Validating THERP: Assessing the scope of a full-scale validation of the Technique for Human Error Rate Prediction

Rachel Benish Shirley, Ohio State University (corresponding author)

Shirley.72@osu.edu, 1557 Meadow Road, Columbus Ohio 43212, (216)-502-1018

Dr. Carol Smidts (Smidts.1@osu.edu), Meng Li, Atul Gupta (Ohio State University)

Abstract

Science-based Human Reliability Analysis (HRA) seeks to experimentally validate HRA methods in simulator studies. Emphasis is on validating the internal components of the HRA method, rather than the validity and consistency of the final results of the method. In this paper, we assess the requirements for a simulator study validation of the Technique for Human Error Rate Prediction (THERP), a foundational HRA method. The aspects requiring validation include the tables of Human Error Probabilities (HEPs), the treatment of stress, and the treatment of dependence between tasks. We estimate the sample size, n , required to obtain statistically significant error rates for validating HEP values, and the number of observations, m , that constitute one observed error rate for each HEP value. We develop two methods for estimating the mean error rate using few observations. The first method uses the median error rate, and the second method is a Bayesian estimator of the error rate based on the observed errors and the number of observations. Both methods are tested using computer-generated data. We also conduct a pilot experiment in The Ohio State University's Nuclear Power Plant Simulator Facility. Student operators perform a maintenance task in a BWR simulator. Errors are recorded, and error rates are compared to the THERP-predicted error rates. While the observed error rates are generally consistent with the THERP HEPs, further study is needed to provide confidence in these results as the pilot study sample size is small. Sample size calculations indicate that a full-scope THERP validation study would be a substantial but potentially feasible undertaking; forty hours of observation would provide sufficient data for a preliminary study, and observing 101 operators for twenty hours each would provide data for a full validation experiment.

1. Introduction

In the thirty years since the Technique for Human Error Rate Prediction (THERP) was introduced in 1983, the challenge of verifying and validating Human Reliability Analysis (HRA) methods remains unanswered [1]. Over the years, researchers and practitioners have completed a variety of benchmarking and validation exercises. Many of the studies benchmark HRA methods against each other (i.e. compare predicted HEPs from different methods against each other) but do not have actual error rate data to validate their findings.

One notable exception is Kirwan's quantitative benchmarking exercise ([2], [3], [4]). In this study, Kirwan compares HRA estimates to Human Error Probabilities extracted from the Computerized Operator Reliability and Error Database (CORE-DATA). Analysts using three HRA methods analyze a series of scenarios to estimate HEPs, which Kirwan et al then compare to the empirical data from CORE-DATA. The methods assessed are THERP, the Human Error Assessment and Reduction Technique (HEART), and the Justification of Human Error Data Information (JHEDI). Emphasis is on skill-and-rule based tasks, a common application for THERP. The study evaluated a total of thirty HEPs, with twenty-three from real events, five from simulator studies, one based expert-judgment, and one taken from

ergonomics experiments. For each of the three techniques evaluated, ten experts assessed the thirty HEPs. Assessors reviewed scenario descriptions associated with the thirty HEPs and estimated error probabilities for each. Results were generally favorable, with significant correlation between estimates and their corresponding true values for all three techniques. On average, researchers found that assessments were within a factor of ten of the reported HEP.

A recent (2008) study conducted at Halden HAMMLAB compares experienced operators' performance in the Halden simulator to HRA assessments conducted using thirteen different HRA methods (Table 1).

Table 1: HRA Methods Assessed in Halden HAMMLAB's International HRA Empirical Study

HRA Methods in the International HRA Empirical Study [5]
Accident Sequence Evaluation Program Human Reliability Analysis Procedure (ASEP)
Technique for Human Error Rate Prediction (THERP)
A Technique for Human Event Analysis (ATHENA)
CBDT + THERP Cause-Based Decision Tree (CBDT) Method
Commission Errors Search and Assessment-Quantification (CESA-Q)
Cognitive Reliability and Error Analysis Method (CREAM)
Decision Trees + ASEP
Enhanced Bayesian THERP
Human Error Assessment and Reduction Technique (HEART)
Korean Human Reliability Analysis method (KHRA)
Methode d'Evaluation des Missions Operateurs pour la Securite (MERMOS)
New Action Plan for the Improvement of the Human Reliability Analysis Method (PANAME)
Standardized Plant Analysis Risk-Human Reliability Analysis (SPAR-H)

In this experiment, fourteen different crews worked through several different scenarios. While their performance provided quantitative information, the sample size was too limited to make the quantitative data particularly useful; the strength of this study is the qualitative data acquired during the exercise. Of particular interest in the International HRA Empirical Study are PSFs, with a focus on the difference between the PSFs observed in the simulator and those predicted by the various HRA methods. HRA analysts were asked to list PSFs, causal factors and any other characterizations specified by their respective methods, as well as to predict what might be difficult for operating crews in each scenario [5]. These elements of the HRA assessment are compared to the experiences of fourteen licensed PWR crews in four scenarios: simple and complex variations on a steam generator tube rupture and a loss of feedwater accident.

What is missing from the studies is a systematic, science-based validation method for the components of various methods based on external data, i.e. observed error rates. We believe that simulators are a rich resource for obtaining this type of data. For example, the THERP method relies on tables of error rates for various types of potential errors. **The pilot project described below explores the requirements for obtaining measured error rates to validate the HEPs in the THERP tables. This represents a first**

step towards developing a science-based method for verifying and validating a wide variety of the components that contribute to HRA methods.

2. Materials and Methods

As THERP is the HRA method selected for the pilot study, we discuss the THERP method and enumerate the components of the method requiring validation. We also discuss the requirements for developing a valid study in a simulator environment. Finally, we discuss a pilot study conducted at The Ohio State University and the lessons learned from this project.

2.1 Technique for Human Error Rate Prediction (THERP)

The pilot study considers the Technique for Human Error Rate Prediction (THERP), a foundational HRA method. THERP was developed by the nuclear power industry in the 1980s [6]. It is most commonly used to assess routine operations and maintenance tasks. THERP analysts perform a task analysis and develop an HRA event tree to model all possible operator errors, then estimate the human error probability (HEP) associated with each task to compute an overall HEP for the scenario or task (Figure 1). The so-called “THERP Handbook,” NUREG/CR-1278 [6], includes tables of error types and associated HEPs. These values come from the available literature and, in many cases, expert judgment. The tables also include Error Factors (EFs), which specify the expected range of the HEP.

THERP analysts adjust the HEP to reflect operating conditions and performance shaping factors (PSFs). Some PSFs are explicitly addressed in the THERP handbook, while others are left to the analyst’s discretion.

Analysts also account for dependence between steps. The THERP handbook employs a positive dependence model (i.e. an error in step A increases the probability of an error in step B). Using the model, analysts categorize the relationship between two consecutive steps as having zero, low, medium,

Figure 1: Sample THERP event tree diagramming the probability of misreading an analog meter. In this example, events are assumed to be independent (i.e. there is *Zero Dependence* between steps).

high, or complete dependence. They then use the formulas provided in the THERP Handbook to adjust the joint HEP to reflect the dependence between steps.

The HEP adjusted for PSFs and dependence is referred to as the modified HEP. The HEP for the entire task, calculated using the modified HEPs in the HRA event tree, is the joint HEP.

2.1.1 Elements Requiring Validation

This analysis focuses on the aspects of THERP that relate to routine operations and maintenance tasks, as this has traditionally been the primary application of the THERP method. There are 103 estimated HEPs in the THERP Handbook. **Restricting analysis to routine operations and maintenance tasks occurring in the control room results in 66 HEPs to validate. Each HEP includes 3 assumptions requiring verification: (1) the median HEP, (2) use of the lognormal distribution to characterize the error rate for each error type, and (3) the standard deviation of each error rate (specified by the Error Factor). In addition, 4 levels of stress, 2 levels of experience, and 5 levels of dependence must be validated.**

THERP expressly addresses three primary performance shaping factors (PSFs): stress, expertise, and tagging system. THERP specifies four stress levels and two experience levels, each of which can be evaluated for step-by-step and dynamic procedures. However, as THERP is applied primarily to maintenance and routine operations in which step-by-step procedures are used, application in dynamic scenarios does not need to be included in an initial validation effort. Similarly, tagging levels are not considered in this assessment as they are assumed to be used primarily outside the control room. This results in eight conditions to test: two levels of experience for each of the four stress levels.

The sufficiency and accuracy of the THERP dependence model must also be addressed. These rules are listed in Table 20-17 of the THERP Handbook and reproduced in Table 19 of this paper.

The 27 THERP tables are summarized in Table 2. We limit our analysis to HEPs and rules related to routine tasks as this is the primary application for THERP, and to activities that occur in the control room, as the experiment design is built around control room simulators. Items that are not addressed include tables relating to screening and diagnosis, treatment of operator response to annunciators, and procedure writing, etc. While these aspects of the THERP method are not discussed in this paper, the techniques and structure developed here could be used to design an experiment to evaluate operator response to multiple, sequential alarms, for example.

Table 2: A summary of the THERP tables of error types. Those that are not considered in this analysis are marked as not applicable (NA) for various reasons.

Category	THERP Table #	THERP Table Topic	Status in Validation
<i>Screening</i>	1	<i>Diagnosis</i>	NA – not procedure-based tasks
	2	<i>Rule-based Actions</i>	NA – not procedure-based tasks
<i>Diagnosis</i>	3	<i>Nominal Diagnosis</i>	NA – not procedure-based tasks
	4	<i>Post Event CR Staffing</i>	NA – not procedure-based tasks
Errors of Omission, Written Materials Mandated	5	<i>Preparation</i>	NA – not control room activity
	6	Administrative Control	5 HEPs to validate (3 additional HEPs not included in this analysis)

	7	Procedural Items	5 HEPs to validate
Errors of Omission, No written materials	8	<i>Oral instruction</i>	<i>NA – not procedure-based tasks</i>
	9	Display selection	4 HEPs to validate
Errors of Commission	10	Displays – Read/Record (Quantitative)	11 HEPs to validate
	11	Displays – Check/read (quantitative)	8 HEPs to validate
	12	Control/MOV selection and use	14 HEPs to validate
	13	Locally Operated Valves - selection	5 HEPs to validate
	14	Locally Operated Valves – stuck valve detection	4 HEPs to validate
PSFs	15	<i>Tagging Levels</i>	<i>NA – outside the control room</i>
	16	Stress/Experience	[addressed separately]
	17	Dependence	[addressed separately]
	18	<i>Dependence – Failure in previous task</i>	<i>NA – Captured in Table 17</i>
	19	<i>Dependence – Success in previous task</i>	<i>NA – Captured in Table 17</i>
Uncertainty Bounds Guidelines	20	<i>Estimated HEPs and Uncertainty Bounds (UCBs)</i>	<i>NA – estimated HEPs not specified in THERP Handbook</i>
	21	<i>Conditional HEPs and UCBs</i>	<i>captured in Table 17</i>
Recovery Factors	22	Errors by Checker	10 HEPs to validate
	23	<i>Annunciated Cues</i>	<i>NA – not routine operations</i>
	24	<i>Annunciated Cues</i>	<i>NA – not routine operations</i>
	25	<i>Control Room Scanning</i>	<i>NA – checks for abnormal situations</i>
	26	<i>Control Room Scanning</i>	<i>NA – checks for abnormal situations</i>
	27	<i>Basic Walk-Around Inspection</i>	<i>NA – outside the control room</i>

2.2 Simulator Study Validity

The proposed experiment validates the 66 HEPs and associated factors in a controlled simulator environment. The first challenge in performing a simulator study is developing the simulator environment. The simulator must be sufficiently similar to the natural setting (in this case, a commercial nuclear power plant control room) to produce real-world effects. At the same time, the simulation must be sufficiently controlled for meaningful data to be collected. A host of factors must be considered to meet or address these two competing goals. Some of the biases expected in an NPP simulator study are listed in Table 3, along with recommended mitigation strategies [7].

