DISS. ETH NO: 27813

A Bayesian integration of simulator data and judgment to develop empirically-based reference values for human reliability

A thesis submitted to attain the degree of
DOCTOR OF SCIENCES of ETH Zürich
(Dr. Sc. ETH Zürich)

presented by
SALVATORE FRANCESCO GRECO

Laurea Magistrale in Ingegneria Energetica e Nucleare, Università di Palermo

born on 22.08.1988
citizen of Italy

accepted on the recommendation of
Prof. Dr. Horst-Michael Prasser
Prof. Dr. Marloes H. Maathuis, Prof. Dr. Katrina M. Groth,
Prof. Dr. Carol Smidts, Dr. Luca Podofillini

2021

To “Nonno Saro” and Letizia,
in loving memory

Acknowledgments

This thesis summarizes my Ph.D. work performed at Risk and Human Reliability group of Paul Scherrer Institut from November 2016 to October 2020. First and foremost, I would like to thank Prof. Horst-Michael Prasser, my thesis advisor at ETH Zürich, for providing his valuable guidance and regularly monitoring my progress throughout my Ph.D. studies. It has been an honor being one of his last Ph.D. students at D-MAVT: his teachings and infinite knowledge have been a precious source of inspiration for my research work.

This dissertation would not have been possible without the invaluable help of Dr. Luca Podofillini, my direct supervisor at PSI, who played a central role during my doctoral studies and provided a significant contribution to both my work and personal growth. Luca took me under his wing since my internship at Risk and Human Reliability group in 2014 and, with his huge experience, has taught me everything I know about Human Reliability and Probabilistic Safety Assessment. He has always encouraged me to think outside-the-box and approach research challenges from multiple perspectives, interpret my results with a critical eye, and systematically accomplish my tasks with the highest degree of accuracy possible. I thank him for the enormous amount of time and energy he dedicated to the supervision of my work and publications, for having believed in my skills and supported me since the beginning of my Ph.D. studies (and lastly, for having shown a Sicilian guy that running in the snowy Swiss forests below zero degrees Celsius is possible). In his twofold guise of supervisor and friend, he made me a better researcher and person: for this and for all I learnt from him, I will be forever grateful.

I would like to thank Dr. Vinh N. Dang, my supervisor and head of the Risk and Human Reliability group, for giving me the opportunity to join his group and for all the precious comments and feedback he gave me during our meetings. He helped me a lot in refining the research objectives and tasks of my Ph.D. work, and his meticulous revisions significantly improved the quality of my Ph.D. deliverables.

My special thanks go to my trusted friends and colleagues Dr. Dhruv Pandya, Dr. Matteo Spada and Lidia Stermieri, whose constant presence during the Ph.D. life has been of fundamental importance for me to go all the way. Dhruv has been the best office mate one could ever have: since my arrival at PSI, he has always supported me with his wise advices and mentorship, and assisted me in quickly adapting to the new Ph.D. dimension in Switzerland. Matteo, with his outstanding scientific background and positive attitude, has been a role model to me during my time at PSI: he has always been available in helping with modelling problems and clearing my doubts about Bayesian data analysis and Monte Carlo Markov Chain methods. Lidia, with her contagious enthusiasm and emotional support, has given me countless advices and always

shown her faith in me. I am very grateful to them for this and for all the memorable (and de-stressing) social moments we had in Zürich in the past years.

I would also like to thank Prof. Mariarosa Giardina and Dr. Pietro Buffa, respectively my former thesis advisor and supervisor at University of Palermo, for having introduced me to the world of academic research. They both played a fundamental role in my professional development during my bachelor and master studies: without their teachings and stimuli, most likely my Ph.D. adventure would never have begun.

My thanks go also to all my former colleagues at Laboratory for Energy Systems Analysis whom I had the privilege to interact with since my internship in 2014, in particular: Dr. Calvin Whealton, Dr. Miltiadis Kyriakidis, Dr. Slavka Prvakova, Dr. Lusine Mkrtchyan, Dr. Edin Alijagic, and Dr. Durga Rao Karanki. They all welcomed me with open arms at PSI and offered me their unconditional assistance during my stay at PSI.

I would like to express my deepest gratitude to my family, especially my parents, who funded my school and university years and provided me with the right tools to build a brighter future. They have always motivated me to never give up and maintain a healthy work-life balance: for this and for all the sacrifices they made for me, I will be forever grateful.

Finally, a special mention goes to my partner Chiara. She has always been there for me in good times and bad, and provided me the mental strength to overcome the hurdles along the Ph.D. journey. She is the most important person in my life and spiritual guide, therefore I dedicate my dissertation work to her.

Funding Acknowledgment

This work was funded by the Swiss Federal Nuclear Safety Inspectorate (ENSI), under contract CTR00690. The views expressed in this work are solely those of the authors.

Abstract

As part of Probabilistic Safety Assessment (PSA), Human Reliability Analysis (HRA) addresses the contribution of human failures to risk in complex technical systems, e.g. nuclear power plants, chemical and aerospace systems. HRA methods support in the identification of the safety-critical tasks performed by the personnel, in the characterization of the contextual factors influencing crew performances, and finally in the assessment of the task failure probabilities (referred as Human Error Probabilities, HEPs).

In most HRA methods, HEP estimation is supported by quantitative models that represent both the operational tasks and the contextual factors via categories (typically, task types and Performance Shaping Factors, PSFs), and relate these categories to HEP values. Reference HEP values and bounds for the task and PSF categories are needed to parametrize a method's quantification model. Since the early developments of HRA models, the data underlying the reference HEP values is generally obtained by combining empirical data from different sources (e.g. licensee event reports, human factors experiments, training in control room simulators) with judgment elicited from domain experts (e.g. as quantitative probability estimates, or qualitative rankings). Due to the general lack of data, its diversity and its often uncertain quality, there is lack of traceability in the aggregation of the various data sources, as well as in their combination with expert judgment. As a result, it is now difficult to determine to what extent the HEP values produced are empirically based. Also, as new data would become available, it is not clear how to incorporate it as new evidence, to progressively obtain solid HEP values. As such, traceable data aggregation models need to be developed to accommodate new evidence and to ensure that in the long-term the empirical basis of HRA models will be strengthened.

In recent years, the collection of human performance data from plant simulators is receiving renewed attention. Recent efforts by HRA community addressed the development and application of data collection protocols and the analysis of the first batches of data. However, gaps remain on how to use this information to quantify HEP values and bounds, and how to eventually incorporate them into HRA models. Framed within this research direction, the present Ph.D. work aims at developing new quantitative models, based on Bayesian statistical methods, traceably integrating simulator data and expert judgment in the production of reference HEP values and bounds for HRA methods' task type and PSF categories.

The Ph.D. work focuses in particular on the following three research gaps in HRA literature: first, as data is collected over different plants, operating crews and in different time spans, characterization and treatment of data variability is required for appropriate statistical inference; second, to further understand the sources of crew performance variability, a structured methodology is needed to identify crew performance drivers from simulator data, and empirically incorporate their effects on the HEP estimates; third, the need for a traceable and transparent integration of expert judgment whenever the latter is used in the development of the reference HEP values.

Addressing the first research gap, the thesis develops a Bayesian variability model able to formally capture the multiple sources of variability (within-category, crew-to-crew) in simulator data, and estimate failure probabilities (with their variability and uncertainty distributions) for various combinations (“constellations”) of task type and PSF categories of existing data collection taxonomies. For the given constellation, the variability model mathematically represents HEP variability stemming from differences in tasks, scenarios, plants and crew characteristics via continuous parametric distributions: in this formulation, the model can be flexibly adapted to address specific variability aspects (e.g. plant-to-plant, scenario-to-scenario, crew-to-crew) according to data availability and the scope of the application. The variability model is first verified on artificially-generated data and then applied to a case study with simulator datasets from literature, showing the effects of modelling variability in HEP estimates to avoid potential overconfidence and biases.

Addressing the second research gap, the Ph.D. work specifically focuses on modelling the crew-to-crew variability component from crew behavioral characteristics manifested in simulator studies. To this end, the thesis introduces the concept of behavioral patterns to categorize the spectrum of crew behavioral characteristics (e.g. in team decision-making communication strategies, adherence to procedures) for a given constellation of task and PSF categories, and represent performance variability over a finite (“discrete”) set of crew behavioral groups. The discrete formulation with behavioral patterns is included in a new Bayesian hierarchical model, to quantitatively capture performance variability across crew behavioral groups and provided with a multi-step methodology, to support the identification of behavioral patterns from data. Both the multi-step methodology and the Bayesian hierarchical model are applied to a case study involving different emergency scenarios from recent simulator studies. Besides demonstrating their feasibility on a practical HRA application, the numerical application shows the effects of incorporating behavioral characteristics of operating crews in HEP estimates, compared to alternative quantitative approaches for simulator data.

Lastly, the thesis addresses how mathematically integrate expert judgment (in the form of quantitative estimates of task failure probability) in an upgraded formulation of the Bayesian variability model for simulator data. This new formulation of the variability model is used as the basis for the development of a Bayesian two-stage model, with the goal to improve the estimation of plant-specific task failure probabilities in presence of limited empirical data. The developed two-stage model is first verified with artificially-generated evidence, to analyze the effects of judgment incorporation on HEP estimates and investigate model sensitivity to biases in expert judgment. Then, the model is applied to a collection of human failure events from the recent HRA Empirical Studies, to show its potential for use in plant-specific PSA applications.

Sommario

Nell’ambito della valutazione probabilistica della sicurezza (Probabilistic Safety Assessment, PSA), l'analisi dell'affidabilità umana (Human Reliability Analysis, HRA) studia il contributo dell’errore umano nel quadro dei rischi associati ai sistemi tecnologici complessi, tra i quali centrali nucleari, sistemi chimici e aerospaziali. I metodi HRA supportano l'identificazione di tasks operazionali critiche in termini di sicurezza, la caratterizzazione dei fattori contestuali che ne influenzano la performance e infine la valutazione della probabilità di errore umano (Human Error Probability, HEP) associata.

Nella maggior parte dei metodi HRA, la stima degli HEPs è supportata da modelli quantitativi i quali rappresentano le tasks degli operatori e i relativi fattori contestuali tramite elementi categorici (tipicamente, tipologie di tasks e i cosiddetti Performance Shaping Factors, PSFs), relazionando tali categorie a valori di HEP. Al fine di parametrizzare il modello quantitativo alla base di un determinato metodo HRA, è necessario ricavare valori di HEP di riferimento per le rispettive categorie di tasks e PSFs. Sin dai primi sviluppi dei modelli HRA, tali valori di riferimento sono stati storicamente ottenuti combinando dati empirici provenienti da diverse fonti (ad es. reports di eventi, studi su fattori umani, simulatori della sala controllo principale) e il giudizio di esperti del settore (ad es. in forma di stime dirette di probabilità d’errore, o di scale qualitative). A causa della diffusa scarsità di dati empirici, nonché della loro diversa natura e talvolta incerta qualità, l'aggregazione delle varie fonti di dati e la loro integrazione con il giudizio degli esperti sono caratterizzate da una generale mancanza di tracciabilità. Di conseguenza, risulta adesso difficile determinare fino a che punto i valori di HEP prodotti dai modelli HRA si fondino su basi empiriche. Inoltre, con la progressiva disponibilità di nuovi dati, non è altrettanto chiaro come incorporare nuova evidenza nei modelli al fine di attualizzare i valori di HEP. Pertanto, è necessario sviluppare nuovi strumenti quantitativi che consentano di trattare ed aggregare dati in modo tracciabile al fine di garantire, nel lungo termine, un graduale miglioramento della base empirica dei modelli HRA.

Il rinnovato interesse nei confronti della raccolta dati nei simulatori di impianti nucleari ha recentemente contribuito alla nascita di diverse iniziative internazionali. Negli ultimi anni, gli sforzi della comunità HRA si sono concentrati sullo sviluppo di nuovi protocolli di raccolta e catalogazione e sull'analisi dei primi pacchetti dati disponibili. Tuttavia, rimangono ancora lacune nell’utilizzo di tali dati per la quantificazione di HEPs e nella loro incorporazione nei modelli HRA. Il presente lavoro di dottorato si inquadra in tale ambito di ricerca con l’obiettivo di sviluppare nuovi modelli quantitativi, basati su metodi statistici bayesiani, in grado di integrare, in modo sistematico e tracciabile, dati empirici provenienti dai simulatori e giudizio degli esperti nella produzione di valori e intervalli di HEP di riferimento per le tipologie di tasks e categorie di PSFs dei metodi HRA.

In particolare, il lavoro di dottorato si concentra su tre lacune in letteratura HRA: in primo luogo, dal momento che i dati vengono raccolti da diversi impianti, operatori e periodi temporali, occorre caratterizzare e trattare la variabilità nei dati in modo appropriato per una corretta inferenza statistica; in secondo luogo, è necessario ricorrere a metodologie strutturate in grado di identificare i fattori principali di variabilità nelle performances degli operatori e incorporare empiricamente i loro effetti nelle stime di HEP; in terzo luogo, ogniqualvolta il giudizio degli esperti venga impiegato nel processo di quantificazione degli HEPs, è indispensabile che esso venga integrato in modo trasparente e distinguibile dai dati empirici.

Nell'affrontare il primo obiettivo sopra citato, la tesi ha sviluppato un modello di variabilità bayesiano in grado di trattare formalmente le varie sorgenti di variabilità (all'interno delle categorie di tasks e PSFs; tra diversi operatori) nei dati provenienti dai simulatori, e di stimare HEPs (con le relative distribuzioni di variabilità e incertezza) per le varie combinazioni ("costellazioni") di tipologie di tasks e categorie di PSFs impiegate dalle tassonomie di raccolta dati esistenti. Per data costellazione, il modello rappresenta la variabilità nella probabilità di errore (legata a differenze tra tasks, scenari, impianti e caratteristiche degli operatori) tramite distribuzioni parametriche continue, con una formulazione matematica adattabile alla tipologia di dati a disposizione e allo scopo dell'applicazione. Il modello di variabilità è stato dapprima verificato con dati artificiali e successivamente applicato a un caso studio tratto dalla letteratura, allo scopo di mostrare gli effetti del trattamento della variabilità sui valori di HEP stimati.

In linea col secondo obiettivo, il lavoro di dottorato si è successivamente concentrato sul modellare esplicitamente la componente di variabilità legata alle caratteristiche comportamentali degli operatori. A tal scopo, la tesi ha introdotto il concetto di patterns comportamentali per classificare lo spettro di caratteristiche comportamentali degli operatori (ad es. nelle strategie di comunicazione, nell'aderenza alle procedure) e rappresentare, per una data costellazione di categorie, la variabilità nelle performances degli operatori tramite un set finito ("discreto") di gruppi comportamentali. Tale formulazione discreta con patterns comportamentali è stata implementata in un nuovo modello gerarchico bayesiano per catturare quantitativamente la variabilità nelle performances dei vari gruppi di operatori, e integrata in una metodologia multi-step in grado di supportare l'identificazione di pattern comportamentali dai dati da simulatore. Al fine di dimostrarne la fattibilità pratica, la metodologia multi-step e il modello gerarchico bayesiano sono stati applicati a un caso studio comprendente diversi scenari di emergenza simulati in studi di letteratura recenti. L'applicazione numerica ha messo in evidenza gli effetti delle diverse caratteristiche comportamentali degli operatori nelle distribuzioni di variabilità di HEP stimate dal modello, confrontando i risultati ottenuti con modelli quantitativi alternativi presenti in letteratura.

Infine, rispondendo al terzo obiettivo, la tesi ha affrontato l'integrazione matematica del giudizio degli esperti (in forma di stime quantitative di probabilità d'errore) nel modello di

variabilità “continuo” originario. La nuova formulazione matematica è stata adottata come base per lo sviluppo di un modello bayesiano a due stadi, con l’obiettivo di migliorare la stima di HEPs per Human Failure Events (HFEs) caratterizzati da scarsità di dati empirici. Il modello a due stadi è stato inizialmente testato con evidenza artificiale, al fine di analizzare l’influenza del giudizio degli esperti sulle distribuzioni di HEP stimate e investigare la sensibilità del modello ad eventuali biases presenti tra gli esperti. Successivamente, a scopo dimostrativo, il modello a due stadi è stato applicato a un problema PSA pratico, per quantificare gli HEPs di una collezione di HFEs tratti dai recenti HRA Empirical Studies.

List of Abbreviations

Abbreviation	**Full form**
ATHEANA	A Technique for Human Error Analysis
BBN	Bayesian Belief Network
CBDT	Cause-Based Decision Tree
CD	Core Damage
CNI	Constrained Non Informative
CPD	Conditional Probability Distributions
CPT	Conditional Probability Table
CREAM	Cognitive Reliability and Error Analysis Method
HAMMLAB	Halden Man-Machine Laboratory
HCR	Human Cognitive Reliability
HEART	Human Error Assessment and Reduction Technique
HEP	Human Error Probability
HFE	Human Failure Event
HMI	Human-Machine Interface
HuREX	HUman Reliability data EXtraction
HRA	Human Reliability Analysis
IAEA	International Atomic Energy Agency
JAGS	Just Another Gibbs Sampler
LOFW	Loss Of Feed Water
MCMC	Monte Carlo Markov Chain
MERMOS	Methode d'Evaluation de la Realisation des Missions Operateur pour la Surete
ORE	Operator Reliability Experiment
PRA	Probabilistic Risk Assessment
PSA	Probabilistic Safety Assessment
PSF	Performance Shaping Factor
PV	Population Variability
PVC	Population Variability Curve
RCS	Reactor Coolant System
SACADA	Scenario Authoring, Characterization, And Debriefing Application
SLOCA	Small break Loss Of Coolant Accident
SPAR-H	Standardized Plant Analysis Risk–Human reliability
SGTR	Steam Generator Tube Rupture
THERP	Technique for Human Error Rate Prediction

Table of Contents

Acknowledgments III

Abstract / Sommario V

List of Abbreviations XI

List of Tables XVII

List of Figures XXI

Chapter 1: Introduction 1

- 1.1: Background and work motivation 1
  - 1.1.1: Modelling variability within the constellations of taxonomy categories ... 4
  - 1.1.2: Lack of a methodology to incorporate crew behavioral characteristics in HRA models 5
  - 1.1.3: Limited traceability of judgment incorporation in existing HRA methods 6
- 1.2: Research goals and objectives 8
- 1.3: Key contribution of the research work 13
  - 1.3.1: First-of-a-kind quantitative model to treat data variability “within the constellation” of task and PSF categories 13
  - 1.3.2: A (model-based) novel approach to identify crew performance drivers from simulator data and empirically include their effects in HEP estimates 14
  - 1.3.3: A (multi-purpose) quantitative framework to systematically and traceably integrate data and judgment in HEP estimation 15
  - 1.3.4: Applicability to other sectors 16
- 1.4: Outline of the thesis 17
- References 19

Chapter 2: Bayesian variability model for simulator data 23

- 2.1: Introduction 25
- 2.2: Uncertainty and variability aspects in HRA and simulator data for HRA 28
  - 2.2.1: Treatment of uncertainty and variability in existing HRA methods 28

2.2.2: Characterization of uncertainty and variability aspects in simulator data for HRA 29
2.3: A Bayesian variability model for simulator data 33
2.3.1: Modelling assumptions 34
2.3.2: Variability model for HEP 35
2.3.3: Development of the Bayesian inference model 37
2.3.4: Use of lognormal probability density functions to represent variability.. 42
2.4: Numerical application 43
2.4.1: Variability model vs lumped-data approach 44
2.4.2: Sensitivity to available data 46
2.4.3: Application to real simulator data from literature 54
2.5: Discussion 57
2.6: Conclusions 58
References 60

Chapter 3: Behavioral patterns to model crew performance variability 65
3.1: Introduction 66
3.2: Concepts: behavioral patterns from simulator data and variability modelling .. 69
3.2.1: Behavioral patterns: definition and relationship with typical HRA quantification 69
3.2.2: Using behavioral patterns in a variability model for HEP 74
3.3: A methodology to incorporate crew behavioral patterns in HEP quantification 76
3.3.1: Derivation of behavioral categories from data collection 77
3.3.2: Grouping crew performance data and HEP quantification 79
3.3.3: Development and implementation of the Bayesian model 82
3.4: Case study from literature data 84
3.4.1: Case study 84
3.4.2: Sensitivity analysis 90
3.5: Discussion 94
3.6: Conclusions 96
References 98

Chapter 4: Traceable integration of data and judgment in HEP estimation 103
4.1: Introduction 106
4.2: Integrated use of data and expert estimates in modelling HEP population variability: concepts and mathematical formulation 108
4.2.1: Concept: data and expert opinions as “mixed evidence” for HEP estimation 108
4.2.2: HEP population variability within ***F***: mathematical formulation and modelling assumptions 112
4.3: The Bayesian two-stage model for HEP quantification 114
4.3.1: General formulation 114
4.3.2: Configuration with lognormal variability distribution and implementation 120
4.3.3: Stage 1: verification with artificial data and sensitivity analysis 122
4.4: Application to case study 131
4.4.1: Case study: set of HFEs and evidence from literature 132
4.4.2: HEP quantification 135
4.5: Discussion 137
4.6: Conclusions 139
References 141

Chapter 5: Conclusions and future works 155
5.1: Overview of research objectives and tasks 155
5.2: Conclusions 157
5.2.1: Bayesian variability model for simulator data (Chapter 2) 160
5.2.2: Behavioral patterns to model crew performance variability (Chapter 3) 161
5.2.3: Traceable integration of data and judgment (Chapter 4) 164
5.3: Future works and recommendations 166
5.4: Publications 169
References 170

Appendix A 173

Appendix B 181

Appendix C 183

Appendix D 185

Appendix E 189

Appendix F 193

Appendix G 197

Appendix H 199

Curriculum Vitae 201

List of Tables

Table 2.1: Sources of uncertainty and variability in HEP estimates by HRA methods (given a constellation of task type and PSF categories). Note our work addresses the first two items of this table. 29

Table 2.2: Hypothetical simulator data used to inform the categorical elements of a generic HRA model for HEP estimation. Categorical elements taken from the SACADA taxonomy [17]. Note that the table reports only few elements of the rich SACADA context characterization. 31

Table 2.3: Comparison between the lognormal variability model and the beta-binomial: numerical results for Cases 1 and 2 (from Figure 2.5). Number of simulated tasks: 200, 500, 1000.......... 47

Table 2.4: Example of HEP estimation for a plant-specific task: prior distribution from lumped-data model (Table 2.3, “lumped posterior”, N_F = 200) and from the variability model (Table 2.3, “Var. model posterior”, N_F = 200). 47

Table 3.1: Grouping hypothetical data from different simulator contexts to inform the set of categories ($\boldsymbol{F}$) of a generic HRA model. Operational contexts and crew observed behaviors are adapted from [13, 43]. 73

Table 3.2: Derivation of behavioral categories from crew observed behaviors classified by team-, person-based factors: examples of application of steps I.2-I.3 in Figure 3.3, adapted from the case study in subchapter 3.4.......... 80

Table 3.3: Matching crew behaviors to behavioral patterns: examples of application of steps II.1-II.2 in Figure 3.3. Note that a predefined list of behavioral categories has to be available prior to the matching, e.g. from the application of steps I.1-I.3 in Figure 3.3.......... 81

Table 3.4: Simulator data used in the case study. 85

Table 3.5: Left: preliminary list of behavioral categories emerging from the empirical data for the case study [13, 43]. The associated team-, person-based metrics [44] used for the preliminary categorization are provided in Table D.1 (Appendix D). Right: compact set after expert-based aggregation, for use in the numerical application (subsection 3.4.1.2).. 87

Table 3.6: Seven crew groups and associated behavioral patterns identified in the case study (note that each group corresponds to a specific behavioral pattern).......... 88

Table 3.7: Numerical results from Figure 3.5 (note that each group corresponds to a specific behavioral pattern).. 89

Table 3.8: Higher level of data aggregation: an example with four behavioral groups (note that each group corresponds to a specific behavioral pattern). 93

Table 4.1: Hypothetical simulator and expert-elicited data for different tasks and operational contexts, for constellation ($\boldsymbol{F}$) (the examples of task and PSF categories are taken from SACADA taxonomy [24] for illustration purpose). 111

Table 4.2: Example of artificially-generated dataset, containing evidence hypothetically collected for ten task/context realizations (index i) relevant to the generic constellation $\boldsymbol{F}$. Sampled from a lognormal HEP population variability distribution with median = 5e-02, mean = 5.46e-02, and error factor = 2. 125

Table 4.3: Numerical results from Figure 4.5. 126

Table 4.4: Set of HFEs from US [54] and International [55-57] Empirical Studies analysed in the case study, and associated constellation of task/PSF categories ($\boldsymbol{F}$) from SACADA taxonomy [24]. 134

Table 4.5: Datasets for Stage 1 (left) and Stage 2 (right) used in the case study. For $E_{J|HFE}$, EF_i = 5 is assigned to each expert estimate. 135

Table 4.6: Scaling guidance used to convert the expert-based HFE difficulty rankings from [54-57] into HFE probability estimates (Table 4.5, last column). Adapted from the qualitative likelihood scale provided in Table 3.8-2 of NUREG-1880 (ATHEANA User's Guide [59]). 136

Table 5.1: General overview of the deliverables of this thesis and their uses in HRA/PSA . 159

Table B.1: Numeric results from sensitivity analysis on choice of priors for the lognormal variability model as shown in Figure 2.7 (Case 1, target HEP variability distribution with median = 5e-2, mean = 5.46e-2, and error factor = 2). 181

Table B.2: Numeric results from sensitivity analysis on choice of priors for the lognormal variability model as shown in Figure 2.9 (Case 2, target HEP variability distribution with median = 5e-3, mean = 6.25e-3, and error factor = 3). 182

Table C.1: Numeric results from the application of the lognormal variability model on real simulator data taken from Groth et al. 2014 [1] shown in Figure 2.11. 183

Table C.2: Numeric results from the application of the lognormal variability model on real simulator data taken from Jung et al. 2018 [2] shown in Figure 2.11. 183

Table D.1: List of teamwork competences and the associated metrics (taxonomy from Skjerve and Holmgren (2016) [1] used to categorize crew behaviors emerging from the case study (descriptions for each behavioral category are provided in Table 3.5). 185

Table E.1: Numerical results from the sensitivity analysis on the hierarchical model with varying number of and degrees of variability across behavioral groups. 190

Table E.2: Alternative variability functions tested for the variability models: lognormal, $p \sim LN(\mu,\sigma^2)$; logistic-normal, $p \sim P(N(\mu,\sigma^2))$. Prior distributions: diffuse $\pi_0(\mu)$ between log(1e-5) and log(1) for the lognormal formulations, between logit(1e-5) and logit(1) for logistic-normal formulations; diffuse $\pi_0(\sigma)$ between 0 and 5 (as recommended in Kelly and Smith, 2011 [1]). 191

Table G.1: Results from the sensitivity analysis to expert biases (subsection 4.3.3.3).......... 197

Table G.2: Numerical results from the application to case study (Section 4.4).................... 198

List of Figures

Figure 1.1: Upper box: research gaps addressed by the Ph.D. work. Lower box: use cases of the developed Bayesian data aggregation models for PSA/HRA. 7

Figure 1.2: Overview of the three research objectives and relevant tasks of the Ph.D. work 9

Figure 2.1: Simplified comparison between lumped-data and variability models (generic distributions shown). Left: probability density as uncertainty on the HEP population average (lumped-data model). Right: probability density as variability and uncertainty on HEP values variable by source (variability model) (given a constellation of task type and PSF categories). 34

Figure 2.2: Sketch of the variability model (generic distributions shown). HEP represented by a population variability distribution, $f_{\boldsymbol{F}}(HEP|\boldsymbol{\theta_F})$, combining variability within-category - ($f_t(p_t|\boldsymbol{\theta}_t)$, on the left - and crew-to-crew - $f_{c|t}(p_{c|t}|p_t^*, \boldsymbol{\theta}_{c|t})$, on the right (see eq. 2.2). The crew-specific HEP variable, $p_{c|t}$, is distributed around the HEP value of a specific realization of the task and PSF constellation (p_t^* in right plot). 36

