Challenge to Collect Empirical Data for Human Reliability Analysis—Illustrated by the Difficulties in Collecting Empirical Data on the Performance-Shaping Factor Complexity

Karin Laumann[1]
Department of Psychology,
Norwegian University of Science and Technology,
Trondheim 7491, Norway
e-mail: karin.laumann@ntnu.no

Martin Rasmussen Skogstad
Department of Psychology,
Norwegian University of Science and Technology,
Trondheim 7491, Norway

This paper discusses the challenges with collecting positivistic empirical data (objective, observable, reliable, replicable, experimental, and true) in human reliability analysis (HRA), and illustrates it by presenting the difficulties in collecting empirical data on the performance-shaping factor (PSF) complexity. The PSF complexity was chosen to illustrate the difficulties with empirically collecting data because it is included in many HRA guidelines and it has been discussed as an important PSF to understand error rates in large accident scenarios. This paper discusses the challenges with collecting empirical data from a pure positivistic paradigm with experiments, as well as from literature reviews and data from event reports, training, and operations. The paper concludes that because of all the challenges with the positivistic empirical data collections methods in HRA, we should discuss whether experts' judgements could be a better approach to obtain HRA data and error rates. In a postpositivistic view, qualitative data or experts' judgments could also be looked at as empirical data if the data were collected in a systematic and transparent way. [DOI: 10.1115/1.4044795]

Human reliability analysis (HRA) has been defined by Swain [1]: "As any method by which human reliability is estimated." Mkrtchyan et al. [2] has more specifically defined human reliability analysis as: "human reliability analysis aims at systematically identifying and analyzing the causes, consequences and contribution of human failure in socio-technical systems (e.g., nuclear power plants, aerospace systems, air traffic control operations, chemical and oil and gas facilities)."

The lack of empirical data has often been described as the largest issue in HRA, for example, by, [1,3–7]. However, the invention and development of HRA methods seemed more to be based on resolving a practical need for estimating human reliability in risk analysis rather than being developed based on empirical research. The quantitative part of HRA methods were mainly developed, with use of the available empirical evidence that existed when the methods were developed. Hence, HRA methods give rather crude, relative, and far from precise estimates of the likelihoods of human errors. Laumann et al. [8] have recently discussed the challenges with data for HRA. This paper is an extension of that paper and will present as an example the challenges to collecting empirical data on the performance-shaping factor (PSF) complexity, so as to further illustrate the difficulties in collecting empirical data in HRA.

It is not easy to define "empirical data." The definitions of "empirical" seem to include words like observable, reliable, replicable, experiments, and truth. "Empirical" seems to be looked at as objective data and as the opposite of subjective data. The demands for empirical data collection seem to fit best into a positivistic research paradigm. Guba and Licoln [9] have described two similar but also different positivistic paradigms: in a pure positivism research paradigm, one assumes that an objective reality exists. This reality is independent from the researcher, it can be found and measured by scientific methods, and true scientific laws can be developed. In this paradigm, only quantitative methods such as experiments are used. In a postpositivism paradigm, one also assumes that an objective reality exists. However, this reality is complex, and it cannot be completely understood. One can never be sure that a true reality has been found, and one can only argue that it is likely that a true reality has been found. Within this perspective, both qualitative and quantitative data can be collected. From a postpositivism paradigm, one can say that empirical evidence and truth is a question about likelihood and degree rather than strictly one or the other.

What counts as empirical data sources in HRA seem to be data from experiments, literature reviews from experiments, and data from real operational data sources, such as event reports. Also, in HRA, several types of empirical data have been collected into databases. Data that do not seem to count as empirical in HRA are experts' judgments, since those are thought of as subjective and nonreplicable (for example, Ref. [4]). In a positivistic paradigm, experts' judgments would not be characterized as empirical data. However, in a post positivism-paradigm, qualitative data could also be looked at as empirical data, and therefore systematic and transparent collections of experts' judgments could, within this paradigm, also, be considered as empirical data.

It seems like the request for empirical data in HRA comes from a purely positivist paradigm, where experts' judgments are not considered as empirical data. This paper will discuss the difficulties in obtaining HRA data on the PSF Complexity and its levels with the so-called objective methods such as experiments, literature reviews (mostly from experiments), and observations from training, real work performance, and accidents and incidents, such

[1]Corresponding author.

Manuscript received December 17, 2018; final manuscript received June 20, 2019; published online November 19, 2019. Assoc. Editor: Raphael Moura.

as those described in the event reports. This paper will briefly discuss experts' judgements as a method to obtain HRA data.

To discuss challenges with empirical data collection, this paper will describe challenges to collecting data on the PSF complexity. Complexity was chosen because it is included in many HRA methods, and newer methods, such as a technique for human error analysis (ATHEANA) [10], claim that this is one of the most important PSFs for understanding error-forcing contexts. Complexity is also among the PSFs, where empirical data collection is most difficult.

Definition of Complexity and Its Expected Effect on Performance in Human Reliability Analysis Methods

To discuss empirical data collection on the PSF Complexity in HRA, we have chosen to present how some HRA methods define and include the PSF complexity: standardized plant analysis risk-human (SPAR-H) [11], petro-HRA [12], human error assessment and reduction technique (HEART) [13,14], and ATHEANA [10]. Technique for human error rate prediction [15] was also considered, since it is also a very well-known method, but it includes very little direct information related to complexity [16], and therefore it was not included here. This paper should not be viewed as a critique of these methods. These methods were included to illustrate the challenges and difficulties with collecting empirical data, not because these methods are particularly poor on these issues. These issues exist for all HRA methods, and these methods were chosen only to illustrate the arguments. Next, the methods will be briefly described with a particularly focus on how they include the PSF complexity in the method.

Standardized Plant Analysis Risk-Human. Standardized plant analysis risk-human [11] included two nominal tasks and a nominal failure rate for each of these. The nominal tasks are an action task with a nominal failure rate 0.001 and a diagnosis task with a nominal error rate of 0.01. Action should be used if the task is a mainly performance task. Diagnosis should be used if the task is a mainly cognitive task. SPAR-H includes eight PSFs. Boring and Blackman [16] define PSF as: "*an aspect of the human's individual characteristics, environment, organization, or task that specifically decrements or improves human performance, thus respectively increasing or decreasing the likelihood of human error*." The PSFs in SPAR-H are: available time; stress/stressors; complexity; experience/training; procedures; ergonomics/human machine interface (HMI); fitness for duty; and work processes.

