Aggregation of Autocalculated Human Error Probabilities from Tasks to Human Failure Events in a Dynamic Human Reliability Analysis Implementation

Ronald L. Boring[a*], Martin Rasmussen[b], Thomas A. Ulrich[a], and Nancy J. Lybeck[a]
[a] Idaho National Laboratory, Idaho Falls, Idaho, USA
[b] NTNU Social Research, Trondheim, Norway

Abstract: Dynamic human reliability analysis (HRA) typically operates at the task or subtask level, allowing a tight coupling between plant evolutions and operator response. However, most HRA methods calculate human error probabilities (HEPs) at the human failure event (HFE) level. This results in a mismatch between the types of HEPs generated for static vs. dynamic HRA. There has been no clear guidance on how to aggregate the higher sampling frequency of dynamic HEPs to match the HEPs in static HRA. Applying available task dependence correction factors, for example, artificially inflates the HEPs, resulting in unrealistically conservative HEPs generated from dynamic HRA. In this paper, we review aggregation techniques that calibrates dynamic task-level HEPs to static HFE-level HEPs. This aggregation allows more direct validation of dynamic HRA results to reference static HRA results.

Keywords: Human Reliability Analysis (HRA), Task Decomposition, Human Error Probability (HEP), Dynamic Human Reliability Analysis, Human Failure Event (HFE).

1. INTRODUCTION

The standard unit of analysis in human reliability analysis (HRA) is the human failure event (HFE). In their probabilistic risk assessment (PRA) standard, the American Society of Mechanical Engineers (ASME) has defined an HFE as, “a basic event that represents a failure or unavailability of a component, system, or function that is caused by human inaction, or an inappropriate action” [1]. Unfortunately, this definition fails to offer precision in terms of the level of task decomposition [2]. Generally, the HFE will encompass many human tasks that together contribute to the failure of a hardware system. The number of tasks and subtasks remains underspecified.

Such underspecification was the source of considerable inter-analyst variability in the European Union’s Human Factors Reliability Benchmark Exercise (HF-RBE) in the 1980s [3]. In fact, it took three iterations of the study before the level of analysis within the HFE could be controlled enough to allow comparison of HRAs performed by different teams. The more recent International HRA Empirical Study [4] carefully defined the HFEs, arguably to a higher level of specificity than might be found in conventional PRA in practice. The highly specified HFEs ensured greater comparability between HRAs without confounds due to analysts looking at different tasks as part of their analysis.

Despite considerable advancement in the field of HRA in the span of time between the two benchmark activities, there remains surprisingly little advancement in the definition or specification of the HFE. In fact, with the move toward streamlined, primarily quantitative HRA methods throughout much of the interval between benchmarks, many methods provide even less guidance on task analysis and HFE definition than earlier methods. For example, while the Technique for Human Error Rate Prediction (THERP) [5] offers extensive guidance on how to decompose human activities and synthesize them for calculation of the human error probability (HEP), newer methods like the Standardized Plant Analysis

[*] ronald.boring@inl.gov

Risk-Human (SPAR-H) method [6] simply guide the analyst on quantifying a predefined HFE. That predefinition comes from the PRA, but the PRA is primarily concerned with system impact rather than task decomposition.

The use of a PRA-defined HFE serves HRA well in practice. The input from the HRA into the PRA needs to be an HEP, and the analyst is skilled at providing a realistic HEP commensurate to the number of tasks encompassed in the HFE. Still, there is considerable skill of craft in scaling the analysis to the right level of detail. It is possible that defining the tasks within the HFE and aligning the HRA method to that task level remains a significant source of inter-analyst variability. The challenge of aligning analysts and methods to the right task remains a concern, even 30 years after the HF-RBE [3].