Table 3: Anticipated biases in a simulator study

Bias	Definition	Mitigation Options
Hypervigilance	Participants are extraordinarily vigilant and fastidious because they know they are being observed or because they expect an event to occur	Assign a distracter task to divert participants' attention
Cavalier behavior	Participants are negligent or cavalier because there are no real world consequences to the outcome of the simulation	Provide incentives for diligent performance
Simulator sickness	Aspects of the simulator make participants physically sick, for example, motion sickness in a driving simulator. Aspects of a digital NPP control room may have similar effects.	Perform an ergonomic assessment and design the simulator accordingly.
Prominence hypothesis bias	Participants bring previously formed hypotheses or beliefs into decision making situations.	Refrain from giving participants too much insight into the study before the simulation begins.
Policy response bias	Participants believe they will benefit from a certain response.	Make the study anonymous.
Context effects	The simulator does not adequately represent attributes of the real world.	Preserve the look and feel of the plant environment as much as possible within the simulator; add a bias correction if necessary.
Incentive effects	Incentives introduced to mitigate other biases are disproportionate or do not perfectly adjust for the bias	Avoid offering incentives
Technology bias	Technology (such as touch screens in a digital NPP simulator) is unfamiliar to participants.	Provide adequate training, recruit operators familiar with newer technologies, or bias results according to participant technical proficiency levels.

In this study, we outline an approach that could be used in any simulator setting, then test the initial experiment design using student operators. Several biases mentioned in Table 3 indicate the need for careful participant selection from a pool of experienced NPP operators. In order to avoid the inevitable waste of time and money that would follow from using experienced operators in pilot experiments, this study employs students who have been trained to respond to select NPP events but who do not have the deep knowledge of the plant expected of professional operators. The validity of using student operators with segmented training as a proxy for professional operators is a question that must be addressed in future studies. We expect that some aspects of HRA method validation can be undertaken using student operators, while other aspects require true expertise. The structure developed and tested in this study can be used with experienced operators for a definitive validation study.

Although the simulator program used in this study is a full-scale digital model of a commercial nuclear power plant, there are several disparities between the original plant and the digital simulator. The simulator used in this study projects images of the analog panels onto a series of computer monitors. Although all the controls are accessible to the operator, they are arranged and accessed differently, and are manipulated through touch screens or a mouse and keyboard. Importantly, the commercial plant's digital displays are not included in the simulator, so operators do not have any of the supplemental information professionals tend to rely on for quick information.

Furthermore, as every NPP in the United States is unique, any operator outside of his or her "home" plant is working in a slightly off-normal environment. Although differences are generally negligible, an operator might reasonably be expected to spend more time locating controls or identifying appropriate set points in an unfamiliar plant.

To address the unanswered questions regarding the validity of simulator studies in the environments listed above, ideally the same experiment would be conducted in the following settings:

- Experienced operators in their home plant's traditional (i.e. analog) panel simulator
- Experienced operators in an unfamiliar plant's traditional simulator
- Experienced operators in a digital simulator
- Novice (student) operators in a digital simulator

This set of experiments would provide a mechanism for determining which aspects of a study must be conducted by experienced operators and which aspects can be tested using a more cost-effective approach.

2.3 Pilot Study at The Ohio State University Nuclear Power Plant Simulator Facility

The objective of the pilot study is to collect data in a realistic setting, under typical performance shaping factors. To make the simulation as realistic as possible, we employed a full-scale digital simulator of a commercial BWR power plant and completed the test using an authentic plant procedure.

We selected a portion of the routine High-Pressure Core Injection (HPCI) system test procedure. The procedure was modified slightly to accommodate the simulator interface, but the procedure steps were not simplified. Pump and valve names, for example, retained their original nomenclature.

The study used a full-scale digital simulator of a commercial nuclear power plant (Figure 2). This digital simulator is modeled after a boiling water reactor currently operating in the United States that has been modified to eliminate resemblance to the reference plant. Operators interact with mock-ups of the control room hard panels displayed on touch-screen computer monitors and use drop-down menus to initiate remote functions that would be performed in other locations in the plant. Although the monitors are touch-screen, many operators preferred to use the mouse to interact with the "soft panel" mimics.

Participants had access to the full HPCI panel, which also included instrumentation and controls for several other systems. Screens in the room also displayed a rudimentary safety parameter display system and a map of the core. Participants did not have access to the other panels because the procedure did not require them to interact with the rest of the control room.

Overall, the selected procedure portion included 39 steps or sub-steps. Of these, 18 were actions and 21 were check/confirmation steps. Three steps were remote functions that were selected from a dropdown list rather than being performed on a soft panel.

Figure 2: The OSU Nuclear Power Plant Simulator Facility

Ten students in a Human Reliability Analysis class were trained to be operators for the experiment [8]. The student operators were upperclassmen and graduate students in the mechanical and nuclear engineering programs at The Ohio State University. Although many of the students were familiar with nuclear systems, they all participated in a brief introduction to boiling water reactors, the emergency core cooling system, and HPCI before being trained in the simulator on how to complete the HPCI test procedure.

One question that remains unanswered following the pilot study is whether trained student operators are adequate substitutes for experienced power plant operators. In a follow up survey administered several months after the study, students were asked how confident they would feel resuming the operator role. Student responses averaged 4 out of 5, indicating they felt fairly confident in their roles as operators. Further tests to validate their competence could be administered in a more extensive trial.

2.3.1 Simulation Procedure

Student operators worked in three teams, each team rotating through three roles in the simulator room: primary operator, secondary (remote function) operator, and observer/data recorder. The observer was tasked with recording errors and deviations from the procedure that were committed by the two operators. Because they were all in the same room, the three students could easily confer with each other throughout the procedure runs. Although the data recorder was technically not part of the operator team, in some cases the observer stepped in to confer with the primary operator or assist the secondary operator. A facilitator also remained in the room during the experiment.

Each team was tasked with completing the HPCI test procedure as many times as possible within their 2.5 hour session in the simulator, but, in an effort to eliminate time pressure, they were not given a specific number of runs to complete. The first team of students completed three runs, the second team completed ten runs, and the third team of students completed six runs.

2.3.2 Data Collection

Error data was collected through the simulator run logs and by the human observer who participated in each run. The simulator run logs recorded every command sent to the simulator, e.g. every valve opening, pump starting, or breaker closing. Relevant simulator parameters such as water level were not tracked, so these logs do not identify errors of commission that occurred when operators performed a

parameter-dependent action before the variable reached the appropriate level. Similarly, many of the steps were simply status checks that could not be recorded by the simulator software. For these steps, data collection depended on the human observer. Observers were asked to confirm that each step was completed correctly or indicate how the operator erred.

2.3.3 Recommended Changes to Data Collection Setup

Two modifications to the initial design are recommended for future iterations of this study. First, additional training should be provided for the students, and competence testing should be administered to verify that students share a common minimum level of expertise. One advantage of the initial study is that students could participate with minimal training, because the selected procedure was entirely rule-based. Determining the optimal level of training for student operators remains a challenge.

Second, the observer should be removed from the simulator room, or interaction between the observer and the operator should be eliminated. In several instances, the student observer stepped in to assist the student operator. This confirms that some student operators were not sufficiently trained to complete the selected procedure. In addition to improved training, a separate observation room or observer via video recordings would eliminate the interference of an outside observer.

2.4 THERP Analysis of the HPCI procedure in the Pilot Experiment

The data collected during the preliminary study are compared to a THERP analysis of the test procedure scenario. The students in the HRA course used the THERP Handbook to identify the number and type of potential errors, dependence between steps, and expected operator stress. Working in groups of two to four students, they completed an initial analysis before performing the test procedure, then modified their analysis after their experience in the simulator. Four reports were submitted.

Figure 3: Error types identified by each analysis group (G1, G2, G3 and G4). Error types are listed by THERP table number followed by the error number. Numbers of opportunities for each error are listed according to each group's analysis. Errors of Omission (Table 7, Number 4) were the most prominent type of possible error. The number of identified opportunities for each error vary widely; where Group 3 (G3) identified 23 possible errors of omission, Group 2 (G2) identified 40 possible errors.

As an experienced HRA analyst would expect, assessments varied significantly between groups (variations in expert application of THERP and other methods are discussed in [4]). Items as fundamental as the types of potential errors identified and the number of opportunities for error were not consistent. Figure 3 shows the potential errors identified by each group and the number of opportunities operators have to commit each error. Group 1 (G1 in Figure 3), for example, identified 39 instances in the procedure in which operators could skip a step, and error defined in line 4 of Table 7 in the THERP Handbook, while Group 3 identified only 23 possible errors of omission. Note that this chart includes all the steps in the procedure, but in the following analysis only independent steps are considered.

We use the weighted Cohen's Kappa index, κ , to quantify the consistency between groups' assessments. $\kappa = 1$ indicates complete agreement, while $\kappa = 0$ indicates no agreement. The Inter-Rater Reliability (IRR) based on number of errors per error type is 0.31, a "fair" IRR rating. The κ for dependence is slightly better, 0.46, deemed "moderate" in the literature [9]. This number represents the IRR based on the number of steps assessed at the same dependence levels between groups.

Figure 4: Dependence on previous step, as assessed by each analysis group.

Figure 4 shows how the four groups assessed the dependence of each step on the previous step. While groups one, two and four used zero dependence as the baseline dependence between steps, group three rated dependence between most steps as low rather than zero. Removing group three from the IRR calculations increases the dependence IRR from 0.46 to 0.73, showing "moderate" inter-rater reliability between groups one, two and four [10].

With only fair consensus between groups, it is not clear which assessment most accurately represents the potential errors associated with the given task. In the interest of completing the demonstration exercise, Group 2's assessment is used in the following analysis. However, it must be remembered that

this is a somewhat arbitrary choice; full execution of this experiment requires a robust assessment developed by several practicing professionals.

Table 4 lists the types and numbers of opportunities for error in the test procedure identified by group two.

2.4.1 Opportunities for Error in the Test Procedure

In Table 4, Columns A, B and C are taken directly from the THERP Handbook. Column A specifies the THERP table and line number that defines the potential error, and Column B lists the definition provided in the THERP Handbook. Column C lists the median error rate for the given error as listed in the THERP Handbook, with the Error Factor included in parentheses. In the THERP paradigm, ninety percent of error rates are expected to fall between the median value divided by the error rate and the median value multiplied by the error rate. For example, for errors of omission (7-4), the median value is 0.01 and the error factor is 3. Therefore, 90% of the error rates for errors of omission are expected to fall between 0.003 and 0.03.

Columns D, E and F are related to the experimental setup of the preliminary study. Column D lists the number of opportunities that operators will have to make a particular error while performing the procedure one time. Importantly, Table 4 includes only the number of zero dependence opportunities for error, i.e. instances in which the error could occur in steps that are judged to be independent of the previous step. Again referring to errors of omission (7-4), Group 2 found 33 opportunities for errors of omission in one execution of the procedure.