In SPAR-H [11], complexity is defined as follows:

> "*Complexity refers to how difficult the task is to perform in the given context. Complexity considers both the task and the environment in which it is to be performed. The more difficult the task is to perform, the greater the chance for human error. Similarly, the more ambiguous the task is, the greater the chance for human error. Complexity also considers the mental effort required, such as performing mental calculations, memory requirements, understanding the underlying model of how the system works, and relying on knowledge instead of training or practice. Complexity can also refer to physical efforts required, such as physical actions that are difficult because of complicated patterns of movements.*"

SPAR-H [11] also includes a model of different complexity factors. These are shown in a figure on page 22 in the SPAR-H manual, but they are not further described there. These factors are (page 22):

> *"Multiple faults, high degree of memorization required, system interdependencies not well define, multiple equipment unavailable, large number of actions required, large number of distractions present, task requires coordination with ex-control room activities, parallel tasks, mental calculation required, low fault tolerance levels, symptoms of one fault mask other faults, misleading or absent indicators, transitioning between multiple procedures, large amount of communication required."*

Levels and multipliers for Complexity in Standardized Plant Analysis Risk-Human. The levels for complexity in SPAR-H [11] are:

> *Highly complex—very difficult to perform. There is much ambiguity in what needs to be diagnosed or executed. Many variables are involved, with concurrent diagnoses or actions (i.e., unfamiliar maintenance task requiring high skill). Multiplier diagnosis and action = 5.*
>
> *Moderately complex—somewhat difficult to perform. There is some ambiguity in what needs to be diagnosed or executed. Several variables are involved, perhaps with some concurrent diagnoses or actions (i.e., evolution performed periodically with many steps). Multiplier diagnosis and action = 2.*
>
> *Nominal—not difficult to perform. There is little ambiguity. Single or few variables are involved. Multiplier diagnosis and action = 1*
>
> *Obvious diagnosis—diagnosis becomes greatly simplified. There are times when a problem becomes so obvious that it would be difficult for an operator to misdiagnose it. The most common and usual reason for this is that validating and/or convergent information becomes available to the operator. Such information can include automatic actuation indicators or additional sensory information, such as smells, sounds, or vibrations. When such a compelling cue is received, the complexity of the diagnosis for the operator is reduced. For example, a radiation alarm in the secondary system, pressurized heaters, or a failure of coolant flow to the affected steam generator are compelling cues. They indicate a steam generator tube rupture (SGTR). Diagnosis is not complex at this point; it is obvious to trained operators. There is no obvious action PSF level assignment available to the analyst. Easy to perform actions are encompassed in the nominal complexity rate. Multiplier diagnosis = 0.1.*

The SPAR-H manual [11] says that it is not possible to describe how much of the complexity factors or how much of a combination of these factors is needed to go from one level to the next. SPAR-H leaves this up to the HRA analyst to decide.

Boring and Blackman [16] describe that the multipliers in SPAR-H for the complexity levels come from a table in technique for human error rate prediction that shows operators' response to simultaneous alarms. It is peculiar that SPAR-H has defined a much broader PSF complexity but uses only one small element of their definition, "numbers of alarms" as the basis for the multiplier. There seems to be little consideration of the overall meaning of the concept (complexity) and the meaning in the table in THEPR (numbers of alarms) from which the multipliers originated.

Petro-Human Reliability Analysis. Petro-human reliability analysis (Petro-HRA) [12] was developed with the purpose of adjusting SPAR-H to the petroleum industry. Petro-HRA has one nominal task and one nominal failure rate, which is 0.01. Petro-HRA includes nine PSFs: time; threat stress: task complexity; experience/training; procedures; human-machine interface; attitudes to safety, work and management support; teamwork; and physical working environment. The task complexity subcomponents and their descriptions of them were developed based on a literature review performed by Rasmussen et al. [17].

Definition of Task complexity in Petro-Human Reliability Analysis. The definition of Task complexity in Petro-HRA is [12]:

> *"Task complexity refers to how difficult the task is to perform in the given context. More complex tasks have a higher chance of human error. Task complexity can be broken down into various complexity factors that alone or together increase the overall complexity of a task. Task complexity factors include goal complexity, size complexity, step complexity, connection complexity, dynamic complexity, and structure complexity.*
>
> *Goal complexity refers to the multitude of goals and/or alternative paths to one or more goals. The complexity of a task will increase with more goals/paths, especially if they are incompatible with each*

other (e.g., parallel or competing goals and no clear indication of the best path/goal).

Size complexity refers to the size of the task and the number of information cues. This also includes task scope, which is the subtasks and spread of faults to other tasks. The complexity of a task will increase as the amount and intensity of information an operator has to process increases. Step complexity refers to the number of mental or physical acts, steps, or actions that are qualitatively different from other steps in the task. Complexity of a task will increase as the number of steps increases, even more so if the steps are continuous or sequential.

Connection complexity refers to the relationship and dependence of elements of a task (e.g., information cue s, subtasks, and other tasks). Task Complexity will increase if the elements are highly connected and it is not clearly defined how they affect each other.

Dynamic complexity refers to the unpredictability of the environment where the task is performed. This includes the change, instability or inconsistency of task elements. Task complexity will increase as the ambiguity or unpredictability in the environment of the task increases.

Structure complexity refers to the order and logical structure of the task. This is determined by the number and availability of rules and whether these rules are conflicting. Task complexity will increase when the rules are many and conflicting or if the structure of the task is illogical."

Levels and multipliers for Task complexity in Petro-Human Reliability Analysis. Petro-HRA [12] has the following levels and multipliers for task complexity:

Very high negative effect on performance. The task contains highly complex steps. One or several of the complexity categories are present and influence performance very negatively. For example, several parallel goals are present, the size of the task is huge with many information cues and many steps, it is unclear which task elements to perform, if an order is relevant, if tasks have any effect on the situation, and the task environment changes. Multiplier 50.

Moderate negative effect on performance. The task is moderately complex. One or several of the complexity categories are present and influence performance negatively. Multiplier 10.

Very low negative effect on performance. The task is to some degree complex. One or several of the complexity categories are to some degree present and are expected to have a low negative effect on performance. Multiplier 2.

Nominal effect on performance. The task is not very complex and task complexity does not affect operator performance. Task complexity has neither a negative nor a positive effect on performance. Multiplier 1.

Low positive effect on performance. The task is greatly simplified and the problem is so obvious that it would be difficult for an operator to misdiagnose it. E.g., detecting a single alarm, or sensory information such as clear visual and auditory cues. Multiplier 0.1.

