SAND2016-5620J

Assessment of HRA Method Predictions against Operating Crew Performance: Part II: Scenario Description, Human Failure Events, Overall Simulator Data, and HRA Method Predictions

Abstract: This is the second in a series of four papers documenting two large-scale human reliability analysis (HRA) empirical studies – the International HRA Empirical Study and the US HRA Empirical Study. The goal of the two studies was to develop an empirically-based understanding of the performance, strengths, and weaknesses of HRA methods by comparing HRA method predictions against actual operator performance in simulated accident scenarios on nuclear power plant (NPP) simulators. The first paper (Part I), provided background information for the studies and an overview of their design and methodology. This paper first briefly describes the scenarios simulated in the studies and the associated human failure events (HFEs) addressed in the HRA analyses. Then, it discusses the overall simulator data followed by observations on the operating crew performance in the scenario simulations. Finally, it presents some quantitative comparisons of the HRA methods' predictions with the simulator data.

Keywords: human reliability analysis, probabilistic risk assessment, simulation, human error probability, human failure event, nuclear power plant

1 Introduction

This is the second in a series of four papers [1-3] discussing two large-scale human reliability analysis (HRA) empirical studies – *the International HRA Empirical Study* (hereafter "the International Study") [4-7], and *the US HRA Empirical Study* (hereafter "the US Study") [8]. The goal of the two studies was to develop an empirically-based understanding of the performance, strengths, and weaknesses of HRA methods by comparing HRA method predictions against actual operator performance in simulated accident scenarios on nuclear power plant (NPP) simulators.

An overview of the study design and methodology has been provided in Paper 1 [1]. This paper first briefly describes the scenarios simulated in the studies and the associated human failure events (HFEs) addressed in the HRA analyses. Then, it discusses the overall simulator data followed by observations on the operating crew performance from the scenario simulations. Finally, it presents some quantitative comparisons of the HRA methods' predictions with the simulator data.

2 Scenario description and human failure events (HFEs)

2.1 The International Study

Two categories of scenarios were developed in the International Study: steam generator tube rupture (SGTR) and loss of feed water (LOFW) scenarios (see [4-7] for more detailed scenario

description). There are two scenario variants in each category, which are referred to as the *base case* scenario and the *complex* scenario. As described in Paper 1, the base case scenarios involve relatively advantageous conditions and are similar to incidents that might be trained on during routine simulator training. The complex scenarios involve relatively adverse conditions complicated by secondary malfunctions.

2.1.1 SGTR scenarios

2.1.1.1 SGTR base scenario

SGTR base scenario is a standard SGTR scenario without added complications. The plant is initially operating at 100% power. At the start of the scenario, a tube rupture occurs in an SG, which is sufficient to cause nearly immediate secondary radiation alarms and other abnormal indications/alarms, such as SG abnormal levels and lowering pressurizer levels. Conditions, while continually degrading, are not enough to cause an immediate automatic scram.

2.1.1.2 SGTR complex scenario

Compared to the SGTR base scenario, the SGTR complex scenario has the following two additional complications:

- The SGTR starts off with a coincidental steamline break, which will cause an immediate automatic scram.
- The Main Steam Isolation Valves (MSIVs) automatically close as expected in response to the steamline break. As part of the simulation design, the valve closure is coincidental to

the failure of secondary radiation indications, but the indication failure is not immediately known nor expected by operators.

2.1.1.3 HFE definitions

Nine HFEs are defined for the two variants of the SGTR scenarios. Related HFEs have the same number identifier while the letter in the identifier is "A" for the base scenario and "B" for the complex scenario.

- HFE 1A and HFE 1B: Failure to identify and isolate the ruptured SG.
- HFE 2A and HFE 2B: Failure to cool down the reactor coolant system (RCS) expeditiously.
- HFE 3A and HFE 3B: Failure to depressurize the RCS expeditiously.
- HFE 4A: Failure to stop the safety injection (SI).
- HFE 5B1: Failure to give a closing order to the PORV block valve associated with the partially open PORV within five minutes of closing the PORV used to depressurize the RCS when the PORV position indication shows "closed."
- HFE 5B2: Failure to give a closing order to the PORV block valve associated with the partially open PORV within five minutes of closing the PORV used to depressurize the RCS when the PORV position indication shows "open."

Note that half of the crews received the PORV position indication showing open (HFE 5B2) and the other half received the PORV position indication showing closed (HFE 5B1).

2.1.2 LOFW scenarios

2.1.2.1 LOFW base scenario

If feedwater cannot be re-established following a total LOFW, Bleed and Feed (B&F) of the RCS should be started when the SG wide range (WR) level is less than 12% in two of three SGs, or the reactor pressure is high due to loss of secondary heat sink. B&F consists of manually starting safety injection pumps and opening the pressurizer relief valves.

2.1.2.2 LOFW complex scenario

The LOFW complex scenario is complicated by the following two equipment malfunctions.

- Since one condensate pump is running, the crew can establish condensate flow to the SGs by depressurizing the SGs to a pressure lower than the discharge pressure of the condensate pump. However, the pump is degraded with a discharge pressure lower than normal; therefore, the condensate flow cannot be established to the SGs before the SGs become empty.
- Two of the three SGs have WR level indicators incorrectly show a steady value somewhat above 12% when the actual level is 0%. The crew has to identify and diagnose the indicator failures, since 12% is the criterion to start B&F, which will never be met when interpreted literally.

2.1.2.3 HFE definitions

Four HFEs are defined for both variants of the LOFW scenarios. Similar to SGTR scenarios, HFEs denoted as "B" are HFEs in the complex scenarios, while "A" represents HFEs in the base scenarios.

- HFE 1A and HFE 1B: Failure to establish/initiate B&F before SG dryout. SG dryout occurs when there is no water left in the SGs, indicated by 0% WR SG level.
- HFE 2A and HFE 2B: Failure to establish/initiate B&F within 25 minutes of SG dryout. HFE 2A and 2B are conditional on HFE 1A and HFE 1B, respectively.
- HFE 1A1: Joint failure of 1A and 2A (i.e. HFE 1A * HFE 2A) in the base scenario variant.
- HFE 1B1: Joint failure of 1B and 2B (i.e. HFE 1B * HFE 2B) in the complex scenario variant.

2.2 The US Study

Three scenarios were developed in the US Study (see [8] for more detailed scenario description).

2.2.1 Scenario 1 – Total loss of feedwater (LOFW) followed by steam generator tube rupture (SGTR)

2.2.1.1 Total loss of feedwater (LOFW)

The plant is initially operating at 100% power. There are three main feedwater (MFW) pumps and four auxiliary feedwater (AFW) pumps. Two minutes into the scenario, all main feedwater pumps are tripped and the start-up feed pump cannot start. If the crew fails to manually trip the reactor, it will automatically trip on low steam generator (SG) level within 50-60 seconds.

Three AFW pumps fail after automatic start. The fourth AFW pump will start automatically and indicate full flow, but this flow will not reach the SGs because a recirculation valve is mis-positioned open. There is no indication of the valve's position in the control room, thus the open recirculation valve will mask the fact that no AFW at all is going into the SGs. The SG levels will go down, and the plant computer will not show a red path on the heatsink status tree because of the indicated flow from the running AFW pump. The crew will have to identify that the indication of AFW flow from the running AFW pump is false and establish Bleed and Feed (B&F) when the wide range (WR) level on any two SGs are less than 50%. All attempts to establish AFW before B&F initiation will fail.

Two HFEs are defined for the LOFW part of Scenario 1.

- HFE 1A: Failure to establish bleed and feed within 45 minutes of the reactor trip, if the crew initiates a manual reactor trip before an automatic reactor trip.
- HFE 1B: Failure to establish feed and bleed within 13 minutes of the reactor trip, if the crew does not manually trip the reactor before an automatic reactor trip occurs.

2.2.1.2 Steam generator tube rupture (SGTR)

After B&F has been established, the crew will be able to establish AFW flow to one or several SGs. As soon as the AFW flow is established, a tube rupture occurs in the first SG that is fed. The tube rupture may be masked by the AFW flow to the SG as long as it is being fed. The leak size of the ruptured tube is about 500 GPM at 100% power, but the flow will depend on the differential pressure between the RCS and the ruptured SG. As a result of B&F, there is initially no secondary radiation because there is only a minimum steam flow, and the blowdown (BD) and sampling is secured because of the SI. The crew may have problems with the RCS integrity status tree when they try to terminate B&F, which will their response to SGTR.

One HFE is defined for the SGTR part of Scenario 1.

- HFE 1C: Failure to isolate the ruptured steam generator and control pressure below the SG PORV setpoint to avoid SG PORV opening. The time window to perform the required actions is estimated to be approximately 40 minutes.

2.2.2 Scenario 2 – Loss of CCW and RCP sealwater

The plant is initially operating at 100% power and CCW Pump B is unavailable. Two minutes into the scenario, Distribution Panel 1201 fails. As a consequence, the crew has to establish manual control of SGs, pressurizer, control rods, and nuclear instrumentation. The feedwater regulation valve on SG A cannot be operated manually and remains fully open feeding the SG. If the crew does not trip the reactor, there will be an automatic turbine trip on high SG level (87%), which causes a reactor trip. When the reactor trips, CCW Pumps A and C and Charging Pumps A and B become unavailable. As a result, there is no injection flow to RCP seals and no

CCW flow to RCP thermal barriers. The failed distribution panel is unrelated to the loss of CCW and sealwater but increases the complexity of the scenario. It masks the status of CCW and sealwater by keeping the crew busy due to the number of alarms.