Table 4: Types and numbers of opportunities for error in the procedure identified by group 2. Column E lists the total number of opportunities for error students encountered during the entire data collection process; Column D shows the number of opportunities for that error in each run through the procedure. Students completed 19 runs of the procedure during the data collection process, so Column E is equal to Column D multiplied by 19.

A	B	C	D	E	F
<i>THERP Reference Table</i>	<i>THERP Table/Error Description</i>	<i>HEP + EF</i>	<i># of zero-dependence Opportunities for error in Test Procedure</i>	<i># of opportunities for error in 19 runs</i>	<i># of observed errors</i>
<i>Table 7</i>	<i>Estimated probability of errors of omission per item of instruction when use of written procedures is specified,</i>				
7-4	<ul style="list-style-type: none"> long list (>10 items), when procedures without checkoff provisions are used 	0.01 (3)	33	627	9
<i>Table 9</i>	<i>Estimated probability of error in selecting unannunciated displays for quantitative or qualitative readings</i>				

9-4	<ul style="list-style-type: none"> From an array of similar-appearing displays identified by labels only 	0.003 (3)	16	304	1
Table 10	<i>Estimated HEPs for errors of commission in reading and recording quantitative information from unannounced displays</i>				
10-1	<ul style="list-style-type: none"> Analog meter 	0.003 (3)	2	38	4
10-2	<ul style="list-style-type: none"> Digital readout (<4 digits) 	0.001 (3)	3	57	0
10-9	<ul style="list-style-type: none"> Recording task, Number of digits or letters to be recorded > 3 	0.001 (3)	1	19	0
Table 12	<i>Estimated probabilities of errors of commission in operating manual controls</i>				
12-2	<ul style="list-style-type: none"> Select wrong control on a panel from an array of similar-appearing controls identified by labels only 	0.003 (3)	8	152	1
12-8	<ul style="list-style-type: none"> Turn a two-position switch in wrong direction or leave it in the wrong setting 	0.0001 (10)	8	152	0
12-9	<ul style="list-style-type: none"> Set a rotary control to an incorrect setting 	0.001 (3)	2	38	0
12-10	<ul style="list-style-type: none"> Failure to complete change of state of a component if switch must be held until change is completed 	0.003 (3)	2	38	0
12-11	<ul style="list-style-type: none"> Select wrong circuit breaker in a group of circuit breakers densely grouped and identified by labels only 	0.005 (3)	2	38	0

Column E records the total number of opportunities for error over the entire preliminary data collection effort. During the experiment, operators completed the procedure nineteen times, meaning that the total number of opportunities for error is nineteen times the number of opportunities for error when completing the procedure once. Hence, operators had 627 opportunities to skip a step (error of omission), 33 per run through the procedure. Column F shows the number of errors that actually occurred during the nineteen runs through the procedure in the data collection period. We see that nine errors were recorded, out of 627 opportunities to commit that error.

3 Calculations: the Scope of a Validation Study for THERP

The pilot study demonstrates that simulator studies are feasible and illustrates some of the challenges related to conducting a successful validation exercise. In order to assess the scope of a full-scale THERP validation project, we must estimate the sample size for the study, i.e. the number of observed error rates necessary to produce statistically significant data. **We develop an approach for estimating the sample size, as well as defining the number of opportunities for error that constitute one sample. As this value turns out to be quite large, we introduce two methods for reducing the number of necessary observations: a median estimator and a Bayesian estimator of the mean error rate.**

3.1 Sample Size for Validating a THERP HEP

The sample is a sample of error rates. The number of error rates, n , required for a statistically significant sample is estimated via the hypothesis test using the Student's t distribution. This requires data to be normal, but error rates are expected to follow a lognormal distribution (characteristic of expert performance). We transform the error rate data from lognormal to normal by taking the natural logarithm of the error rate. By definition, the median of the lognormal distribution is the mean of the normal distribution. **If Θ is the median error rate, we calculate the mean error rate, μ , by taking the natural logarithm of Θ :**

$$\mu = \ln \Theta \quad (1)$$

The null hypothesis is that the true mean, μ , is the THERP mean, μ_T , where $\mu_T = \ln HEP_T$:

$$H_0: \mu = \mu_T$$

We define α as the probability of a Type I error, that is, the probability of rejecting the null hypothesis when the null hypothesis is valid. Rejection will occur when

$$z < \frac{z_{\alpha}}{2} \text{ or } z > z_{1-\frac{\alpha}{2}} \quad (2)$$

where z is the standard deviate. Therefore, the null hypothesis will be rejected when

$$\left| \frac{\bar{x} - \mu_T}{\frac{\sigma}{\sqrt{n}}} \right| > z_{1-\frac{\alpha}{2}} \quad (3)$$

Where $\bar{x}$ is the sample mean, σ is the standard deviation of the transformed error rate distribution, and n is the number of observed error rates in the sample. In order to estimate the necessary sample size, we determine the expected value of σ and limit the accepted range for $\bar{x}$.

3.1.1 Estimating the standard deviation, σ

We estimate σ using the range predicted by the EF, which specifies that 90% of the error rates are expected to fall within the interval $[HEP/EF, HEP * EF]$. This corresponds to $[\mu - \ln EF, \mu + \ln EF]$ when transformed to the normal distribution. Therefore, the lower bound of this interval corresponds to a probability of 0.05 and the upper bound to 0.95. We use this relationship to estimate σ :

$$\frac{(\mu_T + \ln EF) - \mu_T}{\sigma} = z_{0.95} \quad (4)$$

Solving for σ ,

$$\sigma = \frac{\ln EF}{z_{0.95}} \quad (5)$$

Once experimental data has been collected, the Chi-Square test can be used to validate the standard deviation in the observed data sets [11]. T is the critical test value:

$$T = (n - 1) * \left(\frac{s}{\sigma_T}\right)^2 \quad (6)$$

Here n is the sample size, s is the experimental standard deviation of the sample, and σ_T is calculated using the EF ($\sigma_T = \ln EF / z_{0.95}$). We do not reject the hypothesis that $\sigma = \sigma_T$ if

$$X_{\frac{\alpha}{2}, n-1}^2 < T < X_{1-\frac{\alpha}{2}, n-1}^2 \quad (7)$$

The analyses in this paper are based on the assumption that the standard deviation is accurately represented by the THERP EF; if this is false, an alternative approach would need to be developed.

3.1.2 The alternative hypothesis: defining the acceptable range for $\bar{x}$

Figure 5: H_0 is not rejected if $\bar{x}$ falls between c^- and c^+ ; μ_0 is considered accurate if the true mean, μ , falls between $\mu^- (\mu_T - \ln \xi)$ and $\mu^+ (\mu_T + \ln \xi)$. The power of the test is the probability that $\bar{x}$ falls outside $[c^-, c^+]$ when μ is outside $[\mu^-, \mu^+]$.

The objective of the test is to verify that the true error rate is not significantly greater than HEP_T . "Significant difference" depends on the specific objective of the validation exercise. In this analysis, researchers define the accepted range as anywhere within middle thirty percent of the hypothesized error distribution. In other words, the null hypothesis is valid if the true mean error rate falls between the 35th and the 65th percentiles of the error rate distribution, and the alternative hypothesis is that the true median error rate falls outside this range (Figure 5).

Thirty percent is selected to ensure that researchers are able to differentiate between a base HEP and a modified HEP that is two times the base HEP. This criterion reflects the minimum change made to the base HEP due to factors such as stress. To determine the maximum value for μ at which the null hypothesis is acceptable, let ξ represent the multiplier to the base HEP corresponding to the specified range, R .

The null and alternative hypotheses are therefore

$$\begin{aligned} H_0: ER &= HEP_T \\ H_a \begin{cases} H_a^+: ER < HEP_T * \xi \\ H_a^-: ER < HEP_T / \xi \end{cases} \end{aligned} \quad (8)$$

Transforming this into the normal distribution,

$$H_0: \mu = \mu_T$$

$$H_a \begin{cases} H_a^+ : \mu > \mu_T + \ln \xi \\ H_a^- : \mu < \mu_T - \ln \xi \end{cases} \quad (9)$$

We calculate ξ in terms of R and EF :

$$\frac{\mu_T + \ln \xi}{\sigma} = z_R \quad (10)$$

Replacing σ with the definition from Equation 5 and solving for ξ :

$$\xi = \exp \left(z_R * \frac{\ln EF}{z_{0.95}} + \mu_T \right) \quad (11)$$

This value ensures researchers are able to differentiate between a base HEP and a modified HEP that is twice as great as the base HEP. In other words, the null hypothesis must be rejected if the observed error rate is two times HEP_T .

Table 5 lists the multiplier, ξ , over increasing acceptable range values (R), for the two most common EF values, 3 and 10 (which correspond to $\sigma = 0.68$ and $\sigma = 1.40$ from Equation 5).

Table 5: The THERP Multiplier, ξ , as a function of accepted range, R

Range, R	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%
Top percentile	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95
$\xi, EF = 3$	1.00	1.09	1.18	1.29	1.42	1.57	1.75	2.00	2.35	3
$\xi, EF = 10$	1.00	1.19	1.43	1.71	2.08	2.57	3.25	4.27	6.01	10

As Table 5 shows, restricting the HEP to 30% of the expected range corresponds to a multiplier $\xi = 1.29$ for an EF of 3 and $\xi = 1.71$ for an EF of 10. With $R = 30\%$, $z_R = z_{0.65}$, corresponding to the upper limit of the middle 30% of the distribution. $R = 30\%$ is selected so that ξ is less than two for all EF. Although $R = 30\%$ for this paper, other applications might lend themselves to the selection of a broader (or narrower) range. The following analysis can be completed using any value for R , z_R and ξ .

3.1.3 The power of the test

The power of the test is the probability of correctly rejecting the null hypothesis, H_0 , when the alternative, H_a , is true. As β is the probability of accepting the null hypothesis when the alternative is valid, the power of the test is equal to $1 - \beta$.

We specify the desired power of the test in order to estimate the sample size necessary to perform a test of sufficient power. The first step is to calculate β^+ and β^- , the probabilities of accepting the null hypothesis when H_a^+ or H_a^- are valid.

3.1.3a The first alternative, H_a^+

The critical value, c^+ , is the maximum value of $\bar{x}$ for which the null hypothesis will be accepted. β^+ , the probability of accepting H_0 when H_a^+ is valid, is therefore

$$\beta^+ = \Phi \left(\frac{c^+ - (\mu_T + \ln \xi)}{\frac{\sigma}{\sqrt{n}}} \right) \quad (12)$$

Where Φ is the cumulative distribution function of the standard normal distribution (mean = zero, standard deviation = one). We determine c^+ from Equation 3:

$$c^+ = z_{1-\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}} + \mu_T \quad (13)$$

Therefore, β^+ becomes

$$\beta^+ = \Phi \left(z_{1-\frac{\alpha}{2}} - \frac{\ln \xi}{\frac{\sigma}{\sqrt{n}}} \right) \quad (14)$$

Using the definition from Equation 10, this simplifies to

$$\beta^+ = \Phi \left(z_{1-\frac{\alpha}{2}} - z_R * \sqrt{n} \right) \quad (15)$$

3.1.3b The second alternative, H_a^- :

As with c^+ , c^- is the minimum value of $\bar{x}$ for which the null hypothesis will be accepted. Following the same logic (and noting that H_a^- is less than H_0), the probability of accepting H_0 when H_a^- is valid is

$$\beta^- = 1 - \Phi \left(\frac{z_{\alpha}}{2} + z_R * \sqrt{n} \right) \quad (16)$$

As this value is equivalent to β^+ , we use $\beta = \beta^+ = \beta^-$ to determine the power of the test.