While the underspecification of task decomposition in HFEs may still linger quietly as an unresolved issue for human analysts performing HRAs, the problem is not limited to traditional, so-called static HRA. Recent work in dynamic HRA reignites the problem of task decomposition, because dynamic HRA necessarily considers human activities at the task or procedure step level. Dynamic HRA strives for a realistic simulation of human performance through either a virtual operator or virtual analysis model [7]. The virtual operator may be linked to a virtual nuclear plant simulation in a dynamic PRA, such that virtual operator activities are tightly coupled to virtual plant performance [8, 9]. Because human performance is driven by a sequence of perceptions, decisions, and actions, simulations of human performance cannot exist at the aggregate level. The script of human goals and tasks to accomplish them must be played out step by step in order to achieve a realistic representation of that activity. Each step is influenced by context—the nature of the task, the environment in which the task is carried out, and a myriad of psychological and physiological factors that shape the performance of the human. While it would be possible to take a snapshot of a cluster of activities to plug into a dynamic HRA, this big picture would not be tightly coupled to the context. A generic snapshot would fail to be a function of the inputs into the human that ultimately determine the performance outcome.

At present, there is no surrogate or reduced order model for a series of human tasks in a dynamic simulation. The human actions must be modeled at the task (or smaller) level. The necessity of task-level modeling in dynamic HRA means that HEPs are calculated (and preferably autocalculated [9, 10]) at the task level, not the HFE level. In this paper, we overview some strategies to reconcile task-level HEPs with the HFE-level HEPs preferred in PRA in current practice.

2. TASK-LEVEL MODELING

A human error is not necessarily the same as a human failure, and a task is not necessarily the same as a human failure event. An HFE typically consists of several tasks performed toward a common goal. For example, an HFE of *Failure to Initiate Safety Injection* in a nuclear power plant likely spans multiple tasks and corresponding steps in procedures, when available. Each task or step presents an opportunity for success or error (and corresponding task HEP), and thus an HFE may actually entail multiple human errors (and successes) at different steps in the overall process.

The HFE is generally represented as a binary branch of an event tree in the PRA, and the tasks are generally represented as nodes linked by logical *AND* or *OR* gates in a fault tree. The fault tree fails to represent the sequence of tasks, nor does it specify the optimal level of tasks to be modeled, any more than the HFE does in the event tree. While the optimal level of ideal decomposition certainly features prominently in discussions of task analysis (e.g., [11]), it often remains a matter of analyst discretion in HRA. Often, those salient points of physical connection between operators and hardware (e.g., opening a valve) are more completely modeled than cognitive tasks like detection, monitoring, and decision making.

We have already discussed the need for task-level modeling in dynamic HRA. Additional reasons for using task-level HEP calculations include the following:

1. *Recovery:* Just as every task includes the opportunity for success or error, each error likely has opportunity for recovery. The error-recovery pairing is essential to accurate modeling of the HEP for that task. The original HRA method, THERP [5] included its own form of event tree—the HRA event tree—which is no longer widely used. In THERP's HRA event tree, a binary success-failure node representation of all tasks served to link the tasks together explicitly, including the mathematical aggregation of all task HEPs. Each node in the HRA event tree featured an HEP, and each error should explicitly consider the opportunity for recovery. After THERP, the modeling of recovery has gradually shifted to the HFE level, but at this level there is opportunity to overlook the precise pairing of task errors and task recoveries. Recovery occurs at the task level. It may impact the HFE, but it is best understood at the task level.
2. *HRA without PRA:* While PRA and HRA are well entrenched in some industries like nuclear power, there are many safety-critical industries where PRA and HRA are still emerging. The adoption of HRA may follow independently of PRA. For example, in the petroleum industry in the U.S., there are minimal PRA requirements. Yet, there is active work in human factors and human performance improvement. HRA is sometimes being sought as a complementary tool to human factors and human performance in the design of new systems. Identifying and quantifying human errors helps to prioritize the most impactful items in terms of safety when refining the design of a new system. In such cases, there is no PRA to define the HFE, and the HFE is derived from a task analysis [12-13]. For HRA to be useful to such applications, it must be standalone, and it must be able to calculate HEPs meaningfully at the task level, which corresponds to the task analysis level used in human factors.
3. *HRA for Design:* To reiterate a point made in [14], when designing a new system and considering the opportunity for human error in that system, the analysis may benefit from finer granularity. Identifying the opportunity for error is better accomplished by considering each task a user will undertake rather than considering a block of tasks at the HFE level. The individual error matters in the design, because that task-level error provides a specific resolution area rather than a general mitigation requirement. A pinpointed issue to fix is of considerably greater value to the system engineer than a broad issue that may vaguely span multiple aspects of the system.