The following HFE is defined for Scenario 2.

- HFE 2A: Failure to trip the RCPs and start the Positive Displacement Pump (PDP) to prevent RCP seal loss of coolant accident (LOCA). Any RCP that experiences a simultaneous loss of seal injection flow and loss of CCW flow to thermal barriers shall be stopped within 1 minute after determining that both RCS seal injection flow and thermal barrier cooling were lost. The PDP can only be started if the RCP seal temperatures are below 230^{o}F.

2.2.3 Scenario 3 – Steam generator tube rupture (SGTR)

Scenario 3 is a standard SGTR scenario without added complications. The plant is initially operating at 100% power. One minute into the scenario, a tube rupture occurs in an SG. The leak size is about 500 GPM at 100% power.

The following HFE is defined for Scenario 3.

- HFE 3A: Failure to isolate the ruptured steam generator and control pressure below the SG PORV setpoint before SG PORV opening. The time window to perform the required actions is estimated to be 2 to 3 hours.

3 Overall empirical simulator data

This section presents the overall empirical simulator data, which are expressed in terms of success and failure, human failure event (HFE) difficulty ranking, and failure bounds. See [4-8] for detailed simulator results in terms of operational descriptions and performance-shaping factor (PSF) evaluations.

3.1 Failure rates and empirical HEPs

Fourteen crews participated in the International Study and four crews participated in the US Study. The numbers of observations and failures for each HFE are listed in Tables 1a – 1c. As discussed in Section 3.7.1 in Part I, the number of crew observations is large for simulator studies, but they represent a small sample and result in generally low level of statistical significance, considering the HEPs' expected range of values, particularly for those response actions where the HEPs would be expected to be low. Consequently, the empirical HEPs were not estimated using the binomial probability distribution (i.e. failure counts in number of trials). Instead, the 90th percentile uncertainty bounds of the empirical HEP distributions were derived from a Bayesian process, in which the observations were interpreted as evidence for updating a prior distribution (see [4-8] for more information about the updating process).

[Insert Tables 1a – 1c about here]

The 5th and 95th percentiles of the empirical HEP distributions for each of the HFEs obtained from the Bayesian updates are shown in Figures 1a – 1c. On the horizontal axis, the HFEs are ordered by the difficulty ranking with difficulty decreasing from left to right (see more discussion on difficulty ranking in Section 3.2).

[Insert Figures 1a – 1c about here]

- The uncertainty bounds of the posterior empirical HEP distributions are fairly insensitive to the choice of the prior distribution. For example, a minimally-informed prior and the Jeffreys prior resulted in roughly the same confidence bounds for the SGTR scenarios in the International Study. Given the moderate priority given to the quantitative assessment criteria, it is expected that the choice of the prior would basically leave the method assessments unchanged.
- The purpose of the Bayesian update is to obtain the uncertainty bounds of the empirical HEPs rather than to assess the means and medians, because the means or the medians of the posterior distributions are very sensitive to the prior. In addition, the means and the medians are fairly large (between 1E-2 and 1E-1) even for the cases where no failures are observed. This is an effect of the weak evidence due to the small sample – low failure probabilities need to be supported by a large number of observations. The use of the mean or median values would suggest more accuracy in the empirical HEPs than warranted by the limited sample size, especially for cases where the uncertainty bound is broad.

- When several failures are observed, the Bayesian update yields narrow confidence intervals. This evidence is strong (in spite of the sample size) and the posterior becomes comparable to the result obtained from a classical, frequentist statistical analysis (not shown). In contrast, when no failures are observed, the posterior distribution spans about three orders of magnitude. Only the larger HEPs (above 0.1) are outside the confidence interval. As mentioned above, this is an effect of the weak evidence due to the small sample.
- As mentioned in Section 3.2 in Part I, the HFE success criteria used in the studies are in several cases more demanding than the criteria that would be typically used in many PRA studies. As a result, the empirical HEPs found in this study are in some cases comparatively conservative. In any event, the failure counts and HEPs in the studies are larger and not strictly comparable to the values for similar HFEs in a PSA study. It is worth noting that in a number of crew performances where the crew failed to meet the success criteria defined for a given HFE, the crew subsequently performed the required task and successfully managed the initiating event.
- The simulator observations, interpreted as failure counts for joint HFEs, resulted in 0 failures in 10 observations for both joint HFE 1A1 and HFE 1B1 in the LOFW scenario. The corresponding uncertainty bounds for HFEs 1A1 and 1B1 would be the same as for HFE 1A, that is, broad and therefore limited in providing insights. Secondly, the uncertainty bounds do not discriminate between the two HFEs. On the other hand, the difficulty of HFE 1B1 relative to HFE 1A1, considering when B&F is implemented relative to the procedural criteria and qualitative considerations, is unambiguous.

- Both the mean values and the breadth of the confidence intervals obtained in this way are not representative of HEP distributions that could be used in a PRA. For the HRA method assessment, the confidence bounds are indeed appropriate because the method assessment is not based on predicting the empirical HEPs (with their limited statistical significance). Instead the assessment considers whether the mean values obtained in the HRA analysis are consistent with the simulator-based confidence intervals of the empirical HEPs.
- To produce HEP distributions to be used in a PSA, this Bayesian update would be insufficient. One would need a different analytical process, presumably incorporating more expert judgment.

3.2 HFE difficulty rankings

The difficulty rankings of the HFEs are listed in Tables 1a – 1c (see Section 3.3 of Part I for description on the methodology for the ranking). The difficulty ranking is mostly but not fully consistent with the ranking according to the confidence intervals of the empirical HEPs. It is not based solely on the empirical HEPs but is instead an overall, qualitative, and partly subjective assessment of the relative difficulty (a relative failure likelihood) that combines the basis for the empirical HEPs, i.e., the number of failures based on the HFE definitions, with qualitative considerations of the performance. Where the empirical HEPs are derived solely from the number of observed failures and the number of total observations, the qualitative considerations included in the difficulty ranking accounted additionally for other objective evidence from the simulation such as the performance as measured by plant parameters, the amount by which the success criteria were missed (in terms of the time windows defined for the HFEs or the plant

parameters), and the difficulties experienced by the crews (even if these difficulties were surmounted) during the tasks associated with the HFE.

Two HFEs will have the same empirical HEP uncertainty bounds if they have the same sample size and number of failure counts. The qualitative data from the simulator incorporated in the difficulty ranking allow these HFEs to be distinguished, e.g., by considering the performance difficulties associated with the HFE that are observed in the simulator. In determining the difficulty ranking, the qualitative simulator data is emphasized in cases where the significance of the failure count is particularly low due to the sample size. In other words, this emphasis avoids assuming that an HFE with one observed failure has a higher failure likelihood than one with no observed failures. Finally, near-failures can be included in these qualitative considerations.

In the SGTR scenarios in the International Study, for example, there is a tie in the difficulty ranking between the SGTR HFEs 1A, 2A and 2B. It can be seen that this difficulty ranking is not consistent with the confidence intervals of the empirical HEPs, which would instead indicate a tie between SGTR HFEs 3A, 1A, and 2A. While the failure counts were identical for SGTR HFEs 3A and 1A, more performance difficulties were observed for 3A than for 1A. This distinction is not seen in the confidence intervals, which rely on the failure counts, but is seen in the difficulty ranking. Similarly, no failures were observed for HFEs 2B, 5B2, and 4A. However, more performance difficulties were observed for 2B than for 5B2 and for 5B2 than for 4A. While these difficulties may not be accurate predictors of HFE failures, they do indicate an increased degree of difficulty and a potential source of failures [9].

4 Observations from scenario simulation

4.1 Variability of performance

The simulator data showed a large degree of variability in the ways in which different crews responded to the scenarios. As discussed in Section 3.3 in Part I, this was not unexpected because the scenarios were set up so as to evolve in ways as influenced by operators. The comprehensive emergency response guidelines do not restrict crew responses within strict boundaries, simply because they do not cover the situational variations created or prompted by some of the study's scenarios in enough detail. In addition, some scenarios were designed to be challenging so that crew performance variability could be observed. Process variability beyond outcome variability (i.e. variability in performance quality beyond the mere failure counts) was necessary to rate the difficulty of the HFEs.

Some variability arises from differences in initial crew actions in the scenarios, which lead to cascading differences in plant configurations, timing of the cues, and other scenario dynamics. One example is differences in procedure progressions (i.e. different crews move differently through procedures), including differences in procedural paths, progression times, and procedure transfer criteria. Other sources of variability relate to differences in the teams' internal functioning (i.e. how the crew members interact with each other), including differences in task allocation, information communication, and decision making. This type of variability is generally not treated within the scope of most HRAs; however, the empirical observations confirm that the team's functioning is an important underlying cause of differences in crew performance.

Overall, the studies suggest that HRA analysts may not give sufficient attention to variability of scenario development, which is caused by complicating factors, cognitive requirements of procedure-following, and differences in teamwork and expertise.

4.2 Challenges and issues in operator-procedure interaction

This section lists a set of challenges and issues observed in operator-procedure interaction in the simulated scenarios (see [5, 10] for more discussion and specific examples). The observations are based on specific operator behaviors and operational contexts, as well as on the overall tendencies observed in the crew samples. General topics are also inferred from the empirical observations and presented as potential problems. Although these observations are not used for HRA method assessment, they can clarify general aspects of operator-procedure interactions, explain some of the discrepancies between method prediction and actual crew performance, as well as suggest elements for improved qualitative HRA analyses.