3.1.4 Sample size as a function of the power of the test

Again, the power of the test is the probability of rejecting H_0 when H_a is true, i.e. $1 - \beta$. Thus,

$$Power = 1 - \beta^- = \Phi \left(\frac{z_{\alpha}}{2} + z_R * \sqrt{n} \right) \quad (17)$$

In this analysis, we set $\alpha = 0.01$ because we desire a low probability of falsely rejecting H_0 . This value can be changed to match new specifications if desired. **With $\alpha = 0.01$, we find that a power of 0.90 can be obtained with a sample size $n \geq 101$** , while a power greater than 0.99 will be obtained with 162 or more sample error rates. If the THERP-derived estimate of σ is valid, β is independent of error rate and error factor, and this sample size is appropriate for all error types.

3.2 Defining a sample

We have determined how to specify n , the number of observed error rates required for a statistically significant sample given the desired power of the test and specified acceptable range. Next, we define m , the number of observations of opportunities for error required per observed error rate.

For a single data point, the observed error rate is the number of errors, e_i , divided by the number of observations, m_i . As all samples are expected to have the same number of opportunities for error, m_i is simply m . This is illustrated in Table 6.

Table 6: Sample data table. Three error rates are collected, each from ten observed opportunities for error ($n = 3, m_i = 10$ for all three samples).

	i	$m = 10$ opportunities for success (S) or failure (F)	e_i	$x_i = \ln \left(\frac{e_i}{m} \right)$
$n = 3$	1	SSFSSSFSS	$e_1 = 2$	$x_1 = \ln \left(\frac{2}{10} \right) = -1.6$
	2	SFSSSFSSS	$e_2 = 3$	$x_2 = \ln \left(\frac{3}{10} \right) = -1.2$

	3	SSSSSFSSSS	$e_3 = 1$	$x_3 = \ln\left(\frac{1}{10}\right) = -2.3$
				$\bar{x} = -1.7$

Recall that $\mu_T = \ln HEP_T$. In the same way, the experimental data x_i must be transformed from a lognormal to a normal distribution:

$$x_i = \ln\left(\frac{e_i}{m}\right) \quad (18)$$

$\bar{x}$ is the mean of the sample, taken from n trials:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n \ln \frac{e_i}{m} \quad (19)$$

This can be re-written as

$$\bar{x} = \left(\frac{1}{n} \ln(\prod_{i=1}^n e_i)\right) - \ln m \quad (20)$$

Thus the condition for rejection (Equation 3) becomes

$$\left| \frac{1}{\frac{\sigma}{\sqrt{n}}} \left(\left(\frac{1}{n} \ln(\prod_{i=1}^n e_i) - \ln m \right) - \mu_T \right) \right| > z_{1-\frac{\alpha}{2}} \quad (21)$$

Note that if any e_i is zero, the rejection condition is automatically satisfied. **This provides a constraint on m : the number of opportunities for error in one trial should be sufficiently large to provide the expectation that at least one error will occur.**

If the null hypothesis is true and μ_T is the mean error rate, then the expected human error probability is simply the HEP from the THERP handbook, HEP_T . Using this estimator and considering the process of observing errors as repeated Bernoulli trials over m error opportunities, the probability of observing at least one error in the i^{th} set of m opportunities is

$$p(e_i > 0) = 1 - (1 - HEP_T)^m \quad (22)$$

The probability of observing at least one error in each of n sets of m trials is therefore

$$p(\prod_{i=1}^n e_i > 0) = (1 - (1 - HEP_T)^m)^n \quad (23)$$

Let γ represent the minimum acceptable probability of observing at least one error in each of the sets of trials. The constraint on m is therefore:

$$\gamma < (1 - (1 - HEP_T)^n)^m \quad (24)$$

Solving for m in terms of γ, n :

$$m > \frac{\ln\left(1 - \gamma^{\frac{1}{n}}\right)}{\ln(1 - HEP_T)} \quad (25)$$

We calculate m for several HEP_T values at four values of γ with $n = 101$ in Table 7.

Table 7: Number of trials (m) required per sample as a function of γ , the probability of observing at least one error. m increases as γ increases.

HEP_T	$\gamma = 0.5$	$\gamma = 0.8$	$\gamma = 0.9$	$\gamma = 0.99$
0.01	$m = 496$	$m = 609$	$m = 683$	$m = 917$
0.005	$m = 995$	$m = 1,220$	$m = 1,370$	$m = 1,838$
0.003	$m = 1,659$	$m = 2,036$	$m = 2,285$	$m = 3,067$
0.001	$m = 4,983$	$m = 6,113$	$m = 6,863$	$m = 9,211$
0.0001	$m = 49,848$	$m = 61,159$	$m = 68,657$	$m = 92,149$

Although a sample size of 101 may be feasible in a laboratory study, the number of trials required for each sample is formidable, even for the relatively high HEP of 0.01.

3.3 Reducing m , the number of observations per sample

We develop two approaches to reducing m . First, we return to the THERP definition of HEP: the HEP is defined as the expected median error rate for a particular error type. If $e_{i-\text{median}}$ is the median value is a set of n observations of m opportunities for error, an alternative definition of $\bar{x}$ is therefore

$$\bar{x} = \ln\left(\frac{e_{i-\text{median}}}{m}\right) \quad (26)$$

Using this definition, up to half the samples can have no observed errors, as long as the median error rate is greater than zero. The limitation of this approach is that it precludes verifying the standard deviation (Equations 6 and 7) as the Chi Squared Test for Variance can only be applied to normally distributed data, and $\ln 0$ is undefined.

As an alternative, we develop a Bayesian estimator of the error rate. The estimator provides error rates for samples in which no errors are observed. This allows for all the analysis discussed in Section 3.2 to be completed with a significantly reduced m required per sample.

Both approaches yield similar values for m ; these values are approximately fifteen percent of the m required using the initial approach in Table 8.

3.3.1 Method 1: calculating m using the median error rate

Using the definition $\bar{x} = \ln(e_{i-\text{median}}/m)$, a viable estimate of the median error rate can be obtained in a sample in which half of the samples have no observed errors. This is the Median Estimator.

As in Section 3.2, we use the binomial distribution to estimate the probability of observing at least 51 non-zero samples in a set of 101. The probability of success, p , is the probability of observing at least one error in a sample of m opportunities for error. We define $\Gamma_{51,101}$ as the threshold for the probability of observing at least one error in 51 or more of the 101 samples and use the binomial distribution to determine a minimum value for p which can then be used to solve for m in Equation 22.

$$\Gamma < \sum_{i=51}^{101} \binom{101}{i} p^i (1-p)^{101-i} \quad (27)$$

Setting $\Gamma_{51,101} = 0.9$, we find p must be greater than 0.56; if $\Gamma_{51,101} = 0.99$, p increases to 0.62. Solving for m using Equation 25 yields the values in Table 8. An increase in $\Gamma_{51,101}$ from 0.9 to 0.99 corresponds to an eighteen percent increase in m ; these values are twelve to fourteen percent of the m required to obtain 90% confidence that no samples with zero errors will be observed.

Table 8: Number of trials, m , required per sample in order to observe a median error rate greater than zero. $\Gamma_{51,101}$ is the probability of observing at least one error in at least half of the n samples; p is the probability of observing at least one error in a single sample. $\Gamma_{51,101} = 0.92$ (corresponding to $p = 0.57$) is selected to determine m .

m(HEP,p) Number of trials per sample (m) as a function of HEP and probability of observing at least one error (p)						
$\Gamma_{51,101}$	p	HEP = 0.01	HEP = 0.005	HEP = 0.003	HEP = 0.001	HEP = 0.0001
0.9	$p = 0.56$	$m = 82$	$m = 164$	$m = 273$	$m = 821$	$m = 8209$
0.92	$p = 0.57$	$m = 85$	$m = 168$	$m = 281$	$m = 844$	$m = 8439$
0.99	$p = 0.62$	$m = 96$	$m = 193$	$m = 322$	$m = 967$	$m = 9675$
1	$p = 0.999$	$m = 683$	$m = 1370$	$m = 2285$	$m = 6863$	$m = 68657$

One further element must be considered. The sample size corresponding to $\Gamma_{51,101} = 0.9$ is small enough that one observed error would yield an error rate, $\bar{x}$, that is greater than c^+ , leading researchers to reject the null hypothesis. For example, if one error is the median number of errors observed in a set of 101 observations of 82 opportunities each, the median error rate is $1/82$, or 0.012. Employing the hypothesis test outlined above,

$$\frac{\ln 0.012 - \ln 0.01}{\frac{\ln 3 / z_{0.995}}{\sqrt{101}}} = 2.74 > 2.58 \quad (z_{0.995}) \quad (28)$$

As the test value for $\alpha = 0.01$ is 2.58, this result would be rejected. This highlights a final constraint on m by defining a minimum observable error rate, m_{min} :

$$\frac{\left(\ln\left(\frac{1}{m_{min}}\right) - \mu_T\right)}{\frac{\sigma}{\sqrt{n}}} < z_{1-\frac{\alpha}{2}} \quad (29)$$

Keeping $\alpha = 0.01$ and $n = 101$, this yields m_{min} values that result in $p = 0.57$ and $\Gamma = 0.92$ for EF = 3. For EF = 10, $p = 0.50$ and $\Gamma = 0.50$. **For consistency, m corresponding to $\Gamma = 0.92$ is used for all EF values. This is referred to as $m_{\Gamma=0.92}$.** These values are listed in bold in Table 8. Results from sample data generated using $m_{\Gamma=0.92}$ are discussed in Section 4.1.

3.3.2 Method 2: calculating m using a Bayesian estimator

As mentioned above, the limitation of this approach is that it precludes verifying the standard deviation.

As an alternative, we employ Bayesian methods to estimate the error rate from a sample with fewer than adequate observations for error, $\hat{\theta}_i$. We then calculate $\bar{x}$ using Equation 19, replacing (e_i/m) with $\hat{\theta}_i$.

With $\hat{\theta}$ as the estimator and θ representing the true error rate¹, a linear loss function, $L(\theta, \hat{\theta})$, represents the cost of incorrect estimation:

¹ In this notation, the true mean, μ , is equal to the natural log of the error rate, θ : $\ln \theta = \mu$; $\theta = e^\mu$.

$$L(\theta, \hat{\theta}) = \begin{cases} k_0(\theta - \hat{\theta}), \theta > \hat{\theta} & \text{(underestimates error rate)} \\ k_1(\hat{\theta} - \theta), \theta < \hat{\theta} & \text{(overestimates error rate)} \end{cases} \quad (30)$$

We choose $k_0 > k_1$ to ensure that the penalty for underestimating is greater than the penalty for overestimating.

The median value of $\hat{\theta}$ corresponds to the minimum expected loss [12]. To estimate $\hat{\theta}$, we therefore minimize

$$\int_a^b L(\theta, \hat{\theta}) p(\theta|y) d\theta \quad (31)$$

where a and b are the lower and upper bounds of the domain of θ , and $p(\theta|y)$ is the posterior for θ given the evidence, that is, the observations y .