Figure 2.3: The Bayesian hierarchical variability model, from top to bottom: $\pi_0(\boldsymbol{\theta_F})$, prior distributions for model parameters $(\boldsymbol{\theta_F})$; $f_{\boldsymbol{F}}(p_{c|t}|\boldsymbol{\theta}_t, \boldsymbol{\theta}_{c|t})$, the HEP variability distribution, where $f_t(p_t|\boldsymbol{\theta}_t)$ models within-category variability and $f_{c|t}(p_{c|t}|p_t^*, \boldsymbol{\theta}_{c|t})$ models crew-to-crew variability; $Bin(k = k_{ij}|p_{c|t}, N_{ij})$, the binomial distribution of evidence of k_{ij} failures on N_{ij} repetitions of the i-th task by the j-th crew. Generic distributions shown. 38

Figure 2.4: Overall aggregation framework to compare the variability and the lumped data models. Left: artificial data for the constellation $\boldsymbol{F}$ based on the example in Table 2.2. Top right: lognormal variability model, informed by the crew-specific data points (k_{ij}, N_{ij}) and returning as output the posteriors for the HEP variability distribution parameters, i.e. μ_t and $\sigma_{\boldsymbol{F}}$. Bottom right: conjugated beta-binomial model with lumped data ($k_{\boldsymbol{F}}$, $N_{\boldsymbol{F}}$), giving as output the posterior distribution for the single-value HEP (population average). 40

Figure 2.5: Expected mean (filled symbols), median (blank symbols), and 5th – 95th percentiles (whiskers) of $P_{\boldsymbol{F}}(HEP)$ by the lognormal variability model and the lumped-data beta-binomial model, tested against the same simulator datasets (number of simulated tasks: 200, 500, 1000). Datasets are artificially generated from lognormal HEP variability distribution with: median 5e-2 (dotted line), mean 5.46e-2 (dashed line) and error factor 2 (dot-dashed lines at 5th percentile 2.5e-2 and 95th percentile 1.0e-1) for Case 1 (left); median 5e-3 (dotted line), mean 6.25e-3 (dashed line) and error factor 3 (dot-dashed lines at 5th percentile 1.7e-3 and 95th percentile 1.5e-2) for Case 2 (right). 45

Figure 2.6: Data requirements of the lognormal variability model with flat hyper-priors $\pi_0(\mu_t)$ and $\pi_0(\sigma_{\boldsymbol{F}})$. Top: Case 1 (median = 5e-2, mean = 5.46e-2 and error factor = 2. Bottom: Case 2 (median = 5e-3, mean = 6.24e-3, and error factor = 3. For each sample size (x-axis), 100 datasets (dots) are Monte Carlo-sampled from the target distribution. From left to right: expected error factor, mean (log-scale), and median (log-scale) of the $P_{\boldsymbol{F}}(HEP)$'s returned by the model. 48

Figure 2.7: Sensitivity of the lognormal variability model to the choice of prior distributions for the hyper-parameters, i.e. $\pi_0(\mu_t)$ and $\pi_0(\sigma_{\boldsymbol{F}})$, and to the sample size, Case 1. Left: only $\pi_0(\mu_t)$ is informative. Upper/lower bounds of the lognormal distributions: "Good mean", 5e-3/5e-1; "High mean", 5e-2/1; "Low mean", 5e-4/5e-2. Right: both $\pi_0(\mu_t)$ and $\pi_0(\sigma_{\boldsymbol{F}})$ are informative. "With sigma" corresponds to a normal distribution with bounds 1.5/5 (expressed in terms of error factor)....................... 50

Figure 2.8: Behavior of the lognormal variability model with informative priors on both $\pi_0(\mu_t)$ and $\pi_0(\sigma_{\boldsymbol{F}})$ at $N_{\boldsymbol{F}}$ = 50 (target HEP variability distribution as in Case 1: median = 5e-2, mean = 5.46e-2, and error factor = 2). For each option of informative priors in x-axis (Figure 2.7, right plot), 100 datasets (dots) are Monte Carlo-sampled from the target distribution. From left to right, in y-axis: expected error factor, mean (log-scale), and median (log-scale) of the $P_{\boldsymbol{F}}(HEP)$'s provided by the model for each choice of prior (dotted lines: statistics of the target distribution). 50

Figure 2.9: Sensitivity of the lognormal variability model to the choice of prior distributions for the hyper-parameters, i.e. $\pi_0(\mu_t)$ and $\pi_0(\sigma_{\boldsymbol{F}})$, and to the sample size, Case 2. Left: only $\pi_0(\mu_t)$ is informative. Upper/lower bounds of the lognormal distributions: "Good mean", 5e-4/5e-2; "High mean", 5e-3/5e-2; "Low mean", 5e-5/5e-3. Right: both $\pi_0(\mu_t)$ and $\pi_0(\sigma_{\boldsymbol{F}})$ are informative. "With sigma" corresponds to a normal distribution with bounds 1.5/5 (expressed in terms of error factor)....................... 51

Figure 2.10: Behavior of the lognormal variability model with informative priors on both $\pi_0(\mu_t)$ and $\pi_0(\sigma_{\boldsymbol{F}})$ at $N_{\boldsymbol{F}}$ = 200 (target HEP variability distribution as in Case 2: median = 5e-3, mean = 6.25e-3, and error factor = 3). Same considerations as in Figure 2.8. Dotted lines: statistics of the target distribution. 51

Figure 2.11: Results from the application of the lognormal variability model to real simulator data available in literature (datasets in x-axis: left, [21]; right, [27]). On y-axis (in log-scale): expected mean (filled symbols), median (blank symbols), and 5th – 95th percentiles (whiskers) of the $P_{\boldsymbol{F}}(HEP)$'s estimated by both the lognormal variability model (circles) and the lumped-data beta-binomial model (circles) given the same marginal prior distribution on HEP.. 56

Figure 3.1: Relationship between performance influencing factors (taxonomy from IAEA [45]) and behavioral patterns used in this work to represent crew performance variability in HEP quantification.. 69

Figure 3.2: Comparison of the HEP formulations and the associated data aggregation adopted by the proposed variability model (left: "aggregation by groups") and the alternative approaches tested in the case study (center: "lumped-data", with HEP as population average; right: "no aggregation", with HEP as a "continuum" of task-, crew-specific error probabilities). All the p's are intended as conditional on the given set $\boldsymbol{F}$, e.g. $p_{c|\boldsymbol{F}}$, $p_{ij|\boldsymbol{F}}$.. 74

Figure 3.3: Overview of the multi-step methodology to derive and use behavioral patterns in HEP quantification for a generic set $\boldsymbol{F}$... 76

Figure 3.4: Bayesian models for HEP quantification coupled to the three modelling approaches of Figure 3.2. All the p's are intended as conditional on the given set $\boldsymbol{F}$, e.g. $p_{c|\boldsymbol{F}}$, $p_{ij|\boldsymbol{F}}$. 83

Figure 3.5: Results from the numerical application to the case study. In x axis, from left to right: conjugate beta-binomial with lumped data ("Lumped"); continuous variability model with crew-, task-specific data ("No aggregation"); hierarchical beta-binomial model with seven behavioral groups ("7 groups"); and group-specific $P(HEP)$'s ("Groups: 1 to 7"). In y axis (log scale): mean (symbols), 5th and 95th percentiles (whiskers) of the $P(HEP)$... 89

Figure 3.6: Influence of the number of identified behavioral groups ("7 groups" vs "4 groups") on the $P(HEP)$ estimated by the hierarchical beta-binomial model for both the case study (Tables 10 and 12) and the artificially-generated dataset with less performance variability (Table E.1)... 93

Figure 4.1: Concept for use of reference HEP values and its application in this work.......... 110

Figure 4.2: Mathematical formulation of HEP($\boldsymbol{F}$) adopted in this work (adapted from [44] and extended to incorporate expert estimates). Top: continuous parametric distribution for $p_{t|\boldsymbol{F}}$, with unknown parameters $\boldsymbol{\theta_F}$. Bottom left: binomial aleatory process (failures k_i and observations N_i) for the i-th task, governed by the associated HEP value (p_i). Bottom right: probability distribution representing the expert estimate $\tilde{p}_i$ and associate accuracy. 114

Figure 4.3: Flowchart of the inference process in the developed Bayesian two-stage model. Stage I (blue box): estimation of the HEP population variability distribution associated to the given constellation $\boldsymbol{F}$ of task type / PSF ratings, i.e. $P_{\boldsymbol{F}}(p_{t|\boldsymbol{F}})$. Stage II (red box): estimation of the HEP uncertainty distribution for the plant-specific human failure event, i.e. $P_{HFE}(HEP)$. 119

Figure 4.4: Left: Stage 1 configuration with lognormal population variability distribution (PV: population variability). Centre, right: alternative modelling approaches for Stage 1 tested in subsection 4.3.3, respectively based on the lognormal variability model for simulator data presented in [44] and the lumped-data approach as in [25-27]..... 122

Figure 4.5: In y axis (logarithmic scale): expected mean (filled symbols), 5th and 95th percentiles (whiskers) of $P_{\boldsymbol{F}}(p_{t|\boldsymbol{F}})$ returned by three Stage 1 formulations in Figure 4.4, tested against the same datasets (in x axis, number of task realizations: 10, 20, 50, 100). Datasets generated from a known lognormal variability distribution with mean = 5.46e-02 (dashed line) and error factor = 2 (dot-dashed lines at 5th percentile = 2.5e-02 and 95th percentile = 1.0e-01). 125

Figure 4.6: Data requirements of the lognormal population variability models (“Stage 1: only data” and “Stage 1: data and judgment”, in Figure 4.4 left and center). For each sample size in x axis, 100 datasets are Monte Carlo-sampled from the target HEP variability distribution with median = 5e-02 and error factor = 2 (dashed lines). From left to right: expected error factor and median (log-scale) of $P_{\boldsymbol{F}}(p_{t|\boldsymbol{F}})$ returned by the models, in the form of 50% and 90% confidence intervals (filled symbols: average value of the 100 datasets). 128

Figure 4.7: Sensitivity of the lognormal PV-binomial-lognormal model (Figure 4.4, left) to biases in expert estimates (results for “unbiased” case are taken from Figure 4.5). In y axis (logarithmic scale): expected mean (filled symbols), 5th and 95th percentiles (whiskers) of $P_{\boldsymbol{F}}(p_{t|\boldsymbol{F}})$ returned by the model for three different bias cases (“unbiased”, “conservative”, “optimistic”), at 10 and 50 task realizations (x axis). Target HEP variability distribution: lognormal with mean = 5.46e-02 (dashed line) and error factor = 2 (dot-dashed lines at 5th percentile = 2.5e-02 and 95th percentile = 1.0e-01). 129

Figure 4.8: Results from the application of the two-stage model to a subset of HFEs (in x-axis) from the case study (the numeric results for the complete set of HFEs are provided in Table G.2, Appendix G). On y-axis (in log-scale): expected mean (filled symbols), 5th and 95th percentiles (whiskers) of the $P_{HFE}(HEP)$’s returned by the two-stage model (for both Case 1 and Case 2) and the lumped-data approaches from literature sources [53-57].. 137

Figure A.1: Integration of HRA model results in a PSA event tree analysis: examples adapted from literature [17] (CD = Core Damage). 175

Chapter 1: Introduction

This chapter presents an overview of the Ph.D. research work, aimed at developing Bayesian models integrating simulator data and expert judgment in the derivation of reference error probability values for human reliability. Section 1.1 discusses the motivation and the rationale behind this work. Section 1.2 presents specific goals and objectives, focusing on the methodologies and modelling approaches adopted to achieve them. Section 1.3 presents the relevant scientific contributions and deliverables produced by the Ph.D. work. The closing section of this chapter (Section 1.4) provides the outline of the thesis.

The present dissertation is structured as a collection of three journal papers, out of which two have been peer-reviewed and accepted by the journal editors, and one is currently under internal review:

- Greco SF, Podofillini L, and Dang VN. A Bayesian model to treat within-category and crew-to-crew variability in simulator data for Human Reliability Analysis. *Reliab Eng Syst Safe* 2021, 206:107309, ISSN 0951-8320 (Chapter 2).
- Greco SF, Podofillini L, and Dang VN. Crew performance variability in human error probability quantification: a methodology based on behavioral patterns from simulator data. *Proc I Mech E Part O: J Risk and Reliability* 2021, doi:10.1177/1748006X20986743 (Chapter 3).
- Greco SF, Podofillini L, and Dang VN. A Bayesian two-stage approach to integrate simulator data and expert judgment in human error probability estimation. Currently under internal review, expected submission date: June 2021 (Chapter 4).

1.1 Background and work motivation

Human Reliability Analysis (HRA) is the part of Probabilistic Safety Assessment (PSA) addressing the contribution of human failures to the quantification of risk of complex technical systems, typically nuclear power plants, chemical and aerospace systems [1]. HRA methods support the identification of the safety-critical tasks performed by the operating crews, the characterization of contextual factors influencing human performance, and the estimation of error probability values (referred as Human Error Probabilities, HEPs) of postulated human failure events (HFEs). HRA results are typically integrated into PSA studies to quantify the overall frequency of accidental scenarios and support safety-related decision-making of licensees and regulators [2-3] (see Appendix A for further background information on HRA).

In most HRA methods, the estimation of human error probability values is supported by **models that, depending on the method, relate the types of operator tasks and performance**

influencing factors to values of failure probability. As the combination (namely, "constellation") of task and factor categories changes, HRA models provide different HEP estimates, representing the spectrum of performance variability as task characteristics and operational contexts vary. As a common feature, HRA models represent both operational tasks and context-related influencing factors via categorical factors, typically task types and Performance Shaping Factors (PSFs) [1-3]. The definitions and metrics of the task types and PSFs levels (or ratings) depend on method taxonomy, for instance: in the Technique for Human Error Rate Prediction (THERP [4]), task types as "check/reading digital indicators", and PSFs such as training and stress (e.g. "very low", "optimum" stress); in the Human Error Assessment and Reduction Technique (HEART [5-7]), generic task types as "complex task requiring high level of comprehension and skill" and error producing conditions such as "a low signal-noise ratio". A similar use of categorical factors can be found in the majority of existing HRA methods, e.g. in the Cognitive Reliability and Error Analysis Method (CREAM [8]), in the Standardized Plant Analysis Risk-Human reliability (SPAR-H [9-10]), as well as in newer methods as the Integrated Human Event Analysis System (IDHEAS [11]).

Reference HEP values and bounds for the task categories and the PSF effects are needed to parametrize a method's quantification model, both for traditional as well as for advanced models. In traditional HRA models, reference values refer to baseline HEPs, e.g. HEP corresponding to performing a task under optimal/nominal performance conditions (e.g. a routinely trained diagnosis task, supported by procedural guidance). Also, such reference values would be used to assess the effect of variation of performance factors (i.e. of the PSFs) on the probability value, typically as a multiplier to the baseline HEP to reflect the effect of adverse performance conditions. Recently, advanced models such as Bayesian Belief Networks (BBN) have been developed for HRA applications to capture the complex task, PSF, and HEP relationships and to enhance traceability in use of diverse data and judgment [12-16]. Reference values are needed as well for these advanced models, to inform the BBN parameters, i.e. the Conditional Probability Distributions (CPDs).

The data underlying the reference HEP values is generally obtained by combining empirical evidence and expert judgment [1-3]. Since the early developments of HRA models, empirical data has been gathered from a variety of information sources: licensee event reports, retrospective analyses of accidents and operational events, human factors and behavioral science experiments, and training in control room simulators. Judgment is typically elicited from domain experts in different forms, e.g. quantitative probability estimates and/or qualitative statements on the importance of influencing factors [17]. For most of the currently used HRA models, there is **lack of traceability in the aggregation of the various data sources, as well as in their combination with expert judgment**. The result is that it is now difficult to determine to what extent the HEP values produced are empirically based. The underling raw data and the process to feed HRA models are not evident. Also, as new data would become available, it is not clear how to incorporate it as new evidence, to progressively obtain solid HEP values. Indeed, given the complexity of human performance influences, HRA data collection is a long-term process (in practice, continuing). As such, traceable data aggregation models need to be developed to accommodate new evidence, to evaluate relevance of data for sharing, and generally to ensure that in the long-term the empirical basis of HRA models will be strengthened [17-20].

Fresh impetus to HRA data collection came from the International [21] and US [22] HRA Empirical Studies, aimed at assessing the validity of HRA method predictions against data from nuclear power plant main control room simulators. Besides improving HRA practice and methods, these studies resulted in methodological advances in the collection of simulator data for HRA purposes, fostering several recent activities [23-25]. Two notable, ongoing, data collection programs are the HUman REliability data Extraction framework, HuREX [23], and the Scenario Authoring, Characterization, And Debriefing Application, SACADA [24]. With their long-term data collection perspective, these simulator programs are expected to produce a large amount of empirical evidence for use in modern HRA models, more representative of recent operational conditions (e.g. reflecting modern interfaces and procedural guidance).

The majority of recent research activities dealing with the use of simulator data for HRA has addressed the development of protocols and taxonomies to collect data: in particular, the interpretation of performance outcomes in terms of failure or success and the definition of the types of information on crew performance to collect [24, 26-27]. However, **gaps remain on how to use this information to quantify HEP values and bounds, and how to eventually incorporate them into HRA models** [18, 28-33]. Framed within this research direction, this Ph.D. work aims at **developing new Bayesian quantitative models integrating simulator data and expert judgment in the estimation of human error probabilities**. The developed models are intended to **traceably produce empirically-based reference HEP values and bounds** that can be used (Figure 1.1, lower box) to inform HRA methods' task type and PSF categories (or PSF multiplier values, depending on the method), as well as anchoring distributions for parametrizing advanced HRA models (such as the modern BBN-based models [16, 20, 32-34]). In addition, the produced HEP values and bounds can be used as generic population variability distributions to support HEP quantification in plant-specific PSA studies characterized by scarce plant-specific data (see subsection 1.3.3).

This Ph.D. work addresses in particular the following three research gaps (Figure 1.1, upper box). A first, foundational aspect relates to the treatment of data variability: as data is collected over different plants, operating crews, scenarios, tasks and in different time spans, characterization and treatment of data variability is required for appropriate statistical inference (see next subsection 1.1.1). Second, to further understand the sources of performance variability, a structured methodology is needed to identify crew performance drivers from simulator data, and empirically incorporate their effects on the HEP variability distribution (subsection 1.1.2). Third, the need for a traceable and transparent integration of expert judgment whenever the latter is used in the development of the reference HEP values (subsection 1.1.3). The three research gaps are further elaborated in the following subsections.

1.1.1. Modelling variability within the constellations of taxonomy categories

The first research gap deals with variability aspects in simulator data collection. In a similar way as HRA methods, data collection protocols characterize simulator observations through categories related to taxonomies of tasks (e.g. “entering step in procedure” in HuREX [23]), failure mechanisms (e.g. “failure to prioritize” in SACADA [24]), contextual factors (e.g. “overloaded status of alarm board” in SACADA [24]), and the like (both SACADA and HuREX taxonomies are further discussed in Section 2.2).

As the SACADA and HuREX databases are being populated, research on quantification of HEP values from the emerging data is ongoing internationally [28-33, 35] (more details in Section 2.2). These pioneering works have shown the advantages of Bayesian inference models in using the collected simulator data to quantify the HEP (and the associated uncertainty) for the relevant constellation of taxonomy categories, e.g. from the SACADA taxonomy [24]: macrocognitive function “understanding the situation/problem”, given the situational factor “information quality” with level “conflicting”. These works focused on the relationship between the given task, the set of PSFs and the error probability, and investigated performance variability in simulator tasks under different PSF effects, i.e. “**across constellations”**: with respect to the previous example, e.g. when the “information quality” is “misleading” instead of “conflicting”. However, **variability in simulator data exists as well within task and PSF categories, i.e. within the given constellation** (i.e. combination of task type and factors). This stems from the fact that simulator records relevant to the given constellation are collected from different simulator scenarios, different plants, and different realizations of the contextual factors. For instance, taking as reference the categories from SACADA taxonomy [24], the tasks “monitoring trend of steam generator level” in a Steam Generator Tube Rupture (SGTR) scenario and “monitoring trend of pressurizer pressure” in a Small break Loss Of Coolant Accident scenario (SLOCA) correspond to different realizations of the same task type “understanding the situation/problem”. In addition, records are collected from different crews, with different behavioral characteristics and operating styles (e.g. different team dynamics, communication strategies, and task prioritization). Therefore, for a given constellation of taxonomy categories, variability in simulator data has a twofold aspect: on one hand, the variability stemming from the different realizations of the associated task and PSF categories; on the other hand, the variability due to the different crew-specific features. In this thesis, these variability aspects are referred to as **variability "within-category"** and **"crew-to-crew" variability**, respectively. Such variability aspects requires explicit consideration: **not adequately representing the existing sources of variability in simulator data may possibly lead to overconfidence in the HEP estimates** [17, 36-37]. As the on-going data collection efforts will provide more evidence, it becomes important to strengthen the empirical basis of both the population average HEP values, as well as of the HEP spectrum of variability within each constellation. **How mathematically modelling variability within the constellations of simulator taxonomy categories is addressed by the first block of the Ph.D. work** (research objective #1 in Section 1.2).

1.1.2. Lack of a methodology to incorporate crew behavioral characteristics in HRA models

The second research gap concerns the incorporation of crew behavioral characteristics emerging from simulator data in the HEP quantification. Since the early developments of HRA methods, the influence of personal and team characteristics on crew performance has not been explicitly considered as input factors to quantitative HRA models (e.g. as PSFs) but implicitly, within the variability and uncertainty ranges associated to the HEP values. HRA acknowledges that person-, team-related factors can exert an important influence on performance variability and, to some extent, addresses these in the qualitative analysis supporting HEP quantification [4, 17] (this aspect is further discussed in Section 3.2).

The recent HRA Empirical Studies (the International [21] and the US [22]) highlighted **the key importance of several crew behavioral aspects**, such as "team dynamics, work processes, communication strategies, sense of urgency […]" [21], **as main contributors to performance variability in operational tasks**, especially in emergency situations where standard procedure following is challenged by a fast scenario progression and a limited procedural guidance. In such performance conditions, crew characteristics (e.g. in information sharing, task prioritization, adherence to procedural guidance) played a key role in determining not only the pace through the procedures, but also which procedural path to follow [21-22]. More recent studies in the Halden Man-Machine Laboratory (HAMMLAB) simulator [38-39] further underscored that, for emergency scenarios characterized by a procedure-situation mismatch, "the crews that followed the procedures more strictly had lower performance than crews that engaged more in autonomous initiatives and extra-procedural activities". These studies [21-22, 38-39] acknowledged the benefits of using simulator data to investigate the effects of crew behavioral characteristics on performance variability in operational tasks as well as the need to formally incorporate these in the HEP quantification, especially for those "scenarios that exceed the limits of the basic nuclear power plant design" and "include multiple equipment failures" [21]. Indeed, incorporation of some crew variability aspects in HRA is one of the distinctive characteristics of the emerging modern HRA methods, for example by the use of Crew Response Diagrams in the IDHEAS method [11] or Crew Response Trees in the Phoenix method [40].

Besides information on tasks and PSFs, the ongoing large-scale simulator programs [23-24] can collect data on crew behaviors observed in different simulated emergency scenarios. To date, **no methodological approaches have been proposed to formally analyze crew behavioral characteristics emerging from data and model their effects on the HEP**. This has been outside the scope of the existing works with simulator data [28-30]: these exploratory works maintained the traditional HEP formulation as a function of scenario-, task- and context-related factors, and thought of the influence of person-, team-related factors on the HEP values as a statistical population. Characterizing the crew performance drivers for different operational tasks is not straightforward, given the complexity of both human behaviors and emergency scenarios typically addressed in PSA applications. **A structured methodology is therefore required** to support the identification of relevant crew behavioral characteristics that determine variability in task performance and empirically incorporate their effects in the error probability for multiple constellations of task and PSF categories of HRA models. **This need is addressed by the second block of the Ph.D. work** (research objective #2 in Section 1.2).

1.1.3 Limited traceability of judgment incorporation in existing HRA methods

The collection of HRA data in large simulator programs is a long-term, sustained activity. To date, a significant amount of human failure data has been already collected worldwide: for instance, as at September 2018 [31], the SACADA database counts more than 25000 data points distributed across few hundred constellations of task and PSF categories (a portion of the database is publicly available at the US Nuclear Regulatory Commission website [41]). However, current data availability is still not sufficient to derive HEP estimates for the entire spectrum of constellations of HRA model categories. For some of these constellations, the amount of data points can be indeed small, or even null. The issue of scarcity of data is especially relevant for those constellations representative of scenarios and operating contexts that are difficult to reproduce in nuclear power plant simulators: for instance, scenarios involving long time windows (hours) or involving challenges external to the main control room (e.g. natural hazards). Hence, the incorporation of **expert judgment in HRA models still plays an important role**, to fill the information gap and improve the quality of the HEP estimates.

The third research gap addressed by the Ph.D. work deals with a general limitation of existing HRA methods: **the actual process of judgment incorporation in the derivation of reference HEP values and bounds is generally not traceable**. For instance, quoting the THERP Handbook [4], “the data underlying THERP’s model is mostly coming from human factor experiments and field studies (…). The probability values are generally derived data, in the sense that they contain much extrapolation and judgment". In presence of limited data traceability in HRA method documentation, the empirical basis of the HEP estimates cannot be easily distinguished from the judgment-based information, thus **affecting the acceptability of HRA method results** for use in safety-relevant applications. This calls for **new quantitative approaches ensuring a traceable and transparent integration of judgment** whenever the latter is used to complement empirical data in the estimation of reference HEP values and variability bounds for HRA methods' task type and PSF categories, as well as in the production of anchoring distributions for the emerging BBN-based approaches [16, 20, 32-34]. The main thrust is that **a systematic, traceable aggregation of simulator data and judgment will allow feeding HRA models with new data as it becomes available**, progressively replacing judgment and older evidence that may become outdated because of new advances in plant operation and design. Previous works in PSA explored the potential of Bayesian methods in treating expert-elicited probability estimates [42-44] and formally combining these with reliability data (possibly sparse and from diverse sources) in the estimation of reliability measures (e.g. core melt frequency of nuclear power reactors [45], pump failure rate [46]). Building on these works, **the third block of the Ph.D. work addresses how mathematically integrate simulator data and expert-elicited probability estimates in the production of reference HEP values and variability bounds** (research objective #3 in Section 1.2).