- *Misreading or skipping procedure steps* is an example of slips and lapses that operators make in following procedures. It can occur even when the information in the procedure is clear and unambiguous and when operators use human performance tools such as signing off on completed steps. Such slips and lapses can lead to an incorrect plant status assessment or initiation of incorrect response.
- *Verbatim procedure following*. Operators may literally follow procedures without understanding the intent of the procedures and observed plant symptoms. Literal following is also observed even when operators counter well-understood goals, such as

waiting for conditions to worsen to meet a literal condition in a procedure. It seems that some operators cope with challenging and stressful situations by literally following the procedures to reduce their cognitive efforts to a minimum.

- *Premature termination of plant parameter trend assessment*. Observing and assessing plant parameter trends is one aspect of procedure following that is particularly dependent on operators' expectations and evaluations, as time and other boundary conditions are not typically specified. In an emergency situation, plant parameters change dynamically. There are times when a specific parameter satisfies a procedure step entry criterion before the step entry but then changes in the opposite direction after the operators leave the step. An insufficient trend assessment can cause operators to develop an incorrect mental model of the plant status. Problems might arise when the operators have to decide whether a plant's behavior is the result of known actions (manual or automatic) or of a plant fault.
- *Inadequate procedure guidance*. Due to the complex nature of the processes in NPPs, it is difficult to develop EOPs that cover every possible contingency in detail. In some cases, the actions within a procedure step do not fulfill the goal of the step under a scenario-specific situation. In some other cases, the actions of a step are described but the rationale or intent of the step is not explicitly specified. Although operators may feel challenged when there is inadequate procedure guidance, training can help them handle the gap between the EOPs and the real situations.
- *Decisions and judgments based on operator's knowledge*. When facing cognitively challenging situations involving complicating factors, operators sometimes need to rely on their knowledge to differentiate expected from unexpected plant behavior and pursue

all possibilities for the unexpected behavior. Similarly, under circumstances when there is inadequate procedure guidance, operators need to rely on their knowledge to make judgments and take necessary actions outside procedures.

- *Ineffective communication and teamwork*. Crew members have their unique roles and responsibilities at the plant. They sometimes often need to simultaneously work in different procedures or on different tasks in complex dynamic situations. For example, one operator may be assigned to a task at a particular procedure while the rest of the crew continues in the procedure. The division of work increases the requirement for effective communication and teamwork because the operator who works on a separate task needs to update the rest of the crew on the task status but the rest of the crew has a new focus as they continue in the procedure. This may cause a problem if the rest of the crew needs to reevaluate the plant status or perform some actions based on the operator's update but they are distracted by their new focus. In addition, some procedure steps involve actions to be performed both in the control room and in the plant. The completion of such actions require effective crew communication and coordination, especially when the actions are not logically separated and prioritized.
- *Ineffective foldout page use*. The foldout page (reference page) should be read when starting an EOP and kept open, as it includes several continuous conditions (i.e., actions or transitions that are applicable at any step in the procedure body). It has been observed that (1) foldouts are not always read through before starting a procedure; (2) foldouts might be read without conditions being followed; and (3) continuous conditions might not be monitored or enacted when relevant.

- *Execution and procedure following complexity*. The complexity in executing procedure steps has traditionally been associated with structural elements such as language clarity, syntactical complexity (e.g., present of double negatives and passive statements), and number of sub-steps, as well as substantive aspects, such as training and experience of the operators for the task involved or complex behavior of the equipment used.
- *Notes and Cautions.* In the Westinghouse EOPs, notes and cautions contain special information that do not follow the two-column format: notes contain information to support operator action, while cautions inform about potential hazards to equipment and personnel and about actions dependent on changes to plant conditions. Their intended effectiveness might be undermined by some of their own characteristics: (a) presence of continuous actions/verifications; (b) physical/temporal distance from the place/time where/when they became relevant; (c) they are not always read (and are not totally consistent with the overall step-by step logic).

The challenges and issues discussed above consist of both slips and lapses in executing intended actions and difficulties in high-level cognitive processes (e.g., situation assessment and response planning). Compared to slips and lapses, which could be soon detected and corrected with compensatory factors, the difficulties in high-level cognitive processes seemed to have caused relatively more serious consequences. Note that the challenges and issues are coupled with crew performance difficulties as a result of their interaction with particular plant conditions and operational contexts. As discussed in Paper 1, some of the scenarios developed in the studies, while plausible, comprised far more difficulties than those modeled in standard HRA and PRA studies. In other words, those scenarios are rare in reality. Thus, most of the observations above

are unlikely to occur and should not be interpreted as common issues or safety status in current NPPs. Nonetheless, a close look at the observations from the challenging scenarios can shed light on the subtle and complex aspects of operators' procedure following behavior that are not well understood or are even overlooked, and provide useful insights to further improve operator performance and procedure effectiveness.

5 Predicted HEPs vs. empirical HEPs

In this section, the quantitative predictions of all methods are presented and compared as a whole against the 90% uncertainty bounds of the empirical HEP distributions. The detailed comparisons and assessments for each individual method are given in [4-8], while in this section the final results and conclusions are presented.

5.1 The International Study

Figures 2a – 2b show the predicted failure probabilities of all HRA teams against 5th and 95th percentile bounds of the empirical HEPs for the SGTR scenarios and LOFW scenarios of the International Study, respectively. On the horizontal axis, the HFEs are ordered by the difficulty ranking with difficulty decreasing from left to right.

[Insert Figures 2a – 2b about here]

Each figure shows several extreme outlier estimates, which are explainable based on analysts' interpretation of the information provided or the assumptions they made to address missing or incomplete information. When outliers are censored by excluding the maximum value and the minimum value, the method-to-method variability is 1.5 - 2 orders of magnitude for the SGTR scenarios and approximately 1 order of magnitude for the LOFW scenarios. (Note that HFEs 1A1 and 1B1 of the LOFW scenario are not included in this consideration, because they are joint HFEs and need special consideration.) For both scenario categories, the variability is present for both the easy and the difficult HFEs. Furthermore, the variability is not correlated across the HFEs in the sense that the same HRA analysis did not consistently produce the highest (or the lowest) HEP for the set of HFEs. In other words, none of the methods was systematically more conservative or optimistic than the other methods. Additionally, the ranking of the HEPs is not consistent from method to method.

The analyses differentiate among the HFEs to varying degrees. Some analyses appear to be unable to significantly differentiate among HFEs where no failures were observed against those in which a majority of the observed crews failed. The range of the HEPs from these analyses for the set of HFEs is rather narrow, in some cases, less than an order of magnitude. One possible explanation is that this is a reflection of the discriminating power of the method. Methods with more degrees of freedom in choosing the HEPs can, in principle, provide a wider range of possible values. However, even if a method has many degrees of freedom (e.g., different numbers and levels of PSFs), this may not necessarily be exercised and the focus of the analysis may be on a narrow set of PSFs.

There are a number of possible reasons for less variability across the teams in the LOFW scenario as compared to the SGTR scenario, so the reduction should be interpreted cautiously. The LOFW predictions were made after the SGTR predictions and subsequent to the discussion of the crew performance data and predictive-empirical comparisons for the SGTR scenarios; consequently, the HRA teams had more information about the crews than when they performed the analyses of the SGTR HFEs. Additionally, they also had more experience with the study protocols and methodology. On the other hand, an alternative explanation for the reduction could be that the LOFW HFEs better matched the capabilities of the HRA methods. Another possibility is that the HRA methods may have used more similar models for the failure mechanisms associated with the LOFW HFEs; for example, using the THERP dependence model [11] for the LOFW HFEs 2A and 2B (many methods treat dependence using this component of the THERP method or based on THERP's model). Comparing the SGTR and LOFW scenarios, there were significant differences in the types of tasks and their demands. Another cause of the reduction may be that in the LOFW scenarios, the HFEs modeled only one task (feed and bleed) with different success criteria. In the SGTR scenarios, the series of diverse tasks (identification, cooldown, depressurization, etc.) means that the HRA teams' model of the evolution of the scenarios may increasingly diverge.

For the SGTR scenarios,

- The predicted HEPs by most HRA applications for the most difficult HFEs 5B1 and 1B are significantly larger than those for the remaining HFEs. On the other hand, the predicted HEPs for these HFEs are outside the uncertainty bounds and below the 5th

percentile (lower) bound for a notable number of HRA applications. The underestimation seems to be fairly systematic. In contrast, most of the predicted HEPs for the remainder of the HFEs (the last 5 on the right in Figure 2a) fall within the uncertainty bounds. There seems to be some underestimation of HFEs 3B and 3A. These are depressurization actions, covered by the procedures, and for which the crews are well-trained. In general, the HRA analyses were less able to address the execution difficulties associated with controlling the depressurization.

- When compared to the HFE difficulty ranking, the predicted HEPs for the first four HFEs from left to right at the aggregate level are consistent with the empirical evidence of decreasing difficulty. However, the methods do not make clear distinctions in the HEPs (taken as a set) for the last five HFEs. Given that the ranking was established for HFEs 1A, 2A and 2B, the HRA predictions are considered to be mostly correlated with the difficulty ranking.