The evidence consists of e errors observed in m opportunities for error. To find the minimum, we set the derivative of the revised problem to zero:

$$\frac{\partial}{\partial \hat{\theta}} \int_a^b L(\theta, \hat{\theta}) p(\theta|m, e) d\theta = 0 \quad (32)$$

Introducing the linear loss function for $L(\theta, \hat{\theta})$, the problem becomes

$$\frac{\partial}{\partial \hat{\theta}} \left[\int_a^{\hat{\theta}} k_1(\hat{\theta} - \theta) p(\theta|m, e) d\theta + \int_{\hat{\theta}}^b k_0(\theta - \hat{\theta}) p(\theta|m, e) d\theta \right] = 0 \quad (33)$$

Defining the cumulative distribution function for $p(\theta|m, e)$ to be $F(\theta|m, e)$, the minimum with respect to $\hat{\theta}$ occurs when

$$F(\hat{\theta}|m, e) - F(a|m, e) = \frac{k_0}{k_1 + k_0} [F(b|m, e) - F(a|m, e)] \quad (34)$$

A similar derivation is available in [13]. The median estimator for $\hat{\theta}$ therefore depends on four factors:

- $F(\theta|m, e)$, the cumulative distribution of θ
- The loss ratio, k_0/k_1
- The domain of θ , specified by the boundaries a and b
- The experimental data, i.e. the observed errors, e , and the number of opportunities for error, m

These factors are constrained by the α selected for the hypothesis test; the estimator for an error rate that matches the HEP_T must be less than the maximum accepted percentile of the distribution. In this case, with $\alpha = 0.01$, the estimator $\hat{\theta}_{ER=HEP}$ cannot fall in the top 0.5% of the distribution.

3.3.2a Elements of the Bayesian Estimator

To select the optimal parameters for the Bayesian estimator, we define the HEP Multiplier, M , as the value by which the THERP HEP is multiplied to obtain the new estimate, $\hat{\theta}$:

$$\hat{\theta} = M * \theta_T = M * HEP_T \quad (35)$$

The HEP Multiplier simplifies the discussion of how the parameters impact the Bayesian estimate by providing a basis of comparing estimators for different error rates.²

As in Equation 29, $\hat{\Theta}$ is constrained by α . This sets the limit for M when the error rate equals HEP_T :

$$\frac{\ln(M \times HEP_T) - \mu_T}{\frac{\sigma}{\sqrt{n}}} < z_{1-\frac{\alpha}{2}} \quad (36)$$

For $n = 101$, this corresponds to $M < 1.19$ for $EF = 3$ and $M < 1.43$, $EF = 10$. With this constraint in mind, **we recommend using the CDF of the Beta function to represent $F(\Theta|m, e)$, setting the Loss Ratio = 1.5, and restricting the domain of $\hat{\Theta}$ to the lower 80% of the expected range as defined in the THERP handbook.** We address the rationale for each of these parameters in turn.

The assumed distribution of Θ : $p(\Theta|m, e)$ and $F(\Theta|m, e)$

THERP assumes error rates follow a lognormal distribution, which is characteristic of skilled performance. Making use of the beta/binomial conjugate, we specify that

$$p(\Theta|m, e) = B(e + 1, m - e + 1) \quad (37)$$

The beta distribution is selected because it is constrained between zero and one and because it is a flexible distribution that can accommodate many shapes. Using the beta distribution, $F(\Theta|m, e)$ can be taken as the cumulative distribution function of the beta distribution.

The Loss Ratio, k_0/k_1

The loss ratio, k_0/k_1 , represents the *cost of underestimating risk* (k_0) over the *cost of overestimating risk* (k_1). We expect this ratio to be greater than one, reflecting potential high consequences of underestimating the risk of errors in a maintenance task. Letting k represent the loss ratio, we rewrite Equation 34:

$$F(\hat{\Theta}|m, e) - F(a|m, e) = \frac{k}{1+k} [F(b|m, e) - F(a|m, e)] \quad (38)$$

This relationship shows that the estimated median error rate, $\hat{\Theta}$, increases as k increases. Figure 6 shows M as a function of k with an unconstrained range (i.e. $a = 0, b = 1$). These are estimates based on an observed error rate of 0.01; we see one error in one hundred observations (1/100), two errors in two hundred observations (2/200), etc. As the number of trials increases, the estimated error rate decreases.

² Note that, in the Bayesian estimator derived below, M is not constant but increases as e and m increase; see Figure 6.

Figure 6: The estimated error rate, $\hat{\theta}$, increases as the Loss Ratio, k , increases. This chart shows the multiplier to the THERP HEP, M , corresponding to the Bayesian estimated error rate for $HEP_T = 0.01$ with $EF = 3$. While the observed error rate remains constant, as the number of observations increases, the estimated error rate decreases. Here, the range for $\hat{\theta}$ is unconstrained; the estimated error rate may fall anywhere between 0 and 1.

This shows that the number of trials is an important element in selecting k ; with only one hundred observations, a loss ratio of 1.5 yields the multiplier $M = 2$, but as m increases to 1,000, $M = 1.5$.

As we are trying to minimize the observations required per sample, this indicates that a viable estimator must use a low loss ratio.

The acceptable range of θ : (a, b)

The most rigorous approach in a validation study with no prior knowledge about θ is to allow the estimate to fall anywhere between zero and one; however, in a study of error rates, we can reasonably expect θ to be low. We therefore specify b , the maximum allowed value for the estimator. This is determined by an upper bound, the percentile of the distribution that the estimator is restricted to.

$$\frac{\ln b - \mu_T}{\frac{\sigma}{\sqrt{n}}} = Z_{UpperBound} \quad (39)$$

Solving for σ using Equation 5, the domain of θ is therefore

$$\{a = 0, b = \exp\left(\mu_T + \ln EF \times \frac{Z_{UpperBound}}{Z_{0.95}}\right)\} \quad (40)$$

Figure 7 shows how $\hat{\theta}$ varies with a change in the domain of θ using a Loss Ratio $k = 1.5$. This figure shows the HEP Multiplier for two representative EFs: 3 (solid lines) and 10 (dashed lines). The HEP used for these estimates are 0.01 (EF = 3) and 0.0001 (EF = 10). This reflects the values used in the THERP handbook. Note that when $a = 0$ and $b = 1$, M is the same for EF = 3 and EF = 10.

Figure 7: The estimated error rate, $\hat{\Theta}$, increases as the accepted range for the estimator increases. Solid lines correspond to $EF = 3$; dotted lines to $EF = 10$. In this chart, the loss ratio is fixed at 1.5.

The range used in this analysis is the lower 75-80% of the THERP-predicted range. These values were selected based on the simulated data discussed in Section 4.1.

With $a = 0$, the $F(a|m, e)$ is simply the CDF of the beta distribution at zero, which is zero. Therefore, the Bayesian estimator for the error rate can be written as

$$F_{\beta}(\hat{\Theta} | e + 1, m - e + 1) = \frac{k}{k+1} F_{\beta}(b | e + 1, m - e + 1) \quad (41)$$

Again, the recommended value for k is 1.5, and b is determined using Equation 40.

4 Validating the THERP HEPs

Both simulated data and data obtained in the pilot study are analyzed using the two methods discussed above. Data from the pilot study provides insight into the feasibility of a full-scope THERP validation study, and simulated data attest to the validity of the median error rate and Bayesian error rate estimators.

4.1 Computer-generated Data

We generate sample data using the statistical software R to test the utility of the specifications for n and m with various HEP values. We simulate 100 experiments for each HEP value tested. Each experiment consists of 101 sample error rates randomly distributed around a lognormal distribution with μ and σ determined by Equation 5. In each experiment, m is determined by setting $\Gamma_{51,101} = 0.92$; this corresponds to m_{min} for all error rates with $EF = 3$. Taking the median error rates from each simulated

data set, we reject the null hypothesis when the actual error rate is much less than or much greater than HEP_T , but we do not reject the null hypothesis when the error rate matches the actual HEP (Table 9).

Table 9: Results from experiments using computer-generated data show the utility of the median error rate estimator, which correctly distinguishes between data sets with varying error rates. Results in bold indicate that H_0 would not be rejected.

HEP	$HEP = 0.25 \times HEP_T$	$HEP = 0.5 \times HEP_T$	$HEP = 0.75 \times HEP_T$	$HEP = HEP_T$	$HEP = 1.25 \times HEP_T$	$HEP = 1.5 \times HEP_T$	$HEP = 1.75 \times HEP_T$	$HEP = 2.0 \times HEP_T$
EF = 3								
0.01	Reject H_0 in 100 of 100 experiments	Reject H_0 in 100/100	Reject H_0 in 90/100	Reject H_0 in 2/100	Reject H_0 in 70/100	Reject H_0 in 100/100	Reject H_0 in 100/100	Reject H_0 in 100/100
0.005	100/100	100/100	93/100	3/100	76/100	100/100	100/100	100/100
0.003	100/100	100/100	90/100	1/100	82/100	100/100	100/100	100/100
0.001	100/100	100/100	95/100	5/100	82/100	100/100	100/100	100/100
EF = 10								
0.0001	100/100	87/100	6/100	1/100	4/100	14/100	52/100	73/100

We perform a similar test using the Bayesian estimated error rates instead of the median error rates. The selected parameters for the Bayesian estimator of HEPs with $EF = 3$ are:

- Loss Ratio, $k = 1.5$
- Range, $R = 80\%$ of the total range

The results in Table 10 are similar to those in Table 9. The Bayesian estimator consistently rejects H_0 when the error rate is more than 1.75 times the HEP or when the error rate is less than 0.25 times the HEP. The Bayesian estimator does not reject H_0 when the error rate matches the HEP.

The difference between the Bayesian estimator and the median estimator is that the Bayesian estimator does not discriminate as finely as the median estimator; the estimator does not differentiate between an error rate that is half the HEP, while the median estimator differentiates between an error rate that matches the HEP and an error rate that is 0.75 times the HEP.

Table 10: Computer-generated data used to test the Bayesian estimator ($m = m_{\Gamma=0.92}$). While not as powerful as the median estimator, the Bayesian estimator successfully differentiates between HEP_T and higher error rates for $EF = 3$.

HEP	$HEP = 0.25 \times HEP_T$	$HEP = 0.5 \times HEP_T$	$HEP = 0.75 \times HEP_T$	$HEP = HEP_T$	$HEP = 1.25 \times HEP_T$	$HEP = 1.5 \times HEP_T$	$HEP = 1.75 \times HEP_T$	$HEP = 2.0 \times HEP_T$
EF = 3								

0.01	Reject H_0 in 100 of 100 experiments	Reject H_0 in 1/100	Reject H_0 in 0/100	Reject H_0 in 0/100	Reject H_0 in 22/100	Reject H_0 in 93/100	Reject H_0 in 100/100	Reject H_0 in 100/100
0.005	100/100	1/100	0/100	0/100	23/100	96/100	100/100	100/100
0.003	100/100	0/100	0/100	0/100	19/100	97/100	100/100	100/100
0.001	100/100	1/100	0/100	0/100	20/100	98/100	100/100	100/100
EF = 10								
0.0001	100/100	79/100	4/100	0/100	0/100	0/100	0/100	0/100

The Bayesian estimator's discrimination increases as m increases; however, the strength of the Bayesian estimator lies in situations in which analysts wish to verify that the error rate is not greater than HEP_T . Using simulated data with $m = 0.5 * (m_{T=0.92})$ for an HEP of 0.01 and an actual error rate twice the HEP (0.02), the Bayesian estimator rejected H_0 in 90% of the 1,000 simulated experiments. Increasing m to $0.7 * (m_{T=0.92})$ yields 100% rejection of the high error rate. Similar results were obtained for the other HEP values.