As a final point on task-level modeling, it is important to consider that a task may consist of subtasks. In task analysis, a task may be further decomposed into subtasks, each representing a hierarchy of goals and corresponding activities. An HRA method like the Goals-Operators-Methods-Selection rules (GOMS)-HRA method captures this nuance [15]. GOMS-HRA models human activities in terms of task level primitives (TLPs), which are basic units of human activities such as taking physical actions, checking something, retrieving information, issuing or receiving instructions, selecting something, and making decisions. GOMS-HRA also offers procedure level primitives (PLPs), which correspond to procedure steps [16]. The PLPs are at a higher level of task activity than the TLPs. A single procedure step indexed as a PLP may consist of several TLPs. The relationship between PLPs and TLPs therefore is one of task and subtasks, respectively. GOMS-HRA quantifies at the TLP or subtask level. It is essential that GOMS-HRA be able to scale from the subtask to the task and HFE level in terms of quantification. There is no current method to aggregate such subtask HEPs to the HFE level. This represents a fundamental misalignment of GOMS-HRA to the standard level of analysis in HRA. Yet, GOMS-HRA has already proved a versatile simulation method for dynamic HRA. To begin to redress this shortcoming, next we explore different ways to aggregate subtask and task HEPs to the HFE level.

3. SAMPLE DYNAMIC DATA

For our exploration of HEP aggregation, we are using data from a simulation related to a station blackout at a nuclear power plant [17]. The station blackout has three successive phases:

1. *Loss of offsite power (LOOP):* The electric grid fails, at which point the plant can no longer continue to produce power. The plant no longer has access to an external power source, and the plant is no longer producing its own power. It must rely on backup power in the form of emergency diesel generators.

2. *Loss of diesel generator (LODG):* The diesel generators may fail for reasons incident to weather (e.g., flooding or other natural disaster that renders them inoperable) or for reasons related to prolonged use (i.e., running out of available fuel in the event of a prolonged grid disturbance). Human error can also be a contributor, e.g., if the diesel generator is aligned to the wrong power train, resulting in damage to key systems. For light water reactor designs without passive cooling, it is necessary to maintain power to continue cooling the reactor via coolant recirculation pumps following shutdown.
3. *Loss of battery (LOB):* In the event of diesel failure, the plant has large batteries to allow essential plant functions like coolant recirculation for a specified time. This time window may ideally be the period required for safe shutdown or may incorporate the time required to setup or transport secondary backup generators.

A failure of all three levels is extremely rare but was evidenced at the Fukushimi Daiichi event [18]. Awareness of the possible failure of all levels due to natural disaster has led to additional defense-in-depth measures such as emergency mitigation equipment (so-called "flex" gear) to ensure the availability of necessary power to the plant under a wider range of unlikely but high consequence disruptive events.

Our simulation of the event featured a thermal hydraulics-based plant simulation coupled to a dynamic HRA implementation based on GOMS-HRA [17]. GOMS-HRA served to calculate nominal HEPs following the steps in the Post Trip Action and Station Blackout procedures. The nominal HEPs were modified using performance shaping factors (PSFs) from the SPAR-H method [6]. The PSFs were autocalculated based on plant parameter information from the plant simulation [9, 10]. The PSFs correspond to multipliers to increase or decrease the overall HEP. The simulation therefore calculated distinct context-dependent HEPs for each procedure step. Task durations were also approximated using GOMS-HRA timing data [19].

Figure 1: Human Error Probability (HEP) as a Function of Time (t) for Loss of Offsite Power (LOOP) and Loss of Diesel Generator (LODG).