For the LOFW scenarios,

- For the most difficult HFE 1B, the predicted HEPs are consistently high (nearly all above 0.1), consistent with a high expectation of failure. But they seem to be fairly systematically underestimated by many HRA teams when compared to the 5th percentile of the uncertainty bound.
- Although the complex scenario HFEs are predicted as more difficult than the base scenario HFEs, the HEPs predicted for HFE 2A, at the aggregate level, tend to be larger than for HFE 1A, as do the HEPs for HFE 2B, which were predicted to be larger than

those for HFE 1B. This is not consistent with the HFE difficulty ranking, which states that HFE 2 should be easier than HFE 1 for both the base and the complex cases. In addition, the estimates for HFE 2B seem to be quite conservative with many estimates above the 95th percentile of the uncertainty bound. The large predicted HEPs for both 2A and 2B are in part due to the treatment of dependency in some of the HRA methods. HFEs 2A and 2B represent the same task (implementation of feed and bleed) with more time and more cues than the preceding HFEs (1A and 1B, respectively). The potential impact of the relationship between these HFEs on HEPs is represented by HRA dependence. In accordance with normal PRA practice and as recommended in THERP [11], negative dependence was not considered by any teams. (Negative dependence refers to the failure of the first task reducing the failure probability of the subsequent task). The analyses considered zero dependence at most while many analyses accounted for some positive dependence. The observations of these scenarios suggest that the crews' management of the criteria for feed and bleed may have contributed to negative dependence. The crews implemented feed and bleed (for success of HFE 2B) shortly after having failed to implement this action according to the success criteria for HFE 1B. A summary of the differences in this part of the analysis is provided in Part IV [3]. More generally, the study suggests that this area of HRA generally needs to be investigated further.

5.2 The US Study

The predicted mean HEPs of all HFEs from all HRA methods used in the study are presented in Figure 2.c. The figure shows that

[Insert Figure 2.c about here]

- There is significant disagreement among the predicted HEPs from each team for each of the HFEs. Except for HFE 3A, the variability of the HEPs for each HFE provided by the HRA teams is approximately one to one and half orders of magnitude. Interestingly, although all the HRA teams identified HFE 3A to be the easiest HFE with the lowest HEP, this HFE showed the greatest degree of variability, with two predictions significantly deviating from most of the others and leading to approximately three orders of magnitude difference among the HEPs.
- Similar to the International Study, difficult HFEs (e.g., HFEs 2A and 1C) seem to be fairly systematically underestimated when compared to the 5^{th} percentile uncertainty bound. In most cases, the HEP predictions show a decreasing trend that is consistent with the difficulty ranking. Some methods were not as consistent because they produced higher HEPs for HFE 1C than HFE 2A, which was ranked more difficult than HFE 1C. For HFEs 1C, 1A, and 3A, the HEPs from each HRA team generally show good correlations with the difficulty ranking; in a couple cases the analysts did not produce different HEPs for HFEs 1A and 1C. Note that HFEs 2A, 1C and 1A are all rather difficult events. While plausible, they comprise far more difficulties for the crews than many of the HFEs in standard PRA scenarios and therefore may have taxed the ability of some methods to account for the differences in difficulty.
- It is recognized that this study cannot provide any conclusive evidence regarding the consistency or accuracy of the quantitative analysis from the methods since it is based on

only two or three data points (HRA teams) per method. "Consistency" here refers to producing HEPs (a) reflecting the ranking of HFEs and (b) similar HEP values for a given HFE. "Accuracy" refers to producing values reflecting the empirical data in terms of failures observed. Nevertheless, regarding consistency, a review of Figure 2.c suggests that ASEP [12], ATHEANA [13] and CBDT [14] + HCR/ORE [15] yielded somewhat more consistent quantitative results than SPAR-H [16] across the analysis teams that used each of these methods. This is only true for CBDT+ HCR/ORE if one takes into consideration that the low HEP of HFE 1C by CBDT + HCR/ORE produced by Team 2 was due to the team's misunderstanding of the definition of the HFE. Regarding "accuracy," except ASEP, all other methods seemed to have underestimated HFE 2A and several underestimated the difficulty of 1C.

- There were no observations for HFE 1B (i.e., all the crews tripped manually before an automatic trip would have occurred, so the HFE was not relevant). Nevertheless, comparing the predictions of HFE 1B shows that this HFE is even more uniform than the others. Eight of the nine HRA teams' predicted HEPs were well within one order of magnitude. Two teams actually predicted the exact same mean HEP.

Turning to the differences in the diagnosis and execution HEPs predicted by each method, the results are plotted in Figures 3 and 4, respectively, with the exception of ATHEANA which does not calculate a separate HEP for execution, but includes it in estimating the total HEP. As shown in Figure 5, the diagnosis HEPs generally follow the HFE difficulty ranking. Although some teams produced relatively optimistic values for HFEs 2A and 1C, the HEPs differentiate difficult HFEs from easy ones fairly well. For HFEs 2A and 1C, the HEPs associated with diagnosis tend

to dominate the total HEP value (i.e., diagnosis HEPs determine the trend/shape of the HEP curves); the same could be argued for ATHEANA as well based on the discussion of what was driving performance in the qualitative analysis. This result is in line with the simulator data and the study expectation that the complex scenarios would significantly increase the difficulty of diagnosis rather than execution for most HFEs. Similar to the total HEPs, the diagnosis HEPs exhibit some variability for HFE 3A. Regarding the CBDT and HCR/ORE results, Teams 1 and 2 produced a relatively conservative HEP for HFE 3A when compared with the results produced by Team 3. The difference is explained by the fact that the former two teams did not include recovery in their analysis, which, given the available time and conditions and the empirical results, would seem to be unnecessarily conservative. Thus, this would be one contributor to the variability and leaves only the SPAR-H Team 2 as an outlier with respect to the diagnosis HEPs for the teams presented.

[Insert Figure 3 about here]

Regarding the execution HEPs plotted in Figure 4, the values do not show apparent differentiation across the HFEs as seen in the total HEPs, a reasonable result for in-control room actions which are usually straightforward and typically accomplished quickly once the crew determines what is needed to be done. It is interesting to note that most of the teams obtain comparable execution HEPs for HFEs 1A and 3A. This may be partly because these teams used the same data source (e.g., THERP) and the same or similar quantification approaches (e.g., THERP or ASEP). Nevertheless, for HFEs 2A and 1C, the execution HEPs show large variability. One contributor to variability is the high execution HEPs produced by ASEP Team 2

and SPAR-H Team 1 for HFEs 2A and 1C, respectively. However, since the execution HEPs do not dominate the total HEP for the two HFEs, the variability does not affect the total HEPs.

[Insert Figure 4 about here]

It is interesting to note the following when examining Figures 2.c, 3 and 4 together. For methods that calculate final HEPs as the sum of diagnosis and execution HEPs, even though the final HEPs seem to be reasonable with respect to HFE difficulty rankings, this conclusion could be questionable if one considers the relative contribution of diagnosis and execution HEPs. For example, the diagnosis HEP predicted by ASEP Team 2 for HFE 1A is significantly lower than the execution HEP (i.e., execution HEP dominates the total HEP), which is not consistent with what would be expected for a challenging HFE. Similarly, the execution HEPs predicted by ASEP Team 2 for HFE 2A and SPAR-H Team 1 for HFE 1C, seem to be unjustifiably high. Addressing the question of what is the most important contributor to an HEP (diagnosis or execution) is an important aspect of HRA because safety improvements dealing with diagnostic issues could be very different than those dealing with execution issues.

6 Concluding remarks

As described in Paper 1, simulator data was analyzed and aggregated into a high-level representation that correspond to the ways in which the HRA teams were asked to report their predictions. In general, there are two types of simulator data: qualitative and quantitative data.

They correspond to the following two types of comparisons between HRA methods' predictions and simulator data:

- Quantitative comparisons: (a) Predicted HEPs vs. empirical HEPs; (b) Predicted HEP ranking vs. HFE difficulty ranking
- Qualitative comparisons: (a) Predicted performance drivers vs. observed drivers; (b) Predicted difficulties vs. observed difficulties (operational descriptions and PSFs).

As discussed in in Section 3.7.1 of Paper 1 [1] and Section 3.2 of this paper, although the number of crew observations is considered large for simulator studies, they still represent a small sample in statistical terms given the HEPs' expected range of values. The limitation of the small sample size is manifested in the large uncertainties on the empirical HEPs, which makes drawing definitive conclusions is limited. On the other hand, the studies produced a rich set of qualitative data on performance issues. Therefore, the assessment prioritized qualitative comparisons, and the comparisons of HRA method qualitative predictions to qualitative simulator data produced the majority of the insights concerning the methods and the most valuable results of the studies.

The rich qualitative data obtained in the studies is used in three ways. In all three, the qualitative data is used in combination with quantitative data. First, the assessment of quantitative predictions examines the predicted HEP ranking by comparing this ranking to the HFE difficulty ranking. The HFE difficulty ranking combines qualitative observations of the crew performances with the observed failure counts to obtain a qualitative failure likelihood ranking. These observations related to performance issues associated with the HFEs can be used to

differentiate HFEs with similar failure counts, since the differences in the failure counts (especially when the count is zero or low) may not be significant in themselves.

The second way in which the quantitative predictions and data are used is in directing the assessment of the qualitative predictions. The qualitative comparisons integrate into the method assessments and make use of the rich qualitative data collected in the studies. In these qualitative comparisons; the quantitative predictions and data determine the essential predictions to be compared with simulator data on the qualitative level. For instance, the comparisons focused on the qualitative predictions associated with the HFEs with higher (qualitative) likelihood. Conversely, the performance issues associated with the lower likelihood HFEs are viewed as more minor.