As shown in Table 10, the Bayesian estimator for error rate 0.0001 and the EF of 10 is not effective. Although the estimator successfully rejects data from error rates that are twice the HEP, it also rejects data with error rates equal to the HEP. After some trial and error, it was found that reducing the range to 75% and increasing m to $2 * (m_{T=0.92})$ yields satisfactory results. Doubling m also improves the response from the median estimator (Table 11).

Table 11: Bayesian estimator and median error rate tests for EF = 10 indicate that the number of observations required for tests with EF = 10 are twice the value for m computed using $\Gamma = 0.92$.

EF = 10	Bayesian Estimator		Median Estimator	
	$m = m_{\Gamma=0.92}$	$m = 2 * m_{\Gamma=0.92}$	$m = m_{\Gamma=0.92}$	$m = 2 * m_{\Gamma=0.92}$
$HEP = 0.25 * HEP_T$	Reject H_0 in 0 of 1000 experiments	Reject H_0 in 0/1000	Reject H_0 in 1000/1000	Reject H_0 in 1000/1000
$HEP = 0.5 * HEP_T$	1/1000	6/1000	838/1000	1000/1000
$HEP = 0.75 * HEP_T$	35/1000	10/1000	97/1000	854/1000
$HEP = HEP_T$	495/1000	74/1000	3/1000	276/1000
$HEP = 1.25 * HEP_T$	934/1000	433/1000	21/1000	189/1000
$HEP = 1.5 * HEP_T$	994/1000	859/1000	165/1000	522/1000
$HEP = 1.75 * HEP_T$	1000/1000	979/1000	457/1000	834/1000

$HEP = 2.0 \times HEP_T$	1000/1000	998/1000	755/1000	966/1000
--------------------------	-----------	----------	----------	----------

4.1.1 The recommended number of opportunities for error, m , per sample

Based on the computer generated data and the discussion from Section 3.3, the recommended sample size for all validation studies is $n = 101$, corresponding to a power of 90%. The recommended number of opportunities for error, m , are listed in Table 12.

Table 12: The recommended number of observations for error types based on HEP values.

HEP	EF = 3 $m = 0.75 \times m_{\Gamma=0.92}$	EF = 5 $m = m_{\Gamma=0.92}$	EF = 10 $m = 2 \times m_{\Gamma=0.92}$
0.0001	6331	8440	16880
0.0005	1267	1688	3376
0.001	634	844	1688
0.002	317	422	844
0.003	211	281	562
0.005	127	169	338
0.006	106	141	282
0.008	80	106	212
0.01	64	84	168
0.05	13	17	34
0.1	7	9	18
0.2	4	4	8
0.3	3	3	6
0.4	2	2	4
0.5	2	2	4
0.9	1	1	2

These values assume the study objective is a conservative validation, seeking to verify that error rate estimates are sufficiently high. Therefore, $m = 0.7 * m_{\Gamma=0.92}$ for EF =3, corresponding to the minimum sample size for the Bayesian estimator. Greater EF values require increased numbers of observations as they correspond to greater expected variance. For EF = 10, $m = 2 * m_{\Gamma=0.92}$ is sufficiently large to confirm that the error rate is not greater than HEP_T . For EF = 5, verifying that the error rate is not greater than HEP_T is satisfied with $m = m_{\Gamma=0.92}$; in order to differentiate between HEP_T and a lower error rate with EF = 5, $m = 2 \times m_{\Gamma=0.92}$ provides sufficient resolution.

4.2 Experimental data

In the pilot study, students completed a procedure that included ten possible types of errors. These error types are listed in Table 13, along with the THERP HEP, the EF, the number of opportunities for the error to be committed, the THERP-predicted number of errors, the number of times the error was observed, and the observed and estimated error rates.

The expected number of errors observed is defined by HEP_T and the number of opportunities for error, m .

$$\langle e \rangle = HEP_T \times m \quad (42)$$

The expected range is specified by the EF:

$$\left[\left(\frac{HEP_T}{EF} \right) \times m, (HEP_T \times EF) \times m \right] \quad (43)$$

These values are listed in the sixth column of Table 13.

Only one THERP HEP was high enough to predict at least one error: THERP Table 20-7, #4, errors of omission. This error type had both a high HEP (0.01) and a high number of observations (627, more than four times the experiment average). For $HEP_T = 0.01$, the predicted number of errors in 627 opportunities for error is six. Given $EF = 3$, THERP predicts that in 90% of observations of 627 opportunities for an error of omission, observed errors will fall between two and eighteen. Nine errors were observed in the pilot experiment.

Seven of the nine remaining error types have such low m values that even the THERP 90th percentile corresponds to a prediction of no observed errors. Six of these seven error types had no observed errors. The seventh had four recorded errors. Researchers observing the experiment believe that this is a knowledge-based error due to operator misunderstanding of how to perform part of the procedure, which explains why the error rate is higher than the THERP predicted rate.

The remaining two error types with observed errors (20-7, #9 and 20-12, #2) exceeded the expected number of errors but fell well within the THERP-predicted range (< 3 and < 2 respectively).

Table 13: Results from the pilot experiment, including the number of errors observed, the expected number of errors for each error type, and the observed error rate. Note: this table treats the collected data as one observation ($n = 1$). Observations types with high m values could be separated into several sets of observations.

THERP Table	HEP	EF	Total observed opportunities for error m	errors observed e	Expected # of errors $HEP_T \times m$ (range specified by EF)	Expected range in # of errors (specified by EF)	Observed Error Rate, e/m
20-7, #4	0.01	3	627	9	6.27	2-18	0.010668
20-7, #9	0.003	3	304	1	0.912	< 3	0.003289
20-10, #1	0.003	3	38	4	0.114	< 1	0.105263
20-10, #2	0.001	3	57	0	0.057	< 1	0
20-10, #9	0.001	3	19	0	0.019	< 1	0
20-12, #2	0.003	3	152	1	0.456	< 2	0.006579
20-12, #8	0.0001	10	152	0	0.0152	< 1	0
20-12, #9	0.001	3	38	0	0.0380	< 1	0
20-12, #10	0.003	3	38	0	0.114	< 1	0
20-12, #11	0.005	3	38	0	0.190	< 1	0

4.2.1 Median estimator analysis

To test the analysis developed above, we consider error type 20-7, #4. For an HEP of 0.01, m_{min} is 85. We can therefore treat the 627 opportunities for error as 7 samples of 89-90 opportunities each. For analysis, we randomly distribute the nine observed errors over seven observations in Table 14.

Table 14: Pilot study data for error type 20-7 #4. Enough observations were made for seven data points. The number of opportunities for error per data point (m), the number of errors observed in each (e), and the observed and estimated error rates are listed.

Observation, i	1	2	3	4	5	6	7
Opportunities for Error, m	89	89	90	90	90	89	90
Observed Errors, e	1	2	3	0	2	1	0
Error Rate, e/m	0.011	0.022	0.033	0.000	0.022	0.011	0.000
Bayesian estimated error rate, $\hat{\theta}$	0.011311	0.013245	0.014245	0.007017	0.013231	0.011311	0.007017

The median error rate is 0.011. Equation 3 determines whether to reject the hypothesis that the true median is HEP_T . Because runs 4 and 7 have no observed errors, we cannot calculate an experimental standard deviation. Therefore, we use the THERP-derived estimate of σ (Equation 5) in the following calculations.

Setting $\bar{x} = \ln 0.011$, $\mu = \ln 0.01$, and $\sigma = \ln 3 / z_{0.95}$ in Equation 3, we do not reject the null hypothesis based on this result.

Recall that the power of the test refers to the probability of rejecting the null hypothesis when the alternative is valid. To assess the power of this result, we use Equation 17. Here, $\xi = \frac{0.011}{0.01} = 1.124$.

$$Power = 1 - \beta^+ = 1 - \Phi\left(z_{1-\frac{\alpha}{2}} - \frac{\ln \xi}{\frac{\sigma}{\sqrt{n}}}\right) = 1 - \Phi\left(2.58 - \frac{0.117}{0.252}\right) = 0.017 \quad (44)$$

As expected, the power is very low, confirming that this initial study requires significant additional data for decisive results. We can only conclude that, based on the initial data, the 0.01 error rate may be valid.

Following the same process for error type 20-7, #9 yields similar results: we do not reject $HEP_T = 0.003$ for this type of error, but as the power of the test (with $n = 1$) is only 0.007, this data provides no confidence in this result.

4.2.2 Bayesian estimators using experimental data

Calculating the Bayesian estimated error rate for error type 20-7, #4 using the seven data points in Table 14 yields an estimated error rate of 0.0107, slightly lower than the median observed error rate of 0.0112.

The standard deviation of the natural log of the error rates for the seven trials is 0.30; σ_T is 0.67. To determine whether σ_T is an adequate estimate of the experimental standard deviation, s , we calculate the Chi Squared test statistic [11].

$$T = (n - 1) \times \left(\frac{s}{\sigma_T} \right)^2 \quad (45)$$

In this case, $T = 1.20$, which corresponds to 2% on the Chi-Squared distribution. As α is set at 0.01, this is an acceptable variance. Replacing σ_T with s in the calculations for $\hat{\Theta}$ yields an estimated error rate of 0.008 rather than 0.0107.

The estimate for error type 20-7, #9 is 0.00338, slightly greater than the observed error rate, 0.00329. As discussed above, the power of this analysis is very low because of the small number of data points ($n = 1$).

4.3 Sample size for validating the 66 THERP HEPs

Using the analysis above, we develop a framework for validating the THERP HEPs. **Let us define one data point as the number of unique opportunities for error that must be observed in order to measure one of the n error rates for all of the HEPs of interest.** In other words, a single data point is the sum of m for all HEPs of interest.

The HEP values for the 66 error types to be assessed are listed in Table 15, along with the recommended m for each HEP. This results in a recommended 107,943 observations per data point for a full-scope validation.

The total number of observations (the sum of $n \times m$ for all error types) is 107,943. This number is somewhat misleading. Note that several error types present themselves concurrently with other potential errors. For example, consider errors of omission. Every time an operator performs a task, he or she may commit an error of omission. Similarly, the opportunity to select the wrong valve or circuit breaker can be observed simultaneously with the opportunity to turn a valve to the wrong position. Taking this into account reduces the total number of observations per data point to 102,402 (reducing the total by approximately five percent). The revised number of observations required per data point is listed in the rightmost column of Table 15. Furthermore, the bulk of the observations required are for very low HEP values: 0.0001, 0.0005 and HEPs deemed “negligible” in the THERP handbook. **Eliminating the 93,029 observations required for error types with error rates that are “negligible” or have an EF of 10 brings the total number of observations required per data point down to 9,373** (rightmost column in Table 15). A sensitivity analysis of the THERP method should be performed to determine the significance of the HEPs that fall into this category.

Table 15: The number of observations required to test the HEPs for the relevant error types listed in the THERP handbook. The “Modified number of observations required for HEP” accounts for error types that will naturally be tested simultaneously—for example, ‘selecting the wrong valve’ will be observed concurrently with ‘moving the valve to the wrong position.’