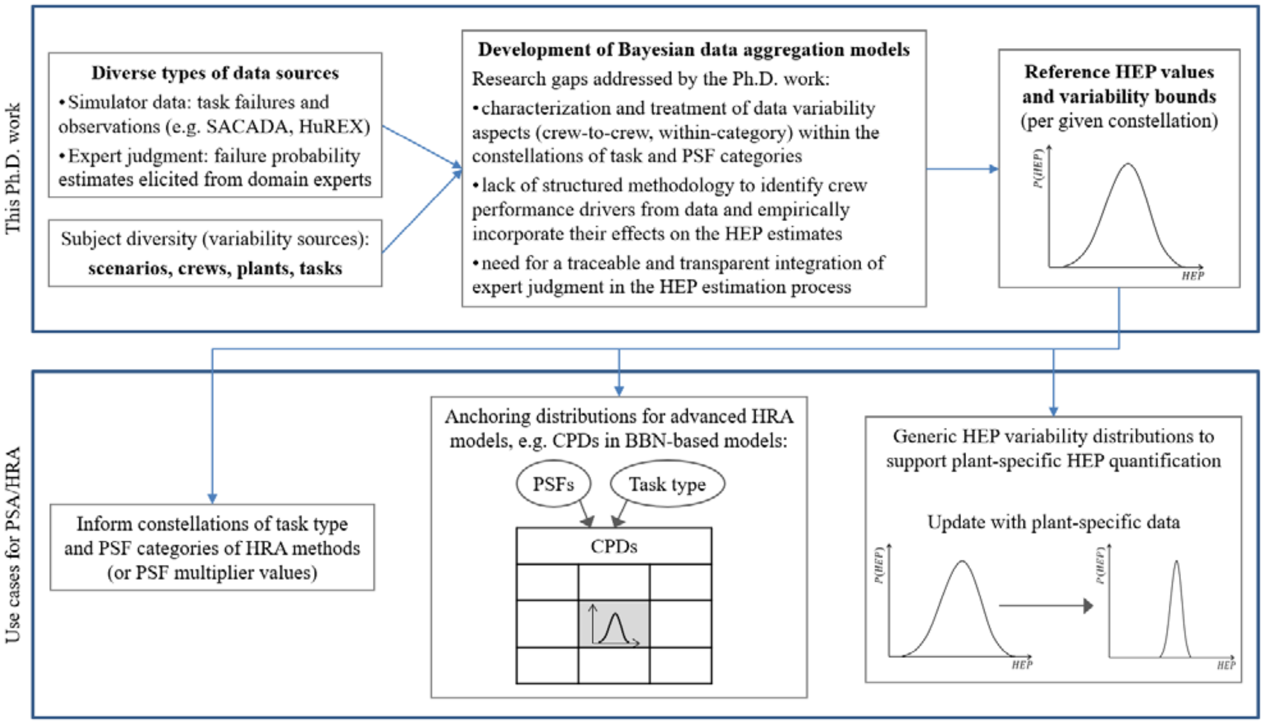

This Ph.D. work
Diverse types of data sources
•Simulator data: task failures and observations (e.g. SACADA, HuREX)
•Expert judgment: failure probability estimates elicited from domain experts
Subject diversity (variability sources): scenarios, crews, plants, tasks
Development of Bayesian data aggregation models
Research gaps addressed by the Ph.D. work:
•characterization and treatment of data variability aspects (crew-to-crew, within-category) within the constellations of task and PSF categories
•lack of structured methodology to identify crew performance drivers from data and empirically incorporate their effects on the HEP estimates
•need for a traceable and transparent integration of expert judgment in the HEP estimation process
Reference HEP values and variability bounds (per given constellation)
P(HEP)
HEP
Use cases for PSA/HRA
Inform constellations of task type and PSF categories of HRA methods (or PSF multiplier values)
Anchoring distributions for advanced HRA models, e.g. CPDs in BBN-based models:
PSFs
Task type
CPDs
Generic HEP variability distributions to support plant-specific HEP quantification
Update with plant-specific data
P(HEP)
HEP
P(HEP)
HEP

Figure 1.1. Upper box: research gaps addressed by the Ph.D. work. Lower box: use cases of the developed Bayesian data aggregation models for PSA/HRA.

1.2 Research goals and objectives

As mentioned in the previous Section 1.1, the overall motivation of this Ph.D. work is to improve the traceability in the aggregation of HRA data sources, as well as in the use of expert judgment, in the production of reference HEP values and bounds for HRA models. The three research gaps discussed in subsections 1.1.1-1.2.3 translate into **three specific research objectives**, aiming at **developing new quantitative models - based on Bayesian statistical methods** - to:

- **formally treat variability aspects (crew-to-crew, within-category)** within the constellations of task and PSF categories of simulator data collection taxonomies (e.g. HuREX [23], SACADA [24]), with **general applicability** to different constellations (research objective #1);
- **support the identification of relevant crew behavioral characteristics** emerging from simulator data (e.g. in team decision-making, communication strategies, adherence to procedures, etc.) that determine performance variability for a given constellation, and **incorporate their effects on the HEP**, jointly with the influence of the set of PSFs (research objective #2);
- **systematically and traceably incorporate expert judgment** in the HEP estimation process, to allow for updates as new empirical evidence becomes available and strengthen the technical basis of HEP estimates (research objective #3).

Besides the development, implementation and verification of the Bayesian models, the three research objectives foresee also the **demonstration of the developed models to a case study** of interest for practical HRA applications. In order to achieve the above stated objectives, the Ph.D. work was structured into **three groups of research tasks**, which are presented in the remainder of this section. Figure 1.2 links the three research objectives and the relevant tasks to the corresponding chapters of this thesis.

Research tasks 1.1-1.4 address the first research objective (red block in Figure 1.2) and concerns the development of a Bayesian variability model to capture the multiple sources of variability (within-category, crew-to-crew) in simulator data (Chapter 2: Bayesian variability model for simulator data). The idea is to build a quantitative tool, able to mathematically aggregate simulator data from nuclear power plants to estimate failure probabilities (with their variability and uncertainty distributions), for the various constellations of task type and PSF categories of data collection taxonomies (e.g. [23-24]). The key feature of the variability model is the mathematical representation of the HEP variability spectrum via continuous parametric distributions (research task 1.2). For a given constellation, the simulator records relevant to different operational tasks, scenarios, plants and crew characteristics are associated to different HEP realizations and treated as part of a population, continuously distributed around the population-average HEP according to the parametric distribution selected by the analyst (e.g. a lognormal probability density function). The variability model is coupled to a Bayesian model to empirically infer the parameters of the HEP variability distribution (e.g. for a lognormal distribution, mean and standard deviation) from the simulator observations relevant to the investigated constellation. Research task 1.3 addresses the verification of the developed Bayesian variability model with artificially-generated data, and aims at investigating model sensitivity to data availability in presence of different types of prior information on the parameters of the variability distributions. Finally, the developed model is applied to a case study involving simulator datasets from literature [28, 30] and compared against the existing conjugate beta-binomial approaches with lumped-data [28-30], to demonstrate the effects of modelling variability on HEP estimates (research task 1.4). The model formulation with continuous parametric distributions is foreseen to be generically applicable to different taxonomies of task and PSF categories (e.g. HuREX [23], SACADA [24]), as well as flexibly adaptable to target specific variability components (e.g. plant-to-plant, scenario-to-scenario, crew-to-crew) according to data availability and the scope of the application (see Section 2.5 for further details on model applicability).

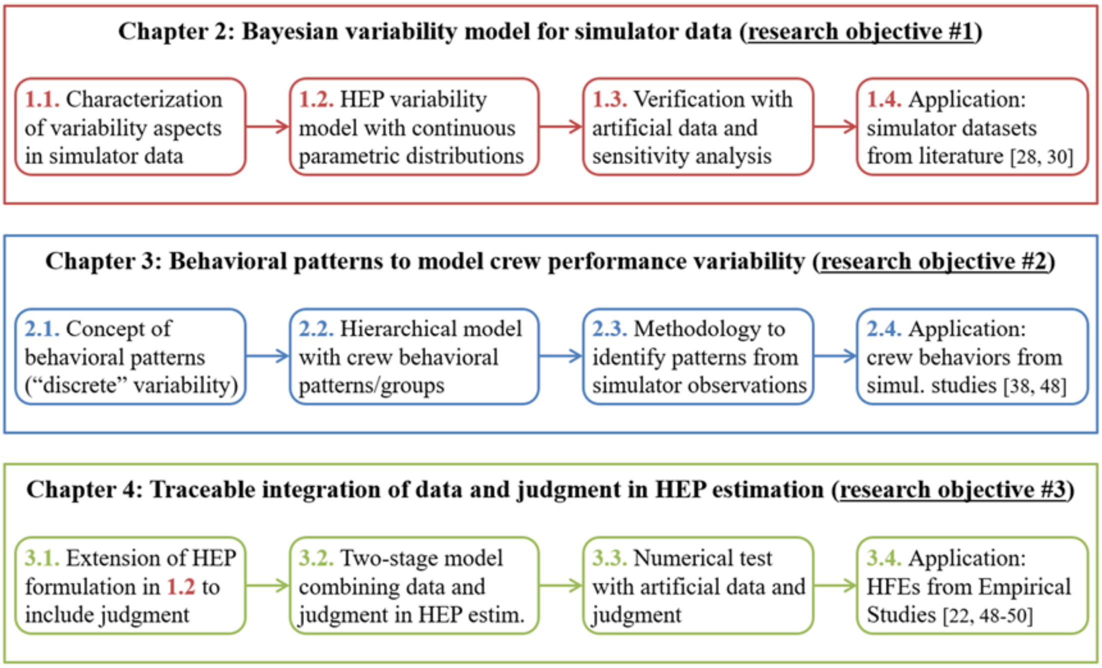

Chapter 2: Bayesian variability model for simulator data (research objective #1)
1.1. Characterization of variability aspects in simulator data
1.2. HEP variability model with continuous parametric distributions
1.3. Verification with artificial data and sensitivity analysis
1.4. Application: simulator datasets from literature [28, 30]
Chapter 3: Behavioral patterns to model crew performance variability (research objective #2)
2.1. Concept of behavioral patterns ("discrete" variability)
2.2. Hierarchical model with crew behavioral patterns/groups
2.3. Methodology to identify patterns from simulator observations
2.4. Application: crew behaviors from simul. studies [38, 48]
Chapter 4: Traceable integration of data and judgment in HEP estimation (research objective #3)
3.1. Extension of HEP formulation in 1.2 to include judgment
3.2. Two-stage model combining data and judgment in HEP estim.
3.3. Numerical test with artificial data and judgment
3.4. Application: HFEs from Empirical Studies [22, 48-50]

Figure 1.2. Overview of the three research objectives and relevant tasks of the Ph.D. work.

Research tasks 2.1-2.4 address the second research objective (blue block in Figure 1.2) and specifically focus on modelling the crew-to-crew variability component from crew behavioral characteristics manifested in simulator studies (Chapter 3: Behavioral patterns to model crew performance variability). For this purpose, research task 2.1 introduces “behavioral patterns” (e.g. “collective” or “non-inclusive” information sharing, “proactive” or “reactive” interpretation of procedures) to represent the spectrum of crew behavioral characteristics empirically observed during task performance, and model their effects on crew performance variability for a given constellation of task/PSF categories. The underlying concept is that crews sharing similar patterns during task performance can be aggregated in the same “behavioral group” and associated the same value of error probability in a “discrete” formulation of the HEP variability model. Compared to the continuous formulation addressed in research task 1.2, the discrete formulation of HEP focuses on explicitly capturing the impacts of different behavioral patterns on the error probability, using a Bayesian hierarchical model (research task 2.2). The core element of the second research objective is the development of a multi-step methodology to support the identification of a finite set (hence, the discrete attributes in the HEP formulation) of crew behavioral patterns from simulator data and their use in HEP quantification (research task 2.3). The methodology is structured in two main blocks. In the first block, crew behavioral characteristics observed during task performance are systematically analyzed adopting teamwork, decision-making and situation awareness taxonomies (e.g. [47]) and classified into a list of “behavioral categories” accordingly, for instance: concerning communication, the frequency with which strategic meetings are held (e.g. “frequent strategic meetings”); concerning work attitudes, the compliance to procedure indications (e.g. “strict procedure following” or “more autonomous initiatives”), and the like. In the second block, crew performances are matched to the corresponding categories and grouped according to “behavioral patterns”, i.e. specific combinations of behavioral categories, representative of the spectrum of performance variability for the given constellation of task type and PSFs. The empirically-identified patterns are then used to inform crew behavioral groups in the Bayesian hierarchical model, to estimate the HEP variability distribution for the given constellation. Finally, research task 2.4 foresees the application of both the multi-methodology and the Bayesian hierarchical model to a case study from literature, involving crew behaviors observed in different emergency scenarios from recent simulator studies [38, 48]. Besides demonstrating the feasibility of the proposed methodology on a practical HRA application, the goal of the numerical application is to show the effects of incorporating behavioral characteristics of operating crews in HEP estimates, compared to alternative quantitative approaches for simulator data (see Section 3.4 for further details).

In order to achieve the third research objective, research tasks 3.1-3.4 (green block in Figure 1.2) aims at systematically and traceably incorporating expert judgment in the estimation of human error probability values and bounds from simulator data (Chapter 4: Traceable integration of data and judgment in HEP estimation). Research task 3.1 addresses how mathematically integrate judgment and simulator data in an upgraded version of the HEP variability model with continuous parametric distributions (research task 1.2). Here, judgment is considered in the form of task failure probability estimates provided by domain experts, via direct elicitation or through the application of an existing HRA method. Each expert estimate is associated to an uncertainty measure (e.g. an error factor[1] on the point estimate) that numerically expresses the confidence level of the HRA analyst on expert accuracy (similarly as in [43, 46]). This new formulation of the HEP variability model is used as the basis for a two-stage Bayesian approach (research task 3.2), developed with the goal to improve the estimation of plant-specific task failure probabilities in presence of limited empirical data (see Section 4.3 for further details). In the first stage, simulator data and expert estimates (with the associated confidence measure) are combined in the upgraded variability model to derive reference HEP values and variability bounds for the constellation of task and PSF categories representative of the plant-specific task of interest. The output of the first stage (i.e. the HEP variability distribution of the representative constellation) is then updated in the second stage with plant-specific failure data and expert estimates, to quantify the error probability of the plant-specific task. Research task 3.3 addresses the verification of the developed two-stage Bayesian model with artificially-generated data and judgment, with the goal to analyze the effects of judgment incorporation on HEP estimates and investigate model sensitivity to biases in expert judgment. Lastly, Research task 3.4 involves the application of the two-stage Bayesian model to quantify the error probabilities of a collection of human failure events from the recent HRA Empirical Studies [22, 48-50].

The research tasks underlying the Ph.D. work and the corresponding chapters of this thesis are summarized in the following:

- **Chapter 2**: **Bayesian variability model for simulator data** (**<u>research objective #1</u>**):
  - 1.1. Characterization of variability aspects (crew-to-crew, within-category) in the constellations of task/PSF categories of data collection taxonomies (Section 2.2).
  - 1.2. Mathematical formulation of HEP variability model with continuous parametric distributions to represent data variability for a given constellation, and development of a Bayesian model to empirically estimate the parameters of the variability distribution from simulator data (Section 2.3).
  - 1.3. Model verification and sensitivity analysis with artificial data, to investigate data

---

[1] In PSA/HRA, the error factor is a commonly-adopted measure of dispersion for characterizing the spread of a lognormal distribution. Typically, the EF is expressed by the square root of the ratio 95th/5th percentiles.

requirements to inform HEP variability in presence of different types of prior information on model parameters (subsection 2.4.2).

  1.4. Application to simulator datasets from literature [28, 30], to demonstrate the effects of modelling variability on HEP estimates (subsection 2.4.3).

- **Chapter 3: Behavioral patterns to model crew performance variability (<u>research objective #2</u>)**:
  - 2.1. Concept of behavioral patterns to explicitly represent the influence of crew behavioral characteristics observed in simulator studies on performance variability: discrete formulation of HEP variability (Section 3.2).
  - 2.2. Development of a Bayesian hierarchical model to capture (from data) performance variability across crew behavioral patterns/groups, and incorporate their effects on the HEP estimate for the given constellation of task/PSF categories (Section 3.3).
  - 2.3. Multi-step methodology to support the identification of crew behavioral patterns from simulator data and their use in HEP quantification (Section 3.3).
  - 2.4. Application to crew behaviors collected from different emergency scenarios in recent simulator studies [38, 48], to demonstrate the effects of empirically incorporating crew behavioral characteristics in the HEP estimates (Section 3.4).

- **Chapter 4: Traceable integration of data and judgment in HEP estimation (<u>research objective #3</u>)**:
  - 3.1. Extension of the HEP variability formulation with continuous parametric distributions (research task 1.2) to mathematically incorporate judgment (Section 4.2).
  - 3.2. Development of a two-stage Bayesian model to formally combine data and judgment in the estimation of HEP values and bounds for constellations of task/PSF categories (first stage), and plant-specific task failure probabilities (second stage) (subsection 4.3.2).
  - 3.3. Numerical test with artificially-generated data and judgment, to analyze the effects of judgment incorporation on HEP estimates and investigate model sensitivity to biases in expert judgment (subsection 4.3.4).
  - 3.4. Application to a collection of human failure events from the recent HRA Empirical Studies [22, 48-50] (Section 4.4).

1.3 Key contributions of the research work

The following subsections separately discuss the key contributions from the three blocks of research.

1.3.1. First-of-a-kind quantitative model to treat data variability “within the constellation” of task and PSF categories

To date, no quantitative approach for HEP estimation has been proposed to address crew-to-crew and within-category variability aspects within the constellations of categories of simulator data collection taxonomies (research objective #1). As mentioned in the background, the existing approaches with lumped data [28-30] focused on the quantification of average HEP values for these constellations, and considered HEP variability only “across constellations” (e.g. with different PSF levels). The Bayesian variability model provided by this thesis (research tasks 1.1-1.4, presented in Chapter 2) represents a first-of-a-kind attempt to formally use emerging simulator data not just to inform the average HEP value for the given constellation, but also the associated variability bounds. The focus of the proposed model is indeed on modelling variability “within the constellations” of taxonomy categories: this new approach entails considering the evidence from different realizations and different crews as multiple pieces of evidence for the given constellation, pertaining to a population of failure probability values.

It is important to note that, depending on the application, it may be advisable to address within-constellation variability or focus on the aggregated effect. For example, an important HRA issue is to investigate PSF effects across different constellations of task and PSF categories. To this end, the effect of changes in one or more elements of the constellation on the HEP may be investigated by focusing on the aggregated effect, i.e. on the population average for the constellation, therefore adopting the typical beta-binomial model with lumped-data [28-30]. On the other hand, when the estimated HEP is used to inform a given constellation of an HRA model, adopting a variability model becomes important to capture the variability aspects (crew-to-crew, within-category) within the constellation and ideally allow for plant-specific HEP values. Indeed, for use in PSA, HEP values need to be plant- and scenario-specific: by lumping the effects of different tasks, scenarios, and crew characteristics in the population-average HEP, the lumped-data approaches [28-30] do not represent the intrinsic variability of the data sources. This can lead to overconfidence in HRA model results, as well as significant biases for plant-specific human error probabilities, as demonstrated in the numerical test with artificially-generated data performed in research task 1.3 (subsection 2.4.2). From this perspective, the proposed variability model is expected to foster the capabilities of future HRA models in treating specific variability aspects (e.g. plant-to-plant, scenario-to-scenario) in simulator data, according to data availability and the scope of the application (see Section 2.5 for further details).

The developed Bayesian variability model is intended for general application to any HRA model for HEP quantification. The currently available HRA models strongly differ in the task and factors considered and in the granularity of their definition. It can be expected that these aspects are strongly connected with the variability that the model shall be able to represent (see examples provided in Section 2.5). As a working hypothesis, it may be reasonable to assume that the coarser the granularity of the model (more macroscopic tasks), the larger the variability corresponding to the within-category variability. Also, the more the task involves decision-making and communication at the crew level, the more crew variability will be relevant, compared for example to execution-related tasks performed by single persons. Finally, it can be expected that variability would also be larger for HRA models with coarser PSF categories, e.g. binary as opposed to multivalued. With the current interest by the community on empirically estimated HEPs, it may be well important that future studies will address the extent to which variability shall be addressed as well as with the goal of develop guidelines to do it (see “Future works and recommendations”, Section 5.3).

1.3.2. A novel (model-based) approach to identify crew performance drivers from simulator data and empirically include their effects in HEP estimates

The methodology based on behavioral patterns produced by research tasks 2.1-2.4 (presented in Chapter 3) represents a novel modelling approach to empirically incorporate crew behavioral characteristics determining performance variability in the estimation of HEP values and bounds from simulator data, for various constellations of task and PSF categories (research objective #2). In the proposed approach, the HEP is still expressed as a function of task-, scenario-, and context-based factors (i.e. the constellation of task type and PSF levels/ratings), as in typical HRA models. On the other hand, crew performance variability is captured by different patterns of crew behavioral categories (in teamwork, decision-making and situation awareness) emerging from simulator observations. In this configuration, behavioral patterns are interpreted as manifestations of the overall spectrum of influences: task, scenario, context, as well as person, team and organizational ones. Therefore, similar to typical HRA quantification models, the HEP is expressed explicitly as a function of task-, scenario-, and context-based factors. Differently, in the proposed concept, crew performance variability is expressed via a model (based on behavioral differences across groups of crews) and estimated from empirical data, whereas in most other HRA models performance variability is not incorporated and not informed by data.

Through the structured use of teamwork, decision-making and situation awareness taxonomies, the proposed methodology supports the HRA analyst in the identification of relevant crew performance drivers driving performance variability and in their categorization via a finite set of behavioral patterns, as demonstrated in the application to case study performed in research task 2.4 (Section 3.4). Besides its use for HEP quantification, the methodology can be applied to highlight those crew behavioral patterns that favor lower failure probability values in a given accidental scenario and, accordingly, suggest safety-enhancing measures to nuclear power plant managers (e.g. support training of operators, implementation of new steps in procedural guidance). This implication acquires further importance for those constellations representative of emergency scenarios where, due to mismatches between the procedural guidance and the current situation, task dynamics diverge from normally-trained operational tasks: in such conditions, the behavioral characteristics of the crew play a key role during scenario progression and can determine large performance variability across the operating crews (as empirically proven in the recent HAMMLAB study [38-39]).

In addition, the modelling approach based on behavioral patterns could be used to support the incorporation of crew-to-crew variability aspects in future, advanced crew performance models (e.g. BBN-based models [16, 20, 32-34]), representing the complex relationships among the spectrum of performance influencing factors (task-, context-, team-, and person-based) and the HEP (this aspect is further elaborated in “Future works and recommendations”, Section 5.3).

1.3.3. A (multi-purpose) quantitative framework to systematically and traceably integrate data and judgment in HEP estimation

As discussed in Section 1.1, improving the traceability in the use of expert judgment (research objective #3) is an important requirement for using HRA results to inform regulatory and operational decisions: decisions with safety-related implications need to be based on an empirically sound basis. To this end, the quantitative framework produced by research tasks 3.1-3.4 (presented in Chapter 4) accomplishes a systematic and traceable integration of diverse information sources (simulator data, expert-elicited probability estimates and plant-specific failure data) throughout the HEP estimation process.

The developed quantitative framework has been structured as a two-stage Bayesian model, to address different purposes. On the one hand, the first stage of the model can produce reference HEP values and bounds to parametrize HRA models, in particular those constellations of task/PSF categories for which current availability of simulator data is still not sufficient to derive statistically significant information. In this regard, the systematic incorporation of judgment in the first stage (see Section 4.3) enables for continuous updates of the HEP estimates as new empirical evidence becomes available (e.g. from the long-running data collection programs [23-24]), progressively replacing the judgment-based information and hence reducing the subjective component in the reference HEP values and bounds underlying HRA models.

Similarly, the first stage can be used also to produce anchoring information for future HRA models (e.g. the CPDs in the emerging BBN-based approaches [16, 20, 32-34]): in this regard, the developed framework can contribute to the advancement in the empirical foundation of future HRA models (see "Future works and recommendations", Section 5.3). On the other hand, the combined use of data and judgment in the second stage of the Bayesian model can be used to improve the quality of HEP estimates (i.e. reduce the uncertainty on the estimated HEP values) for those human failure events characterized by scarce empirical observations, as demonstrated in the practical application performed in research task 3.4 (Section 4.4). Given its flexibility, the developed two-stage Bayesian model represents a valid, ready-to-use tool for the quantification of HEPs (and the associated uncertainty) in plant-specific PSA (e.g. to inform the HFEs of PSA event trees, as in the example provided in Appendix A). In this regard, the more transparent incorporation of judgment in the HEP estimation is expected to increase the acceptability of HRA results for use in risk-relevant applications.

1.3.4. Applicability to other sectors

Within the risk analysis field, the attention of the scientific community to Bayesian models is increasing. Therefore, the deliverables of this thesis are expected to receive large interest from the scientific community as well as HRA practitioners operating in sectors (e.g. industrial plants, experimental nuclear facilities, healthcare, safety of critical infrastructures) sharing similar modelling needs as those addressed by the research objectives of this Ph.D. work (Section 1.2), e.g.: the need to consider data variability in parameter estimation; the incorporation of expert judgment to compensate scarce empirical data. In principle, the modelling solutions proposed by this thesis can be transferable to other sectors, but require adaptation in order to meet sector-specific requirements and data availability. For instance, the proposed variability models (with continuous, Section 2.2, and discrete, Section 3.2, variability formulations) can be used in other risk analysis applications where the data sources are characterized by multiple layers of variability. For this purpose, the variability functions could be selected as to appropriately represent variability aspects relevant to the specific application (e.g. vendor-to-vendor, plant-to-plant variability for industrial components; patience-to-patience in healthcare HRA). Similarly, the proposed methodology based on behavioral patterns (Section 3.3) could be applied to other HRA sectors where crew behavioral characteristics play an important role in determining performance variability in operational tasks (e.g. resilience engineering, natural hazardous events in critical infrastructures). For instance, the lists of behavioral categories identified in the application of the methodology to case study (research task 2.4, in Section 3.4) contains subsets of categories that could be reasonably applicable to operational tasks other than in nuclear power plants. In particular, categories in "progress through procedures" and "flexibility in dealing with procedures and cues" can be relevant to domains where task performance is guided by procedures. In this regard, the taxonomy of teamwork competences [47] adopted for the application to case study (Section 3.4) specifically addressed metrics that are relevant to nuclear power plant operational tasks: thus, it may not suffice to cover the spectrum of person-, team-related metrics that are relevant to other sectors, with the set of identified categories likely requiring to be complemented with information from sector-specific human factor studies.

1.4 Outline of the thesis

The remainder of this dissertation is divided into four chapters, which are described in the following paragraphs.

Chapter 2 addresses the development of the Bayesian model to treat variability aspects (crew-to-crew, within-category) within the constellations of task and PSF categories of simulator data collection taxonomies. First, the chapter discusses how uncertainty and variability aspects in error probabilities have been addressed by existing HRA methods, with a focus on the analysis of simulator data (subsection 2.2.1). Then, the chapter presents the mathematical formulation of the HEP variability model, using continuous parametric distributions to capture crew-to-crew and within-category variability in the given constellation of task and PSF categories (subsection 2.2.2). The variability model is coupled to a Bayesian inference model, to derive the parameters of the variability distributions from simulator data relevant to the given constellation (Section 2.3). The chapter also addresses the verification of the developed model (configured with lognormal variability distributions) with artificially-generated data (subsection 2.4.1), and investigates model sensitivity to data availability in presence of different types of prior information on the parameters of the variability distributions (subsection 2.4.2). Lastly, the chapter presents the numerical application of the model to simulator datasets from literature, to demonstrate the effects of modelling variability on HEP estimates (subsection 2.4.3). The results from the case study, as well as the applicability of the model with respect to the ongoing data collection protocols, are further discussed at the end of the chapter (Section 2.5).

Chapter 3 presents the modelling approach based on behavioral patterns, focusing on the identification of crew performance drivers from simulator data and empirically incorporate their effects on the HEP variability distributions. The chapter first introduces the concept of behavioral patterns to categorize crew behavioral characteristics emerging from simulator observations and represent the spectrum of performance variability over a finite (“discrete”) set of crew behavioral groups, for a given constellation of task and PSF categories (Section 3.2). The chapter then presents the multi-step methodology to identify behavioral patterns from simulator data (subsection 3.3), and use these to inform the Bayesian hierarchical model developed to quantitatively capture performance variability across the identified set of groups (subsection 3.3.3). Finally, the chapter demonstrates the multi-step methodology on a case study from literature, involving crew behaviors observed in different emergency scenarios from recent simulator studies (Section 3.4). The chapter concludes discussing the benefits (as well as the limitations) of the proposed methodology with respect to HRA applications (Section 3.5).