Third, comparing the predicted HEPs with the empirical HEP uncertainty bounds identified which HRA predictions could be considered relative outliers. The HRA analyses that produced such outliers were scrutinized in terms of the features of the methods and of their implementation in the analyses that contributed to these outlier predictions. Examining these quantitative outliers provided important insights concerning HRA methods and practice.

Finally, it is worth noting that the limitations on the statistical analysis and quantitative simulator data are in many ways inherent to HRA. Human performance is known to be situation-specific. Moreover, HRA data and analysis must consider not only average performance (aggregating data from different contexts) but also the impacts of the situational context factors on performance in specific scenarios. The qualitative simulator data of the studies produced valuable evidence

concerning these specific impacts and the associated mechanisms through which the factors influence performance. It represents a strong basis for the broad insights on the HRA methods obtained in the studies. Through these insights on HRA methods, the studies have shown a useful way in which quantitative and qualitative comparisons and criteria can be combined in HRA method assessments.

7 References

[1] Liao H, Forester J, Dang VN, Bye A, Chang, YHJ, Lois E. Assessment of HRA Method Predictions against Operating Crew Performance: Part I: Study Background, Design, and Methodology. Reliability Engineering & System Safety, 2016 (under review).

[2] Liao H, Forester J, Dang VN, Bye A, Chang, YHJ, Lois E. Assessment of HRA Method Predictions against Operating Crew Performance: Part III: Insights from Intra-Method Comparisons. Reliability Engineering & System Safety, 2016 (under review).

[3] Liao H, Forester J, Dang VN, Bye A, Chang, YHJ, Lois E. Assessment of HRA Method Predictions against Operating Crew Performance: Part IV: Conclusions and Achievements. Reliability Engineering & System Safety, 2016 (under review).

[4] Lois E, Dang VN, Forester J, Broberg H, Massaiu S, Hildebrandt M, Braarud PØ, Parry G, Julius J, Boring R, Männistö I, Bye A. International HRA Empirical Study—Phase 1 Report: Description of Overall Approach and Pilot Phase Results from Comparing HRA

Methods to Simulator Data. NUREG/IA-0216, Vol. 1. US Nuclear Regulatory Commission, Washington, DC; 2009.

[5] Bye A, Lois E, Dang VN, Parry G, Forester J, Massaiu S, Boring R, Braarud PØ, Broberg H, Julius J, Männistö I, Nelson P. International HRA Empirical Study—Phase 2 Report: Results from Comparing HRA Method Predictions to Simulator Data from SGTR Scenarios. NUREG/IA-0216, Vol. 2. US Nuclear Regulatory Commission, Washington, DC; 2011.

[6] Dang VN, Forester J, Boring R, Broberg H, Massaiu S, Julius J, Männistö I, Liao H, Nelson P, Lois E, Bye A. The International HRA Empirical Study - Phase 3 Report: Results from Comparing HRA Methods Predictions to HAMMLAB Simulator Data on LOFW Scenarios. NUREG/IA-0216, Vol. 3, US Nuclear Regulatory Commission, Washington, DC; 2011.

[7] Forester J, Dang VN, Bye A, Lois E, Massaiu S, Broberg H, Braarud P, Boring R, Männistö I, Liao H, Julius J, Parry G, Nelson P. The International HRA Empirical Study—Final Report: Lessons Learned from Comparing HRA Methods Predictions to HAMMLAB Simulator Data. NUREG-2127. US Nuclear Regulatory Commission, Washington, DC; 2013.

[8] Forester JA, Liao H, Dang VN, Bye A, Presley M, Marble J, Broberg H, Hildebrandt M, Lois E, Hallbert B, Morgan T. Assessment of HRA Method Predictions Against Operating Crew Performance on a US Nuclear Power Plant Simulator. NUREG-2156. US Nuclear Regulatory Commission, Washington, DC; 2016.

[9] Dang VN, Massaiu S, Bye A, Forester JA. Quantitative Results of the HRA Empirical Study and the Role of Quantitative Data in Benchmarking. The 10th International Probabilistic Safety Assessment and Management Conference, (PSAM 10), Seattle, USA; 2010.

[10] Liao H, Hildebrandt M. Empirical Insights on Operators' Procedure Following Behavior in Nuclear Power Plants. The 15th International Conference on Human-Computer Interaction 2013, July 2013, Las Vegas, NV, USA; 2013

[11] Swain A, Guttman H. Handbook of Human Reliability Analysis with Emphasis on Nuclear Power Plant Applications. NUREG/CR-1278-F, U.S. Nuclear Regulatory Commission, Washington, DC; 1983.

[12] Swain A. Accident Sequence Evaluation Program Human Reliability Analysis Procedure. NUREG/CR-4772, Sandia National Laboratories for the U.S. Nuclear Regulatory Commission, Washington, DC; 1987.

[13] Technical Basis and Implementation Guidelines for A Technique for Human Event Analysis (ATHEANA), NUREG-1624, Rev. 1, USNRC, Washington, DC; 2000.

[14] Parry G, et al. An Approach to the Analysis of Operator Actions in PRA, EPRI TR-100259, Electric Power Research Institute, Palo Alto, CA; 1992.

[15] Spurgin A, et al. Operator Reliability Experiments Using Power Plant Simulators. EPRI NP-6937, Vol. 1, Electric Power Research Institute, Monterey, CA, 1990.

[16] Gertman, D., Blackman, H., Marble, J., Byers, J., Haney, L., & Smith, C. (2005): "The SPAR-H Human Reliability Analysis Method," NUREG/CR-6883. U.S. Nuclear Regulatory Commission, Washington, DC; 2005.

Table 1a. Crew Failure Rates and HFE Difficulty Ranking for the SGTR Scenarios in the International HRA Empirical Study

HFE	No. of observations	No. of failures	Difficulty Ranking
HFE 5B1	7	7	5 (Very difficult)
HFE 1B	14	7	4 (Difficult)
HFE 3B	14	2	3.5 (Somewhat difficult)
HFE 3A	14	1	3 (Somewhat difficult)
HFE 1A	14	1	2.5 (Easy to Somewhat difficult)
HFE 2A	14	1	2.5 (Easy to Somewhat difficult)
HFE 2B	14	0	2.5 (Easy to Somewhat difficult)
HFE 5B2	7	0	2 (Easy)
HFE 4A	14	0	1 (Very easy)

Table 1b. Crew Failure Rates and HFE Difficulty Ranking for the LOFW Scenarios in the International HRA Empirical Study

HFE	No. of observations*	No. of failures	Difficulty Ranking**
HFE 1B	10	7	5 (Very difficult)
HFE 2B	7	0	3.5 (Somewhat difficult to difficult)
HFE 1A	10	0	2.5 (Easy to somewhat difficult)
HFE 1B1	10	0	
HFE 1A1	10	0	

* 14 crews participated, as in the SGTR runs, but only 10 crews were analyzed due to simulator problems.
** HFEs 1A1 and 1B1 are not included in the difficulty ranking, because their definitions overlap those of HFEs 1A and 2A, and 1B and 2B, respectively (ranking part vs. joint HFEs). HFE 1B1 is considered relatively more difficult than HFE 1A1, considering when B&F is implemented relative to the procedural criteria and qualitative considerations.

Table 1c. Crew Failure Rates and HFE Difficulty Ranking for the Scenarios in the US HRA Empirical Study*

HFE	No. of observations	No. of failures	Difficulty Ranking
HFE 2A	4	4	1 (Very difficult)
HFE 1C	4	3	2 (Difficult)
HFE 1A	4	0	3 (Fairly difficult to difficult)
HFE 3A	3	0	4 (Easy)

* Empirical data were not available for HFE 1B.

Figure Captions

Figure 1a. Empirical HEP distribution for the SGTR Scenarios in the International HRA Empirical Study

Figure 1b. Empirical HEP distribution for the LOFW Scenarios in the International HRA Empirical Study

Figure 1c. Empirical HEP distribution for the Scenarios in the US HRA Empirical Study

Figure 2a. Predicted HEPs vs. Empirical HEPs for the SGTR Scenarios in the International HRA Empirical Study

Figure 2b. Predicted HEPs vs. Empirical HEPs for the LOFW Scenarios in the International HRA Empirical Study

Figure 2c. Predicted HEPs vs. Empirical HEPs for the Scenarios in the US HRA Empirical Study

Figure 3. Predicted Diagnosis HEPs by HRA Methods for the Scenarios in the US HRA Empirical Study (Note: ATHEANA Teams are not shown in this figure as ATHEANA does not calculate diagnosis and execution HEPs separately.)

Figure 4. Predicted Execution HEPs by HRA Methods for the Scenarios in the US HRA Empirical Study (Note: ATHEANA Teams are not shown in this figure as ATHEANA does not calculate diagnosis and execution HEPs separately.)

Figure 1a.