EF	HEP	# of Errors with HEP	Recommended m	Total # of observations required for HEP	Modified # of observations required for HEP
3	0.001	10	634	6340	5072

	0.002	3	317	951	951
	0.003	10	211	2110	1266
	0.005	3	127	381	254
	0.006	2	106	212	212
	0.008	1	80	80	80
	0.01	6	64	384	320
5	0.001	1	844	844	844
	0.01	3	84	252	252
	0.05	5	17	85	68
	0.1	4	9	36	36
	0.2	1	4	4	4
	0.3	1	3	3	3
	0.5	5	2	10	10
	0.9	1	1	1	1
10	0.0001	1	16880	16880	0
	0.0005	3	3376	10128	0
	0.001	1	1688	1688	0
	0.05	1	34	34	0
	(negligible)	4	16880	67520	0
Totals		66		107,943	9,373

5 Validating stress, experience, and dependence

In addition to validating the HEPs listed in the THERP tables, a full-scale validation must include treatment of performance shaping factors (PSFs) and dependence between steps. Each topic is addressed in turn.

5.1 Validating the THERP performance shaping factors: stress and experience

The primary PSFs addressed in THERP are stress and experience.³ PSF multipliers are applied to HEP values; either the joint HEP will be adjusted or individual HEPs will be adjusted if the PSF only applies to certain HEP values.

THERP specifies four levels of stress: low, optimum, moderately high, and extremely high. Low, optimum, and moderately high stress correspond to light, optimal, and heavy workload; extremely high stress represents accident conditions.

Within each level, the multiplier depends on the type of task (step-by-step or dynamic) and the experience of the operator. Skilled operators have at least six months' experience, while novice operators have less than six months' experience. For dynamic tasks performed under high stress, THERP does not provide a multiplier. Instead, the HEP is assumed to be 0.25 for skilled workers and 0.5 for novice operators. Table 16 lists the THERP multipliers for the various conditions.

³ THERP also addresses three tagging levels, i.e. three approaches to tagging equipment undergoing maintenance, etc. As this is more applicable to operator tasks throughout the plant rather than tasks localized within the control room, we do not address tagging as part of the simulator study. However, similar experiments could be devised to test and validate the impact of tagging schemes on performance.

Table 16: Stress and experience multipliers listed in the THERP Handbook. *bHEP* refers to the base HEP, and *aHEP* refers to the adjusted HEP, that is, the expected HEP under the specified stress condition.

Stress Level	Step-by-step tasks		Dynamic Processes	
	<i>Skilled Operator</i>	<i>Novice Operator</i>	<i>Skilled Operator</i>	<i>Novice Operator</i>
Low Stress (very low task load)	aHEP = 2*bHEP	aHEP = 2*bHEP		
Optimum Stress (Optimum task load)	aHEP = bHEP	aHEP = bHEP	aHEP = bHEP	aHEP = 2*bHEP
Moderately High Stress (Heavy task load)	aHEP = 2*bHEP	aHEP = 4*bHEP	aHEP = 5*bHEP	aHEP = 10*bHEP
Extremely High Stress (Threat stress)	aHEP = 5*bHEP	aHEP = 10*bHEP	aHEP = 0.25	aHEP = 0.5

If an operator has a heavy task load, the adjusted HEP of a step-by-step task with a base HEP of 0.005 will therefore be adjusted to 0.01 for an experienced operator, and 0.02 if the operator is a novice. Under accident conditions, this estimate will increase to 0.025 and 0.05 respectively. For a dynamic task with a nominal HEP of 0.005, the adjusted HEP for an experienced operator would be 0.025 under moderate stress, and 0.25 in high stress conditions.

THERP analysts are encouraged to use a similar approach for any other factors that may impact performance. Multipliers are left to the analyst's discretion; analysts must identify additional PSFs and estimate their impact with limited guidance from the THERP handbook.

These multiplying factors can be tested using the same processes developed to test the base HEPs. For a very low task load, for example, the base HEP is expected to double. Table 17 provides m for testing various multipliers with various base HEPs; n remains the same (101 for a power of 0.9) and, for consistency, we use $m_{\Gamma=0.92}$ for all values.

Table 17: Number of observations (m) required to test all possible HEP multipliers for stress and experience.

	aHEP = bHEP	aHEP = 2*bHEP	aHEP = 4*bHEP	aHEP = 5*bHEP	aHEP = 10*bHEP
bHEP = 0.01	m = 84	m = 42	m = 21	m = 17	m = 9
bHEP = 0.005	m = 169	m = 84	m = 42	m = 34	m = 17
bHEP = 0.003	m = 281	m = 141	m = 70	m = 56	m = 28
bHEP = 0.001	m = 844	m = 422	m = 211	m = 169	m = 84
bHEP = 0.0001	m = 8440	m = 4220	m = 2110	m = 1688	m = 844

As stress and experience are assumed to have a blanket effect, observations could be made using error types with an HEP of at least 0.01 to reduce the number of opportunities for error required to be observed. The assumption is that the bulk of the validation would be performed under optimal workload using experienced operators. Given this assumption, the five additional conditions that must be tested and the number of observations required are listed in Table 18.

Table 18: Conditions to test for stress and experience, along with the number of observations required per sample (m) for each test. Skilled operator for optimum task load is not listed as this condition will be used for the base HEP validations.

Task Load	Operator	HEP	m
Low task load	Skilled Operator	$HEP = 2 \times bHEP$	$m = 43$
	Novice Operator	$HEP = 2 \times bHEP$	$m = 43$
Optimum task load	Novice Operator	$HEP = bHEP$	$m = 85$
Heavy task load	Skilled Operator	$HEP = 2 \times bHEP$	$m = 43$
	Novice Operator	$HEP = 4 \times bHEP$	$m = 22$
			Total = 236

Table 18 shows that 236 additional opportunities for error must be observed per data point to test experience and stress. A controlled research environment provides the opportunity to manipulate an operator's workload and assess the subsequent impact on HEPs. Further studies could manipulate other elements of operator experience to develop a more sophisticated approach to analyzing PSFs, rather than leaving this element of the analysis up to the analyst's discretion.

5.2 Validating the THERP dependence model

A second factor to consider in a full scope THERP validation is the dependence between tasks. THERP employs a positive dependence model, modeling only dependence between steps in which failure in step $j - 1$ increases the probability of failure in step j (or success in step $j - 1$ increases the probability of success in step j). Negative dependence (failure in $j - 1$ increases the probability of success in j) is not modeled, as this is believed to reflect an overall conservative approach.

THERP specifies five levels of dependence between steps: zero dependence (ZD), low dependence (LD), medium dependence (MD), high dependence (HD) and complete dependence (CD). These levels are defined by the equations in Table 19, which are designed to be relatively insensitive to variations in HEP values less than 0.01.

Table 19: The THERP equations for the conditional probabilities associated with each level of dependence between steps or tasks, and the calculated conditional probabilities for five sample HEP values.

Dependence Level	Equation for the probability of failure in step j given failure in step $j - 1$ with the specified dependence	Dependent HEP Value				
		0.01	0.005	0.003	0.001	0.0001
Zero Dependence (ZD)	$P(F_j F_{j-1} ZD) = HEP_T$	0.01	0.005	0.003	0.001	0.0001
Low Dependence (LD)	$P(F_j F_{j-1} LD) = \frac{1 + 19 * HEP_T}{20}$	0.059	0.055	0.053	0.051	0.050
Medium Dependence (MD)	$P(F_j F_{j-1} MD) = \frac{1 + 6 * HEP_T}{7}$	0.151	0.147	0.145	0.1437	0.143

High Dependence (HD)	$P(F_j F_{j-1} HD) = \frac{1 + HEP_T}{2}$	0.505	0.503	0.502	0.500	0.500
Complete Dependence (CD)	$P(F_j F_{j-1} CD) = 1$	1	1	1	1	1

For low HEPs, likelihoods of failure associated with each level of dependence are distributed using a somewhat logarithmic scale, with conditional probabilities corresponding to approximately 0.05 for LD, 0.15 for MD, 0.50 for HD and 1.0 for CD (for ZD, the conditional probability is simply the base HEP). The insensitivity of the conditional probability to the base HEP simplifies the matter of testing the validity of the dependence relationships, as errors in any steps that share the same level of dependence on a previous step can be assessed as a group. However, to test this assertion, base HEP values should be evenly distributed in the dependence test groups so that no single error type or HEP value dominates the experiment. HEP values greater than 0.01 must be addressed separately; this is discussed in Section 5.2.2.

5.2.1 Sample size calculations for Dependence with low HEPs ($HEP \leq 0.01$)

We specify a range for each level of dependence using the midpoints between the natural logarithms of the conditional probabilities. The range is listed in Table 20. To determine the validity of the conditional probability, we employ the approach used in other sample size calculations. Here, D is the THERP-specified conditional probability and μ_D is the natural logarithm of D . μ_D^+ is the maximum of the range for that level of dependence, and μ_D^- is the minimum of the range.

The null and alternative hypotheses are therefore

$$\begin{cases} H_0: \mu = \mu_D \\ H_a^+: \mu > \mu_D^+ \\ H_a^-: \mu < \mu_D^- \end{cases}$$

Here, β^+ and β^- are not necessarily symmetrical. They are calculated using the critical values c^+ and c^- , where

$$\left\{ c^+ = z_{1-\frac{\alpha}{2}} * \left(\frac{\sigma}{\sqrt{n}} \right) + \mu_D, \quad c^- = z_{\frac{\alpha}{2}} * \left(\frac{\sigma}{\sqrt{n}} \right) + \mu_D \right\} \quad (46)$$

β^+ and β^- are therefore

$$\left\{ \beta^+ = \Phi \left(\frac{c^+ - \mu_D^+}{\frac{\sigma}{\sqrt{n}}} \right), \quad \beta^- = 1 - \Phi \left(\frac{c^- - \mu_D^-}{\frac{\sigma}{\sqrt{n}}} \right) \right\} \quad (47)$$

Because the range for each level of dependence is not symmetrical, we use the greater value of β to determine sample size. This is β^- for all dependence levels except MD and ZD (which has no corresponding β^-).

Setting $\alpha = 0.01$, we calculate n for each level of dependence with power set to 0.9 and 0.99. Again, σ is estimated using Equation 5 and m is $m_{\Gamma=0.92}$. As expected from previous sample size calculations, m decreases as n and D increase.

Table 20: Sample size (n) and number of observations per sample (m) for each level of dependence with the statistical powers of 0.9 and 0.99 ($\beta = 0.1$ and $\beta = 0.01$).

Dependence	ZD	LD	MD	HD	CD
	aHEP = 0.01 Range: $\mu_{ZD}^- = 0.00$ $\mu_{ZD}^+ = 0.02$	aHEP = 0.05 Range: $\mu_{LD}^- = 0.02$ $\mu_{LD}^+ = 0.11$	aHEP = 0.25 Range: $\mu_{MD}^- = 0.11$ $\mu_{MD}^+ = 0.35$	aHEP = 0.50 Range: $\mu_{HD}^- = 0.35$ $\mu_{HD}^+ = 0.70$	aHEP = 1.0 Range: $\mu_{CD}^- = 0.70$ $\mu_{CD}^+ = 1$
Power = 0.9	$n_{ZD} = 11$	$n_{LD} = 11$	$n_{MD} = 56$	$n_{HD} = 56$	$n_{CD} = 56$
	$p_{ZD} = 0.7$	$p_{LD} = 0.7$	$p_{MD} = 0.605$	$p_{HD} = 0.605$	$p_{CD} = 0.605$
	$m_{ZD} = 120$	$m_{LD} = 24$	$m_{MD} = 4$	$m_{HD} = 2$	$m_{CD} = 1$
Power = 0.99	$n_{ZD} = 17$	$n_{LD} = 17$	$n_{MD} = 90$	$n_{HD} = 90$	$n_{CD} = 90$
	$p_{ZD} = 0.665$	$p_{LD} = 0.665$	$p_{MD} = 0.58$	$p_{HD} = 0.58$	$p_{CD} = 0.58$
	$m_{ZD} = 109$	$m_{LD} = 22$	$m_{MD} = 4$	$m_{HD} = 2$	$m_{CD} = 1$

This is all fairly straightforward. Given a procedure with 4 steps that have a medium dependence on the preceding step, observing that procedure being completed 90 times should provide the data necessary to ascertain whether 0.25 is an appropriate value for the conditional probability for steps with medium dependence.