Building on the Bayesian variability model for simulator data presented in Chapter 2, Chapter 4 addresses how formally integrate simulator data and expert judgment in the estimation of human error probabilities. The chapter first discusses how judgment (in the form of expert estimates on task failure probability) is mathematically combined with simulator data in the upgraded formulation of the HEP variability model with continuous parametric distributions (Section 4.2). Then, the chapter presents the development of the Bayesian two-stage model (Section 4.3), to systematically and traceably integrate data and judgment in the derivation of reference HEP values and variability bounds for various constellations of task and PSF categories (first stage), as well as in the estimation of failure probabilities for plant-specific tasks (second stage). The developed two-stage Bayesian model was first verified with artificially-generated evidence, to analyze the effects of judgment incorporation on HEP estimates and investigate model sensitivity to biases in expert judgment (subsection 4.3.4). Finally, the chapter discusses the application of the two-stage model to a collection of human failure events from the recent HRA Empirical Studies (Section 4.4). The insights from the application are further discussed at closure (Section 4.5).

Lastly, Chapter 5 presents the conclusions of this dissertation, summarizes the key achievements of the Ph.D. work and opens towards future works.

References

1. Kirwan B. *A guide to practical Human Reliability Assessment*. CRC press: Boca Raton, FL, USA, 1994.
2. Spurgin AJ. *Human Reliability Assessment – theory and practice*. CRC press: Boca Raton, FL, USA, 2010.
3. Podofillini L. Human Reliability Analysis. In: Moller N, Hansson SO, Holmberg JE, and Rollenhagen C. (eds) *Handbook of Safety Principles*. Wiley, 2017, pp.565-592.
4. Swain AD and Guttman HE. Handbook of human reliability analysis with emphasis on nuclear power plant applications. NUREG/CR-1278, U.S. Nuclear Regulatory Commission, Washington DC, USA, 1983.
5. Williams JC. HEART – A Proposed Method for Assessing and Reducing Human Error. In: *9th Advance in Reliability Technology Symposium*, University of Bradford, 1986.
6. Williams JC. A data-based method for assessing and reducing human error to improve operational performance. In: *Proceedings of the IEEE Fourth Conference on Human Factors and Power Plants*, Monterey, California, 5–9 June, pp. 436–450, 1988.
7. Williams JC. HEART – a proposed method for achieving high reliability in process operation by means of human factors engineering technology. *Saf Reliab* 2015, 35(3):5–25.
8. Hollnagel E. *Cognitive Reliability and Error Analysis Method* (*CREAM*). Oxford: Elsevier Science Ltd, 1998.
9. Gertman DI, Blackman HS, Marble JL, et al. The SPAR-H Human Reliability Analysis Method. NUREG/CR-6883, U.S. Nuclear Regulatory Commission, Washington DC, USA, 2005.
10. Whaley AM, Kelly DL, Boring RL, et al. *SPAR-H step-by-step guidance*. INL/EXT-10-18533, Idaho National Labs, Idaho Falls, Idaho 83415, 2011.
11. Xing J, Parry G, Presley M, et al. An Integrated Human Event Analysis System (IDHEAS) for Nuclear Power Plant Internal Events At-Power Application. NUREG-2199 Vol.1, U.S. Nuclear Regulatory Commission, Washington DC and Electric Power Research Institute, Palo Alto CA, USA, 2017.
12. Groth KM and Mosleh A. Deriving causal Bayesian networks from human reliability analysis data: a methodology and example mode. *Proc Inst Mech Eng*, *Pt O: J Risk Reliab* 2012, 226(4): 361-379.
13. Mkrtchyan L, Podofillini L and Dang VN. Bayesian belief networks for human reliability analysis: A review of applications and gaps. *Reliab Eng Syst Saf* 2015, 139:1-16.
14. Sundarmurthi R and Smidts C. Human reliability modelling for Next Generation System Code. *Ann Nucl Energy* 2013, 137-156.
15. Zhao Y and Smidts C. A method for systematically developing the knowledge base of reactor operators in nuclear power plants to support cognitive modeling of operator performance. *Reliab Eng Syst Saf* 2019, 186:64-77.
16. Shirley RB, Smidts C and Zhao Y. Development of a quantitative Bayesian network mapping objective factors to subjective performance shaping factor evaluations: An example using student operators in a digital nuclear power plant simulator. *Reliab Eng Syst Saf* 2020, 194:106416.
17. Hallbert B and Kolaczkowski A. The Employment of Empirical Data and Bayesian

Methods in Human Reliability Analysis: A Feasibility Study. NUREG/CR-6949, pp. 1-4, INL/EXT-06-11670, Washington, D.C.: U.S. Nuclear Regulatory Commission, 2007.

18. Mosleh A and Smith C. The Feasibility Of Employing Bayesian Techniques And Other Mathematical Formalisms In Human Reliability Analysis, in The Employment of Empirical Data and Bayesian Methods in Human Reliability Analysis: A Feasibility Study, NUREG/CR-6949, pp. 5-15, INL/EXT-06-11670, Washington, D.C.: U.S. Nuclear Regulatory Commission, 2007.
19. Mosleh A and Chang YH. Model-based human reliability analysis: prospects and requirements. *Reliab Eng Syst Saf* 2004, 83: 241–253.
20. Groth KM, Smith R and Moradi R. A hybrid algorithm for developing third generation HRA methods using simulator data, causal models, and cognitive science. *Reliab Eng Syst Saf* 2019, 191:106507.
21. Forester J, Dang VN, Bye A, et al. The International HRA Empirical Study Lessons Learned from Comparing HRA Methods Predictions to HAMMLAB Simulator Data. NUREG-2127, US Nuclear Regulatory Commission, Washington DC, USA, 2014.
22. Forester J, Liao H, Dang VN, et al. The US HRA Empirical Study - Assessment of HRA Method Predictions against Operating Crew Performance on a US Nuclear Power Plant Simulator. NUREG-2156, US Nuclear Regulatory Commission, Washington DC, USA, 2016.
23. Park J, Jung W, Kim S, et al. A guideline to collect HRA data in the simulator of nuclear power plants. KAERI/TR-5206, Korea Atomic Energy Research Institute, Republic of Korea, 2013.
24. Chang JY, Bley D, Criscione L, et al. The SACADA database for human reliability and human performance. *Reliab Eng Syst Saf* 2014, 125: 117-133.
25. Liao H, Forester J, Dang VN, et al. Assessment of HRA method predictions against operating crew performance: Part III: Conclusions and achievements. *Reliab Eng Syst Saf* 2019, 191: 106511.
26. Hallbert B, Morgan T, Hugo J, et al. A Formalized Approach for the Collection of HRA Data from Nuclear Power Plant Simulators. NUREG/CR-7163; INL/EXT-12-26327, US Nuclear Regulatory Commission, Washington DC, USA and Idaho National Laboratories, Idaho, USA, 2013.
27. Kim Y, Park J and Jung W. A classification scheme of erroneous behaviors for human error probability estimations based on simulator data. *Reliab Eng Syst Saf* 2017, 163: 1-13.
28. Groth KM, Smith CL, and Swiler LP. A Bayesian method for using simulator data to enhance human error probabilities assigned by existing HRA methods. *Reliab Eng Syst Saf* 2014, 128 (Supplement C): 32-40.
29. Azarm MA, Kim IS, Marks C, et al. Analyses methods and pilot applications of SACADA database. In: *14th Probabilistic Safety Assessment and Management, PSAM 14 2018*, UCLA Meyer & Renee Luskin Conference Center, Los Angeles, California.
30. Jung W, Park J, Kim Y, et al. HuREX – A framework of HRA data collection from simulators in nuclear power plants. *Reliab Eng Syst Saf* 2020, 194: 106235.
31. Chang JY and Franklin C. SACADA Data for HEP Estimates. In: *14th Probabilistic Safety Assessment and Management, PSAM 14 2018*, UCLA Meyer & Renee Luskin Conference Center, Los Angeles, California.

32. Nelson PF and Grantom CR. Methodology for Supporting the Determination of Human Error Probabilities from Simulator Sourced Data. In: *14th Reliability Safety Assessment and Management, PSAM 14 2018*, UCLA Meyer & Renee Luskin Conference Center, Los Angeles, California.
33. Groth KM. A framework for using SACADA to enhance the qualitative and quantitative basis of HRA. In: *14th Reliability Safety Assessment and Management, PSAM 14 2018*, UCLA Meyer & Renee Luskin Conference Center, Los Angeles, California.
34. Mkrtchyan L, Podofillini L and Dang VN. Methods for building Conditional Probability Tables of Bayesian Belief Networks from limited judgment: An evaluation for Human Reliability Application. *Reliab Eng Syst Saf* 2016, 151: 93-112.
35. Kim Y, Park J, Jung W, et al. Estimating the quantitative relation between PSFs and HEPs from full-scope simulator data. *Reliab Eng Syst Saf* 2018, 173: 12-22.
36. Siu NO and Kelly DL. Bayesian parameter estimation in probabilistic risk assessment. *Reliab Eng Syst Saf* 1998, 62(1): 89-116.
37. Kelly DL and Smith CL. *Bayesian Inference for Probabilistic Risk Assessment: A Practitioner's Guidebook*. London, UK: Springer-Verlag, 2011.
38. Massaiu S and Holmgren L. Diagnosis and Decision-Making with Emergency Operating Procedures in Non-Typical Conditions: A HAMMLAB Study with U.S. Operators. HWR-1121. Halden, Norway: OECD Halden Reactor Project, 2014.
39. Massaiu S and Holmgren L. The 2013 Resilient Procedure Use Study with Swedish Operators: Final Results. HWR-1216. Halden, Norway: OECD Halden Reactor Project, 2017.
40. Ekanem NJ, Mosleh A, and Shen SH. Phoenix – A model-based Human reliability analysis methodology: Qualitative analysis procedure. *Reliab Eng Syst Saf* 2015, 145: 301-315.
41. US Nuclear Regulatory Commission (USNRC). NRC's High-Value Datasets: Human Reliability Analysis, https://www.nrc.gov/data (2019, accessed 21 September 2020).
42. Mosleh A and Apostolakis G. The assessment of probability distributions from expert opinions with an application to seismic fragility curves. *Risk Anal* 1986, 6(4): 447-461.
43. Mosleh A. Bayesian modeling of expert-to-expert variability and dependence in estimating rare event frequencies. *Reliab Eng Syst Saf* 1992, 38(1-2).
44. Podofillini L and Dang VN. A Bayesian Approach to Treat Expert-Elicited Probabilities in Human Reliability Analysis Model Construction. *Reliab Eng Syst Saf* 2013, 117: 52-64.
45. Apostolakis G and Mosleh A. Expert Opinion and Statistical Evidence: An Application to Reactor Core Melt Frequency. *Nucl Sci Eng* 1979, 70(2):135-149.
46. Droguett EL, Groen F and Mosleh A. The combined use of data and expert estimates in population variability analysis. *Reliab Eng Syst Saf* 2004, 83(3): 311-321.
47. Skjerve AB and Holmgren L. An investigation of Teamwork Competence Requirements in Nuclear Power Plant Control-Room Crews across Operational States – a Field Study. HWR-1107. Halden, Norway: OECD Halden Reactor Project, 2016.
48. Lois E, Dang V, Forester J, et al. International HRA Empirical Study - Phase 1 Report: Description of Overall Approach and Pilot Phase Results from Comparing HRA Methods to Simulator Performance Data. NUREG/IA-0216 Vol. 1, U.S. Nuclear Regulatory Commission, Washington DC, USA, 2009.
49. Bye A, Lois E, Dang VN, et al. International HRA Empirical Study – Phase 2 Report:

Results from Comparing HRA Method Predictions to Simulator Data from SGTR Scenarios. NUREG/IA-0216 Vol. 2, US Nuclear Regulatory Commission, Washington DC, USA, 2011.
50. Dang VN, Forester J, Boring R, et al. International HRA Empirical Study – Phase 3 Report: Results from Comparing HRA Methods Predictions to HAMMLAB Simulator Data on LOFW Scenarios. NUREG/IA-0216 Vol 3, US Nuclear Regulatory Commission, Washington DC, USA, 2014.

Chapter 2: Bayesian variability model for simulator data

This chapter reproduces the author's article submitted to Reliability Engineering and System Safety journal (see "Publication details" below). The article presents a Bayesian model to treat variability aspects (crew-to-crew, within-category) within the constellations of task and PSF categories of simulator data collection taxonomies (research objective #1 in Section 1.2).

The chapter first discusses how uncertainty and variability aspects in error probabilities have been addressed by existing HRA methods, with a focus on the analysis of simulator data (research task 1.1, in Section 1.2). Then, the chapter presents the mathematical formulation of the HEP variability model, using continuous parametric distributions to capture crew-to-crew and within-category variability in data collection (research task 1.2). The variability model is coupled to a Bayesian inference model, with the goal to derive the parameters of the variability distributions from simulator observations relevant to the given constellation of taxonomy categories. The chapter also addresses the verification of the developed model (configured with lognormal variability distributions) with artificially-generated data, and investigates model sensitivity to data availability in presence of different types of prior information on the parameters of the variability distributions (research task 1.3). Lastly, the chapter presents the numerical application of the model to simulator datasets from literature, to demonstrate the effects of modelling variability on HEP estimates (research task 1.4). The results from the case study, as well as the applicability of the model with respect to the ongoing data collection protocols, are further discussed at the end of the chapter.

Publication details

This article is reproduced with permission from: **Greco SF**, Podofillini L and Dang VN. A Bayesian model to treat within-category and crew-to-crew variability in simulator data for Human Reliability Analysis. *Reliab Eng Syst Safe* 2021, 206:107309 (ISSN 0951-8320).

Additional information relevant to this chapter

- The numerical results from the sensitivity analysis are reported in Appendix B.
- The numerical results from the application to case study are reported in Appendix C.

Abstract

The models adopted in Human Reliability Analysis (HRA) characterize personnel tasks and performance conditions via categories of task and influencing factors (e.g. task types and Performance Shaping Factors, PSF). These categories cover the variability of the operational tasks and conditions affecting performance, and of the associated Human Error Probability (HEP). However, variability exists as well within such categories, for example because of the different scenarios and plants in which data is collected, as well as of the operating crew differences (within-category and crew-to-crew variability). This chapter presents a Bayesian model to mathematically aggregate simulator data to estimate failure probabilities, explicitly accounting for the specific tasks, scenarios, plants and crew behavior variability, within a given "constellation" (i.e. combination) of task and factor categories. The general aim of the proposed work is to provide future HRA with reference data with stronger empirical basis for failure probability values, both for their nominal values as well as for their variability and uncertainty. Numerical applications with both artificially-generated data and real simulator data are provided to demonstrate the effects of modelling variability in HEP estimates, to avoid potential overconfidence and biases. The applicability of the proposed model to ongoing simulator data collection programs is also investigated.

Nomenclature

E:	evidence of the Bayesian model, expressed as set of pairs {(k_{ij}, N_{ij}).
$\boldsymbol{F}$:	set of taxonomy categories (e.g. task type and PSF levels/ratings), referred as "constellation".
$f_{\boldsymbol{F}}(HEP)$:	parametric variability distribution, representing the overall spectrum of variability within a given constellation $\boldsymbol{F}$.
$f_{c\|t}(p_{c\|t}\|p_t^*, \boldsymbol{\theta}_{c\|t})$:	"crew-to-crew" variability term of $f_{\boldsymbol{F}}(HEP)$, modelling the variability across the crews performing the specific task/context realization (characterized by the crew-generic error probability value p_t^*) within the constellation $\boldsymbol{F}$.
$f_t(p_t\|\boldsymbol{\theta}_t)$:	"within-category" variability term of$f_{\boldsymbol{F}}(HEP)$, modelling the variability across the task/context realizations within the constellation $\boldsymbol{F}$.
$k_{\boldsymbol{F}}$, $N_{\boldsymbol{F}}$:	total number of failures and observations for the constellation $\boldsymbol{F}$ ("lumped data").
(k_{ij}, N_{ij}):	number of failures observed on N_{ij} repetitions of the i-th task performed by the j-th crew. $i = \{1, 2 \ldots, m\}, j = \{1, 2 \ldots, n\}$, where m: total number of tasks in the dataset; n: total number of crews performing the i-th task.
$L(E\|\boldsymbol{\theta})$:	likelihood function of the Bayesian model, i.e. the probability density that evidence E is observed.
$N(\ldots)$:	normal distribution
$p_{c\|t}$:	crew-specific HEP variable.
$P_{\boldsymbol{F}}(p_{c\|t})$:	estimated HEP variability distribution for the constellation $\boldsymbol{F}$.

p_t:	task-, context-specific HEP variable (crew-generic).
p_t^*:	specific numerical value (i.e. a realization) of p_t.
t:	index for the task/context realization within the constellation $\boldsymbol{F}$.
(z_t, $z_{c\|t}$):	normally-distributed auxiliary variables associated to p_t and $p_{c\|t}$.
(α, β):	shape parameters of the beta prior distributions.
$(\mu_t, \sigma_{\boldsymbol{F}})$:	parameters of the lognormal variability distribution (mean and standard deviation) used in the numerical application.
$\boldsymbol{\theta}_{\boldsymbol{F}}$:	set of (unknown) parameters of the variability distribution $f_{\boldsymbol{F}}(HEP)$.
$\boldsymbol{\theta}_t$:	set of (unknown) parameters of the within-category variability term (subset of $\boldsymbol{\theta}_{\boldsymbol{F}}$).
$\boldsymbol{\theta}_{c\|t}$:	set of (unknown) parameters of the crew-to-crew variability term (subset of $\boldsymbol{\theta}_{\boldsymbol{F}}$).
$\pi_0(\boldsymbol{\theta})$:	prior distribution of the Bayesian model, representing the knowledge on the set of parameters, (e.g. $\boldsymbol{\theta}_t$, $\boldsymbol{\theta}_{c\|t}$), before collecting the evidence *E*.
$\pi(\boldsymbol{\theta}\|E)$:	posterior distribution of the Bayesian model, representing the knowledge on the set of parameters, i.e. $\boldsymbol{\theta}_t$or $\boldsymbol{\theta}_{c\|t}$, after collecting the evidence *E*.

2.1. Introduction

Human Reliability Analysis (HRA) is the part of Probabilistic Safety Assessment (PSA) addressing the human contribution to the quantification of risk of complex technical systems, typically nuclear power plants, chemical and aerospace systems [1-2]. HRA aims to identify the safety-critical tasks performed by the personnel, to characterize the contextual factors influencing human performance, and to quantify the probability of failures.

To derive the human failure probability values (also referred to as Human Error Probabilities, HEPs), HRA methods characterize the personnel tasks and the factors deemed to influence task performance, the so-called Performance Shaping Factors (PSFs), e.g. adequacy of procedural guidance, of the human-machine interface, time available to accomplish the task, etc. HRA models characterize tasks and factors as categorical elements, with taxonomies and metrics dependent on the method. For instance, the Human Error Assessment and Reduction Technique (HEART [3-4], newly issued in [5]) identifies nine generic task types (e.g. “complex task requiring high level of comprehension and skill”) together with thirty-eight error producing conditions (e.g. “a low signal-noise ratio”). The Technique for Human Error Rate Prediction (THERP [6]) characterizes tasks at a lower level of decomposition (e.g. “set a rotary control to an incorrect setting”, “check/reading digital indicators”) and PSFs such as training and stress (e.g. “Very low”, “Optimum” stress). A similar use of categorical elements appears in all HRA methods, e.g.[7-9]. Recently, advanced models such as Bayesian Belief Networks (BBN) have been developed for HRA applications to capture the complex task, PSF, and HEP relationships and to enhance traceability in use of diverse data and judgment [10-11].

Reference data for the task categories and the PSF effects is needed to parametrize a method’s quantification model, both for traditional as well as for advanced models. The data is generally obtained by combining empirical data and expert judgment [12]. Since the early developments of HRA, empirical data has been mainly gathered from human factor studies, data collection campaigns in main control room simulators, retrospective analyses of accidents, near misses and operational events [1, 13]. An important turning point for HRA came from the International [14] and US HRA [15] Empirical Studies, aimed at assessing the validity of HRA method predictions against data from nuclear power plant main control room simulators. Besides improving HRA practice and methods, these studies resulted in methodological advances in the collection of simulator data for HRA purposes, with important implications on several recent activities [16-18]. Two notable, ongoing, data collection programs are the HUman REliability data Extraction framework, HuREX [16], and the Scenario Authoring, Characterization, And Debriefing Application, SACADA [17]: with their long-term data collection perspective, these are expected to produce a large amount of empirical evidence for new HRA reference data, more representative of recent operational conditions, e.g. reflecting modern interfaces and procedural guidance.

The majority of recent research activities dealing with the use of simulator data for HRA has addressed the development of protocols to collect data: notably, the interpretation of performance outcomes in terms of failure or success and the definition of the types of information on crew performance to collect [17, 19-20]. Open issues remain for how to use this information to quantify HEP values and how to eventually incorporate them into HRA methods, with various approaches being investigated [21-27].

Similarly to HRA methods, the data collection protocols characterize simulator observations through categories related to taxonomies of tasks (e.g. “entering step in procedure” in [16], failure mechanisms (e.g. “failure to prioritize” in [17]), contextual factors (e.g. “overloaded status of alarm board” in [17]), and the like. The data associated to these categories is collected from different simulator scenarios, different plants, from crews with different behavioral styles, and different realizations of the contextual factors. Research on quantification of HEP values from the emerging data is ongoing internationally. A number of pioneering works [21, 25, 27] have shown the advantages of Bayesian inference models in using the collected simulator data to quantify the HEP (and the associated uncertainty) for multiple “constellations” (i.e. combinations) of taxonomy categories, e.g. from the SACADA taxonomy [17]: macrocognitive function “understanding the situation/problem”, given the situational factor “information quality” with level “conflicting”. These works focused on the relationship between the given task, the set of PSFs and the error probability, and investigated performance variability in simulator tasks under different PSF effects, i.e. “across constellations”: with respect to the previous example, e.g. when the “information quality” is “misleading” instead of “conflicting”. However, variability in simulator data exists as well within task and PSF categories, i.e. “within the constellation", for instance, due to the different scenarios and plants in which data is collected as well as to operating crew differences (we refer to it as "within-category" and "crew-to-crew" variability, respectively). Such variability requires explicit consideration: the simple approach of lumping all data relevant to a given constellation of categories would focus on the "population average"-HEP of the constellation. However, it may not adequately represent the existing sources of variability, and may possibly lead to overconfident results [13, 28-29].

The present chapter proposes an inference model to derive HEP estimates from simulator data that explicitly addresses within-category and crew-to-crew variability aspects within a given constellation of task type and PSF categories. The first aspect stems from differences across simulator scenarios and plant-specific realizations of the contextual factors associated to the same categories; the latter from differences across the operating crews, e.g. different problem-solving styles, communication strategies, modality of information sharing, team coordination (e.g. tendency to prioritize tasks). The emerging simulator data is used to inform both the average HEP value as well as the associated variability bounds (hence, the focus on within-category and crew-to-crew variability). The main idea is to produce reference HEP values that can be used to inform HRA methods task type and PSF categories (or PSF multiplier values, depending on the method) as well as anchoring values for parametrizing advanced HRA models, such as BBNs. The parameters of the model are inferred via a Bayesian hierarchical framework, generally applicable to diverse taxonomies of task and PSF categories familiar to the HRA community. Because of the limited data available, most of the established HRA models (e.g. THERP [6]; the Standardized Plant Analysis Risk–Human reliability, SPAR-H [7-8]; the Cognitive Reliability and Error Analysis Method, CREAM [9]) assess data variability by expert judgment: as the running simulator campaigns will produce new data, it becomes important that data variability be formally incorporated in the HEP estimates, decreasing (and eventually replacing) the judgment.

The adoption of variability models is well established in PSA to consider source-to-source variability in parameter estimation problems: plant-to-plant variability in the estimation of component failure rates [30-31) and other reliability measures [32-33]; expert-to-expert variability in the estimation of rare event frequencies [34] and in HRA model construction [35]; combination of statistical data with expert estimates [36] and reliability data [37].

The chapter is structured as follows. Section 2.2 discusses uncertainty and variability aspects in HRA and in simulator data collection. Section 2.3 presents the developed Bayesian variability model and the underlying modelling assumptions. In Section 2.4, numerical applications with artificially generated data show the effects of modelling variability in HEP estimates and investigate the data requirements of the proposed model. In addition, an application to real simulator data from two different data sources (Halden project data from [21] and HuREX data from [27] is presented. The results are further discussed in Section 2.5, along with insights and recommendations on the applicability of the model. Conclusions are given at closure.

2.2. Uncertainty and variability aspects in HRA and simulator data for HRA

The results of HRA methods support risk-relevant decisions; an important requirement is to ensure that the uncertainties of HEP estimates are appropriately quantified [38]. The next subsections discuss how uncertainty and variability have been treated in existing HRA methods (subsection 2.2.1) and in the analysis of simulator data (subsection 2.2.2).

2.2.1. Treatment of uncertainty and variability in existing HRA methods

HRA quantification methods aim at representing the relationships between HEPs and PSFs, taking into account as well the interactions among PSFs. Tasks and contexts are typically characterized via constellations of categories (e.g. of task types and PSFs). As the constellation of these category changes, HRA models provide different Human Error Probability (HEP) estimates, representing the spectrum of performance conditions variability. The models produce estimated HEPs and characterize the uncertainty associated with these estimates, in the form of uncertainty distributions or bounds. For a given task type, a set of PSFs ratings yields a specific HEP distribution. Our work deals with the assessment of these distributions, which represent different aspects of uncertainty and variability [6, 39], as summarized in Table 2.1.

Depending on the methods, bounds and distributions are derived in different ways. As discussed in Chapter 7 of the Handbook [6], THERP assumes a lognormal distribution of the HEPs to account for the various sources of uncertainty and variability associated to HEP values (such as those listed in Table 2.1). For each failure included in its database, THERP provides a nominal HEP (the median of the uncertainty distribution) as well as an Error Factor (EF). These uncertainty bounds, exclusively derived by expert judgment, are meant to reflect the THERP's analysts "judgment regarding the likelihood of various values of HEPs" (from [6]) associated to a task. Different from THERP, HEART's HEP values and bounds are obtained by aggregating empirical evidence on human performances from diverse information sources in the human factor literature ([3-4], and the recently consolidated HEART version from [5]). In particular, for each generic task type, the author used the log-geometric mean of the set of data to derive the HEP central value and the log-standard deviation from the central value to calculate the HEP bounding values (in the form of 5th/95th percentiles). As a further example, the SPAR-H method adopts beta distributions (CNI, Constrained Non-Informative priors, by [40]) to determine uncertainty on HEP because the beta distribution can mimic both normal and lognormal distributions, with the advantage that it is defined from 0 to 1 [7]. As a general conclusion, except for HEART for which uncertainty is derived empirically, expert judgment is the dominant source for all other HRA methods.

2.2.2. Characterization of uncertainty and variability aspects in simulator data for HRA

The usefulness of simulator studies to inform human reliability models is recognized widely [12, 41-44], along with the need for the models to represent the variability of human performance in response to emergency conditions. For example, the Human Cognitive Reliability (HCR) model [45-46] and the Operator Reliability Experiment (ORE) [47] from the early 1980s were aimed at generating time reliability curves based on the variability of operating crew response time to emergency conditions, observed in simulator studies.

Table 2.1. Sources of uncertainty and variability in HEP estimates by HRA methods (given a constellation of task type and PSF categories). Note our work addresses the first two items of this table.