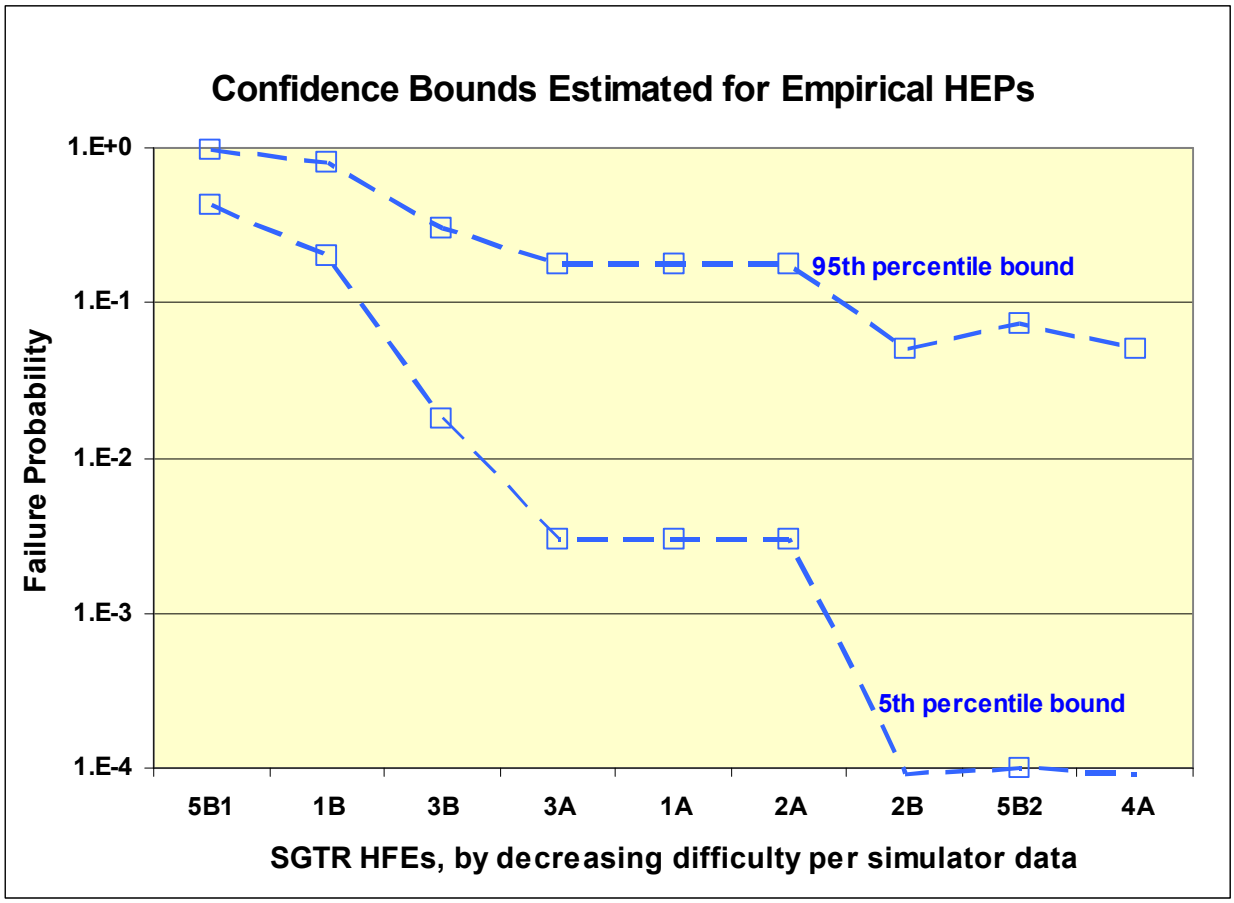

Confidence Bounds Estimated for Empirical HEPs
1.E+0
1.E-1
1.E-2
1.E-3
1.E-4
Failure Probability
95th percentile bound
5th percentile bound
5B1
1B
3B
3A
1A
2A
2B
5B2
4A
SGTR HFEs, by decreasing difficulty per simulator data

Figure 1b.

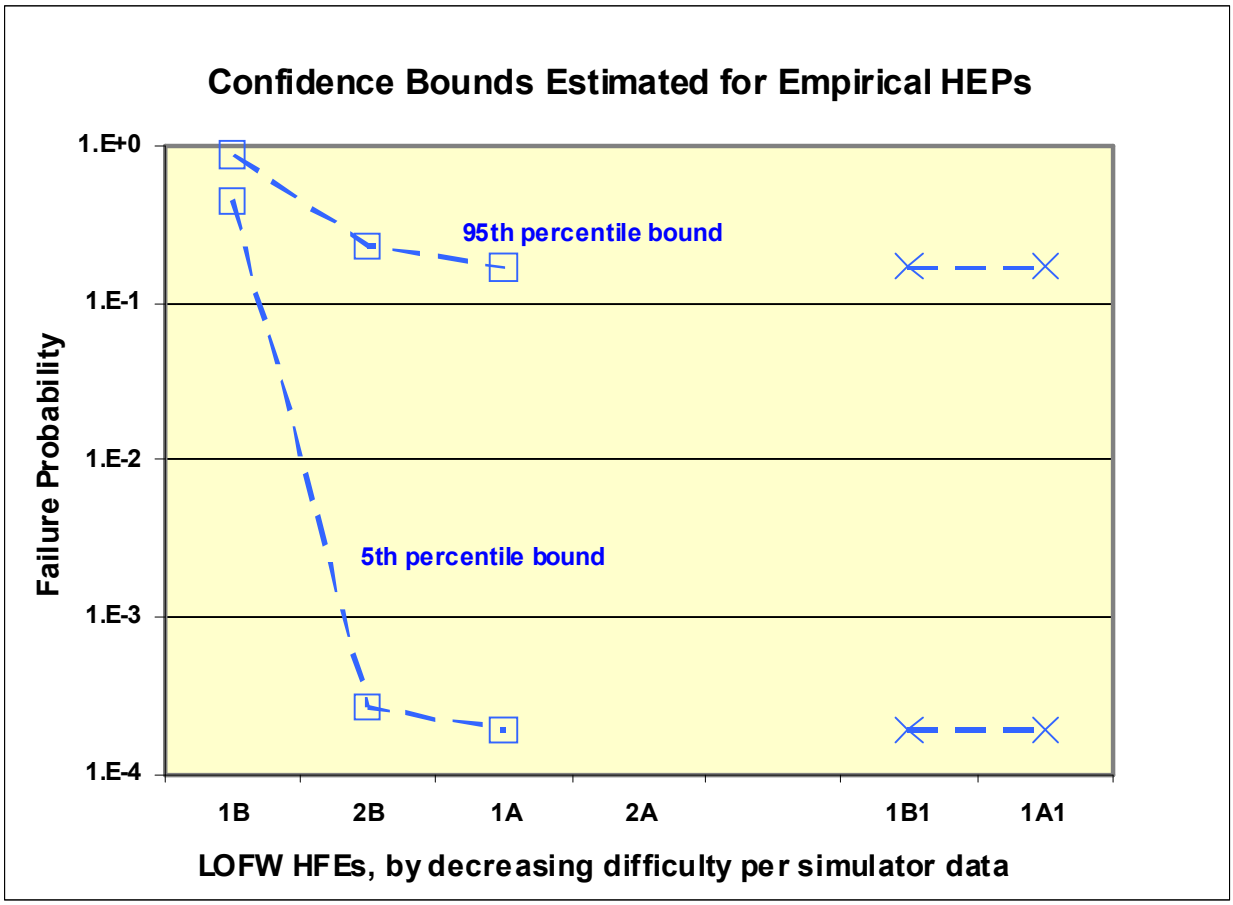

Confidence Bounds Estimated for Empirical HEPs
1.E+0
1.E-1
1.E-2
1.E-3
1.E-4
Failure Probability
95th percentile bound
5th percentile bound
1B
2B
1A
2A
1B1
1A1
LOFW HFEs, by decreasing difficulty per simulator data

Figure 1c.

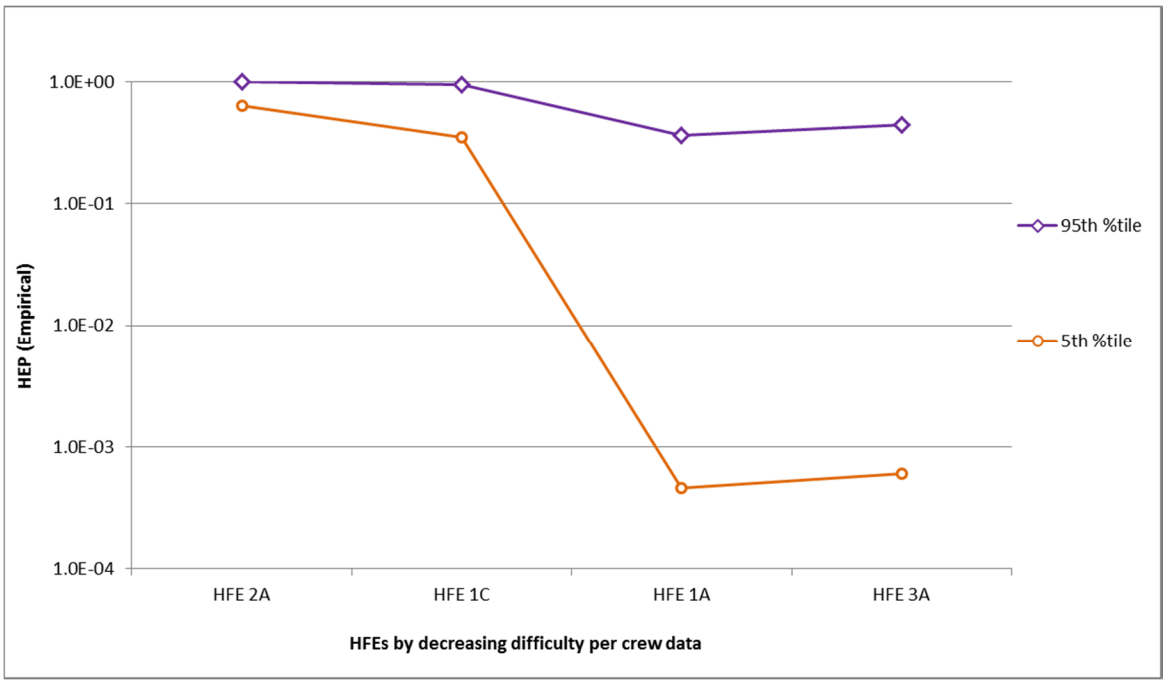


Figure 1c

Figure 2a.