Also, petro-HRA [12] describes the levels for task complexity on an overall complexity level. It also does not describe how much of each of the subcomponent or how much combination of the subfactors that are needed to be at the different complexity levels.

The petro-HRA guideline [12] does not describe the basis for the multipliers for task complexity. However, Rasmussen et al. [17] describe why the multipliers in petro-HRA where chanced from SPAR-H:

"The PSF levels in PetroHRA are not the same as the levels in SPAR-H. The nominal and the obvious levels are very similar with the same multipliers and similar descriptions. Both Petro-HRA and SPAR-H have two levels for tasks where complexity Influences performance in a way that increases HEP. SPAR-H has a moderate level with a multiplier of two (leading to a HEP of 0.02 in diagnosis tasks and 0.002 in action tasks if all other PSFs are nominal) and a high level with a multiplier of five (leading to a HEP of 0.05 in diagnosis tasks and 0.005 in action tasks if all other PSFs are nominal). While having two levels with low multipliers as low as SPAR-H has in the "Complexity" PSF gives nuance in the evaluation of tasks where complexity is not the critical factor, we have chosen to have levels with higher multipliers to credit complexity in tasks where it is the critical factor. Petro-HRA has a moderate level with a multiplier of 10 (leading to a HEP of 0.1 if all other PSFs are nominal) and a high level with a multiplier of 50 (leading to a HEP of 0.5 if all other PSFs are nominal). A PSF should only be set to something other than nominal if it has a significant contribution to a chance of success or failure in a task; therefore, we have chosen not to include levels with multipliers as low as two and five."

A Technique for Human Error Analysis. A technique for human error analysis [10] is an HRA method developed to search for error-likely situations and error forcing contexts and to produce estimates of human error probability. ATHEANA [10] uses an expert elicitation process to perform quantification, which is based on data collection and inspection in the specific contexts that are under analysis. ATHEANA [10] was developed based on the belief that: "human failure events occur when the operators are placed in an unfamiliar situation where their training and procedures are inadequate or do not apply, or when some other unusual set of circumstances exist." Much information in ATHEANA [10] deals with complexity. Here only the most specific information about complexity presented in ATHEANA [10] will be included.

Definition of the Performance-Shaping Factor Complexity in a Technique For Human Error Analysis. The name of this PSF in ATHEANA [10] is: "*Complexity of the required diagnosis and response, the need for special sequencing and the familiarity of the situation.*" Complexity has been defined in ATHEANA [10] as

"This factor attempts to measure the overall complexity involved in the situation and action of interest (e.g., the same operator must perform many steps in rapid succession vs. one simple skill-of-the-craft action). Many of the other PSFs bear on the overall complexity, such as the need to decipher numerous indications and alarms, the presence of many and complicated steps in a procedure, or poor HSI. Nonetheless, this factor should also capture measures such as the ambiguity associated with assessing the situation or executing the task, the degree of mental effort or knowledge involved, whether it is a multi-variable or single-variable associated task, whether special sequencing or coordination is required in order for the action to be successful (especially if it involves multiple persons in different locations), or whether the activity may require very sensitive and careful manipulations by the operator. The more these measures describe an overall complex situation, this PSF should be found to be a negative influence. To the extent these measures suggest a simple, straightforward, unambiguous process (or one that the crew or individual is very familiar with and skilled at performing), this factor should be found to be nominal or even ideal (i.e., positive influence)."

A technique for human error analysis [10] also includes 14 scenario characteristics and 15 parameter characteristics that could be particularly challenging for the operators. These factors are similar to complexity factors in other methods, so they could also be considered as complexity. The scenario characteristics are [10]: "garden path problems, situation that change requiring revised situation assessment, missing information, misleading information, masking activities, multiple lines of reasoning, side effects, impasses, late changes in the plan, dilemmas, trade-offs, double binds, high tempo, multiple tasks, need to shift focus of attention." A description of these characteristics will not be included here. For purposes of illustrating how they are described, the description of the first one, "garden path problems" is included here [10]: "Conditions start out with the scenario appearing to be

a simple problem (based on strong but incorrect evidence) and operators react accordingly. However, later correct symptoms appear, which the operators may not notice until it is too late." The 15 parameter characteristics are [10]:

> "No indication, Small change in parameter, Large change in parameter, Lower or higher than expected value of parameter, Slow rate of change of parameter, High rate of change in parameter, Change in two or more parameters in a short time, Delays in changes in two or more parameters, One or more false indications, Direction of change in parameter(s) over time is not what would be expected (if the nominal scenario was operative vs. the deviant), Direction of change in parameter over time relative to each other, is not what would be expected (if the nominal scenario was operative vs. the deviant), Relative rate of change in two or more parameters is not what would be expected (if the nominal scenario was operative vs. the deviant), Behavior of apparently relevant parameters is actually irrelevant and misleading, Parameters indicate response for which insufficient resources are available or indicate more than one response option."

A technique for human error analysis also includes a description of each of these parameter characteristics, which will not be included here. The ATHEANA [10] manual does not give any information about expected error rates on the complexity PSF or the challenging scenario and parameter characteristics. The quantification is done purely with expert's elicitations.

Human Error Assessment and Reduction Technique. The 1992 version of HEART [13,14] describes nine generic task types (GTTs) with proposed nominal human reliability values for each of them, and also suggested bounding values. HEART also includes short descriptions of 38 error-producing conditions (EPC), where each has a maximum multiplier amount, which is the maximum amount that the nominal error rate might be suggested to change based on this condition. EPC are defined by Williams [14]: "error producing factors are factors that can affect human performance, making in less reliable than it would otherwise be." The analyst should also assessing the proportion of "affect" for the EPCs or how much each EPC contributes to the overall unreliability of the task under consideration. HEART does not give much advice on how to perform this part of the analysis.

The sources for HEARTS GTTs proposed nominal human reliability values and EPC maximum multipliers are the human factors literature. Williams and Bell [18–20] have in the last years conducted new literature reviews to critically evaluate the data and evidence for the GTTs and EPCs in HEART. The conclusion from these literature reviews was that it confirmed 32 of the 38 original EPC concepts and multipliers, six EPCs were slightly revised and two new ones were incorporated into HEART. The proposed nominal human reliability values for the GTTs were slightly modified based on the literature review.