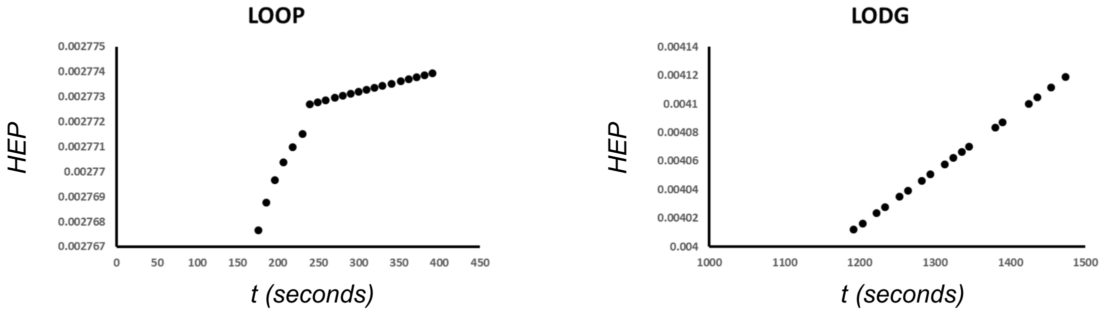

Figure 1 presents HEPs calculated for LOOP and LODG activities. The two data sets represent different types of curves. The LOOP curve, occurring earlier in the event progression when there are multiple changes as the plant cools down, is curve fitted by a two-order polynomial function:

$$HEP = e - 10t^2 - 7t + 0.0028, R^2 = 0.951, \tag{1}$$

where *HEP* is the human error probability at that point in time and *t* is the time of the activity. R^2 is a measure of the goodness of fit. For the interval of time representing actions taken in response to the LODG event, the HEP function is considerably more linear:

$$HEP = 4e - 7t + 0.0036, R^2 = 0.999. \tag{2}$$

The HEPs are higher for Equation 2, because the severity of the event and the factors influencing operator performance (e.g., available time and complexity) are rendered worse as the event progresses.

Equations 1 and 2 are, of course, not the functions originally used to autocalculate the HEPs. They do, however, show the dynamic nature of HEPs across time and tasks. A limitation of this paper is that these data sets do not begin to cover the variety of HEP possibilities over an event progression, but they provide a working sample of the types of data that an actual dynamic HRA approach has generated. A more hypothetical and complete HEP example can be found in [20]. Note that the HEPs as presented here do not account for dependence between tasks and the types of HEP adjustments that might be performed to account for error escalation once an initial error occurs.

4. HUMAN ERROR PROBABILITY AGGREGATION

4.1 Introduction

Let us assume that LOOP and LODG represent HFEs associated with activities where the human contributes to a *Failure to Prepare Plant for Shutdown* for LOOP and a *Failure to Initiate Backup Power* for LODG. Over fifty procedure steps were modeled for the station blackout scenario, such that HFE_{LOOP} and HFE_{LODG} each include approximately 25 separate operator tasks (i.e., procedure level primitives in the parlance of GOMS-HRA) that map to even more task level primitives. The two HFEs produced error curves as shown in Figure 1. The question remains how to translate the individual task-level HEPs to overall HEPs for each of the HFEs. Put another way, how should we map a continuous HEP function from dynamic HRA into a discrete HEP aligned to the HFE used in static HRA?

4.2 Conservative HEP

Some HRAs may forsake realism for conservatism. Especially screening analyses tend to consider near worst case scenarios and result in higher likelihood HEPs. The reasoning for such an approach is typically to test the risk significance of the HFE—if a high HEP has little consequence to the overall outcome in the HRA or PRA, the HFE is screened out.

A conservative interpretation of the error curves for HFE_{LOOP} and HFE_{LODG} might simply entail assuming the worst case HEP, which would be the maximum value:

$$HEP_{conservative} = \max_{HEP\in P} f(HEP), \tag{3}$$

where *f(HEP)* is the set of HEPs calculated for that HFE. The maximum function is not a good reflection of the evolution of the series of HEPs for that HFE unless there is little variability in the HEPs. Applying Equation 3, the conservative HEP is 2.773E-3 for HFE_{LOOP} and 4.118E-3 for HFE_{LODG}. Note that both HFEs have narrow ranges between the minimum and maximum values in the series. HFE_{LOOP} has a range of 6.288E-06, while HFE_{LODG} spans 1.070E-4. With such small ranges, the conservative HEPs for both HFEs may prove reasonable estimations of the aggregate HEPs.

4.3 Central Tendency HEP

The standard measures of central tendency are the median and mean. (For the present discussion, we do not consider the mode.) The median is the true midpoint of the data series, while the mean is the average value. The mean assumes a normal distribution and is therefore susceptible to outliers; however, most human performance is assumed to fall on a normal distribution and would be appropriate for treatment by the mean.