Source of uncertainty and variability	Description	Example
Crew characteristics	Inherent performance variability across people and crews, due to different behavioural characteristics, abilities, attitudes, etc.	Both crews A and B perform exactly the same task in the exact same context. Crew A fails, crew B succeeds. Also inherent randomness of certain human behaviour: same person/crew performs the same task under the same performance conditions: sometimes fails, sometimes succeeds.
Contextual factors	Variability (aleatory) across the different realizations of the contextual factors described by the same category of factor taxonomy	Variability within PSF “time pressure” due to variability in time and sequence of events within the same scenario (dynamic change). Variability within “indications of conditions” PSF due to different indications and/or designs, all can be characterized as “misleading”
Assessment of PSF ratings	Uncertainty on the assessed PSF states for the investigated context. Can also manifest as inter-analyst / rater variability.	It is not possible to state with certainty whether “time pressure” during performance should be considered “moderate” or “high”, due to inherent imprecision of contextual factor descriptions and different subjective interpretation of the PSF category
Model limitations	Uncertainty (epistemic) due to inherent, fundamental limitations of HRA models	Incompleteness of PSFs to represent a specific context of operations, limitation of underlying cognitive models to fully represent cognitive processes, lack of representation of safety culture, organizational and cross-organizational influences.
Scarcity of data	Uncertainty (epistemic) due to the limited knowledge of human performance in specific combinations of scenario/context of operation	Low-probability events (medium Loss Of Coolant Accident, with High Pressure Injection system failing to operate)

More recently, the International [14] and US [15] Empirical Studies were carried out to assess strengths and weaknesses of HRA methods, by comparing HRA predictions to observations of real operational crew responding to simulated accidents. Among various lessons learned, significant performance variability was observed. As a result of team dynamics, work processes, communication strategies, sense of urgency, and willingness to take knowledge-based actions, the observed performances differed not only in terms of the rate of progress through the procedures but also in terms of paths through the procedure or even the applied procedures. Subsequent studies on simulator data further analyzed the variability of crew strategies to make decisions and solve conflicts, especially in cases of complex simulated emergencies that involve non-typical conditions with multiple malfunctions [48-49]. These studies provided important insights on the characterization of crew performance, error identification and analysis, and characterizations of procedures and interfaces; capturing this variability is necessary for the design of HRA databases, as well as when analysing specific failure events [48-49].

Recently, two important simulator data collection initiatives have been initiated: SACADA [17] and HuREX [16]. In a similar way as HRA methods, these data collection protocols operate over taxonomies of categorical factors. SACADA characterizes the context via the “situational factors” (e.g. “information quality”, with the levels: “missing”, “misleading”, and “conflicting”), associated to high-level categories of individual and team cognitive functions (namely, “macrocognitive functions”, e.g. “monitoring/detecting”, “deciding/response planning”). Crew performance in a simulated scenario is evaluated according to a discrete rating classification (e.g. “satisfactory”, “unsatisfactory”, etc.) and the issues that negatively influenced the performance are classified in terms of both failure modes (e.g. “key alarms not detected or not responded to”) and error causes (e.g. “multiple simultaneous alarms”). Similarly to SACADA, the HuREX protocol classifies performance failures in simulator data collection (namely, “unsafe acts”) according to a categorical taxonomy based on cognitive activities (e.g. “situation interpreting”), generic task types (e.g. “measuring parameter - reading simple value”, “transferring step in procedure”), error modes (e.g. “error of commission”), and contextual information relevant to the simulated scenario (e.g. “procedure conformity”, “task familiarity”).

An example of collected data tailored to the SACADA taxonomy is given in Table 2.2 (note that the table reports only few elements of the rich SACADA context characterization). It considers hypothetical data collected on the task type “understanding the situation/problem”, where the alarm board of the Human Machine Interface (HMI) shows one status indication conflicting with critical alarms (“information quality: conflicting” in Table 2.2), the diagnosis of the latter being procedure-driven (“diagnosis basis: procedure” in Table 2.2). Table 2.2 includes failure/success data gathered in different plants (therefore with different HMIs, procedures, training programs) from different operating crews performing in two different simulated scenarios: for instance, 50 observations for Small Loss of Coolant Accident (SLOCA) scenario where the operators have to diagnose the SLOCA following a drop in pressurizer pressure, and 50 observations for Steam Generator Tube Rupture (SGTR) scenario where the operators have to diagnose the SGTR based on an anomalous variation of steam generator water level, given that in both situations a conflicting status indication is displayed (e.g. in SGTR scenario, “one level indication in steam generator stuck low” as in Table 2.2).

As the SACADA and HuREX databases are being populated, on-going research addresses the use of the collected data to inform HRA models. For example, reference [27] derives HEP values for the task categories addressed by the HuREX taxonomy [16], e.g. “directing manipulation”, “entering step in procedure”. Reference [22] uses logistic regression analysis to estimate the quantitative relationships between PSFs and HEP values from a set of 10000 HuREX observations. In all these works, the relevant HEP values are estimated via a Bayesian update (e.g. the conjugated beta-binomial model): the HEP value associated to each taxonomy category is modeled as a unique value (i.e. the HEP population average), to be estimated based on the simulator data evidence.

Table 2.2. Hypothetical simulator data used to inform the categorical elements of a generic HRA model for HEP estimation. Categorical elements taken from the SACADA taxonomy [17]. Note that the table reports only few elements of the rich SACADA context characterization.

Categorical elements of HRA models (“constellation *F*”):

Task type: understanding the situation/problem
Information quality: conflicting
Diagnosis basis: procedure

Data from specific simulator contexts

Scenario	**Realization of contextual factors**	**Task realization**	**Plant**	**Crews**	**Failures**
SGTR	One level indication in steam generator stuck low	Transfer to SGTR procedure	A	5	0
SGTR	One level indication in steam generator offset	Transfer to SGTR procedure	B	6	1
SGTR	One level indication in steam generator indicates zero	(…)	(…)	(…)	(…)
			Total	50	3
SLOCA	One indication on pressurizer pressure stuck high	Transfer to SLOCA procedure	A	5	1
SLOCA	One indication on pressurizer pressure indicates zero	Transfer to SLOCA procedure	B	6	2
SLOCA	Offset indication on pressurizer pressure	(…)	(…)	(…)	(…)
			Total	50	7

Returning to the example data in Table 2.2, when using this data to inform a quantitative HRA model on the considered “constellation” (i.e. combination) of task type and PSF ratings, the lumped-data approach would aggregate all observations as a single piece of evidence of 10 failures over 100 trials. This approach lumps together a number of variability aspects. Indeed, the dataset contains observations of tasks performed in different scenarios and different plants (e.g. “monitoring trend of steam generator level” in SGTR scenario, third column in Table 2.2), corresponding to different realizations of the associated task type (e.g. “understanding the situation/problem” in Table 2.2). The context of operation presents specificities that vary from plant to plant: in the example provided in Table 2.2, the HMI design of the alarm board in plant A is different from the one installed in plant B (e.g. different design and position of the alarms on screen; different number of simultaneous alarms); also, the specific procedural guidance and training program can vary between plant A and B. These plant-specific differences correspond to different realizations (second column in Table 2.2) of the associated contextual factors (e.g. “information quality: conflicting” and “diagnosis basis: procedure” in Table 2.2). Then, different crews are involved with crew-specific behavioral styles (e.g. different team dynamics, communication strategies, etc.).

A similar modelling approach with lumped data was adopted in a previous work by [21], where simulator observations from the US Empirical Study [15] were used in a Bayesian conjugate beta-binomial model with the goal to improve the reference HEP values of the SPAR-H method [7-8].

Concerning SACADA data, a number of feasibility studies have addressed the use of the collected data to inform HRA models [23-26], all based on variants of Bayesian approaches. Reference [25] proposes a multi-step methodology to identify critical situational factors for each macrocognitive function addressed by SACADA taxonomy [17] and uses a conjugate beta-binomial model to estimate HEP distributions for different combinations of these factors. Similarly to [21] and [27], the Bayesian estimates in [25] lump the data available for the relevant factor combination. Reference [25] acknowledges the presence of residual variability (e.g. plant-to-plant, crew-to-crew), but the authors average it out since the current amount of SACADA data does not allow a complete treatment of all sources of uncertainty. Other works adopt more advanced modelling techniques, specifically Bayesian Belief Networks (BBNs), to provide a richer characterization of the task, scenario, and context factors and of their relationships. Reference [26] uses BBNs to model the relationships between situational factors and error modes per each macrocognitive function of SACADA data collection taxonomy, and produce HEP estimates conditional on the set of situational factors. Reference [24] proposes a comprehensive framework combining SACADA data, taxonomies of performance influencing factors, causal BBNs, and Bayesian parameter updating to improve both the qualitative and quantitative basis of HRA models.

The BBN-based approaches [24, 26, 50] resort to a flexible framework to represent different variability aspects into the conditional probability distributions of the node categories and propagate this information through the BBN model. For instance, crew-to-crew variability nodes could be devised to explicitly represent the influence of different crew behavioral styles on the HEP. This calls for approaches to formally incorporate data variability (crew-to-crew, within-category) into the BBN conditional probability distributions. In this direction, the present work could support the development of empirically-based anchor information (i.e. reference HEP values and associated variability bounds) for multiple constellations of node categories of emerging BBN-based HRA models.

To summarize the above discussion, observations in simulator data collection (for a given constellation of task type and PSF categories) bring two aspects of variability into the HEP estimates: on one hand, the variability stemming from the different realizations of the associated constellation of factors of the HRA model (namely, "within-category" variability); on the other hand, the variability due to the different crew-specific features (namely, "crew-to-crew" variability). As formally presented in the next Section 2.3, modeling variability entails considering the evidence from different realizations and different crews as multiple pieces of evidence, pertaining to a population of failure probability values. Figure 2.1 illustrates the difference between the lumped (left) and the population variability (right) models with reference to the simplified data collection example of Table 2.2. It is important to note that in the lumped approach, the probability density function associated to the HEP value represents the uncertainty about the assumed unique value of the HEP itself (i.e. the population average). In the population variability approach, the function represents both the variability of the HEP value within the population and the uncertainty about the population parameters. For use in PSA, HEP values need to be plant- and scenario-specific; therefore, from Figure 2.1, focusing on the population average, the lumped approach may not represent the intrinsic variability of the sources.

2.3. A Bayesian variability model for simulator data

This section presents the mathematical model to account for the two variability aspects relevant for HRA data collection from simulators: within the categories of the data collection taxonomy and crew-to-crew. After discussing the underlying modelling assumptions (subsection 2.3.1), the variability model (subsection 2.3.2) is then coupled to a hierarchical Bayesian model (subsection 2.3.3) to infer from data on the parameter of the HEP variability distribution.

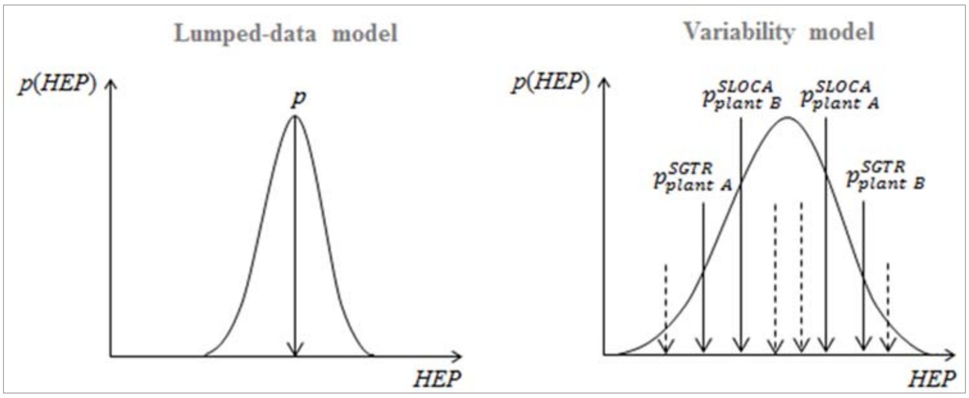


Figure 2.1. Simplified comparison between lumped-data and variability models (generic distributions shown). Left: probability density as uncertainty on the HEP population average (lumped-data model). Right: probability density as variability and uncertainty on HEP values variable by source (variability model) (given a constellation of task type and PSF categories).

2.3.1. Modelling assumptions

The idea is to build a general quantitative tool, able to mathematically aggregate simulator data from nuclear power plants to estimate failure probabilities (with their variability and uncertainty distributions), for constellations of categorical elements (e.g. task type, set of PSF ratings) of a data collection taxonomy (e.g. SACADA, HuREX). The quantity of interest for the developed model is the HEP value associated to the given constellation, $\boldsymbol{F} = \{F_1, F_2 \dots, F_\delta\}$:

$$HEP = f(F_1, F_2 \dots, F_\delta) \tag{2.1}$$

where $\boldsymbol{F}$ is the set of δ categorical elements used by the taxonomy to represent the simulator data record (e.g. in Table 2.2, F_1 represents the task type “understanding the situation/problem”, F_2 the PSF “information quality: conflicting”, and F_3 the PSF “diagnosis basis: procedure”). Each F_i can be expressed as a binary (e.g. present / not present; adequate / not adequate) or a multi-valued (e.g. rating) variable, depending on the particular taxonomy.

Evidence on human performance from simulator data may come in different forms, depending on the aims of the simulator program, its scope, and the intended use of the data. In this study, we focus on data from large-scale simulator programs, in the form of records of failure/successes, while operators perform tasks under a specific combination of PSF states.

The proposed inference model is intended for general application to any HRA model for HEP quantification (the applicability is further discussed in Section 2.5). The following list briefly restates the key terminology used in Sections 2.1-2.2, in order to support the understanding of model development in the remainder of this section:

- “categories”: refers to the taxonomy of task types and PSF levels adopted by the given data collection protocol (e.g. SACADA, HuREX) or HRA method. For instance, task type “diagnosis”, or PSF “time available” with level “barely adequate”;

- "constellation" (set $\boldsymbol{F}$ in this work): refers to a combination of the aforementioned categories, e.g. $\boldsymbol{F}$: {task type = diagnosis, with PSFs: "time available" = "barely adequate", "diagnosis basis" = "procedure-directed check", etc.}. Generally, HRA models provide HEP estimates as function of these constellations: accordingly, the goal of the proposed model is to infer the HEP uncertainty distribution for a given constellation, from simulator data;
- "within-category" variability: refers to variability aspects stemming from the different scenario-specific tasks associated to the same task type (e.g. different realizations of the category "diagnosis"), as well as from the different plant-specific operational contexts associated to the same set of PSF levels (e.g. different realizations of "barely adequate time" for PSF "time available"). Hence the term "within-category", since the same category (i.e. a task type or a PSF level) envelopes different realizations, according to the data collection protocol;
- "crew-to-crew" variability: refers to variability aspects stemming from the different behavioral characteristics (e.g. different problem-solving styles, communication strategies etc.) of the operating crews.

2.3.2. Variability model for HEP

The core of the variability model is the formulation of the HEP as an inherently variable quantity, represented by a probabilistic variability distribution (the population variability), $HEP \sim f_{\boldsymbol{F}}(HEP)$. The distribution function, $f_{\boldsymbol{F}}(HEP)$, is assumed known (e.g. lognormal) and reflects both variability aspects in HEP estimates discussed earlier: within-category as well as crew-to-crew variability. The quantity to infer from evidence is the set of (unknown) parameters of the variability distribution, as opposed to the "lumped-data" approach, where the unknown quantity is the unique HEP value (the population average).

The variability model, shown in Figure 2.2, is based on the following concepts:

- each realization of a constellation of categorical elements of the taxonomy is characterized by a unique HEP, p_t. With reference to Table 2.2, one such realization is the task of transferring to the SGTR procedure, in case one level indication in the steam generator is stuck low, following the procedures of plant A, for instance with associated HEP p_t^*. Basically, a realization defines the simulator scenario and the specific task to be performed by the crew. In this interpretation, Table 2.2 includes six realizations of the same constellation "understanding the situation/problem" in case of "conflicting information quality", associated to six different values of p_t^*. Different plants determine different realizations, because, although enveloped by the same constellation, the PSF manifestations may be different (different procedures, different HMI interfaces, and so forth). Variable p_t is continuous, distributed according to a known distribution f_t with vector of unknown parameters $\boldsymbol{\theta}_t$: $p_t \sim f_t(p_t|\boldsymbol{\theta}_t)$. p_t is intended as the failure probability to perform the specific task manifestation in the specific context manifestation, defined by the simulator run design (hence, the pedix *t*, for "task").

- Crew variability manifests as a crew-specific HEP variable $p_{c|t}$, that models the failure probability of a specific crew given the task performed in the specific simulator scenario, i.e. in a realization of the constellation $\boldsymbol{F}$ (e.g. from Table 2.2, the failure probability of one of the five crews from plant A performing the task "monitoring trend of steam generator level" in the corresponding SGTR scenario). It is assumed that the $p_{c|t}$ is a continuous variable distributed around each p_t^* according to a known distribution $f_{c|t}$ with unknown parameters $\boldsymbol{\theta}_{c|t}$: $p_{c|t} \sim f_{c|t}(p_{c|t}|p_t^*, \boldsymbol{\theta}_{c|t})$. Crew variability is modeled as variability of HEP values across different crews for the same task.

According to this formulation, the "HEP" variable in eq. (1) is represented by $p_{c|t}$, the probability of failure of a specific crew, given a specific task/context constellation.

Combining within-category and crew variability effects, the variability function $f_{\boldsymbol{F}}(HEP = p_{c|t})$ can be expressed as:

$$f_{\boldsymbol{F}}(p_{c|t}|\boldsymbol{\theta}_{\boldsymbol{F}}) = f_F(p_{c|t}|\boldsymbol{\theta}_t, \boldsymbol{\theta}_{c|t}) = \int f_t\,(p_t^*|\boldsymbol{\theta}_t) \cdot f_{c|t}(p_{c|t}|p_t^*, \boldsymbol{\theta}_{c|t})\; dp_t^* \tag{2.2}$$

where $\boldsymbol{\theta}_{\boldsymbol{F}} = (\boldsymbol{\theta}_t, \boldsymbol{\theta}_{c|t})$ is the vector of the unknown parameters of the overall HEP variability distribution.

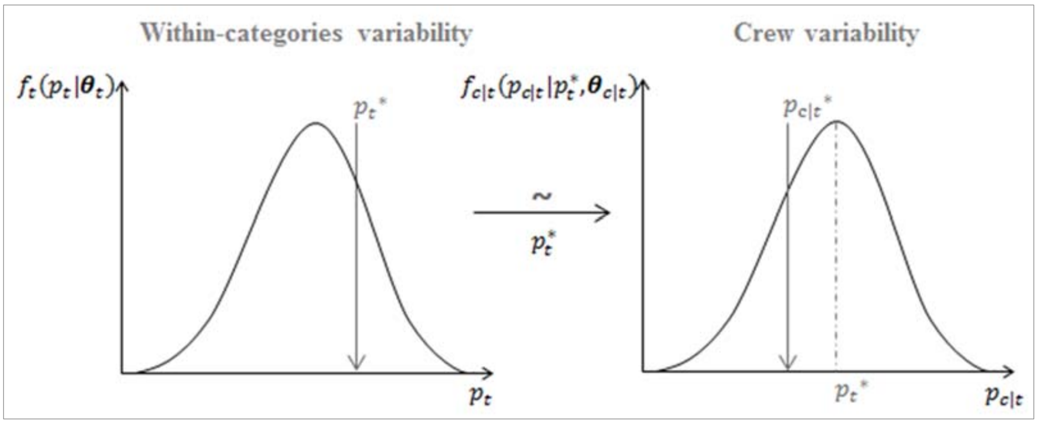


Figure 2.2. Sketch of the variability model (generic distributions shown). HEP represented by a population variability distribution, $f_{\boldsymbol{F}}(HEP|\boldsymbol{\theta}_{\boldsymbol{F}})$, combining variability within-category - ($f_t(p_t|\boldsymbol{\theta}_t)$, on the left - and crew-to-crew - $f_{c|t}\left(p_{c|t}|p_t^*, \boldsymbol{\theta}_{c|t}\right)$, on the right (see eq. 2.2). The crew-specific HEP variable, $p_{c|t}$, is distributed around the HEP value of a specific realization of the task and PSF constellation (p_t^* in right plot).

It is important to stress that the model considers $p_{c|t}$ as a crew-specific HEP value (given the specific task and context realization corresponding to the simulator run). This means that the model foresees that the crew performance of a task in response to a specific simulator run (e.g. one of the scenarios in Table 2.2) is not deterministic. The probability value $p_{c|t}$ associated to a specific crew represents two aspects. On the one hand, it represents the fact that it is not possible to exactly foresee the crew behavior because of the complexity of the factors involved and of intrinsic limitations of human performance models (i.e. "model limitations" in Table 2.1). On the other hand, it represents the intrinsic variability of human performance, even in presence of the same crew in response to the same simulator run (e.g. response times, level of attention, alertness of the same person/crew vary over time, "crew characteristic" in Table 2.1). These two aspects are presented separately to ease the discussion, but of course are closely linked: some crew characteristics are considered as aleatory because of model limitations to foresee them.

Both $p_{c|t}$ and the variability function in eq. 2.2 reflect the aleatory uncertainty elements from Table 2.1. Epistemic (state-of-knowledge) uncertainty comes in the uncertainty associated to the parameters of the variability distribution ($\boldsymbol{\theta}_{\boldsymbol{F}}$). Ideally, as more data is collected, $\boldsymbol{\theta}_{\boldsymbol{F}}$ would be progressively better estimated, with the epistemic component progressively decreasing, and consequently the expected $p_{c|t}$ distribution would get closer to the true (unique) HEP variability distribution for the constellation $\boldsymbol{F}$; the limiting case, with infinite data available, would be that the expected distribution only represents the inherent variability of the HEP. This aspect highlights a significant difference with the lumped approach, where a unique HEP (i.e. the population average) is the unknown parameter. In the lumped configuration, with increasing evidence, the uncertainty distribution will narrow to the unique estimate.

The hierarchical Bayesian model is implemented to update the analyst's degree of belief on the set $\boldsymbol{\theta}_{\boldsymbol{F}}$ and finally derive the estimated uncertainty distribution of $p_{c|t}$.

2.3.3. Development of the Bayesian inference model

Figure 2.3 gives an overview of the hierarchical Bayesian model. The general structure of the model is based on the formulation of the Bayes theorem as follows [39, 51]:

$$\pi(\boldsymbol{\theta}|E) = A^{-1}L(E|\boldsymbol{\theta})\pi_0(\boldsymbol{\theta}) \tag{2.3}$$

where:

- $\boldsymbol{\theta}$ is the set of unknown parameters of the inference problem;
- π_0 and π are the prior and posterior probability functions for $\boldsymbol{\theta}$, modelling the state of knowledge of the analyst on the set of investigated parameters respectively before and after the evidence E is collected (top level in Figure 2.3);

- $L(E|\boldsymbol{\theta})$ is the likelihood term, interpreted as the probability density that the evidence is observed (second and third levels in Figure 2.3);
- E is the set of evidence from the available information sources (bottom level in Figure 2.3);
- $A^{-1} = \int L(E|\boldsymbol{\theta})\pi_0(\boldsymbol{\theta})d\boldsymbol{\theta}$, the denominator of eq. 2.3, normalizes function π to a probability density function.

For the variability model in subsection 2.3.1, $\boldsymbol{\theta_F} = (\boldsymbol{\theta}_t, \boldsymbol{\theta}_{c|t})$ is the set of unknown parameters of the parametric variability function $f_{\boldsymbol{F}}(p_{c|t}|\boldsymbol{\theta}_t, \boldsymbol{\theta}_{c|t})$.

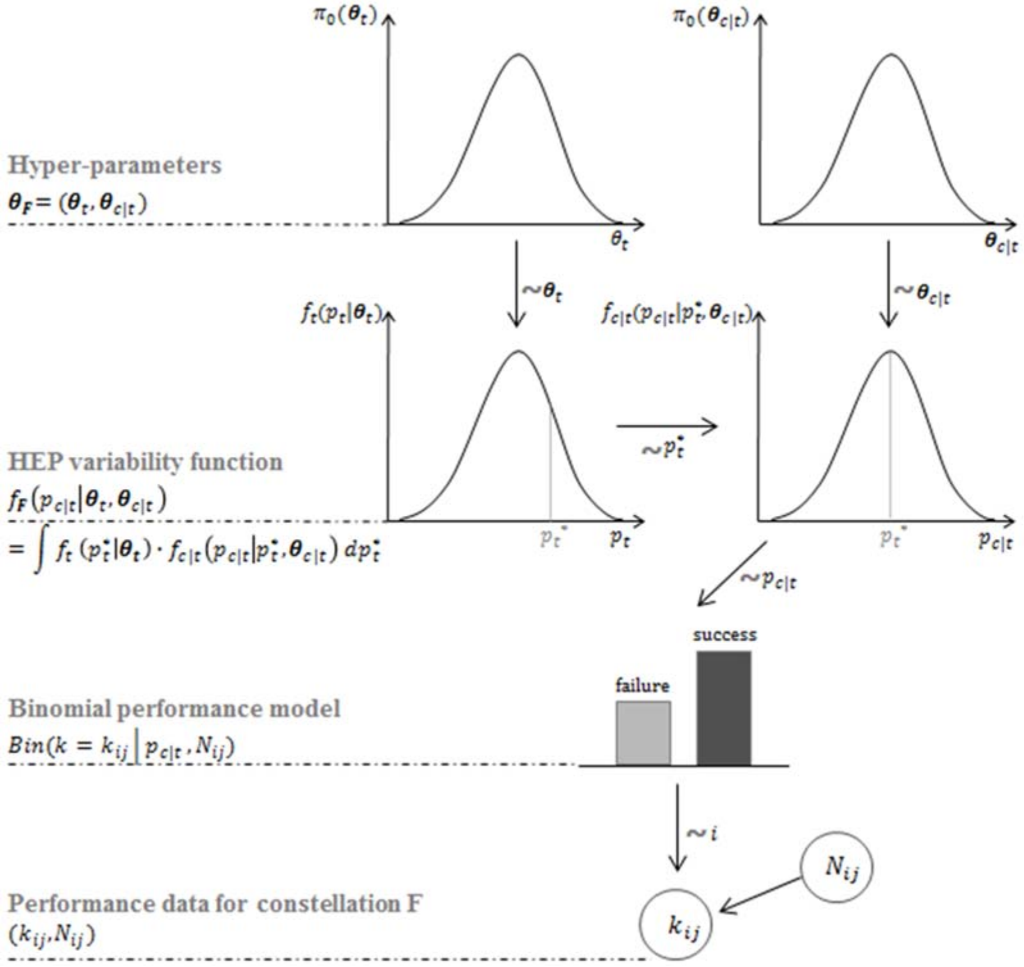

$\pi_0(\boldsymbol{\theta}_t)$
$\pi_0(\boldsymbol{\theta}_{c|t})$
Hyper-parameters
$\boldsymbol{\theta_F} = (\boldsymbol{\theta}_t, \boldsymbol{\theta}_{c|t})$
$\boldsymbol{\theta}_t$
$\boldsymbol{\theta}_{c|t}$
$\sim\boldsymbol{\theta}_t$
$\sim\boldsymbol{\theta}_{c|t}$
$f_t(p_t|\boldsymbol{\theta}_t)$
$f_{c|t}(p_{c|t}|p_t^*, \boldsymbol{\theta}_{c|t})$
$\sim p_t^*$
HEP variability function
$f_F(p_{c|t}|\boldsymbol{\theta}_t, \boldsymbol{\theta}_{c|t}) = \int f_t(p_t^*|\boldsymbol{\theta}_t)\cdot f_{c|t}(p_{c|t}|p_t^*, \boldsymbol{\theta}_{c|t})\, dp_t^*$
p_t^*
p_t
p_t^*
$p_{c|t}$
$\sim p_{c|t}$
success
failure
Binomial performance model
$Bin(k = k_{ij}\,|\,p_{c|t}, N_{ij})$
$\sim i$
N_{ij}
Performance data for constellation F
(k_{ij}, N_{ij})
k_{ij}

Figure 2.3. The Bayesian hierarchical variability model, from top to bottom: $\pi_0(\boldsymbol{\theta_F})$, prior distributions for model parameters $(\boldsymbol{\theta_F})$; $f_{\boldsymbol{F}}(p_{c|t}|\boldsymbol{\theta}_t, \boldsymbol{\theta}_{c|t})$, the HEP variability distribution, where $f_t(p_t|\boldsymbol{\theta}_t)$ models within-category variability and $f_{c|t}(p_{c|t}|p_t^*, \boldsymbol{\theta}_{c|t})$ models crew-to-crew variability; $Bin(k = k_{ij}|p_{c|t}, N_{ij})$, the binomial distribution of evidence of k_{ij} failures on N_{ij} repetitions of the i-th task by the j-th crew. Generic distributions shown.