Human error assessment and reduction technique [13,14] has a very different structure than the other methods included here, so it is difficult to decide which elements in HEART seem to involve complexity. GTT C seems to be related to complexity. The description of GTT C is [14]: "Complex task requiring a high level of understanding and skill (, e.g., diagnosing a fault or a car engine) this task is mission oriented and could involve a great many discrete elements or actions but would normally only involve one basic activity." This GTT had a proposed nominal human unreliability of 0.16 (uncertainty bounds 0.12–0.28) in 2017 version 0.17 (0.05–0.6) [18].

Some of the EPCs in HEART might be related to complexity. However, a significant amount of interpretation is needed to decide on how similar the EPCs are to the concept of complexity in the other methods. Some suggestions of EPCs in HEART [14] that have similarities to complexity could be:

In comparison of SPAR-H with other methods, they have included EPC 10 as the one that they think is most similar to SPAR-H definition of complexity [11]. EPC10 is: "The need to transfer specific knowledge from task to task without loss," *Maximum multiplier of 5.5*. [14]

Other EPCs in HEART [14] that could be viewed as similar to the complexity PSFs in other methods are:

> *EPC3: a low signal-noise ratio. Maximum multiplier of 10.*
>
> *EPC8: A channel capacity overload, particularly one caused by simultaneous presentation of nonredundant information. Maximum multiplier = 6.*
>
> *EPC18 A conflict between immediate and long-term objectives. Maximum multiplier = 2,5.*
>
> *EPC 24 A need for absolute judgments which are beyond the capability of the operator. Maximum multiplier = 1.6.*
>
> *EPC26 No obvious way to keep track of progress during an activity. Maximum multiplier = 1.4.*
>
> *EPC28 Little or no intrinsic meaning in a task. Maximum multiplier = 1.4.*

Since HEART's sources for GTTs proposed nominal human unreliability and the EPCs maximum multipliers are literature reviews, HEART's approach will be further discussed under the heading: "Difficulties in getting empirical data on Complexity from literature studies."

Challenges With Collecting Empirical Data on the Performance-Shaping Factor/Error-Producing Condition Complexity in Human Reliability Analysis

In this section, the challenges with collecting positivistic empirical data on the PSF/EPC complexity from experiments, literature reviews, incidents, and event reports, as well as training and operation data, will be discussed. These are the sources that have been suggested as sources for empirical data collection (for example, Ref. [4]).

Challenges With Collecting Empirical Data on Complexity With Experiments. Experiments to collect HRA data would usually occur in some form of simulator. Boring et al. [3] describe three types of simulators: "a simplified microworld simulator using unskilled student operators; a full-scope control room simulator using skilled student operators; and a full-scope control room simulator using licensed commercial operators." Boring et al. [3] say that because of the costs of the experiments, little data has been collected with full-scope control room simulators using licensed commercial operators and that there exist some challenges in generalizing data collected in simulators with students to nuclear power plant operators. However, there also exist other challenges in collecting data with experiments. Next, the challenges to collecting HRA data in experiments with any simulator will be described.

Challenges With the Definitions of Complexity and Complexity Subcomponents in the Different Human Reliability Analysis Methods. As can be seen in the definitions of complexity in SPAR-H [11], petro-HRA [12], and ATHEANA [10], complexity is considered to consist of many subcomponents, which probably have different effects on performance. The different methods also include different subcomponents of complexity.

To do an experiment, a researcher first has to decide on which complexity taxonomy he or she will use in his/her investigation. It would have made research easier if HRA could have agree on which subcomponents complexity consists of. As it stands, different studies have to be performed based on the different methods to test the different error rates, for the different subcomponents the different methods include. Also Liao et al. [6] discuss the

challenges with a lack of a common theoretical base for data collection in HRA.

Challenges With Developing Levels and Measuring Scales for the Subcomponents of Complexity. To measure the effect of complexity on performance, a measuring scale that measure, the amount of complexity for each subcomponents has to be developed. The levels or strength of complexity in SPAR-H and petro-HRA refer to the overall complexity and not to the subcomponents. It is not described in the method which subcomponents or "how much" of the subcomponent is necessary to assign different levels, multipliers, or error rates. In SPAR-H and petro-HRA, levels are described at the overall task level, such as very complex, moderately complex, and nominally complex, with (abstract) examples of what these could mean. Ideally, there should be "objective" scales where the amount of the different subcomponents could be measured, and these scales should show interrater reliability. For some of the subcomponents, it might be possible to develop an objective scale, such as for step-complexity, where it is possible to count the steps that need to be taken. However, for other subcomponents such as dynamic complexity, it is more difficult to be objectively defined, because it involves more unspecified information, which is more difficult to observe, such as from Rasmussen et al. [17] ambiguity, change, or stability of the task or system's characteristics over time.

Challenges With Development of Experimental Tasks That Fit With the Subcomponents Descriptions and the Description of Levels. To design experiments, to investigate how much complexity affects performance, the most ideal experiments would be to manipulate each subcomponent separately, at different levels, and subsequently measure subject's failures and successes in the task. To manipulate each subfactor for complexity, tasks (scenarios) need to be developed that represent a specific level of that subcomponent. It can be very challenging to develop experimental tasks/scenarios (manipulations) that fit with the general theoretical descriptions of the subcomponents in the HRA method and with a specific level; for example, see "connection complexity" in petro-HRA [12].

It might also be difficult to develop experimental tasks that include only one of the subcomponents of complexity, since these subcomponents are correlated. For example, size complexity and step complexity very often will co-exist, and they are difficult to separate.

Challenges With Separating the Different Performance-Shaping Factors in the Experimental Manipulations. To measure how much different complexity levels affect performance, complexity needs to be manipulated and the other PSFs should be nominal. Complexity is a particularly difficult concept to manipulate in an experiment because it is difficult to manipulate complexity and have the nominal level for the other PSFs. For example, a task might be experienced as less complex with training. It is difficult to manipulate complexity with a nominal level of training because the nominal level of training is usually not well described. In SPAR-H [11], nominal experience/training is defined as more than 6 months of training and experience. Petro-HRA [12] defines the nominal level of experience and training as: "the operator(s) has experience and/or training on the tasks(s) in this scenario and has the necessary knowledge and experience to be prepared for and to do the tasks in the scenario. Experience and training does not reduce nor to a large degree improve performance." The SPAR-H [11] definition is quite objective and defines the nominal levels as "more than 6 month of experience and or training." However, in an experiment (with students or operators), no one would train, for more than 6 month on the task(s), and probably, if there were more than 6 months of training on the task, almost no task would be complex any longer. In petro-HRA [12], it is difficult to objectively define how much training one should give on an experimental task to be at the nominal level. It is also difficult to separate complexity and procedures, because more complex tasks usually also have more complex procedures.