The median HEP values were 2.773E-3 for HFE_{LOOP} and 4.066E-3 for HFE_{LODG}. The median HEP is identical to the conservative HEP for HFE_{LOOP}, owing to the narrow range of autocalculated HEP values for that HFE. The median HEP is lower than the conservative HEP for HFE_{LODG}, which reflects the broader range of data.

The average value, also known as the expected value (E), is calculated as the average value of the HEP function over the HFE interval. This may be represented as:

$$E(HEP) = \int_{t_0}^{t_n} \frac{HEP(t)dt}{t_n - t_0}, \tag{4}$$

where t_0 is the starting time of the HFE interval and t_n is the ending time. The expected or average HEP is 2.772E-3 for HFE_{LOOP}, and 4.065E-3 for HFE_{LODG}. These values are quite similar to the median HEPs for the two HFEs, suggesting minimal skew in the data.

5. CONCLUSION

Both the median and mean (Equation 4) HEP estimates provide more representative single-point echoes of the HEP dynamic function than does the conservative (Equation 3) HEP calculation. While the range of numbers for each HFE is limited, an index of central tendency will prove less prone to outliers than a conservative calculation. Unless the purpose of the analysis is screening, the risk of excessive conservatism limits the generalizability of the results from Equation 3.

Additional measures of mathematical and functional (i.e., task analytic) aggregation should be considered across more diverse data sets. The present simple functions serve to demonstrate that it is possible to map dynamic HEPs at the task level to an aggregate HFE level. Such aggregation serves as a necessary step to calibrating the results of autocalculated HEPs to those produced by human analysts. Human analysts—prone to considerable intra- and inter-analyst variability—may not be the best benchmark for accurate HEPs. Rather than calibrate dynamic HEPs to human analysts, it may prove more fruitful to validate dynamic HEPs to empirical data sources.

HRA must not only be useful at a level that is defined by vague interactions with hardware systems. Yet, such is the case with HFEs. In fact, human error always occurs at the task level. Some task-level human errors are consequential, while others are not. It is the consequential ones that trickle up to affect the HFE, but the HFE is rarely the most suitable level of analysis to understand those underlying human tasks. Human reliability analysts may employ considerable skill in accounting for the relevant tasks when considering the HFE, but such expertise remains largely unformalized. A virtual analytic tool like dynamic HRA cannot deduce handpicked tasks that may have guided a human analysis. It is therefore essential that HRA consider the task—not the HFE—as the fundamental unit of analysis. To do so requires compatibility between HEPs generated for HFEs vs. tasks. This paper is a first attempt to bridge task and HFE quantification. Despite this beginning, additional aggregation techniques should be explored. These become important not only to calibrate to HFEs as treated in HRA in current practice but to allow HRA to operate outside the bounds of HFEs. The groupings of HFEs may ultimately prove less useful than the high consequence tasks amid a series of human activities. Such tasks are likely the main drivers of human errors that cause hardware failures. These tasks are not, however, discrete events. They are part of a continuum of activities, all of which should be considered for the sake of modeling completeness.

Disclaimer

The opinions expressed in this paper are entirely those of the author and do not represent official position. This work of authorship was prepared as an account of work sponsored by Idaho National Laboratory, an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees makes any warranty, express or implied, or assumes any

legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately-owned rights. Idaho National Laboratory is a multi-program laboratory operated by Battelle Energy Alliance LLC, for the United States Department of Energy under Contract DE-AC07-05ID14517. This research was funded through the Laboratory Directed Research and Development program at Idaho National Laboratory.

References

[1] American Society of Mechanical Engineers. (2013). Addenda to ASME/ANS RA-Sb–2013, Standard for Level 1/Large Early Release Frequency Probabilistic Risk Assessment for Nuclear Power Plant Applications, ASME/ANS RA-Sb-2013. New York: American Society of Mechanical Engineers. [2] Boring, R.L., and Joe, J.C. (2014). Task decomposition in human reliability analysis. Joint Probabilistic Safety Assessment and Management and European Safety and Reliability Conference.

[3] Poucet, A. (1989). Human Factors Reliability Benchmark Exercise, Synthesis Report, EUR 12222 EN. Luxembourg: Office for Official Publications of the European Communities.