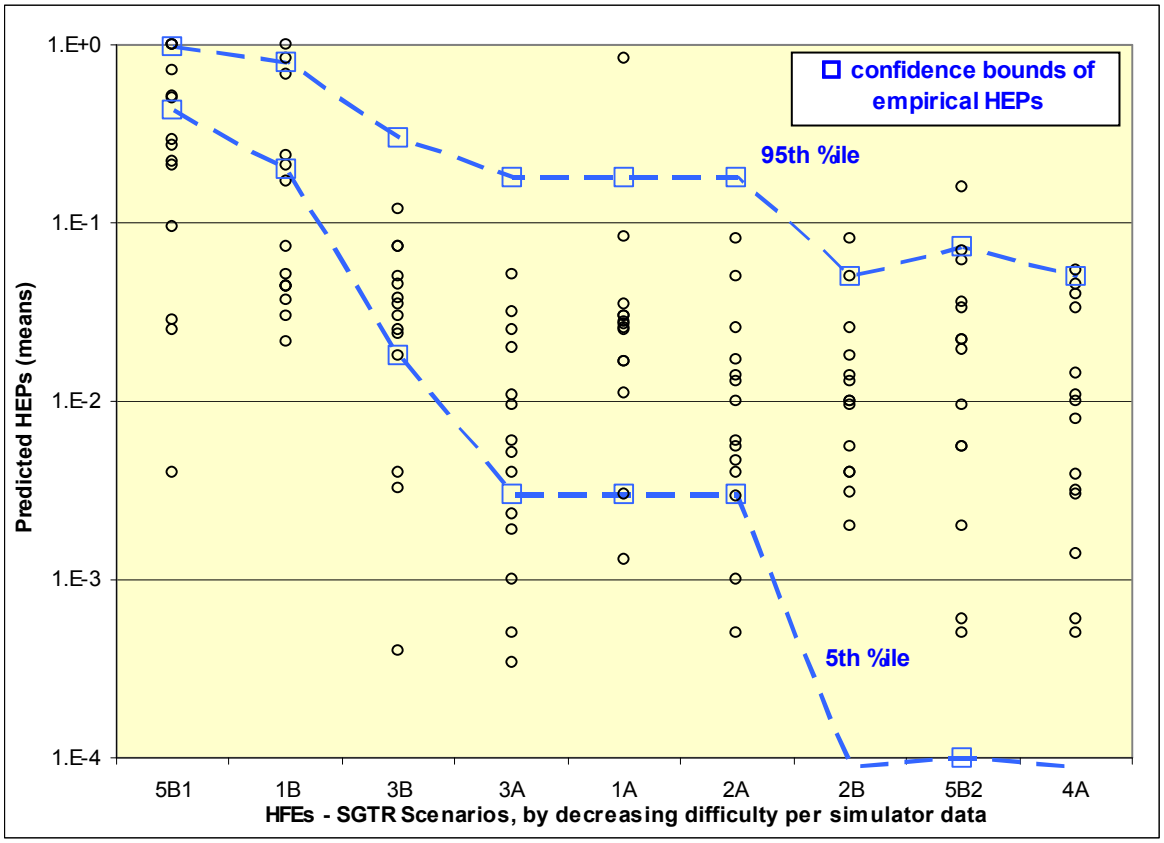

confidence bounds of empirical HEPs
95th %ile
5th %ile
Predicted HEPs (means)
1.E+0
1.E-1
1.E-2
1.E-3
1.E-4
5B1
1B
3B
3A
1A
2A
2B
5B2
4A
HFEs - SGTR Scenarios, by decreasing difficulty per simulator data

Figure 2b.

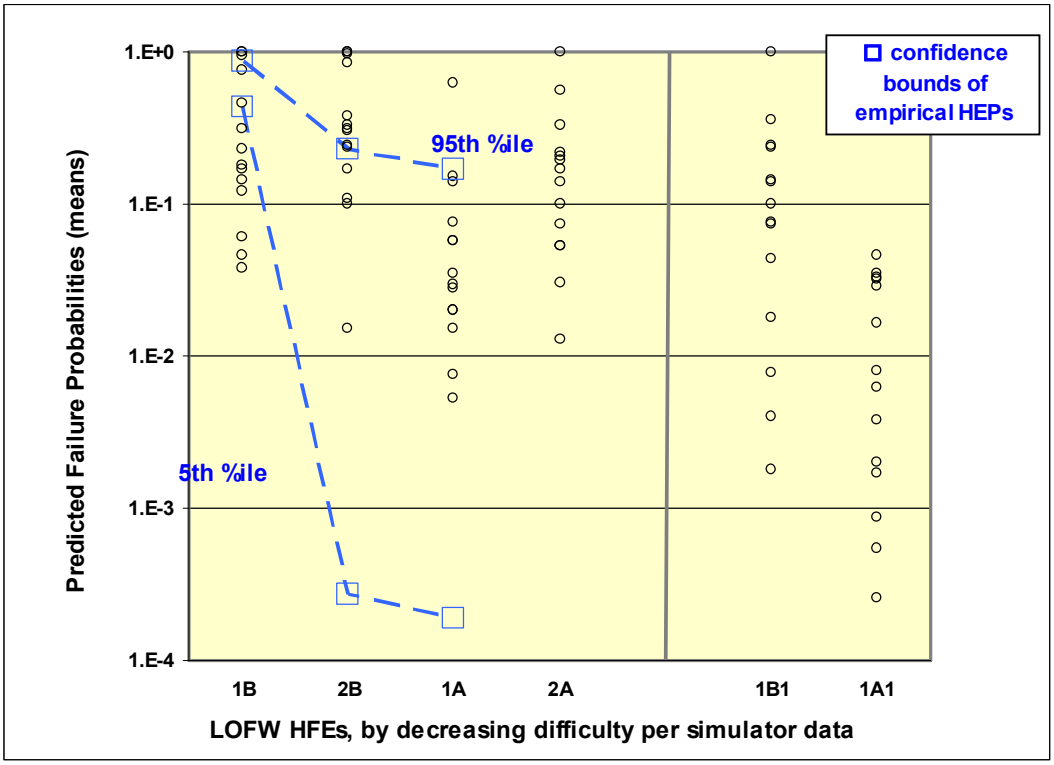

confidence bounds of empirical HEPs
95th %ile
5th %ile
Predicted Failure Probabilities (means)
1.E+0
1.E-1
1.E-2
1.E-3
1.E-4
1B
2B
1A
2A
1B1
1A1
LOFW HFEs, by decreasing difficulty per simulator data

Figure 2.c.

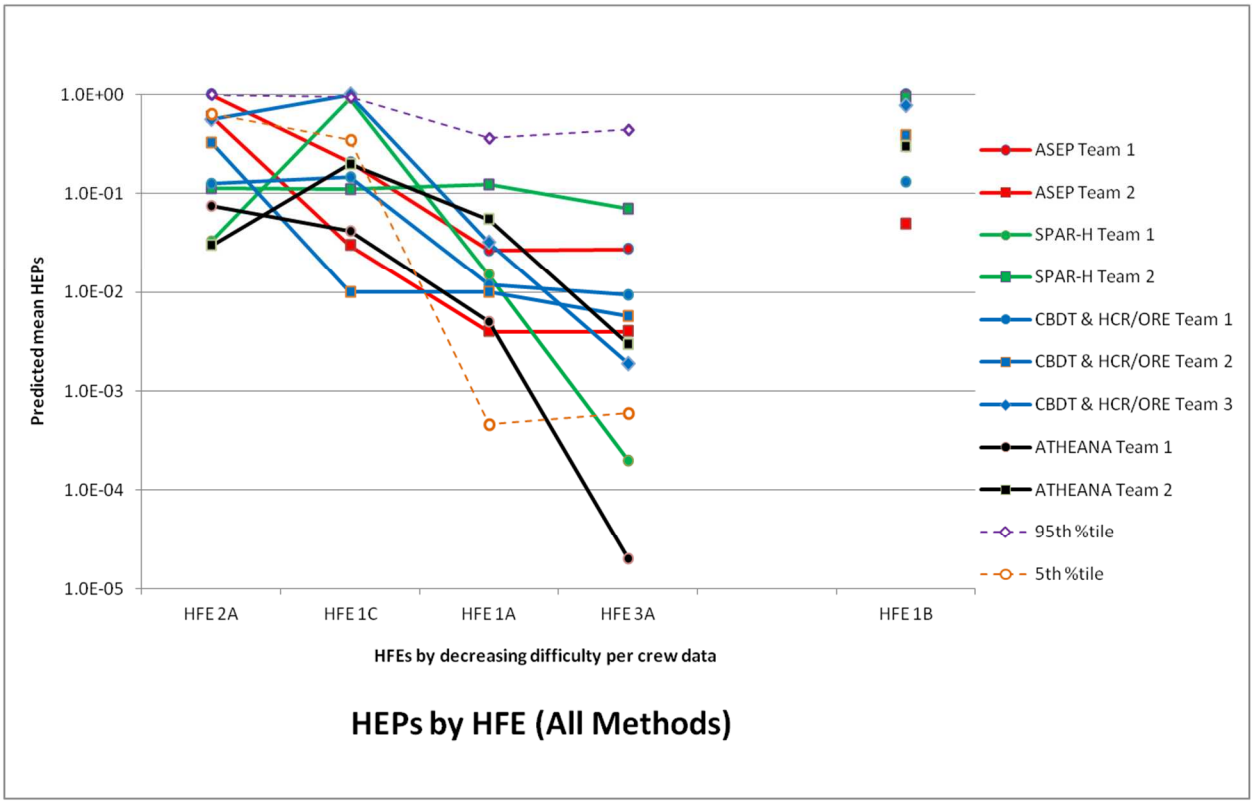

1.0E+00
1.0E-01
1.0E-02
1.0E-03
1.0E-04
1.0E-05
Predicted mean HEPs
HFE 2A
HFE 1C
HFE 1A
HFE 3A
HFE 1B
HFEs by decreasing difficulty per crew data
ASEP Team 1
ASEP Team 2
SPAR-H Team 1
SPAR-H Team 2
CBDT & HCR/ORE Team 1
CBDT & HCR/ORE Team 2
CBDT & HCR/ORE Team 3
ATHEANA Team 1
ATHEANA Team 2
95th %tile
5th %tile
HEPs by HFE (All Methods)

Figure 3.