It is also difficult to manipulate the complexity subcomponents with control over other PSFs. Some are very difficult to control (fatigue, illness) and some would be challenging to control such as experience/training and procedures [8].

Challenges With the Numbers of Manipulations Needed to Test the Combinations of the Subcategories for Complexity. In most cases in a HRA analysis, complexity would be combinations of the subcategories, where, for example, two subcomponents are at a high level, or it could be all of them at a high level. To test all the combinations of the complexity subcomponents and their levels would require an extreme amount of experimental manipulations, and it seems unlikely that anyone would run that large of an amount of experimental manipulations.

Challenge With the Numbers of Participants Needed in These Experiments. For some of the levels of complexity, HRA methods expect error rates of 1 of 100 or 1 of 1000. Even if students are used as participants, the numbers of persons needed for the experiments only to test the nominal levels are unrealistically high and very difficult to perform.

Challenges With Measuring the Outcome Variable or Numbers of Errors per Trial. When performing experiments where tasks with different characteristics are the manipulations, it might be challenging to define what should count as a task, a scenario, or as an error. What counts as an error is depending on the decompositions of the tasks: ATHEANA seems to analyze an entire scenario, so in ATHEANA, a failure would be something that negatively affects the outcome of a scenario. SPAR-H distinguishes between action tasks and diagnosis tasks. Petro-HRA leaves it up to the analyst to select the decomposition level for the analysis. HEART says that the task should be decomposed so that it fits with the GTT description level. To do this kind of experiment, it is important to have a good definition of what counts as a task and what count as an error on a task. Most of the methods are vague on the description of what should count as a task or an error. The error rates might be very different depending on what counts as a task [21].

Challenges With Generalization of Findings. If we want to manipulate high complexity with all the other PSFs at a nominal level, this might become a highly artificial task, because so many considerations have to be taken into account about the task. For example, in an experiment in a simulated control room for nuclear power plant operators, one has to first develop as a baseline, a scenario, or tasks that are nominal on all the PSFs. Then, as the experimental manipulation, one should develop a scenario that is high on complexity, but nominal on all the other PSFs such as HMI, procedures, and experience and training. These factors are correlated in the real world, and usually when the scenario becomes more complex, the procedures get more complex and there is less training on the scenario. To develop a scenario that is only high in complexity and nominal on all the other PSFs is almost impossible, and if it were developed it might become a very artificial scenario, which might not be generalizable to more real tasks in the control room.

Challenges in Getting Empirical Data on Complexity From Literature Studies. One way to collect HRA data is to use information from literature that could say something about the error rates in the HRA methods. Laumann et al. [8] have described the difficulties in matching the information in the general research literature to the information in HRA methods, such as: (a) difficulty in matching the descriptions of the concepts in these studies to the concepts in the HRA methods; (b) the studies usually

include only one or two levels of a PSF/EPC and this level is usually not well described, since that often is not needed for the purpose of the study; and (c) it is difficult to see which PSFs/EPCs might be present during task performance, since this is usually not described in these studies; and, This makes it difficult to know how much of the percentage of errors are caused by different GTT and PSFs/EPCs.

As an example of this, we have looked at half of the studies that Williams and Bell [18] included as a basis for HEART's GTT C, that we thought resembled most of the PSF complexity included in the other methods presented here. To look at which types of studies were considered as a source for this GTT, we investigated the first 18 of the 37 studies that Williams and Bell [18] refer to as a basis for the nominal error rate for this GTT. Williams and Bell [18] do not present any arguments for why each of these studies or tasks in these studies where considered to be a basis for GTT C nominal value. They just state that these studies where sorted under this GTT.

First, we looked at how well the description of the tasks in the studies fitted with the description of the GTT. We have presented the reference and a short description of the task in Table 1. It was difficult to see why Williams and Bell [18] thought these studies and task descriptions in them fitted with the description of the GTT. What these studies seem to have in common is that the tasks seem to involve some kind of demanding cognitive activity. However, it is difficult to see how they match, for example, with the subcomponents for complexity described by SPAR-H, petro-HRA, and ATHEANA. It would have helped if Williams and Bell had described in more detail how they think the task information in these studies fit with their GTT C and/or defined complexity more in HEART so that we could see whether they have another definition of complexity that lies under the selection of studies.

Also, GTT C is defined as: "complex task requiring a high level of understanding and skill." In most of the studies Williams and Bell [18] refer to, the subjects are students and we find it hard to see how the tasks in the studies that Williams and Bell [18] refer to: "require a high level of understanding and skill." The subjects in most of the studies also seem to receive little to no training on the tasks.

The tasks included in the 18 studies are very different and involve different kinds of cognitive demands. We have presented a short description of the publication titles of the papers and descriptions of the tasks in Table 1. The percentage of errors or success in the tasks varies considerably in these studies. We wonder how good a geometric mean represents likely error rates in this task type, since the results are so varied. It might have been a better argument for a task type if the error rates were much more similar. It is a bit questionable, whether a task type is a meaningful concept if the error rates within this task type are so variable. It is also difficult to see which error rates Williams and Bell [18] selected from each study, since the studies usually had more than one manipulation.

Another issue with these studies is that tasks in these studies seem to include many different EPCs/PSFs, such as, for example, high time pressure (for some tasks, the stimuli and or response rate were milliseconds) or high/low training (some studies also

Table 1 Shows publication title, reference and description of tasks in the first 18 of the 37 studies that Williams and Bell [18] refer to as a basis for the nominal error rate on general task type C, that we thought where most similar to complexity as described in the other methods

