ratios of error rates. As such, base PIFs are designed to not exceed 1, while modifier PIFs can. Given that modifier PIF weights are multiplied directly times the collective base probability, the potential for predicted HEPs to exceed 1 exists. PNNL investigated the extent to which this violation occurs.

For each CFM, PNNL identified three modifiers with the most severe weight values. The limited the number of modifiers were limited to three to align with how analysts typically choose two to three PIFs to attribute to a particular HEP for a task. From there, we varied the base HEP values to get a sense of how many scenarios induced error probabilities greater than 1 and by how much these violations exceeded 1. There were 3,671 scenarios analyzed for each CFM. Table B.1 shows the percentage of these 3,671 scenarios that produced probabilities that exceeded 1 and the maximum resulting HEP. The number of violations is not uniform across CFMs and the severity ranges from slightly above 1 to near 50.

Table B.1. Percentage of Severe Modifier Scenarios with HEPs that Exceed a Probability of 1.0

CFM	% of HEP>1	Maximum HEP
Detection (D)	19.86	*18.74*
Decision Making (DM)	80.20	*26.11*
Execution (E)	28.44	*19.92*
Team Coordination (T)	1.96	*1.80*
Understanding (U)	85.89	*47.96*

Figure B.1 depicts the error probabilities across the CFMs. The dashed line corresponds to an HEP of 1. Predicted HEPs are not continuous, and gaps exist per CFM. Discontinuities in HEP increases may represent scenarios that are extreme.

To address the violations of 1 that currently exist in the model, PNNL recommends applying a sigmoidal transformation of the data that preserves ordering of HEPs derived from IDHEAS-DATA [5], matches on base HEPs where modifiers are not present, and aligns with severity ratings provided by IDHEAS-DATA. Figure B.2 depicts the transformed HEP predictions from PNNL simulations. This transformation has the benefits of maintaining severity ordering and being informed by cognitive literature around numerical cognition and severity judgements.

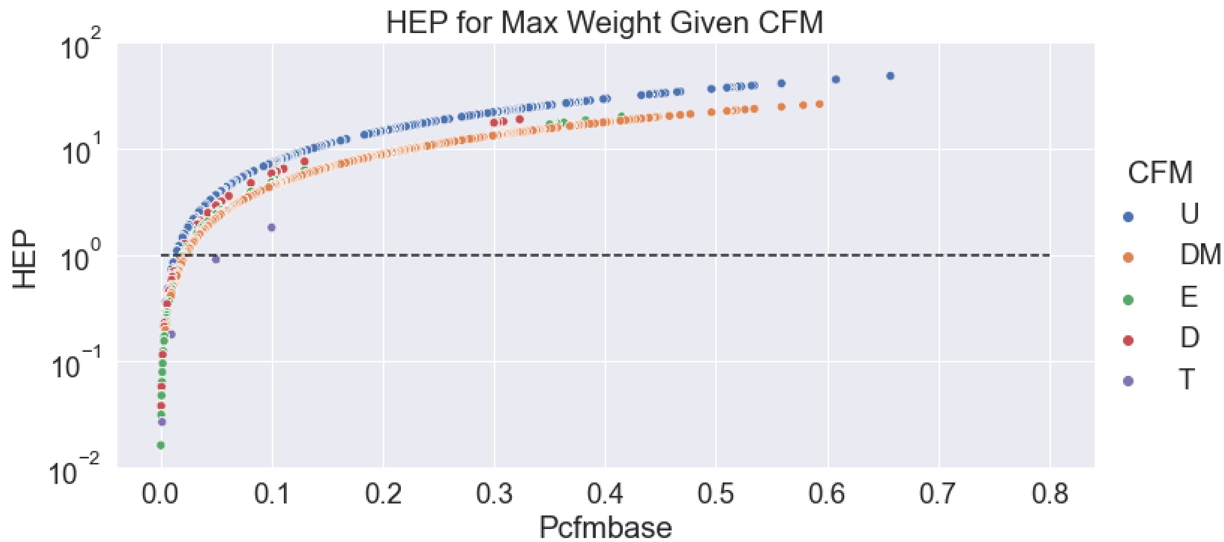

Figure B.1. HEPs for Maximum Weights on a Given CFM

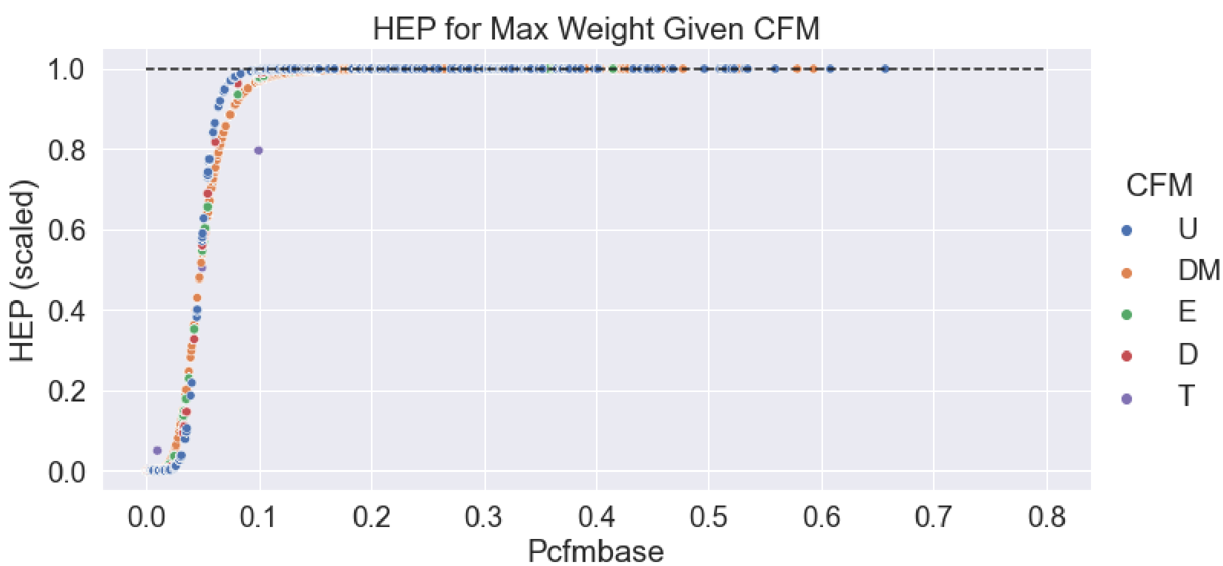

Figure B.2. HEPs for Maximum Weights on a Given CFM After Sigmoidal Transformation