Empirical evidence comes in the form of failure data (i.e. number of failures on number of task repetitions) collected on crew performance on simulator scenarios characterized by the same constellation $\boldsymbol{F}$. It is assumed that data was collected concerning *m* different task/context realizations within constellation $\boldsymbol{F}$, and n_i crews that performed the *i-th* task. Evidence *E* is represented as the set of pairs $\{(k_{ij}, N_{ij}), i = 1, 2 \ldots, m, j = 1, 2 \ldots, n)\}$, where k_{ij} is the number of failures observed on N_{ij} repetitions of the *i-th* task performed by the *j-th* crew (Figure 2.4, left, columns "Repetitions" and "Failures"). This type of datasets enters the likelihood term of the Bayesian model as evidence to update the prior degree of belief of the analyst on the parameters of the HEP variability model for the constellation $\boldsymbol{F}$ (Figure 2.4, right). Note that in the numerical examples, k_{ij} is set equal to 1 (see Figure 2.4, column "Repetitions"), recognizing that it would be very difficult to aggregate performances on the exact same task by the exact same crew (this aspect will be further discussed in Section 2.4 and in Section 2.5).

The construction of the likelihood term $L(E|\boldsymbol{\vartheta}_t, \boldsymbol{\vartheta}_{c|t})$ requires to express the probability of observing k_{ij} failures on N_{ij} repetitions of the specific *i-th* task. For the generic piece of simulator evidence, (k_{ij}, N_{ij}), the likelihood term can be written as:

$$L_{ij}(k_{ij}\ |\boldsymbol{\theta}_t, \boldsymbol{\theta}_{c|t}, N_{ij}) = \int_{p_{c|t}} f_{\boldsymbol{F}}(p_{c|t}|\boldsymbol{\theta}_t, \boldsymbol{\theta}_{c|t}) Bin(k = k_{ij}|p_{c|t}, N_{ij}) dp_{c|t} \tag{2.4}$$

By substituting eq. 2.2 into eq. 2.4, the likelihood term becomes:

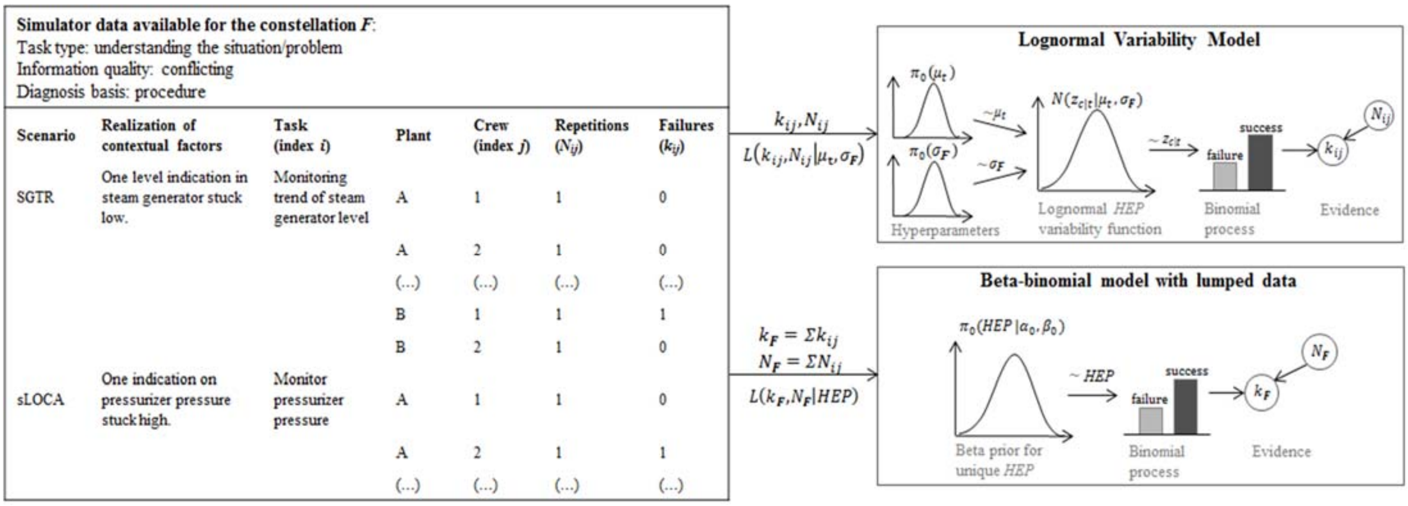
$$L_{ij}(k_{ij}\ |\boldsymbol{\theta}_t, \boldsymbol{\theta}_{c|t}, N_{ij}) =$$

$$= \int_{p_t}\int_{p_{c|t}} f_t(p_t|\boldsymbol{\theta}_t)\ f_{c|t}(p_{c|t}|p_t, \boldsymbol{\theta}_{c|t}) Bin(k = k_{ij}|p_{c|t}, N_{ij}) dp_{c|t} dp_t \tag{2.5}$$

where:

- the probability density that the failure probability of the *i-th* specific task is p_t, i.e. p_t is one realization of the possible within-category variability, is modeled by $\mathrm{f_t}(p_t|\boldsymbol{\theta}_\mathrm{t})$;
- the probability density that the crew-specific HEP value would manifest as $p_{c|t}$ (i.e. one realization of the possible crew-to-crew variability) is modeled by $\mathrm{f_{c|t}}(p_{c|t}|p_t, \boldsymbol{\theta}_{\mathrm{c|t}})$. The task-specific HEP, p_t, constitutes the reference probability value around which $p_{c|t}$ is distributed;
- the probability of observing k_{ij} failures in N_{ij} repetitions of the *i-th* task if the failure probability for the single repetition $p_{c|t}$ is described by the binomial distribution $Bin(k = k_{ij}|p_{c|t}, N_{ij})$.

Each probability value p_t and $p_{c|t}$ is one possible value within their variability; therefore, the expression $f_t(p_t|\boldsymbol{\theta}_t) f_{c|t}(p_{c|t}|p_t, \boldsymbol{\theta}_{c|t}) Bin(k = k_{ij}|p_{c|t}, N_{ij})$ is averaged (integrated) on the variability distributions for p_t and $p_{c|t}$.



Simulator data available for the constellation *F*:
Task type: understanding the situation/problem
Information quality: conflicting
Diagnosis basis: procedure

Scenario	**Realization of contextual factors**	**Task (index *i*)**	**Plant**	**Crew (index *j*)**	**Repetitions (N_{ij})**	**Failures (k_{ij})**
SGTR	One level indication in steam generator stuck low.	Monitoring trend of steam generator level	A	1	1	0
			A	2	1	0
			(...)	(...)	(...)	(...)
			B	1	1	1
			B	2	1	0
sLOCA	One indication on pressurizer pressure stuck high.	Monitor pressurizer pressure	A	1	1	0
			A	2	1	1
			(...)	(...)	(...)	(...)

Figure 2.4. Overall aggregation framework to compare the variability and the lumped data models. Left: artificial data for the constellation ***F*** based on the example in Table 2.2. Top right: lognormal variability model, informed by the crew-specific data points (k_{ij}, N_{ij}) and returning as output the posteriors for the HEP variability distribution parameters, i.e. μ_t and $\sigma_{\boldsymbol{F}}$. Bottom right: conjugated beta-binomial model with lumped data ($k_{\boldsymbol{F}}$, $N_{\boldsymbol{F}}$), giving as output the posterior distribution for the single-value HEP (population average).

When the *i-th* task is performed by n_i crews, the evidence takes the form of the number of failures observed for each crew: (k_{i1}, N_{i1}), (k_{i2}, N_{i2}), … (k_{in_i}, N_{in_i}). The likelihood term L_i relevant to the *i-th* task becomes:

$$L_i(k_{i1}, k_{i2}, \dots, k_{in_i} | \boldsymbol{\theta}_t, \boldsymbol{\theta}_{c|t}, N_{i1}, N_{i2}, \dots, N_{in_i}) =$$

$$= \int_{p_t} f_t(p_t | \boldsymbol{\theta}_t) \prod_{j=1}^{n_i} \int_{p_{c|t}} f_{c|t}(p_{c|t} | p_t^*, \boldsymbol{\theta}_{c|t}) \, Bin(k = k_{ij} | p_{c|t}, N_{ij}) dp_{c|t} \, dp_t \tag{2.6}$$

Note that in the expression above the probability density of observing the evidence (k_{i1}, N_{i1}), (k_{i2}, N_{i2}), … (k_{in_i}, N_{in_i}) given the within-category reference probability p_t is written as:

$$\prod_{j=1}^{n_i} \int_{p_{c|t}} f_{c|t}(p_{c|t} | p_t, \boldsymbol{\theta}_{c|t}) \, Bin(k = k_{ij} | p_{c|t}, N_{ij}) dp_{c|t}$$

Since all crews are carrying out the same specific task, the crew-to-crew variability effect is expressed for all crews conditional on the same reference HEP value, p_t. Then, the probability density of observing each (k_{ij}, N_{ij}) is multiplied because, given p_t, each crew's behavior is independent (the effect of the PSFs common for all crews is represented in the variable p_t).

Extending eq. 2.6 to the entire set of *m* task realizations in the constellation ***F***, the likelihood term is then:

$$L(E|\boldsymbol{\theta}_t,\boldsymbol{\theta}_{c|t}) = L(k_{ij}, i=1,\dots,m; j=1,\dots,n_j\ |\boldsymbol{\theta}_t,\boldsymbol{\theta}_{c|t},N_{ij}) =$$

$$= \prod_{i=1}^{m} L_i(k_{ij}, j=1,\dots,n_j\ |\boldsymbol{\theta}_t,\boldsymbol{\theta}_{c|t},N_{ij}) =$$

$$= \prod_{i=1}^{m} \int_{p_t} f_t(p_t|\boldsymbol{\theta}_t) \prod_{j=1}^{n_i} \int_{p_{c|t}} f_{c|t}(p_{c|t}|p_t,\boldsymbol{\theta}_{c|t})\, Bin(k=k_{ij}|p_{c|t},N_{ij}) dp_{c|t}\, dp_t \tag{2.7}$$

Eq. 2.7 assumes that the failure observations across the different tasks are independent. This implies that crew variability effects on the crew-specific HEP variable, $p_{c|t}$, do not replicate across different tasks: in other words, no systematic effects of crew under-performance (i.e. crew-specific HEP value consistently above average) or over-performance (i.e. crew-specific HEP value consistently below average) are modeled.

The posterior degree of belief on the unknown parameters of the HEP variability distribution for a generic constellation $\boldsymbol{F}$ of task and PSF categories is then expressed as follows:

$$\pi(\boldsymbol{\theta}_t,\boldsymbol{\theta}_{c|t}|E) = \frac{L(E|\boldsymbol{\theta}_t,\boldsymbol{\theta}_{c|t})\pi_0(\boldsymbol{\theta}_t,\boldsymbol{\theta}_{c|t})}{\iint L(E|\boldsymbol{\theta}_t,\boldsymbol{\theta}_{c|t})\pi_0(\boldsymbol{\theta}_t,\boldsymbol{\theta}_{c|t}) d\boldsymbol{\theta}_t\, d\boldsymbol{\theta}_{c|t}} \tag{2.8}$$

where the final formulation can be derived by substituting the likelihood term of eq. 2.7 in eq. 2.8.

The posterior probability distribution of eq. 2.8 can be subsequently used to compute the estimated HEP variability distribution for the constellation $\boldsymbol{F}$, $P_{\boldsymbol{F}}(p_{c|t})$:

$$P_{\boldsymbol{F}}(p_{c|t}) = \int_{\boldsymbol{\theta}_{\boldsymbol{F}}} f_{\boldsymbol{F}}(p_{c|t}|\boldsymbol{\theta}_{\boldsymbol{F}})\pi(\boldsymbol{\theta}_{\boldsymbol{F}}|E) d\boldsymbol{\theta}_{\boldsymbol{F}} =$$

$$= \int_{\boldsymbol{\theta}_t}\int_{\boldsymbol{\theta}_{c|t}}\int_{\boldsymbol{p}_t} f_t(p_t|\boldsymbol{\theta}_t)\, f_{c|t}(p_{c|t}|p_t,\boldsymbol{\theta}_{c|t})\pi(\boldsymbol{\theta}_t,\boldsymbol{\theta}_{c|t}|E) dp_t d\boldsymbol{\theta}_{c|t} d\boldsymbol{\theta}_t \tag{2.9}$$

Formally, $P_{\boldsymbol{F}}(p_{c|t})$ is derived by weighting the parametric distribution, adopted as variability model for HEP, by the posterior distribution of the unknown HEP distribution parameters computed by the Bayesian model.

Within this mathematical framework, the incorporation of further empirical evidence can be accomplished in subsequent steps in a traceable and reproducible way. This feature is of key importance, considering that data collection process from simulators is a long-term program. Indeed, the posterior distributions of HEP computed by the model can be used as prior state of knowledge in future analyses and then updated as new observations become available.

Finally note that the "lumped-data" approaches, e.g. of [21, 27], entail aggregating the evidence to inform a unique HEP value for the constellation $\boldsymbol{F}$ (i.e. the population average), i.e.:

$$k_{\boldsymbol{F}} = \sum_{i=1}^{m}\sum_{j=1}^{n} k_{ij}\,,\ N_{\boldsymbol{F}} = \sum_{i=1}^{m}\sum_{j=1}^{n} N_{ij} \tag{2.10}$$

where $k_{\boldsymbol{F}}$ and $N_{\boldsymbol{F}}$ are respectively the total number of failures and observations aggregated for the constellation $\boldsymbol{F}$ (Figure 2.4, bottom right). In references [21, 27], the pair ($k_{\boldsymbol{F}}$, $N_{\boldsymbol{F}}$) enters a conjugate beta-binomial model to update the prior state of knowledge on the population-average HEP, represented by a beta distribution with shape parameters α_0 and β_0. The update with lumped-data,

$$\alpha = \alpha_0 + k_{\boldsymbol{F}}, \beta = \beta_0 + N_{\boldsymbol{F}} - k_{\boldsymbol{F}} \tag{2.11}$$

yields the posterior distribution of the beta-binomial model (again a beta distribution, with parameters α and β), representing the final uncertainty on the population-average HEP.

2.3.4. Use of lognormal probability density functions to represent variability

This subsection presents the model in case lognormal distributions are used to represent both variability terms in eq. 2.2, within-category and crew variability, f_t and $f_{c|t}$, respectively (Figure 2.4, top right) – this configuration will be used in the applications in Section 2.4. The adoption of lognormal functions as population variability curves has been a common practice when developing hierarchical Bayesian models for PSA applications [28, 30, 52].

Considering a generic constellation of categorical elements $\boldsymbol{F}$, in this configuration both variability terms embodied in $f_F(p_{c|t}|\boldsymbol{\vartheta_F})$ as in eq. 2.2 (within-category and crew-to-crew variability) are distributed accordingly to lognormal probability density functions, therefore:

$$ln(p_t) = z_t \sim N(z_t|\mu_t, \sigma_t);\ ln(p_{c|t}) = z_{c|t} \sim N(z_{c|t}|z_t, \sigma_{c|t}) \tag{2.12}$$

where z_t and $z_{c|t}$ are the normally-distributed auxiliary variables associated to p_t and $p_{c|t}$, respectively (the letter *N* is used in eqs. 2.12-2.14 and Figure 2.4 to denote normal distributions). In this case, the set of unknown parameters to be determined by the Bayesian inference model is then $\boldsymbol{\theta_F} = (\boldsymbol{\theta}_t, \boldsymbol{\theta}_{c|t}) = (\mu_t, \sigma_t, \sigma_{c|t})$. Subsequently, the likelihood term for the generic piece of simulator evidence (eq. 2.5) can be expressed as follows:

$$L_{ij}(k_{ij}\ |\mu_t, \sigma_t, \sigma_{c|t}, N_{ij}) =$$

$$= \int_{z_t}\int_{z_{c|t}} N(z_t|\mu_t, \sigma_t)\ N(z_{c|t}|z_t, \sigma_{c|t})\ Bin(k = k_{ij}|e^{z_{c|t}}, N_{ij}) dz_{c|t} dz_t \tag{2.13}$$

Rearranging the right-side member of the equation:

$$L_{ij}(k_{ij}\ |\mu_t, \sigma_t, \sigma_{c|t}, N_{ij}) =$$

$$=\ \int_{z_{c|t}} Bin(k = k_{ij}|e^{z_{c|t}}, N_{ij})\ (\int_{z_t} N(z_t|\mu_t, \sigma_t)\ N(z_{c|t}|z_t, \sigma_{c|t}) dz_t)\ dz_{c|t} =$$

$$= \int_{z_{c|t}} Bin(k = k_{ij}|e^{z_{c|t}}, N_{ij}) N(z_{c|t}|\mu_t, (\sigma_t^2 + \sigma_{c|t}^2)^{1/2}) dz_{c|t} =$$

$$= \int_{z_{c|t}} Bin(k = k_{ij}|e^{z_{c|t}}, N_{ij})\ N(z_{c|t}|\mu_t, \sigma_{\boldsymbol{F}}) dz_{c|t} \tag{2.14}$$

The last relationship exploits the fact that the convolution of the two normal distributions of z_t and $z_{c|t}$ is again a normal distribution, with mean μ_t and standard deviation $\sigma_{\boldsymbol{F}} = (\sigma_t^2 + \sigma_{c|t}^2)^{1/2}$. According to eq. 2.14, the final set of unknown parameters for the inference problem becomes $\boldsymbol{\theta_F} = (\mu_t, \sigma_{\boldsymbol{F}})$, which respectively represent the mean and the standard deviation of the HEP variability distribution in the logarithmic space. The extension of eq. 2.14 to the entire set of simulated observations relevant to $\boldsymbol{F}$ (see eqs. 2.6-2.7), as well as the specialization of the posterior formula to the new set of unknown parameters (see eq. 2.8), are done as in subsection 2.3.2.

The last step of the Bayesian model development entails the definition of appropriate prior distributions for the parameters of the lognormal variability model, namely $\pi_0(\mu_t)$ and $\pi_0(\sigma_{\boldsymbol{F}})$ (usually referred to in Bayesian literature as the "hyper-priors" of a hierarchical model, see [51]). In the model application presented in Section 2.4, both diffuse and informative priors are used for the hyper-parameters of the Bayesian model, μ_t and $\sigma_{\boldsymbol{F}}$. For the case of diffuse priors, as suggested in reference [52] for lognormal variability distributions in lack of information, uniform distributions are adopted for both the natural logarithm of the mean, $\pi_0(log(\mu_t))$, defined between natural log(1E-5) and 0 (corresponding to the upper limit HEP= 1), and the standard deviation, $\pi_0(\sigma_{\boldsymbol{F}})$, defined between 0.1 and 4 (corresponding to error factors of 1.18 and 720.54, respectively). These ranges have been defined to cover values of interest for HRA applications. More information on the development of proper prior distributions can be found in literature [28, 51].

For all applications, an algorithm has been developed for the R programming environment [53] for the numerical solution of the various equations. The developed R code is available on request to the authors.

2.4. Numerical application

After a first comparison of the proposed variability model with a lumped data model (subsection 2.4.1), the present section addresses the model sensitivity to data availability, both in presence of diffuse (subsection 2.4.2.1), as well as of informed priors (subsection 2.4.2.2). Artificial data is used, i.e. data generated with known characteristics (e.g. median, mean, percentiles of the underlying data distributions): this allows investigating the Bayesian update process, for which the known values become target values. An application to simulator data from literature [21, 27] is presented later (subsection 2.4.3).

Concerning the generated data, two cases of target HEP variability distribution are considered, both lognormal:

- Case 1: median = 5e-2, mean = 5.46e-2, and error factor = 2

- Case 2: median = 5e-3, mean = 6.25e-3, and error factor = 3

The two cases represent HEP ranges of practical interest for HRA, with relatively high (Case 1) and moderate (Case 2) HEP values. The case of lower HEP values (e.g. median 5e-4 or lower) is not considered in this chapter because, as it will become clear later in the result presentation, the use of the proposed model would require a very large amount of simulator data, of questionable practicality.

Each data element is generated by first sampling a possible HEP value from the variability distribution for Case 1 or 2. Recalling from subsection 2.3.2, this HEP value is crew-specific. Then, the realization of the number of observed failures, k_{ij}, on N_{ij} repetitions (by the same crew) is sampled from a Binomial distribution, obtaining the data element (k_{ij}, N_{ij}). Different couples (k_{ij}, N_{ij}) are generated from different HEP values, based on the total number of task realizations relevant to the constellation *F* assumed to be available from the simulator data collection (referred as N_F in subsection 2.3), and constitute the evidence against which the variability model has been tested. For the applications in this chapter, N_{ij} is set to 1: each crew performs the same task only once in the dataset. This corresponds to the lowest possible amount of information on the variability in HEP. Ideally, as simulator data is accumulated over the years, evidence on multiple repetitions may be available (for example some simulator scenarios are trained recurrently by the same crew). This aspect will be returned to in the discussion. To investigate the data requirements, different sample sizes are considered, from relatively small sets (e.g. $\boldsymbol{N_F}$ = 10÷50) to larger sets (e.g. $\boldsymbol{N_F}$ = 200÷1000), to reflect possibly different data availability in the long-term. Note that while N_{ij} refers to crew-specific evidence, $\boldsymbol{N_F}$ refers to the whole data accumulated for the constellation $\boldsymbol{F}$ from different plants, crews, as well as realizations of the task types and PSFs defined by $\boldsymbol{F}$: this justifies the possibility to accumulate evidence on the order of 1000 data points for the estimation of the parameters $\boldsymbol{N_F}$ of the variability function.

2.4.1. Variability model vs lumped-data approach

With reference to the two Cases 1 and 2, this subsection presents the numerical differences between the proposed variability model and a beta-binomial model representative of the lumped-data approach. Figure 2.5 and Table 2.3 show the results. In both Cases 1 and 2, the expected mean, median, 5th and 95th percentiles of the $P_{\boldsymbol{F}}(p_{c|t})$ estimated by the lognormal variability model are compared with the respective statistics provided by a beta-binomial model, with increasing sample sizes (200, 500, and 1000 observations, in x-axis). Consistently with the variability model, the beta-binomial model (eq. 2.11) uses a diffuse prior, in particular the CNI prior, as in [21] (with parameters: α_0= 0.5, β_0= 8.66 for Case 1; α_0= 0.5, β_0= 79.5 for Case 2).

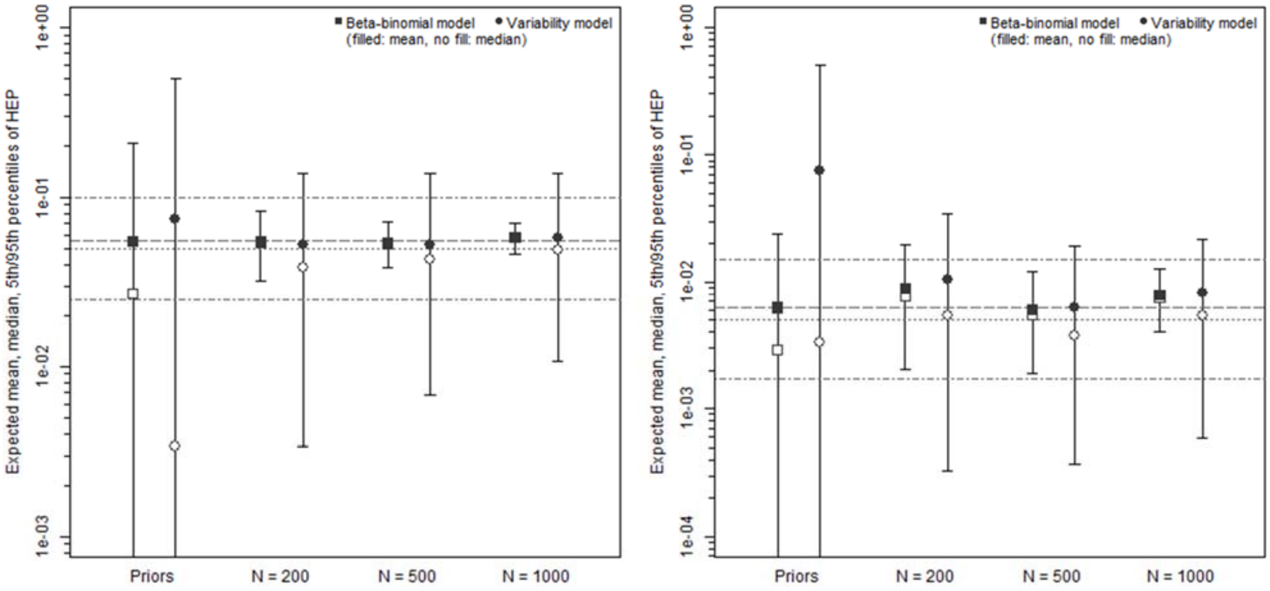


Figure 2.5. Expected mean (filled symbols), median (blank symbols), and 5th – 95th percentiles (whiskers) of $P_{\boldsymbol{F}}(HEP)$ by the lognormal variability model and the lumped-data beta-binomial model, tested against the same simulator datasets (number of simulated tasks: 200, 500, 1000). Datasets are artificially generated from lognormal HEP variability distribution with: median 5e-2 (dotted line), mean 5.46e-2 (dashed line) and error factor 2 (dot-dashed lines at 5th percentile 2.5e-2 and 95th percentile 1.0e-1) for Case 1 (left); median 5e-3 (dotted line), mean 6.25e-3 (dashed line) and error factor 3 (dot-dashed lines at 5th percentile 1.7e-3 and 95th percentile 1.5e-2) for Case 2 (right).

Comparing the expected error factors, the beta-binomial model provides a $P_{\boldsymbol{F}}(p_{c|t})$ that is overly-narrow with respect to the target HEP variability distribution, with values of error factor significantly smaller than the target one (Table 2.3, with target values of 2 and 3 for Case 1 and Case 2, respectively). On the other hand, the lognormal variability model provides broader $P_{\boldsymbol{F}}(p_{c|t})$'s, with error factors larger than the target values and tending to decrease to the target error factor with increasing sample sizes. While still larger that the target values, at 1000 observations the error factors reach the values of about 5 (Table 2.3), which starts to be of practical use for PSA applications (see analysis in the next subsection 2.4.2). Indeed, the larger error factors from the variability model compared to the beta-binomial as well as the decreasing tendency are not surprising: the important point for the practical application of the proposed model is to investigate the model data requirements for practical applications. This will be the goal of the next subsection 2.4.2. Concerning the estimated mean and median, both models tend to converge to the target values, as expected with slower convergence for Case 2.

To show the practical implications if variability is not modelled, assume plant-specific data is collected to infer the plant-specific HEP of a PSA operator action, with plant data from ten operating crews (Table 2.4). Assume also that data is available from simulator databases on the corresponding constellation (e.g. the case $N_{\boldsymbol{F}}$ = 200, Table 2.3). The data can be used as prior, then updated by the plant-specific data. Table 2.4 shows the difference in the posterior estimates depending on whether the prior distribution for the HEP is constructed with the lumped data model (Table 2.3, “lumped posterior”, N_F = 200) or the variability model (Table 2.3, “Var. model posterior”, N_F = 200). Three hypothetical data outcomes are considered, with increasing number of observed failures across the ten crews (Table 2.4, first column: 0, 1, and 2 failures). Given the plant-specific nature of the task (i.e. same scenario, same context of operation: no within-category variability in data), the observations from the ten different crews are all treated as “lumped”, neglecting the underlying crew-to-crew variability aspects in performance, and entered as unique data point in a simple beta-binomial model. Depending on the data outcome, the posterior distribution may become very different. In general, the variability model is more sensitive to the new data as compared to the lumped one. For the considered example, as the number of observed failures increases, the posterior mean for the variability model moves closer to the frequentist estimate (0.1, 0.2 for the 1 and 2 failure cases, respectively). Intuitively, this is due to the fact that the prior for the variability model represents larger variability of performance conditions and crew behaviours, which may also include those characteristic of the plant under consideration. On the other hand, the lumped data prior is narrowed to the population average, which may represent a biased initial value for the specific plant. Mathematically, as the evidence deviates from the population average, the likelihood of the evidence is multiplied by a smaller likelihood value for the lumped data prior (more peaked) compared to the variability model prior (more diffuse).