Title and author	Task
An empirical study of end user behavior in spreadsheet error detection and correction [22]	Detect and correct spreadsheet errors
The role of visual attention in multiple object tracking: evidence from ERPs [23]	Multiple object-tracking task.
Oops! I did it again. An ERP and Behavioral study of double errors [24]	Stroop task. Subjects were shown three color words ("red," "green," and "blue") presented in either red or green font on a computer monitor. The subject were instructed to press the right of left mouse button in response to the color of the words
Relating semantic and episodic memory systems [25]	Subject performed both encoding and retrieval tasks in the MRI scanner. The task is remembering of nouns.
Mistakes that affect others: an fMRI study on processing on own errors in a social context [26]	Participants performed a computerized task, the canon shooting game. The aim of the task was to stop a horizontally mowing cannon (triangle) by a button press, precisely lining up with a stationary target in order to shoot the target.
The role of meaning and familiarity in mental transformation [27]	Mentally rotated meaningful and meaningless objects. Complex and simple object
Passport officers error in face matching [28]	A face matching task. Matching photo and person
Visual selection mediated by location: selecting successive visual object [29]	Each stimulus consists of a series of frame, each containing a target digit of one color and a distractor digit of another color. The task is to name the highest digit of one color and a distractor digit of another code.
Effect on stimulus complexity on mental rotation rate of polygons [30]	Mental rotation of polygons
Shape discrimination of three dimensional objects depend on the number and location of bends [31]	Subjects made shape discrimination of three-dimensional objects differing in orientation, number of bends, and location of bends.
The costs of crossing paths and switching tasks between audition and vision [32]	Participant discriminated either the identity or spatial location of objects presented either visually or acoustically.
Probability effects in stop-signal paradigm: The insula and the significance of failed inhibition [33]	Subjects respond to go stimuli (typically a visual choice reaction time task) but must withhold their response when a second stimulus or stop signal is presented.
Evaluating models of object-decision priming: evidence from event-related potential repetition effects [34]	Subjects should decide on structural possible or structural impossible objects.
Attention capture with various distractor and target types [35]	Detection tasks with various distractors and target types.
Dual-task interference in visual working memory: A limitation in storage capacity but not in encoding or retrieval [36]	Visual working memory task.
Debugging: an analysis of Bug-location strategies [37]	Generation of and location of erroneous line of code in programming
Hitting the wall: errors in developing and code inspecting a "simple" spreadsheet model [38]	Task is to build and code a spreadsheet model.
Visual inspection reliability for precision manufactured parts [39]	Visual inspection of manufactured parts used in nuclear weapons (accept or reject correct and defect parts).

manipulated this). This was also not discussed by Williams and Bell [18].

Difficulties in Obtaining Empirical Data on Complexity From Event Reports, Training and Operation

It has been suggested to collect empirical data from event reports, operational data, and training. Collecting data from event reports, training, and operation have the same challenges, with definition of complexity, complexity subcategories, complexity levels/measuring scales, and tasks that are described for experiments.

Several trials have been performed to include event reports in databases to produce HRA data. This has not been successful since little HRA data has come from this effort (Ref. [8]). The difficulty of obtaining data from event report is that these usually only report when an error occurs. It is difficult to know if some particular PSFs/EFCs, such as complexity, are present when the error occurred compared to normal operations [8,40]. A particular challenge in analyzing complexity in event reports is that the event reports are usually not very detailed and it is very difficult to understand which subcomponents of complexity are involved. Boring et al. [3] and Liao et al. [6] also say that the events occur so infrequently that they do not give much data for HRA. Also, another challenge in using event reports to collect HRA data for complexity is that when large accidents happen, such as, for example, the three mile island incident, ATHEANA [10] claims that these accidents are particularly complex. If this is correct, it is not possible to obtain data on that form for complexity in the larger accidents that are usually analyzed in HRA from the event reports (on smaller event and incidents), since those event reports probably contain less complexity or another form for complexity.

Training in simulators in nuclear power plants has also been suggested as a way to obtain HRA data. To obtain data on complexity from training, the researcher must first conduct an HRA analysis to define the complexity subcomponents and their levels, so this data collection will also include the challenges with the definitions as described under experiments.

It is also challenging to obtain data from training because training by its nature is developed to be training and not to test all subcomponents of complexity apart from other PSFs. Training usually includes several PSFs/EPCS and it is therefore difficult to separate the effect on complexity from the other PSFs.

Operational data would often include normal daily work activities in different industries or control room. This has been investigated for some PSFs, such as Fatigue (see for a summary [41]). Also, for operational data, it is difficult to separate complexity subcomponents from other PSF/EPCs and other kinds of influences that might have an effect on performance and error rate.

Summary and Conclusion

This paper shows that a pure positivism paradigm of empirical data collection for HRA data is very challenging and that it might not be possible to collect empirical evidence in the positivism paradigm for the quantification parts of the HRA methods. There are too many challenges, such as, for example:

(a) Challenges with definitions—The HRA methods are different in how they define the content of their methods. Complexity is defined differently by the different HRA methods. Within the methods, there is not an objective description of the levels/proportion of affect for the different PSFs/EPCs. This is subjectively determined. To measure the effect of PSFs/EPCs on performance, we need objectively measuring scales of the levels and these are challenging to develop. It is also difficult to know what is considered a task in the different methods, which makes it difficult to decide which errors should be considered as errors. The problems with definitions in HRA make all kinds of empirical data collection difficult.
(b) Challenges in creating experimental manipulations (scenarios) based on the definitions. The information in the HRA methods about complexity is abstract and it is difficult to transform the information into scenarios or tasks that can be tested.
(c) The amount of experiments and experimental manipulations needed—There are many complexity subcomponents and to collect all the data on all of them seems to be unmanageable.
(d) The amount of subjects needed for experiments is unrealistic high.
(e) It is difficult to match information in the general research literature to the information needed in HRA.
(f) Events happen so infrequently that it is difficult to get data on complexity from events. It might also not be the same kinds of complexity in smaller events that occur in large accident scenarios, which are analyzed in the HRA methods.
(g) In incidents and events reports, in training, and in operations, it is difficult to separate complexity and other PSFs/EPC or other conditions that might affect error rates.
(h) Generalization or transferability of results from one task or setting to another task or settings. Ideally, data should be collected from several tasks and several settings to investigate the transferability of the data. In HRA, it is difficult to collect these data from any context, so it is not likely that the data will be collected in many contexts.

Pure positivistic empirical data for HRA have been demanded since the development of HRA. Since collecting HRA data is such a large, time-consuming, and difficult task to conduct, we should consider whether these types of pure positivistic empirical data will ever be collected.

There are significant differences in the methods on the degree to which they assume that complexity would affect performance. The highest multiplier you could set in SPAR-H for complexity alone is 5, and for a diagnosis task the highest estimated error rate would be 0.05. ATHEANA seems to say that complexity could make a task very difficult to perform for most operators, which implies a much higher error rate. The reasons for the differences are that the methods are based on very different sources and literature reviews.

Bayesian statistics have been suggested as a way to increase the confidence in HRA data [4]. Hallbert et al. [4] say; "the Bayesian framework accommodates different forms of 'evidence' including indirect observations, model predictions and expert judgment as well as actual data." However, it is questionable whether we actually get a good estimate if we include all kinds of evidence on complexity, with all the different definitions of complexity, its different levels, and different multipliers from the HRA methods in to a Bayesian analysis. To be meaningful in these types of analyses, one has to be careful to only include data that has been measured on the same concepts.