The bulk of the HEP validations are assumed to be conducted using procedure steps with zero dependence on the previous step, leaving only LD, MD, HD and CD to be tested separately. **As the maximum n_D is 56, dependence data only needs to be collected in 56 of the 101 data points.** The total number of additional opportunities for error that must be observed, m_D , is the sum of the number of observations required at each dependence level:

$$m_{D,HEP<0.01} = m_{LD} + m_{MD} + m_{HD} + m_{CD} = 24 + 4 + 2 + 1 = 31$$

for a power of 0.9. In other words, in order **to test the THERP dependence model for low HEPs, an additional 31 opportunities for error must be observed in 56 of the 101 data points collected.**

5.2.2 Sample size calculations for Dependence with high HEPs ($HEP > 0.01$)

For HEPs greater than 0.01, the adjusted HEP depends on the base HEP. Table 21 shows the conditional probabilities for each level of dependence, as well as the calculated m_D for each level of dependence and the associated values for μ^+ and μ^- .

With such high probabilities, n decreases significantly; $n = 5$ is sufficient to test any dependence level in Table 21. Revising Equation 27 for $n = 5$, we find $m_{\Gamma=0.92}$ corresponds to $p = 0.78$; solving for m in equation with $p = 0.78$ yields the values for m in Table 21.

Table 21: Conditional probabilities for HEPs greater than 0.01, along with the number of observations per sample (m) for each level of dependence with the statistical power of 0.9 ($\beta = 0.1$).

HEP	0.05	0.1	0.2	0.3	0.5	0.9
ZD	0.05	0.10	0.20	0.30	0.50	0.90
m_{ZD}	30	15	7	5	3	1
<i>$[\mu_{ZD}^-: \text{no value for } \mu_{LD}^- \text{ as any low HEP is categorized as ZD}]$</i>						
μ_{ZD}^+	0.07	0.12	0.22	0.32	0.51	0.90
LD	0.10	0.15	0.24	0.34	0.53	0.91
m_{LD}	15	10	6	4	3	1
μ_{LD}^-	0.07	0.12	0.22	0.32	0.51	0.90
μ_{LD}^+	0.13	0.18	0.27	0.37	0.55	0.91
MD	0.19	0.23	0.31	0.40	0.57	0.91
m_{MD}	8	6	5	3	2	1
μ_{MD}^-	0.13	0.18	0.27	0.37	0.55	0.91
μ_{MD}^+	0.31	0.35	0.43	0.51	0.65	0.93
HD	0.53	0.55	0.60	0.65	0.75	0.95
m_{HD}	3	2	2	2	2	1
μ_{HD}^-	0.31	0.35	0.43	0.51	0.65	0.93
μ_{HD}^+	0.72	0.74	0.77	0.81	0.87	0.97
CD	1.00	1.00	1.00	1.00	1.00	1.00
m_{CD}	1	1	1	1	1	1
μ_{CD}^-	0.72	0.74	0.77	0.81	0.87	0.97
<i>$[\mu_{CD}^+: \text{no value for } \mu_{CD}^+ \text{ as any high HEP is categorized as CD}]$</i>						
Total m_D	57	34	21	15	11	5

The total number of additional observations required to validate dependence for the HEPs that are greater than 0.01 is 143, the sum of m_D for each high HEP:

$$m_{D,HEP>0.01} = 57 + 34 + 21 + 15 + 11 + 5 = 143$$

Remember that these 143 opportunities for error must only be observed in 5 of the 101 data points collected.

5.2.3 Assigning dependence: analyst agreement

The real challenge in validating dependence lies in determining (or assigning) the level of dependence between steps in a procedure or operation. This is determined by the analyst assessing the procedure, and it touches on yet another aspect of validation: consistency in how the method is applied.

One approach to ensuring analyst consensus is to measure the inter-rater reliability between analysts, typically using Cohen's Kappa to assess the unity of the analyses. The only steps in a procedure that should be included in a dependence validity test must have a high IRR rating—typically 0.61 or greater [9]. We recommend obtaining assessments from at least three experienced professionals and using only those procedures with a high IRR for the validity test described above.

Significant discrepancies between assessments might be indicative of a need to revise the THERP dependence paradigm. For example, if some assessors determine there is low dependence between a pair of procedure steps, while other assessors assign medium dependence, perhaps there should be a “medium-low dependence” category in the method. A sensitivity analysis to investigate the impact and utility of such a change would provide additional insight into this modification. Zio's approach [14] might be useful here.

Cohen's Kappa also could be used to evaluate the consensus between analysts on other aspects of the THERP method. Further discussion of this aspect of the validation project is outside the scope of this paper.

6 Conclusions

From Table 15 we estimate the number of observations for a single data point is 9,373, excluding PSFs and dependence.

From Section 5.2, we add 31 observations per data point to validate dependence for low HEPs for 56 data points, and 143 observations per data point for five data points.

Evaluating stress and experience adds 236 data points (see Table 18).

The total number of observations required for each of the 101 data points is therefore the 9,373 required to test the bulk of the error types listed in the THERP tables, plus an additional 236 observations to assess stress and experience, and finally 31 observations for low HEPs and 143 observations for high HEPs. **This yields a maximum of 9,783 observed opportunities for error per data point in the study described above.**

Once these data are obtained, the average error rate, $\bar{x}$, can be estimated using either the median error rate estimator or by averaging the Bayesian estimated error rates for each set of observations. The estimator selected will depend on the objective of the study and previous data that have been collected.

The challenges in data collection will include (1) identifying appropriate opportunities for error within existing, familiar procedures, (2) obtaining consensus on the dependence of these error opportunities, and (3) determining task load. Most of the data obtained must be collected from skilled operators working under optimum task load conditions and performing procedure steps that have zero dependence on previous steps. A high inter-rater reliability concerning the THERP analysis of the selected procedures must be obtained before the data collection can begin [4].

Data collection must be performed in high-fidelity simulators, with consideration for the biases discussed in Section 2.2. A robust method for observing and recording an operator's success or failure for each error opportunity must be developed.

With these factors taken into consideration, a full- or partial-scale validation study is feasible. Given the 100 reactors operating in the United States today, if each plant were to perform one or two data collection efforts, the 101 samples required for a statistically significant analysis would be obtained.

Assuming ten opportunities can be observed in one minute (6 seconds per task), one set of 9,783 observations could be collected in approximately sixteen hours. If each reactor were to dedicate twenty operator hours to the validation effort, a full-scale validation study would be complete.

Before attempting to validate all of the aspects of the THERP method, a more useful first step would be a study validating the HEPs for a few key error types in conjunction with THERP's treatment of stress and dependence. For example, consider a study limited to validating the HEP associated with making a simple arithmetic miscalculation (THERP Table 20-10, #10). HEP_T for this error type is 0.01 with an EF of 3, and recommended m of 64. This single task could be used as a test for stress by observing expert operators making 64 calculations at the three task loads, for a total of 192 observations per data point. This study would provide a baseline for assessing THERP's PSF approach. Similarly, an additional 31 observations with various levels of dependence on some previous step would provide data for a statistically significant test of THERP's dependence model. **Obtaining 101 observations of 223 opportunities for error would provide significant insight into the THERP model with approximately two percent of the effort needed to validate the entire model. Again estimating ten observed opportunities for error per minute, all 101 sets of observed error rates could be collected in one dedicated, forty hour work week.** Such an experiment would demonstrate the validity of this approach, provide an initial assessment of two of the fundamental mechanisms of the THERP model, and guide further THERP validation work.

Acknowledgements

This research is being performed using funding received from the DOE Office of Nuclear Energy's Nuclear Energy University Programs. The pilot experiment was conducted as part of the HRA class developed at The Ohio State University through an educational grant from the NRC.

References

- [1] R. Boring, "Issues in benchmarking human reliability analysis methods: a literature review," *Reliability Engineering and System Safety*, vol. 95, pp. 591-605, 2010.
- [2] B. Kirwan, "The validation of three human reliability quantification techniques - THERP, HEART and JHEDI: Part I - technique descriptions and validation issues," *Applied Ergonomics*, vol. 27, no. 6, pp. 359-73, 1996.
- [3] B. Kirwan, "The validation of three Human Reliability Quantification techniques - THERP, HEART and JHEDI: Part II--Results of validation exercise," *Applied Ergonomics*, vol. 28, no. 1, pp. 17-25, 1997.
- [4] B. Kirwan, "The validation of three human reliability quantification techniques - THERP, HEART and JHEDI: part iii - Practical aspects of the usage of the techniques," *Applied Ergonomics*, vol. 28, no. 1, pp. 27-39, 1997.
- [5] S. Massaiu, A. Bye, P. O. Braarud, H. Broberg, M. Hildebrandt, V. D. Dang, E. Lois and J. A. Forester, "International HRA Empirical Study, Overall Methodolgy, and HAMMLAB Results," in *Simulator-Based Human Factors Studies Across 25 Years*, London, Springer-Verlag, 2011, p. 253.
- [6] A. D. Swain and H. E. Guttman, "Handbook of Human Reliability Analysis with Emphasis on Nuclear Power Plant Applications Final Report," United States Department of Energy, Albuquerque, NM, 1983.
- [7] A. Gupta, "Simulator biases," in *Development of Boiling Water Reactor Nuclear Power Plant Simulator for Human Reliability Analysis Education and Research*, Columbus, OH: Ohio State University Master's Thesis, 2013, p. 122.
- [8] A. Gupta, R. Benish, B. Hajek and C. Smidts, "Hands-on HRA: Developing a Human Reliability Course," in *Transactions of the American Nuclear Society*, 107, San Diego, CA, 2012.
- [9] R. T. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, vol. 33, no. 1, pp. 159-174, Mar 1997.
- [10] R. Benish, M. Li, A. Gupta and C. Smidts, "Validating THERP: An Approach to Experimentally Validating the Human Error Prediction Rates in the THERP Tables," in *Transactions of the American Nuclear Society*, 107, San Diego, CA, 2012.
- [11] "NIST/SEMATECH e-Handbook of Statistical Methods," April 2012. [Online]. Available: <http://www.itl.nist.gov/div898/handbook/eda/section3/eda358.htm>. [Accessed 28 May 2014].
- [12] D. W. Stroock, *Probability Theory: An Analytic View*, New York, NY: Cambridge University Press, 1999.

- [13] D. Poirier, Intermediate Statistics and Econometrics, MIT Press, 1995.
- [14] E. Zio, P. Baraldi, M. Librizzi, L. Podofillini and V. N. Deng, "A fuzzy set-based approach for modeling dependence among human errors," *Fuzzy sets and systems*, vol. 160, pp. 1947-64, 2009.