2.4.2. Sensitivity to available data

The collection of simulator data is resource-intensive and requires important time and money investments [23]: it becomes important to investigate the amount of data required such that the estimates produced by the model are of practical use (i.e. the associated uncertainties are not too large). In this subsection, for Case 1 and Case 2, convergence of the posterior statistics is followed as the available sample size increases. The error factor is particularly important for practical applications: too large error factors (e.g. 10, meaning a factor of 100 between the 95th and the 5th percentiles) entail diffuse posterior estimates of limited practical use. The aim of this subsection is to investigate the sample size required to obtain error factors comparable to those typical for HRA, e.g. around 5. Indeed, this sample size depends on the variability distribution of the HEP to be estimated. As already mentioned, the two cases 1 and 2 are deemed as representative of the range of interest for practical applications: larger HEP values (e.g. ~ 0.1) can be expected to be less problematic to estimate, while smaller values (e.g. below 0.001) may require too large data sizes for being of practical interest (at least with the model presented in this chapter).

Table 2.3. Comparison between the lognormal variability model and the beta-binomial: numerical results for Cases 1 and 2 (from Figure 2.5). Number of simulated tasks: 200, 500, 1000.

Case 1 - target statistics: median = 5e-2, mean = 5.46e-2, and EF = 2						
	Model (pdf)	**Mean**	**Median**	**5th perc**	**95th perc**	**EF**
	Lumped (CNI prior)	5.50e-02	2.69e-02	2.36e-04	2.06e-01	29.54
	Variability model (prior)	7.44e-02	3.35e-03	2.01e-05	4.98e-01	157.39
N_F=200, 11 failures	Lumped (posterior)	5.50e-02	5.36e-02	3.18e-02	8.31e-02	1.62
	Var. model (posterior)	5.24e-02	3.85e-02	3.35e-03	1.38e-01	6.43
N_F=500, 27 failures	Lumped (posterior)	5.40e-02	5.34e-02	3.86e-02	7.14e-02	1.36
	Var. model (posterior)	5.29e-02	4.33e-02	6.73e-03	1.38e-01	4.53
N_F=1000, 58 failures	Lumped (posterior)	5.80e-02	5.77e-02	4.64e-02	7.05e-02	1.23
	Var. model (posterior)	5.75e-02	4.86e-02	1.07e-02	1.38e-01	3.59
Case 2 – target statistics: median = 5e-3, mean = 6.25e-3, and EF = 3						
	Model (pdf)	**Mean**	**Median**	**5th perc**	**95th perc**	**EF**
	Lumped (CNI prior)	6.25e-03	2.87e-03	2.48e-05	2.39e-02	31.07
	Variability model (prior)	7.44e-02	3.35e-03	2.01e-05	4.98e-01	157.39
N_F=200, 2 failures	Lumped (posterior)	8.93e-03	7.79e-03	2.06e-03	1.97e-02	3.09
	Var. model (posterior)	1.05e-02	5.34e-03	3.27e-04	3.43e-02	10.24
N_F=500, 3 failures	Lumped (posterior)	6.03e-03	5.48e-03	1.87e-03	1.21e-02	2.54
	Var. model (posterior)	6.25e-03	3.76e-03	3.68e-04	1.92e-02	7.22
N_F=1000, 8 failures	Lumped (posterior)	7.87e-03	7.57e-03	4.02e-03	1.27e-02	1.78
	Var. model (posterior)	8.09e-03	5.34e-03	5.86e-04	2.15e-02	6.06

Table 2.4. Example of HEP estimation for a plant-specific task: prior distribution from lumped-data model (Table 2.3, “lumped posterior”, N_F = 200) and from the variability model (Table 2.3, “Var. model posterior”, N_F = 200).

	Prior from lumped-data model			**Prior from variability model**			
Evidence	**Mean**	**Median**	**EF**	**Mean**	**Median**	**EF**	**Δ% mean**
Priors	5.50e-02	5.36e-02	1.62	5.24e-02	3.85e-02	6.43	+ 5%
0 failures, 10 trials	5.25e-02	5.11e-02	1.62	3.81e-02	3.00e-02	4.94	+ 38%
1 failures, 10 trials	5.71e-02	5.57e-02	1.58	6.54e-02	5.76e-02	3.09	- 13%
2 failures, 10 trials	5.81e-02	6.03e-02	1.55	9.26e-02	8.53e-02	2.48	- 37%

2.4.2.1 Diffuse priors

Figure 2.6 shows the posterior estimates by the variability model, set up with flat hyper-priors $\pi_0(\mu_t)$ and $\pi_0(\sigma_{\boldsymbol{F}})$ as a function of the sample size $N_{\boldsymbol{F}}$ (from $N_{\boldsymbol{F}}$ = 50 to 1000) for Cases 1 (Figure 2.6, top) and 2 (Figure 2.6, bottom). From left to right, the figures report the estimated posterior error factor, mean, and median. For each sample size, 100 datasets are sampled to represent the spread of the posterior estimates (each estimate represented by a dot in the figures).

From Figure 2.6, the expected statistics of $P_{\boldsymbol{F}}(p_{c|t})$ across the different datasets tend to converge to the target statistics as the sample size increases. The expected mean and median, averaged over the Monte Carlo samples, get close to their target values, at $N_{\boldsymbol{F}} \approx 200$ for Case 1 and at $N_{\boldsymbol{F}} \approx 250$ Case 2. Indeed, for Case 1 at $N_{\boldsymbol{F}} \approx 200$, the average expected mean is 5.3e-2, with 50% confidence interval (25th - 75th percentiles) of (4.2e-2, 6.2e-2), and the average expected median is 3.9e-2, with 50% confidence interval of (2.9e-2, 4.9e-2); for Case 2 at $N_{\boldsymbol{F}} \approx$ 250, the average expected mean is 7.3e-3, with 50% confidence interval (25th - 75th percentiles) of (4.6e-3, 8.4e-3), and the average expected median is 3.7e-3, with 50% confidence interval of (2.1e-3, 4.2e-3).

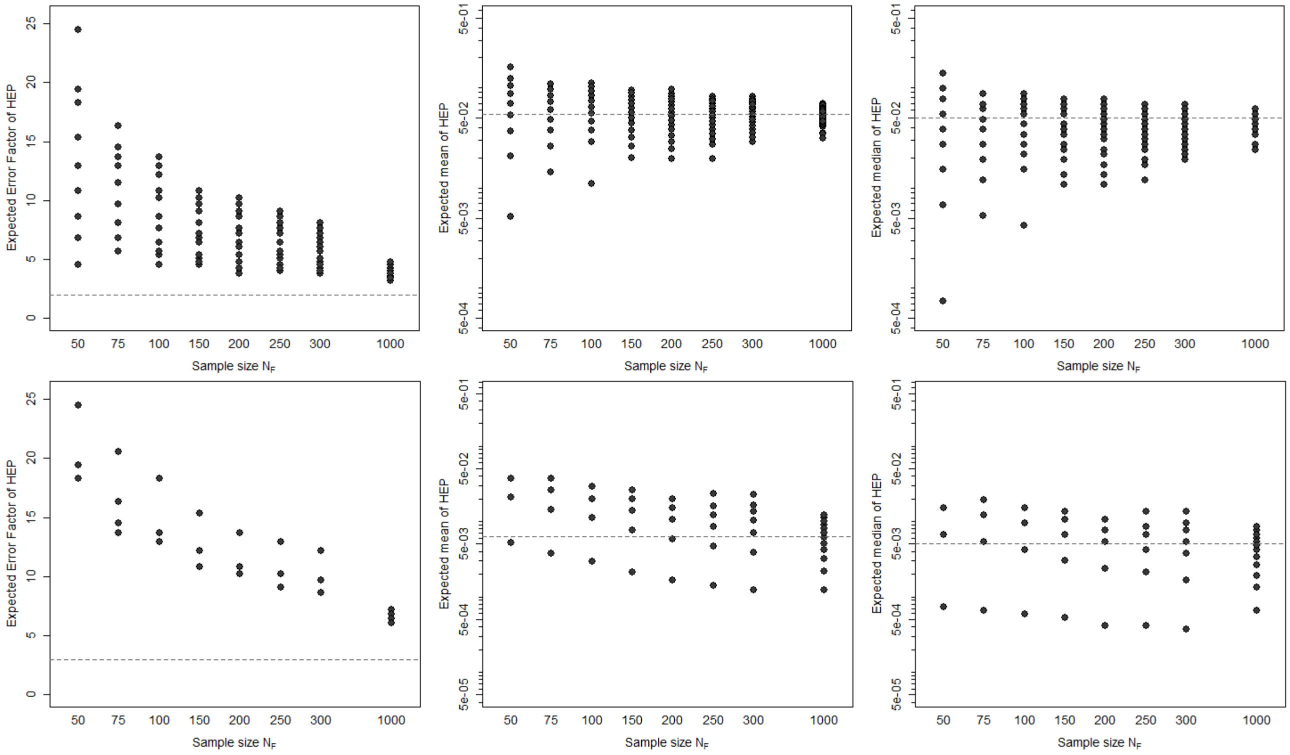


Figure 2.6. Data requirements of the lognormal variability model with flat hyper-priors $\pi_0(\mu_t)$ and $\pi_0(\sigma_{\boldsymbol{F}})$. Top: Case 1 (median = 5e-2, mean = 5.46e-2 and error factor = 2. Bottom: Case 2 (median = 5e-3, mean = 6.24e-3, and error factor = 3. For each sample size (x-axis), 100 datasets (dots) are Monte Carlo-sampled from the target distribution. From left to right: expected error factor, mean (log-scale), and median (log-scale) of the $P_{\boldsymbol{F}}(HEP)$'s returned by the model.

The speed of convergence of the expected error factors is lower compared to the mean and median. For instance, for Case 1, 300 observations are approximately needed to observe an average expected error factor close to 5, i.e. 5.5 at $N_{\boldsymbol{F}} \approx 300$, with 50% confidence interval (4.8, 6.1). For Case 2, with $N_{\boldsymbol{F}} \approx 1000$, the average expected error factor is 6.1, with 50% confidence interval (6.1, 6.1, note the 25th and 75th percentiles match because of numerical discretization). Indeed, the speed of convergence to the target values depends on the amount of evidence at disposal. As the HEP values progressively decrease, fewer failure are observed (i.e. Monte Carlo sampled): as anticipated, for cases with lower HEP values (e.g. below 0.001), the model would require an impracticably large data size (e.g. above 10^4 data points).

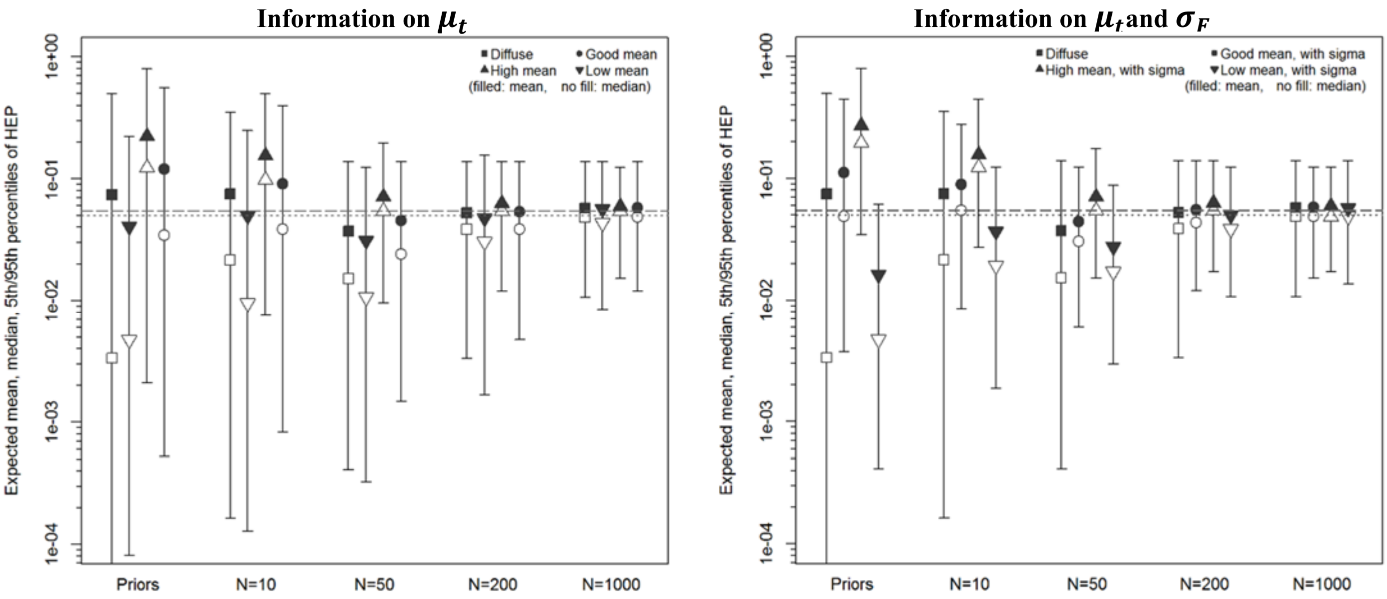
In conclusion, this sensitivity analysis shows that for constellations $\boldsymbol{F}$ characterized by HEP values in the range ~ 0.1 ÷ 0.001, the variability model with diffuse hyper-priors can provide results of practical value for HRA applications with few hundred data points. The latter data requirement are met by the current availability of data points for many constellations $\boldsymbol{F}$ in SACADA [23] and HuREX [27]. When lower HEP values are involved (e.g. HEP ~ 0.001 and below), the adoption of informative prior distributions may be a viable option to decrease the data requirements, as presented in the next subsection 2.4.2.2.

2.4.2.2 Informative priors

This subsection investigates how much data requirements can be reduced with informative hyper-priors for both parameters μ_t and $\sigma_{\boldsymbol{F}}$. Case 1 and Case 2 are addressed in Figures 2.7-2.8 and Figures 2.9-2.10, respectively. Values are reported in Tables B.1-B.2 in Appendix B.

In Figure 2.7, two configurations can be distinguished: only the mean HEP is informed (left plot), both mean and standard deviation are informed (right plot). Both plots show the effect of different combinations for $\pi_0(\mu_t)$ and $\pi_0(\sigma_{\boldsymbol{F}})$ on the posterior HEP estimates as the number of simulator runs increases (in x-axis). The prior information may be available from HRA methods or generic failure databases. The considered prior distributions for the mean, $\pi_0(\mu_t)$, are (left plot):

- "Diffuse": flat distributions for the parameters of the lognormal, mean and standard deviation, same as for subsection 2.4.2.1;
- "Good mean": prior distribution informed around the correct median HEP value for Case 1 (lognormal, with median = 5e-2, 5th percentile = 5e-3, 95th percentile = 5e-1);
- "Low mean" and "High mean": prior distributions with median shifted by one order of magnitude below and above the correct median HEP value for Case 1, respectively (for "Low mean": lognormal, with median = 5e-3, 5th percentile = 5e-4, 95th percentile = 5e-2; for "High mean": lognormal, with median = 5e-1, 5th percentile = 5e-2, 95th percentile = 1).


Information on $\boldsymbol{\mu_t}$

Information on $\boldsymbol{\mu_t}$ and $\boldsymbol{\sigma_F}$

Figure 2.7. Sensitivity of the lognormal variability model to the choice of prior distributions for the hyper-parameters, i.e. $\pi_0(\mu_t)$ and $\pi_0(\sigma_{\boldsymbol{F}})$, and to the sample size, Case 1. Left: only $\pi_0(\mu_t)$ is informative. Upper/lower bounds of the lognormal distributions: “Good mean”, 5e-3/5e-1; “High mean”, 5e-2/1; “Low mean”, 5e-4/5e-2. Right: both $\pi_0(\mu_t)$ and $\pi_0(\sigma_{\boldsymbol{F}})$ are informative. “With sigma” corresponds to a normal distribution with bounds 1.5/5 (expressed in terms of error factor).

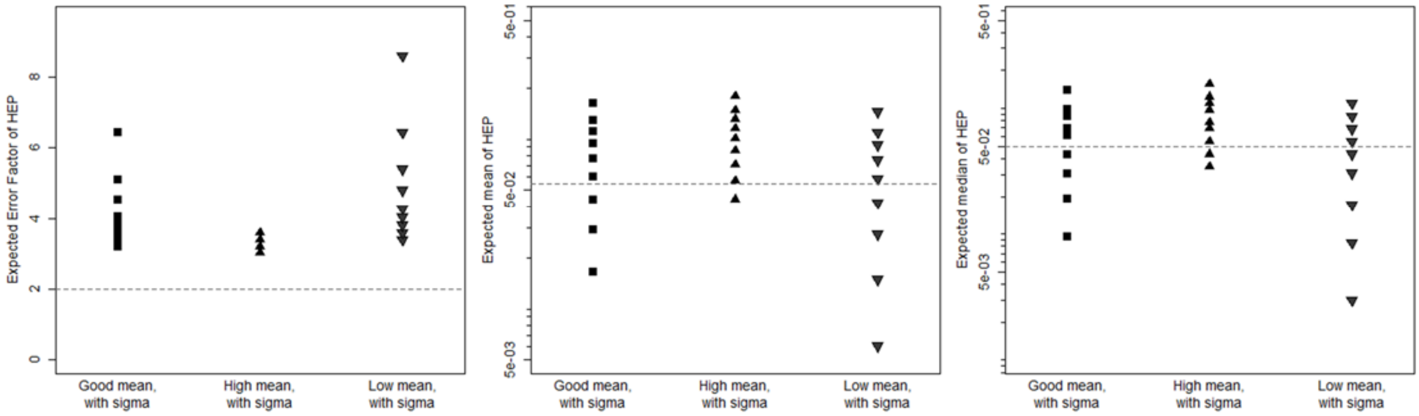


Figure 2.8. Behavior of the lognormal variability model with informative priors on both $\pi_0(\mu_t)$ and $\pi_0(\sigma_{\boldsymbol{F}})$ at $N_{\boldsymbol{F}}$ = 50 (target HEP variability distribution as in Case 1: median = 5e-2, mean = 5.46e-2, and error factor = 2). For each option of informative priors in x-axis (Figure 2.7, right plot), 100 datasets (dots) are Monte Carlo-sampled from the target distribution. From left to right, in y-axis: expected error factor, mean (log-scale), and median (log-scale) of the $P_{\boldsymbol{F}}(HEP)$’s provided by the model for each choice of prior (dotted lines: statistics of the target distribution).

The “Good mean” prior assumes that the information at disposal is correct in the order of magnitude of the HEP range, with two orders of magnitude between the 5th and the 95th percentiles. The “Low mean” and “High mean” priors assume the presence of biases of one order of magnitude. Additional information on the standard deviation, $\pi_0(\sigma_{\boldsymbol{F}})$, is modelled by a normal distribution (“with sigma”) with 5th and 95th percentiles corresponding to error factors of 1.5 and 5, respectively (Figure 2.7, right plot). Limiting values for error factor close to 5 are commonly accepted in establishing confidence intervals for HRA applications [6].

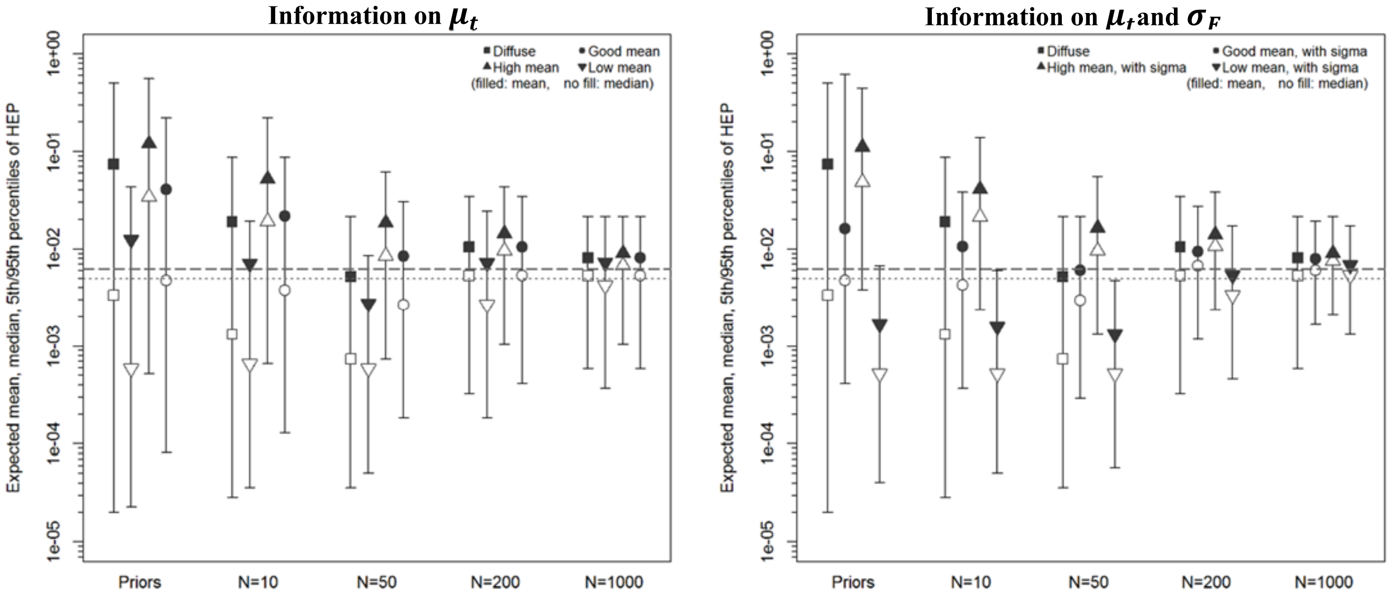


Figure 2.9. Sensitivity of the lognormal variability model to the choice of prior distributions for the hyper-parameters, i.e. $\pi_0(\mu_t)$ and $\pi_0(\sigma_F)$, and to the sample size, Case 2. Left: only $\pi_0(\mu_t)$ is informative. Upper/lower bounds of the lognormal distributions: "Good mean", 5e-4/5e-2; "High mean", 5e-3/5e-2; "Low mean", 5e-5/5e-3. Right: both $\pi_0(\mu_t)$ and $\pi_0(\sigma_F)$ are informative. "With sigma" corresponds to a normal distribution with bounds 1.5/5 (expressed in terms of error factor).

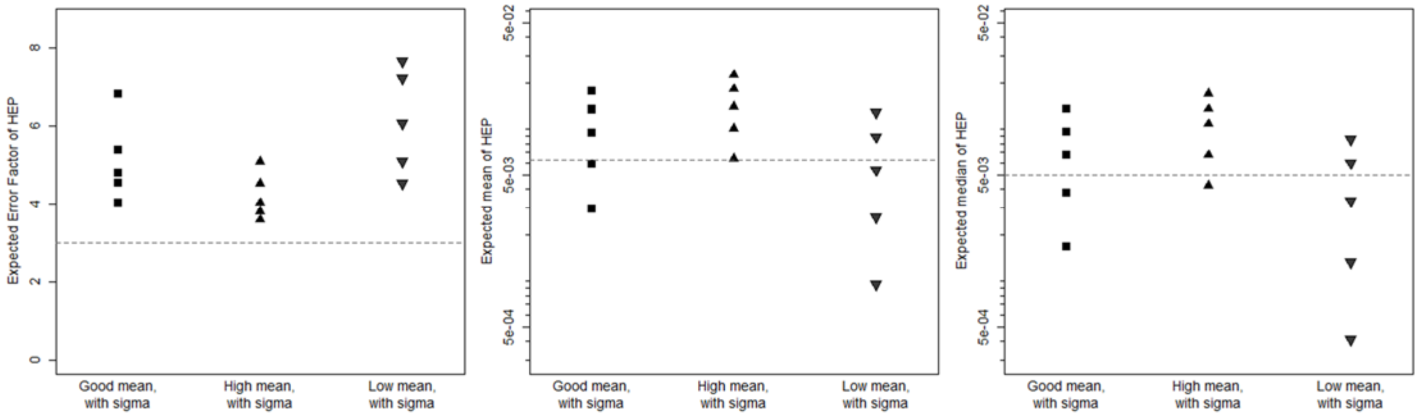


Figure 2.10. Behavior of the lognormal variability model with informative priors on both $\pi_0(\mu_t)$ and $\pi_0(\sigma_F)$ at N_F = 200 (target HEP variability distribution as in Case 2: median = 5e-3, mean = 6.25e-3, and error factor = 3). Same considerations as in Figure 2.8. Dotted lines: statistics of the target distribution.

With informative $\pi_0(\mu_t)$ (Figure 2.7, left plot), when the information on μ_t is not biased ("Good mean"), it is possible to achieve reasonable approximations of the target mean (5.5e-2) and median (5e-2) already at N_F = 200. Indeed, at N_F = 200 the "Good mean" error factor is 16% lower than the one obtained with "Diffuse" prior (see Table B.1). In case of biased information on μ_t, sensible overestimation ("High mean") or underestimation ("Low mean") of the expected mean and median can be observed for all datasets, of course tending to decrease with the amount of data available.

Data requirements can be significantly reduced if both $\pi_0(\mu_t)$ and $\pi_0(\sigma_F)$ are informative

(Figure 2.7, right). If the information on μ_t is not biased ("Good mean, with sigma"), it is possible to obtain good approximations of the expected mean and median, as well as acceptable error factors, already in the range N_F = 10÷50. For instance, at N_F = 50, the model with "Good mean, with sigma" prior returns an expected error factor approximately 4 times lower than the value provided with the "Diffuse" prior, and significantly closer to the target value for Case 1 (error factor = 2). Still at N_F = 50, the biased hyper-priors reflect in biased HEP estimates (Table B.1 and Figure 2.7, right), however the correct values lie within the 90% confidence bounds (5th and 95th percentiles). As the data set increases, the effect of the prior information is progressively reduced, as shown by the statistics for N_F = 200 and 1000, very close to the target values.

To further investigate the possible reduction in data requirements, Figure 2.8 further examines the sample size of N_F = 50, a size reasonably achievable by current simulator data collection programs aggregating multiple plants. Figure 2.8 shows the results for 100 Monte Carlo-sampled datasets relevant to Case 1 at N_F = 50. The results confirm that such size is well enough for "Good mean, with sigma": average expected mean of 5.8e-2 (50% confidence: 4.4e-2, 7.7e-2), average expected median of 4.3e-2 (50% confidence: 3.1e-2, 6.1e-2), average expected error factor = 4.4 (50% confidence: 3.8, 4.5). The Monte Carlo samples show that the biased estimates are not usable, because the correct values lie outside the 50% confidence interval: for "High mean, with sigma", average expected mean 8.3e-2 (50% confidence: 7.1e-2, 1.0e-1), average expected median 6.6e-2 (50% confidence: 5.5e-2, 7.7e-1); for "Low mean, with sigma", average expected mean 4.2e-2 (50% confidence: 2.7e-2, 5.8e-2), average median = 2.9e-2 (50% confidence: 1.7e-2, 4.3e-2). It is however important to mention that the potential bias may be relatively easy to identify a posteriori. For example, from the Monte Carlo samples at N_F= 50, the expected change in marginal prior medians (see Table B.1) after the evidence is:

- for "Good mean, with sigma" between 24% and 36% of the marginal prior median (= 4.9e-2);
- for "High mean, with sigma" between 72% and 285% of the marginal prior median (= 2e-1);
- for "Low mean, with sigma" between 254% and 796% of the marginal prior median (= 4.8e-3).

Indeed, large deviations of the posterior median from the marginal prior median could be used as indicators of an initial bias.