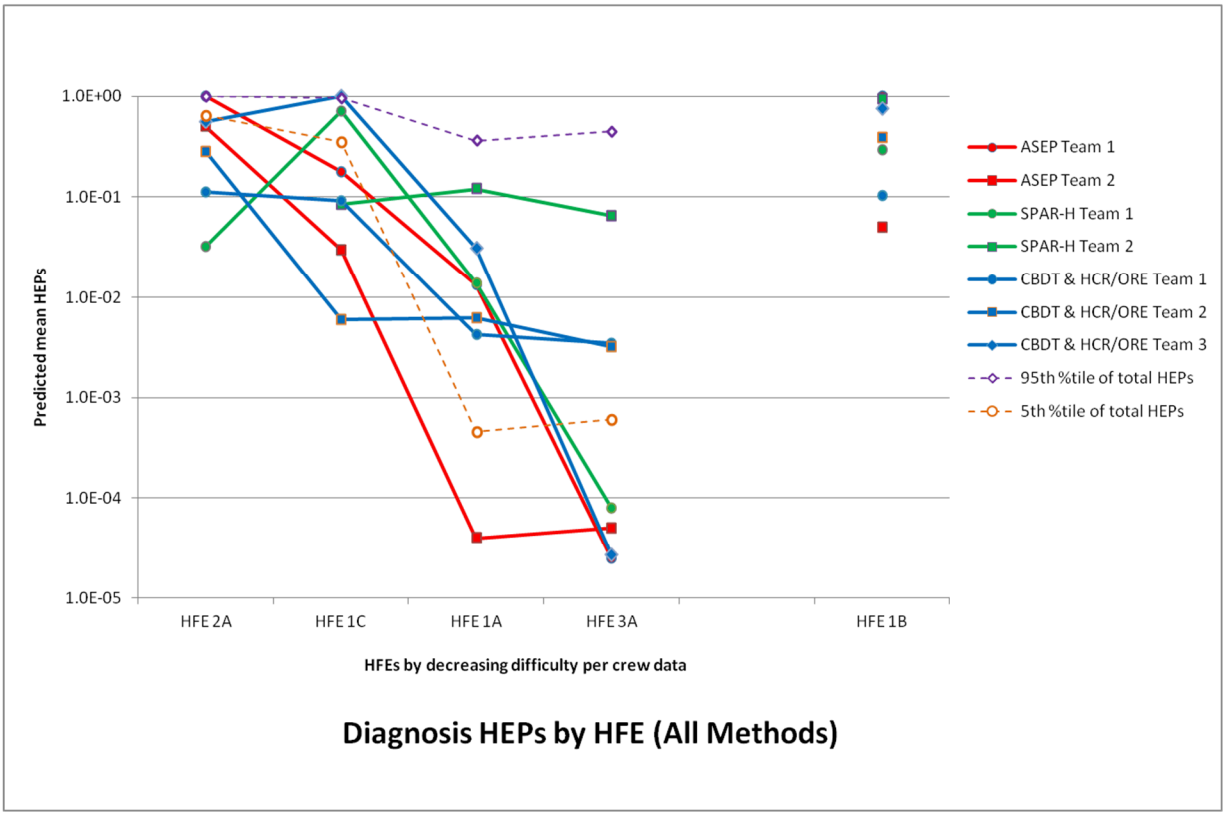
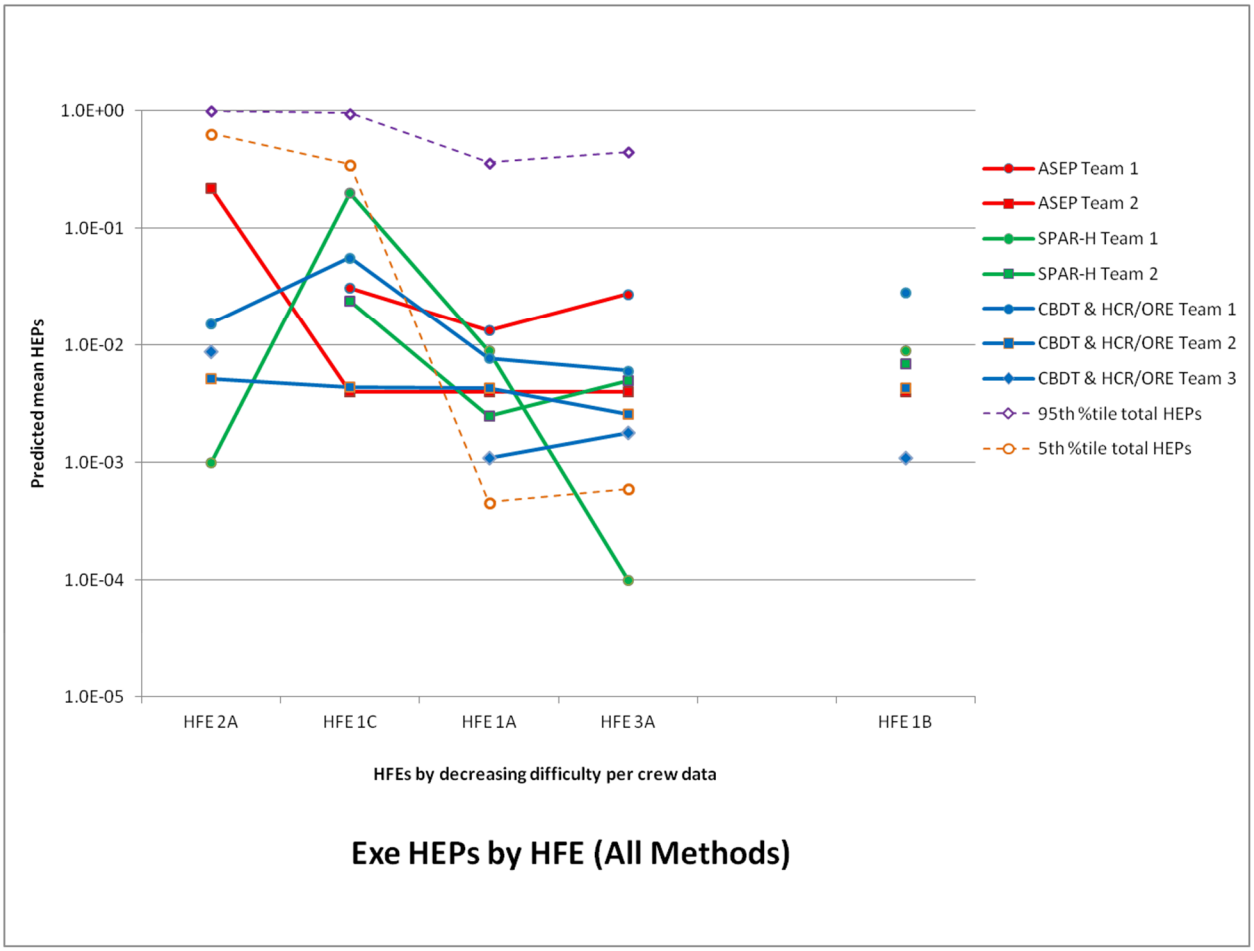

1.0E+00
1.0E-01
1.0E-02
1.0E-03
1.0E-04
1.0E-05
Predicted mean HEPs
HFE 2A
HFE 1C
HFE 1A
HFE 3A
HFE 1B
HFEs by decreasing difficulty per crew data
ASEP Team 1
ASEP Team 2
SPAR-H Team 1
SPAR-H Team 2
CBDT & HCR/ORE Team 1
CBDT & HCR/ORE Team 2
CBDT & HCR/ORE Team 3
95th %tile of total HEPs
5th %tile of total HEPs
Diagnosis HEPs by HFE (All Methods)

Figure 4.


1.0E+00
1.0E-01
1.0E-02
1.0E-03
1.0E-04
1.0E-05
Predicted mean HEPs
HFE 2A
HFE 1C
HFE 1A
HFE 3A
HFE 1B
HFEs by decreasing difficulty per crew data
ASEP Team 1
ASEP Team 2
SPAR-H Team 1
SPAR-H Team 2
CBDT & HCR/ORE Team 1
CBDT & HCR/ORE Team 2
CBDT & HCR/ORE Team 3
95th %tile total HEPs
5th %tile total HEPs
Exe HEPs by HFE (All Methods)