[4] Lois, E., Dang, V.N., Forester, J., Broberg, H., Massaiu, S., Hildebrandt, M., Braarud, P.Ø., Parry, G., Julius, J., Boring, R., Männistö, I, and Bye, A. (2009). International HRA Empirical study—Phase 1 Report, Description of Overall Approach and Pilot Phase Results from Comparing HRA methods to Simulator Performance Data, NUREG/IA-0216, Vol. 1. Washington, DC: US Nuclear Regulatory Commission.

[5] Swain, A.D., & Guttman, H.E. (1983). Handbook of human reliability analysis with emphasis on nuclear power plant applications. Final report. NUREG/CR-1278. Washington, DC: US Nuclear Regulatory Commission.

[6] Gertman, D., Blackman, H., Marble, J., Byers, J., & Smith, C. (2005). The SPAR-H Human Reliability Analysis Method, NUREG/CR-6883. Washington, DC: US Nuclear Regulatory Commission.

[7] Rasmussen, M., Boring, R. L., Ulrich, T., & Ewing, S. (2017). The virtual human reliability analyst. Advances in Intelligent Systems and Computing, 589, 250-260.

[8] Boring, R., Mandelli, D., Rasmussen, M., Herberger, S., Ulrich, T., Groth, K., & Smith, C. (2016). Human Unimodel for Nuclear Technology to Enhance Reliability (HUNTER): A framework for computational-based human reliability analysis. 13th International Conference on Probabilistic Safety Assessment and Management (PSAM 13), Paper A-531, pp. 1-7. [9] Boring, R., Rasmussen, M., Smith, C., Mandelli, D., & Ewing, S. (2017). Dynamicizing the SPAR-H method: A simplified approach to computation-based human reliability analysis. Proceedings of the 2017 Probabilistic Safety Assessment Conference, 1024-1031.

[10] Rasmussen, M., & Boring, R.L. (2016). Implementation of complexity in computation-based Human Reliability Analysis. Risk, Reliability and Safety: Innovating Theory and Practice, Proceedings of the European Safety and Reliability Conference, pp. 972-977.

[11] Kirwan, B., and Ainsworth, L.K. (1992). A Guide to Task Analysis. London: Taylor and Francis.

[12] Boring, R.L. (2015). Aligning task analysis with human reliability analysis. Proceedings of the 59th Annual Meeting of the Human Factors and Ergonomics Society, pp. 416-420.

[13] Boring, R.L. (2015). Adapting human reliability analysis from nuclear power to oil and gas applications. Proceedings of 2015 European Safety and Reliability (ESREL) Conference, pp. 2853-2860.

[14] Boring, R.L. (2010). How many performance shaping factors are necessary for human reliability analysis? Proceedings of the 10th International Probabilistic Safety Assessment and Management Conference.

[15] Boring, R.L., & Rasmussen, M. (2016). GOMS-HRA: A method for treating subtasks in dynamic Human Reliability Analysis. Risk, Reliability and Safety: Innovating Theory and Practice, Proceedings of the European Safety and Reliability Conference, pp. 956-963.

[16] Boring, R. L., Rasmussen, M., Ulrich, T., Ewing, S., & Mandelli, D. (2017). Task and procedure level primitives for modeling human error. Advances in Intelligent Systems and Computing, 589, 30-40.

[17] Boring, R., Mandelli, D., Rasmussen, M., Herberger, S., Ulrich, T., Groth, K., & Smith, C. (2016). Integration of Human Reliability Analysis Models into the Simulation-Based Framework for the Risk-Informed Safety Margin Characterization Toolkit, INL/EXT-16-39015. Idaho Falls: Idaho National Laboratory.

[18] American Nuclear Society. (2012). Fukushima Daiichi: American Nuclear Society Committee Report. LaGrange Park: American Nuclear Society.

[19] Ulrich, T., Boring, R., L., Ewing, S., & Rasmussen, M. (2017). Operator timing of task level primitives for use in computation-based human reliability analysis. Advances in Intelligent Systems and Computing, 589, 41-49.

[20] Boring, R.L. (2015). A dynamic approach to modeling dependence between human failure events. Proceedings of the 2015 European Safety and Reliability (ESREL) Conference, pp. 2845-2851.