The quantitative data in the HRA methods that exist are not objective empirical data. In this paper, complexity was discussed as an example, and it shows that the error rates and multipliers on the PSF complexity are not always logical and that they differ significantly between HRA methods. Based on this, we should discuss whether we would obtain better HRA data if systematic experts' judgments (expert elicitation or expert estimation) were used as a basis for error rates in HRA. Boring [42] defines expert estimation or elicitation as; "expert elicitation involves polling subject matter experts to produce a probability of human error or hardware failure." In HRA, experts' judgments have been used in two different ways: (a) where experts evaluate the evidence or the qualitative data in one particular analysis and do an expert estimation of the failure rates, as done in ATHEANA [10]; and (b) expert's judgments could be used to define the error rates and the quantitative elements (nominal failure rate, multipliers, etc.) in HRA methods.

In HRA, there has been much focus on how to mathematically combine data from different sources and different experts (for example, Ref. [43]). Less attention (with exception of Ref. [42])

has been given do defining the qualitative part of an expert's judgment and defining who are the experts and how to get good estimates from the experts, such as, for example, which questions to ask the experts or which information to present to the experts. These topics should be further researched.

Another part of HRA that should be further developed is that HRA methods as they are today are not user-friendly for the analysts or the researcher. HRA descriptions of task types and PSFs come from academic psychological research. The descriptions are abstract and generic, which make them transferable to many different contexts and tasks, but it also makes it difficult to transfer the information in the HRA method about the PSFs and PSF levels to real tasks and scenarios under analysis or study. The descriptions in HRA methods also include overlapping PSFs and inconsistent information (for a discussion see, for example, Ref. [44]). It is also difficult to transfer the information in the HRA methods on PSF/EPCs to experimental manipulations. HRA method leaves a lot up to the analyst to decide on and there are few practical examples of how to transfer information in the HRA methods to the actual tasks. The HRA analyses I have seen have also not described their comparison of the information in the HRA methods to the information about the actual task. If more of these types of qualitative information could be collected, and if we could agree on how the data from real tasks should (ideally) be analyzed with a particular HRA method, we might make the analyses more consistent and the methods easier to use both in practice and in research.

References

[1] Swain, A., 1990, "Human Reliability Analysis: Need, Status, Trends and Limitations," Reliab. Eng. Syst. Saf., **29**(3), pp. 301–313.

[2] Mkrtchyan, L., Podofillini, L., and Dang, V. N., 2015, "Bayesian Belief Networks for Human Reliability Analysis: A Review of Applications and Gaps," Reliab. Eng. Syst. Saf., **139**, pp. 1–16.

[3] Boring, R., Kelly, D., Smidts, C., Mosleh, A., and Dyre, B., 2012, "Microworlds, Simulators, and Simulation: Framework for a Benchmark of Human Reliability Data Sources," Joint Probabilistic Safety Assessment and Management and European Safety and Reliability Conference (PSAM), Helsinki, Finland, Paper No. 16BTu5-5.

[4] Hallbert, B., Gertman, D., Lois, E., Marble, J., Blackman, H., and Byers, J., 2004, "The Use of Empirical Data Sources in HRA," Reliab. Eng. Syst. Saf., **83**(2), pp. 139–143.

[5] Kim, Y., and Park, J., 2018, "Suggestions of HRA Method Improvement for the Practical Assessment of Human Reliability," J. Ergon. Soc. Korea, **37**(3), pp. 229–241.

[6] Liao, H., Groth, K., and Stevens-Adams, S., 2015, "Challenges in Leveraging Existing Human Performance Data for Quantifying the IDHEAS HRA Method," Reliab. Eng. Syst. Saf., **144**, pp. 159–169.

[7] Williams, J. C., 1985, "Validation of Human Reliability Assessment Techniques," Reliab. Eng., **11**(3), pp. 149–162.

[8] Laumann, K., Blackman, H., and Rasmussen, M., 2018, "Challenges With Data for Human Reliability Analysis," Safety and Reliability–Safe Societies in a Changing World, Proceedings of ESREL 2018, CRC Press, Trondheim, Norway, June 17–21, pp. 315–321.

[9] Guba, E. G., and Lincoln, Y. S., 1994, "Competing Paradigms in Qualitative Research," *Handbook of Qualitative Research*, K. Denzin, and Y. S. Lincoln, eds., Sage Publications, Thousand Oaks, CA, pp. 105–117.

[10] Forester, J., Kolaczkowski, A., Cooper, S., Bley, D., and Lois, E., 2007, "ATHEANA User's Guide-Final Report," U.S. Nuclear Regulatory Commission, Washington DC, Report No. NUREG-1880.

[11] Gertman, D., Blackman, H., Marble, J., Byers, J., snd., and Smith, C., 2005, "The SPAR-H Human Reliability Analysis Method," U.S. Nuclear Regulatory Commission, Washington, DC, Report No. NUREG/CR-6883.

[12] Bye, A., Laumann, K., Taylor, C., Rasmussen, M., Øie, S., van de Merwe, K., Øyen, K., Boring, R., Paltrini, N., Wærø, I., Massaiu, S., and Gould, K., 2017, "The Petro-HRA Guideline," Institute for Energy Technology, Halden, Norway.

[13] Williams, J., 1988, "A Data-Based Method for Assessing and Reducing Human Error to Improve Operational Performance," Conference Record for 1988 IEEE Fourth Conference on Human Factors and Power Plants, Monterey, CA, June 5–9, pp. 436–450.

[14] Williams, J. C., 1992, "A User Manual for the HEART Human Reliability Assessment Method," Nuclear Electric plc, Gloucester, UK, DNV Technical Report No. C2547.

[15] Swain, A. D., and Guttmann, H. E., 1983, "Handbook of Human-Reliability Analysis With Emphasis on Nuclear Power Plant Applications," Sandia National Laboratories, Albuquerque, NM, Report No. NUREG/CR-1278.

[16] Boring, R. L., and Blackman, H. S., 2007, "Origins of the SPAR-H Method's Performance Shaping Factor Multipliers," IEEE Eighth Human Factors and Power Plants and HPRCT 13th Annual Meeting, Monterey, CA, Aug. 26–31, pp. 177–184.

[17] Rasmussen, M., Standal, M. I., and Laumann, K., 2015, "Task Complexity as a Performance Shaping Factor: A Review and Recommendations in Standardized Plant Analysis Risk-Human Reliability Analysis (SPAR-H) Adaption," Saf. Sci., **76**, pp. 228–238.

[18] Williams, J. C., and Bell, J. L., 2016, "Consolidation of the Generic Task Type Database and Concepts Used in the Human Error Assessment and Reduction Technique (HEART)," Saf. Reliab., **36**(4), pp. 245–278.

[19] Williams, J. C., and Bell, J. L., 2015, "Consolidation of the Error Producing Conditions Used in the Human Error Assessment and Reduction Technique (HEART)," Saf. Reliab., **35**(3), pp. 26–76.

[20] Bell, J. L., and Williams, J. C., 2017, "'Evaluation and Consolidation of the HEART Human Reliability Assessment Principles," *Advances in Human Error, Reliability, Resilience, and Performance, AHFE 2017. Advances in Intelligent Systems and Computing*, R. Boring, ed., Vol. 589, Springer, Cham, Switzerland, pp. 3–12.

[21] Rasmussen, M., and Lauman, K., 2016, "Decomposition Level of Quantification in Human Reliability Analysis," *Risk, Reliability and Safety: Innovating Theory and Practice*, L. Walls, M. Revie, and T. Bedford, eds., Taylor & Francis Group, London, pp. 997–1002.

[22] Bishop, B., and McDaid, K., 2007, "An Empirical Study of End-User Behaviour in Spreadsheet Error Detection & Correction," European Spreadsheet Risks Interest Group, pp. 165–176.

[23] Doran, M. M., and Hoffman, J. E., 2010, "The Role of Visual Attention in Multiple Object Tracking: Evidence From ERPs," Attention, Perception, Psychophysics, Vol. 72, 1, pp. 33–52.

[24] Hajcak, G., and Simons, R. F., 2008, "Oops!. I Did It Again: An ERP and Behavioral Study of Double-Errors," Brain Cognit., **68**(1), pp. 15–21.

[25] Menon, V., Boyett-Anderson, J. M., Schatzberg, A. F., and Reiss, A. L., 2002, "Relating Semantic and Episodic Memory Systems," Cognit. Brain Res., **13**(2), pp. 261–265.

[26] MiRadke, S., De Lange, F. P., Ullsperger, M., and De Bruijn, E. R. A., 2011, "Mistakes That Affect Others: An fMRI Study on Processing of Own Errors in a Social Context," Exp. Brain Res., **211**(3–4), pp. 405–413.

[27] Smith, W., and Dror, I. E., 2001, "The Role of Meaning and Familiarity in Mental Transformations," Psychon. Bull. Rev., **8**(4), pp. 732–741.

[28] White, D., Kemp, R. I., Jenkins, R., Matheson, M., and Burton, A. M., 2014, "Passport Officers' Errors in Face Matching," PloS One, **9**(8), p. e103510.

[29] Cave, K. R., and Pashler, H., 1995, "Visual Selection Mediated by Location: Selecting Successive Visual Objects," Percept. Psychophysics, **57**(4), pp. 421–432.

[30] Folk, M. D., and Luce, R. D., 1987, "Effects of Stimulus Complexity on Mental Rotation Rate of Polygons," J. Exp. Psychol.: Hum. Percept. Perform., **13**, pp. 395–404.

[31] Hall, D. L., and Friedman, A., 1994, "Shape Discriminations of Three-Dimensional Objects Depend on the Number and Location of Bends," Percept. Psychophysics, **56**, pp. 288–300.

[32] Murray, M. M., De Santis, L., Thut, G., and Wylie, G. R., 2009, "The Costs of Crossing Paths and Switching Tasks Between Audition and Vision," Brain Cognit., **69**(1), pp. 47–55.

[33] Ramautar, J. R., Slagter, H. A., Kok, A., and Ridderinkhof, K. R., 2006, "Probability Effects in the Stop-Signal Paradigm: The Insula and the Significance of Failed Inhibition," Brain Res., **1105**(1), pp. 143–154.

[34] Soldan, A., Mangels, J. A., and Cooper, L. A., 2006, "Evaluating Models of Object-Decision Priming: Evidence From Event-Related Potential Repetition Effects," J. Exp. Psychol.: Learn., Memory, Cognit., **32**(2), pp. 230–248.

[35] Chastain, G., and Cheal, M., 2001, "Attentional Capture With Various Distractor and Target Types," Percept. Psychophysics, **63**(6), pp. 979–990.

[36] Fougnie, D., and Marois, R., 2009, "Perception, and Psychophysics, Dual-Task Interference in Visual Working Memory: A Limitation in Storage Capacity but Not in Encoding or Retrieval," Atten., Percept. Psychophysics, **71**(8), pp. 1831–1841.

[37] Katz, I. R., and Anderson, J. R., 1987, "Debugging: An Analysis of Bug-Location Strategies," Hum.-Comput. Interact., **3**(4), pp. 351–399.

[38] Panko, R. R., 1996, "Hitting the Wall: Errors in Developing and Debugging a "Simple" Spreadsheet Model," Proceedings of HICSS-29, 29th Hawaii International Conference on System Sciences, Vol. 2, pp. 356–363.

[39] See, J. E., 2015, "Visual Inspection Reliability for Precision Manufactured Parts," Hum. Factors, **57**(8), pp. 1427–1442.

[40] Kim, Y., Park, J., Jung, W., Jang, I., and Seong, P. H., 2015, "A Statistical Approach to Estimating Effects of Performance Shaping Factors on Human Error Probabilities of Soft Controls," Reliab. Eng. Syst. Saf., **142**, pp. 378–387.

[41] Rasmussen, M., and Laumann, K., 2018, "The Evaluation of Fatigue as a Performance Shaping Factor in the Petro-HRA Method," Reliab. Eng. Syst. Saf. (in press).

[42] Boring, R. A., 2007, "A Review of Expertise and Judgment Processes for Risk Estimation. Reliability and Societal Safety, Volume 2: Thematic Topics," European Safety and Reliability Conference (ESREL 2007), Taylor & Francis, London, UK, pp. 1901–1907.

[43] Podofillini, L., and Dang, V. N., 2013, "A Bayesian Approach to Treat Expert-Elicited Probabilities in Human Reliability Analysis Model Construction," Reliab. Eng. Syst. Saf., **117**, pp. 52–64.

[44] Laumann, K., and Rasmussen, M., 2016, "Suggested Improvements to the Definitions of Standardized Plant Analysis of Risk-Human Reliability Analysis (SPAR-H) Performance Shaping Factors, Their Levels and Multipliers and the Nominal Tasks," Reliab. Eng. Syst. Saf., **145**, pp. 287–300.