Figure 2.9 and Table B.2 present the results relevant to Case 2 and Figure 2.9 further explores the influence of informative priors at N_F = 200:

- "Good mean": lognormal, with median = 5e-3, 5th percentile = 5e-4, 95th percentile = 5e-2;

- "Low mean": lognormal, with median = 5e-4, 5th percentile = 5e-5, 95th percentile = 5e-3;
- "High mean": lognormal, with median = 5e-2, 5th percentile = 5e-3, 95th percentile = 5e-1.

Compared to Case 1, Case 2 is characterized by a "weaker" evidence of failure (note that HEP ~ 0.001 in Case 2): this aspect influences the efficiency of informative priors in reducing the data requirements of the model. With informative $\pi_0(\mu_t)$, from the cross-comparison with Case 1 results (left plots in Figures 2.7 and 2.9; Tables B.1 and B.2), the model tends to return significantly higher values of the expected error factor in Case 2: this suggests that informing only μ_t is not sufficient to achieve good approximation of the target mean (6.3e-3) and median (5e-3) with acceptably low $N_{\boldsymbol{F}}$ (e.g. already at $N_{\boldsymbol{F}}$ = 200 as for Case 1).

When informing both $\pi_0(\mu_t)$ and $\pi_0(\sigma_{\boldsymbol{F}})$ without bias ("Good mean, with sigma" in Figure 2.9, right), good approximations of the expected mean and median, as well as acceptable error factors, can be achieved in the range $N_{\boldsymbol{F}}$ = 50÷200 (note the increased data requirements compared to range $N_{\boldsymbol{F}}$ = 10÷50 for Case 1). For instance, at $N_{\boldsymbol{F}}$ = 200, the model with "Good mean, with sigma" prior returns an expected error factor approximately two times lower than the value provided with the "Diffuse" prior and closer to the target value for Case 2 (error factor = 3). Still at $N_{\boldsymbol{F}}$ = 200, however, the biased hyper-priors ("Low mean, with sigma" and "High mean, with sigma") reflect in biased HEP estimates (Table B.2 and Figure 2.9, right), however the correct values lie within the 90% confidence bounds. Figure 2.10 shows the results for 100 Monte Carlo-sampled datasets relevant to Case 2 at $N_{\boldsymbol{F}}$ = 200. The analysis confirms that "Good mean, with sigma" performs efficiently at this sample size: average expected mean = 7.0e-3 (50% confidence: 5.8e-3, 9.4e-3), average expected median = 4.8e-3 (50% confidence: 3.8e-3, 6.7e-3), and average expected error factor = 5.4 (50% confidence: 4.8, 5.4). On the other hand, for the configurations with biased priors, the correct values of the statistics (target mean = 6.3e-3 and target median = 5e-3) lie outside the 50% confidence interval: for "High mean, with sigma", average expected mean = 1.1e-2 (50% confidence: 1.0e-2, 1.4e-2), average expected median = 8.0e-3 (50% confidence: 6.7e-3, 1.1e-2); for "Low mean, with sigma", average expected mean 3.7e-3 (50% confidence: 2.6e-3, 5.3e-3), average median = 2.2e-3 (50% confidence: 1.3e-3, 3.4e-3). As for Case 1, the potential bias in Case 2 can be easily identified by the observed large deviations of the posterior median from the marginal prior median (see Table B.2) across the different configurations, e.g. at $N_{\boldsymbol{F}}$ = 200:

- for "Good mean, with sigma" between 20% and 39% of the marginal prior median (= 4.8e-3);
- for "High mean, with sigma" between 78% and 86% of the marginal prior median (= 4.9e-2);

- for “Low mean, with sigma” between 150% and 554% of the marginal prior median (=5.2e-4).

In conclusion, the analysis highlighted the following two aspects. First, for a given acceptable level of approximation of the target error factor, unbiased informative priors on both the mean and the standard deviation of HEP distribution are effective in reducing the overall data requirements of the lognormal variability model. Secondly, especially for constellations $\boldsymbol{F}$ characterized by lower orders of magnitude of HEP or limited performance data $N_{\boldsymbol{F}}$ (or both), biased informative priors have a strong influence on the HEP uncertainty distribution estimated by the model. Following on this, reducing as much as possible the bias in informative priors becomes of key importance. Besides the approach adopted for the purposes of this numerical application, different techniques (e.g. posterior predictive checks) are available in Bayesian literature to assist the analyst in selecting adequate prior distributions and reduce the initial bias [51].

2.4.3. Application to real simulator data from literature

The proposed variability model is applied to failure data of operating crews in nuclear power plants available in the literature (Halden project data from [21], and HuREX data from [27]). Both references [21] and [27] use the simulator data to inform HEPs of constellations of task type and PSF levels. Reference [21] addresses constellations of SPAR-H PSFs (e.g. “complexity”, “stressors”), while [27] addresses the HuREX framework for different combinations of cognitive activities (e.g. “situation interpreting”, “execution”) and generic task types (e.g. “verifying state of indicator”; “directing manipulation”). In particular, reference [21] addresses five contexts (for the sake of brevity, only SPAR-H’s PSFs with ratings different than “nominal” are reported; see [8] for further information on PSF definitions):

- Context A: Time = extra; Complexity = moderate; Procedures = available but poor.
- Contexts B_a, B_b: Time = barely adequate; Stressors = high; Complexity = moderate; Procedures = available but poor.
- Context C: Time = inadequate; Stressors = high; Complexity = high; Procedures = available but poor; Work processes = poor.
- Context D: Time = extra.

For reference [27] the following operator activities are considered:

- RP-manipulation: cognitive activity = response planning; task type = directing manipulation.
- RP-procedure: cognitive activity = response planning; task type = transferring procedure.

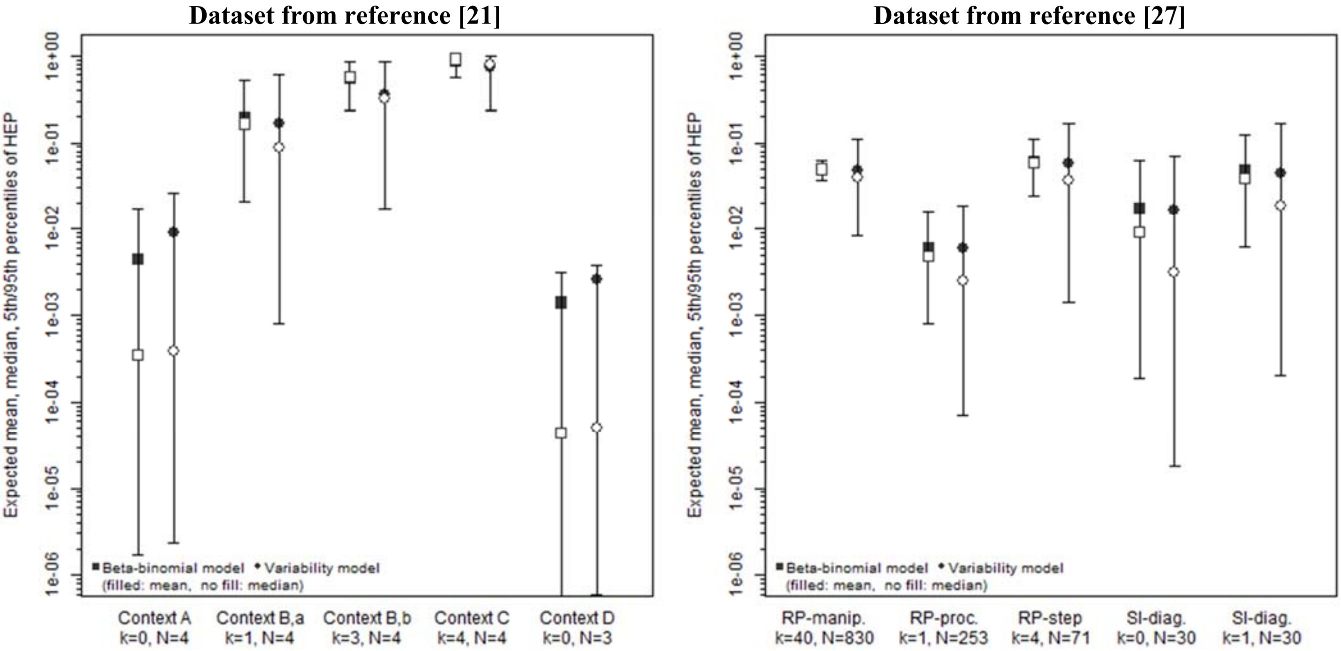
- RP-step: cognitive activity = response planning; task type = transferring step procedure.
- SI-diagnosis: cognitive activity = situation interpreting; task type = diagnosing.

In [27], the authors adopted a conservative assumption consisting of adding a fictitious recorded failure for all those constellations $\boldsymbol{F}$ where actually no failures have been observed. For instance, this was the case of RP-step dataset. In this application, the latter has been treated in two different configurations: the conservative dataset as used by the authors (with one postulated failure: $k_{\boldsymbol{F}}$ = 1, $N_{\boldsymbol{F}}$ = 30), and the real dataset (with zero failures observed: $k_{\boldsymbol{F}}$ = 0, $N_{\boldsymbol{F}}$ = 30).

Both references [21] and [27] adopt the lumped approach, with the conjugated beta-binomial model. Concerning the prior, reference [21] uses the CNI prior (from [40]), built on the basic HEP provided by SPAR H in correspondence of the context. Reference [27] adopts the Jeffreys non-informative distribution, a beta distribution with both shape parameters (i.e. α_0 and β_0 in eq. 2.11) equal to 0.5.

An important difference between the datasets of [21, 27] concerns their size. Reference [21] addresses rather small data sets, four data points on average, including very challenging tasks. Reference [27] addresses significantly larger datasets, because of the different granularity of the data collection taxonomy and because of the larger number of crews from which data is collected. This difference allows comparing the performance of the variability and the beta-binomial models (with lumped data) under very different data availability conditions.

The expected statistics (mean, median, and 5th /95th percentiles) of the HEP posterior distributions estimated by both variability and lumped-data models are shown in Figure 2.11 (y-axis, in log-scale), for each of the datasets used in the application (x-axis, left: [21]; right: [27]). A summary of the numerical results is given in Tables C.1-C.2 in Appendix C. Note that the results for lumped-data models in Tables C.1-C.2 and Figure 2.10 are slightly different from the numerical values in [21, 27], since the prior distributions adopted by these works (the CNI for [21]; the Jeffreys for [27]) were adapted in this application to ensure a fair comparison with the variability model. In particular, for the results to be comparable, the literature models and the variability model should start from the same expected HEP distribution ($P_{\boldsymbol{F}}(p_{c|t})$ from eq. 2.9 for the variability model). To do this, the mean of the lognormal variability model (i.e. μ_t) was assigned the literature priors, i.e. CNI prior for $\pi_0(\mu_t)$ for the comparison with [21]; Jeffreys prior for $\pi_0(\mu_t)$ for the comparison with [27]. Then, the expected HEP distribution from the variability model, i.e. the lognormal parametric distribution weighted by the joint hyperprior $\pi_0(\mu_t, \sigma_{\boldsymbol{F}})$, was derived (for $\pi_0(\sigma_{\boldsymbol{F}})$, the diffuse prior mentioned in subsection 2.3.4 was used). Finally, the lumped-data priors were re-calculated such that the corresponding expected HEP distribution would fit the one from the variability model.


Dataset from reference [21]

Dataset from reference [27]

Figure 2.11. Results from the application of the lognormal variability model to real simulator data available in literature (datasets in x-axis: left, [21]; right, [27]). On y-axis (in log-scale): expected mean (filled symbols), median (blank symbols), and 5th – 95th percentiles (whiskers) of the $P_F(HEP)$'s estimated by both the lognormal variability model (circles) and the lumped-data beta-binomial model (circles) given the same marginal prior distribution on HEP.

From the comparison of result, a general tendency can be observed: overall, the lumped-data beta-binomial models tend to return narrower posterior distributions if compared to the variability model. This tendency replicates across all the tested datasets, with a magnitude that depends on the amount of evidence available (i.e. the sample size and the observed failures).

In particular, for [21] (Figure 2.11, left), the differences in the two models are small for "Contexts A" and "Context D": the corresponding datasets are characterized by few observations and zero failures. As the number of observed failures increases (e.g. "Contexts B$_a$" and "Context B$_b$"), the differences between the posteriors become larger, see the expected error factor in "Contexts A/D, B$_a$ and B$_b$" in Table C.1 (e.g. for "Context B$_b$", the variability model returns an expected error factor 3.7 times higher than the lumped-data model).

A similar trend can be observed for the data-rich application [27]. The differences in the expected error factors become more evident with progressively increasing the number of observed failures in the dataset (e.g. see the different spreads in the HEP uncertainty distributions from "SI-diagnosis, k = 0" to "SI-diagnosis, k = 1", in Figure 2.11, right); the differences persist at very high numbers of observed failures (e.g. for "RP-manipulation" dataset, the error factor estimated by the lumped-data model is approximately 2.8 times lower than the variability model).

2.5. Discussion

The application to simulator data in Section 2.4 has demonstrated the large impact on the estimated HEP distribution of considering the underlying variability in the HRA data. As presented in Section 2.3, the two models reflect two different interpretations of the target HEP. The variability model considers the HEP as a quantity that is specific for a crew and for a realization of the constellation; correspondingly, the HEP variability reflects the variability of the crews and of the realizations. The beta-binomial model considers the HEP as a unique quantity for a given constellation, aggregating all variability aspects in its value.

It is important to note that there is no right or wrong interpretation of the HEP quantity: it depends on the application at hand. For example, an important HRA issue is to investigate PSF effects across different constellations. The effect on the HEP of changes in one or more elements of the vector $\boldsymbol{F}$ in eq. 2.1 may be investigated by focusing on the aggregated effect, i.e. on the population average across crews and within-constellation, therefore adopting the typical beta-binomial model. On the other hand, as presented in Section 2.2, when the estimated HEP is used to inform a given constellation of an HRA model, adopting a variability model becomes important to capture the variability elements discussed in Section 2.2 and ideally allow for plant-specific HEP values (as demonstrated in Section 2.4).

The model presented here supports a first investigation of the need for modelling variability. The interpretation of the HEP as a crew-specific quantity strongly limits the possibility to aggregate the data to inform HEP values. As shown by Figure 2.4, the data informing the HEP variability distribution are only 0's and 1's because of the constraint that one crew only performs the exact same task only once. An alternative would be to consider the HEP values as dependent on particular crew features or styles (e.g. of communication or decision-making), as opposed to being just crew-specific. This approach would not consider each crew being characterized by a different HEP value: each crew feature or style would be connected with an HEP value. Numerically, this would allow aggregating more evidence on the single HEP realization (the number of task repetitions in Figure 2.4 would be per crew feature or style, and not per single crew). On the other hand, this may allow analysis of crew features and styles on the HEP, opening to additional applications to inform crew training. Current work by the authors is addressing definitions of appropriate features and styles as well as the associated adaptations to the model.

As presented in Section 2.3, the inference model is intended for general application to any HRA model for HEP quantification. The currently available HRA models strongly differ in the task and factors considered and in the granularity of their definition. It can be expected that these aspects are strongly connected with the variability that the model shall be able to represent. For instance, the simulator data used in subsection 2.4.3 (Halden in [21]; HuREX in [27]) correspond to constellations at very different granularity. [21] uses the SPAR-H factor taxonomy on an operator task definition close to what would be used for PSA applications (e.g. “isolate the ruptured steam generator and control pressure”). On the other hand, HuREX in [27] operates at a more microscopic granularity level (e.g. “determine the condition of Adverse Containment”, “check if the three Reactor Coolant Pumps should be stopped”). As a working hypothesis, it may be reasonable to assume that the coarser the granularity of the model (more macroscopic tasks), the larger the variability corresponding to the within-category variability. Also, the more the task involves decision-making and communication at the crew level, the more crew variability will be relevant, compared for example to execution-related tasks performed by single persons. Finally, it can be expected that variability would also be larger for HRA models with coarser PSF categories, e.g. binary as opposed to multivalued. With the current interest by the community on empirically estimated HEPs, it may be well important that future studies will address the extent to which variability shall be addressed as well as with the goal of develop guidelines to do it.

HRA research is addressing advanced modelling techniques, in particular Bayesian Belief Networks, to represent the complex relationships among influencing factors as well as to formally incorporate a diversity of data sources. Indeed, within-category variability can be incorporated in these models via appropriate conditional probability distributions. BBNs can incorporate crew-to-crew variability as well, either implicitly, into the BBN internal distributions, as well as explicitly, as dedicated nodes [10, 24). The work presented in this chapter can be used to enhance the empirical basis of the BBN distributions, e.g. as anchoring distributions to populate the model relationships via filling algorithms such as those in [50].

2.6. Conclusions

Due to lack of data, judgments are currently the main source of information to assess the uncertainty and variability in the error probability estimates produced by HRA models. With the on-going large data collection activities, it becomes important that uncertainty and variability be empirically based, along with the associated point estimates.

This chapter presents a Bayesian hierarchical model that addresses the HEP variability due to operating crew differences as well as variability within the categories of task type and performance factors. Such models are typically used to consider source-to-source variability of failure probability estimates for hardware components: this chapter presents their formulation and use for human failure data from simulators.

The presented case studies demonstrate the significant overconfidence in the HEP estimates if variability is not considered, e.g. if all data is lumped to feed a beta-binomial Bayesian model (as typically done in most HRA applications). Also, this may results in significant biases for plant-specific human error probabilities.

Empirically informing variability requires a large amount of data: therefore, numerical applications have investigated the practical applicability of the proposed model. For moderately high HEP values (in the range of 1e-2), estimates of practical use can be obtained with few hundred, say below 500, data points (i.e. simulator runs). This is already achievable by current simulator programs depending on the constellation of tasks and performance factors. Prior information on the model parameters, e.g. from available HRA methods, can reduce the data requirements. For HEP values in the range of 1e-2, about 50 data points are demonstrated to become enough. For lower HEP values, in the range of 1e-3, estimates of practical use become achievable with few hundred data points. Of course, biases in the prior distributions may result in biases in the posterior estimates. However, this chapter has shown that a simple check of the change between the prior and posterior estimates may reveal the presence of the initial bias. Data requirements for further low HEP ranges, i.e. below 1e-3, may be impractical for many operator tasks with the proposed model.

The proposed model treats variability as a continuum. Especially when considering crew-to-crew variability, it may be important to identify relevant crew features that play a role in determining the failure probability. Besides allowing aggregating data from different crews on the basis of their common traits, this may support training of operators on the crew skills that allow lower failure probability values. Work by the authors is ongoing along this direction.

This work is part of a larger effort to derive empirically-based reference HEP values to strengthen the technical basis of HRA methods. The long-term aim is to develop a framework to process diverse data sources, e.g. simulator data, data from existing HRA methods, operational experience data, and evidence from human factor studies. The main thrust is that a mathematical, traceable aggregation of these sources will allow to feed with new data as it becomes available, progressively replacing older evidence that may become outdated because of new advances in plant operation and design.

Acknowledgments

The work was funded by the Swiss Federal Nuclear Safety Inspectorate (ENSI), under contract Nr. 101163. The views expressed in this work are solely those of the authors.

References

1. Kirwan B. *A guide to practical Human Reliability Assessment*. CRC press: Boca Raton, FL, USA, 1994.
2. Podofillini L. Human Reliability Analysis. In: Moller N, Hansson SO, Holmberg JE, and Rollenhagen C. (eds) *Handbook of Safety Principles*. Wiley, 2017, pp.565-592.
3. Williams JC. HEART – A Proposed Method for Assessing and Reducing Human Error. In: *9th Advance in Reliability Technology Symposium*, University of Bradford, 1986.
4. Williams JC. A data-based method for assessing and reducing human error to improve operational performance. In: *Proceedings of the IEEE Fourth Conference on Human Factors and Power Plants*, Monterey, California, 5–9 June, pp. 436–450, 1988.
5. Williams JC. HEART – a proposed method for achieving high reliability in process operation by means of human factors engineering technology. *Saf Reliab* 2015, 35(3): 5–25.
6. Swain AD and Guttman HE. Handbook of human reliability analysis with emphasis on nuclear power plant applications. NUREG/CR-1278, U.S. Nuclear Regulatory Commission, Washington DC, USA, 1983.
7. Gertman DI, Blackman HS, Marble JL, et al. The SPAR-H Human Reliability Analysis Method. NUREG/CR-6883, U.S. Nuclear Regulatory Commission, Washington DC, USA, 2005.
8. Whaley AM, Kelly DL, Boring RL, et al. *SPAR-H step-by-step guidance*. INL/EXT-10-18533, Idaho National Labs, Idaho Falls, Idaho 83415, 2011.
9. Hollnagel E. *Cognitive Reliability and Error Analysis Method* (*CREAM*). Oxford: Elsevier Science Ltd, 1998.
10. Groth KM and Mosleh A. Deriving causal Bayesian networks from human reliability analysis data: a methodology and example mode. *Proc Inst Mech Eng, Pt O: J Risk Reliab* 2012, 226(4): 361–79.
11. Mkrtchyan L, Podofillini L and Dang VN. Bayesian belief networks for human reliability analysis: A review of applications and gaps. *Reliab Eng Syst Saf* 2015, 139: 1–16.
12. Spurgin AJ. *Human Reliability Assessment – theory and practice*. CRC press: Boca Raton, FL, USA, 2010.
13. Hallbert B and Kolaczkowski A. The Employment of Empirical Data and Bayesian Methods in Human Reliability Analysis: A Feasibility Study. NUREG/CR-6949, pp. 1-4, INL/EXT-06-11670, Washington, D.C.: U.S. Nuclear Regulatory Commission, 2007.
14. Forester J, Dang VN, Bye A, et al. The International HRA Empirical Study Lessons Learned from Comparing HRA Methods Predictions to HAMMLAB Simulator Data. NUREG-2127, US Nuclear Regulatory Commission, Washington DC, USA, 2014.
15. Forester J, Liao H, Dang VN, et al. The US HRA Empirical Study - Assessment of HRA Method Predictions against Operating Crew Performance on a US Nuclear Power Plant Simulator. NUREG-2156, US Nuclear Regulatory Commission, Washington DC, USA, 2016.
16. Park J, Jung W, Kim S, et al. A guideline to collect HRA data in the simulator of nuclear power plants. KAERI/TR-5206, Korea Atomic Energy Research Institute, Republic of Korea, 2013.

17. Chang JY, Bley D, Criscione L, et al. The SACADA database for human reliability and human performance. *Reliab Eng Syst Saf* 2014, 125: 117-133.
18. Liao H, Forester J, Dang VN, et al. Assessment of HRA method predictions against operating crew performance: Part III: Conclusions and achievements. *Reliab Eng Syst Saf* 2019, 191: 106511.
19. Hallbert B, Morgan T, Hugo J, et al. A Formalized Approach for the Collection of HRA Data from Nuclear Power Plant Simulators. NUREG/CR-7163; INL/EXT-12-26327, US Nuclear Regulatory Commission, Washington DC, USA and Idaho National Laboratories, Idaho, USA, 2013.
20. Kim Y, Park J and Jung W. A classification scheme of erroneous behaviors for human error probability estimations based on simulator data. *Reliab Eng Syst Saf* 2017, 163: 1-13.
21. Groth KM, Smith CL, and Swiler LP. A Bayesian method for using simulator data to enhance human error probabilities assigned by existing HRA methods. *Reliab Eng Syst Saf* 2014, 128 (Supplement C): 32-40.
22. Kim Y, Park J, Jung W, et al. Estimating the quantitative relation between PSFs and HEPs from full-scope simulator data. *Reliab Eng Syst Saf* 2018, 173: 12-22.
23. Chang JY and Franklin C. SACADA Data for HEP Estimates. In: *14th Probabilistic Safety Assessment and Management, PSAM 14 2018*, UCLA Meyer & Renee Luskin Conference Center, Los Angeles, California.
24. Groth KM. A framework for using SACADA to enhance the qualitative and quantitative basis of HRA. In: *14th Reliability Safety Assessment and Management, PSAM 14 2018*, UCLA Meyer & Renee Luskin Conference Center, Los Angeles, California.
25. Azarm MA, Kim IS, Marks C, et al. Analyses methods and pilot applications of SACADA database. In: *14th Probabilistic Safety Assessment and Management, PSAM 14 2018*, UCLA Meyer & Renee Luskin Conference Center, Los Angeles, California.
26. Nelson PF and Grantom CR. Methodology for Supporting the Determination of Human Error Probabilities from Simulator Sourced Data. In: *14th Reliability Safety Assessment and Management, PSAM 14 2018*, UCLA Meyer & Renee Luskin Conference Center, Los Angeles, California.
27. Jung W, Park J, Kim Y, et al. HuREX – A framework of HRA data collection from simulators in nuclear power plants. *Reliab Eng Syst Saf* 2020, 194: 106235.
28. Siu NO and Kelly DL. Bayesian parameter estimation in probabilistic risk assessment. *Reliab Eng Syst Saf* 1998, 62(1): 89-116.
29. Kelly DL and Smith CL. Bayesian inference in probabilistic risk assessment - The current state of the art. *Reliab Eng Syst Saf* 2009, 94(2): 628-643.
30. Apostolakis G, Kaplan S, Garrick BJ, et al. Data specialization for plant specific risk studies. *Nucl Eng Des* 1980, 56(2): 321-329.
31. Kaplan S. On a two-stage Bayesian procedure for determining failure rates. *IEEE Trans Power Apparatus Syst* 1983, 102(1): 195–262.
32. Droguett E, Groen F and Mosleh A. Bayesian assessment of the variability of reliability measures. *Pesquisa Operacional* 2006, 26:109-127.
33. Yue M and Chu T-L. Estimation of failure rates of digital components using a hierarchical Bayesian method. In: *8th Probabilistic Safety Assessment and Management, PSAM 8 2006*, New Orleans, Louisiana.

34. Mosleh A. Bayesian modeling of expert-to-expert variability and dependence in estimating rare event frequencies. *Reliab Eng Syst Saf* 1992, 38(1-2).
35. Podofillini L and Dang VN. A Bayesian Approach to Treat Expert-Elicited Probabilities in Human Reliability Analysis Model Construction. *Reliab Eng Syst Saf* 2013, 117: 52-64.
36. Droguett EL, Groen F and Mosleh A. The combined use of data and expert estimates in population variability analysis. *Reliab Eng Syst Saf* 2004, 83(3): 311-321.
37. VanDerHorn E and Mahadevan S. Bayesian model updating with summarized statistical and reliability data. *Reliab Eng Syst Saf* 2018, 172: 12-24.
38. Mosleh A and Chang YH. Model-based human reliability analysis: prospects and requirements. *Reliab Eng Syst Saf* 2004, 83: 241–253.
39. Mosleh A and Smith C. The Feasibility Of Employing Bayesian Techniques And Other Mathematical Formalisms In Human Reliability Analysis, in The Employment of Empirical Data and Bayesian Methods in Human Reliability Analysis: A Feasibility Study, NUREG/CR-6949, pp. 5-15, INL/EXT-06-11670, Washington, D.C.: U.S. Nuclear Regulatory Commission, 2007.
40. Atwood CL. Constrained noninformative priors in risk assessments. *Reliab Eng Syst Saf* 1996, 53(1): 37-46.
41. Hallbert B, Gertman D, Lois E, et al. The use of empirical data sources in HRA. *Reliab Eng Syst Saf* 2004, 83: 139-143.
42. NEA/CSNI. Research on Human Factors in New Nuclear Plant Technology. NEA/CSNI/R(2009)7, Nuclear Energy Agency, 2009.
43. Prvakova S and Dang VN. A review of the current status of HRA data. In: *Proceedings of the European Safety and Reliability Conference, ESREL 2013*, Amsterdam, Netherlands.
44. Skjerve AB and Bye A. *Simulator-based Human Factor Studies across 25 years*. Springer-Verlag London, 2011.
45. Hannaman GW, Spurgin AJ and Lukic Y. A Model for Assessing Human Cognitive Reliability in PRA studies. In: *IEEE Third Conference on Human Factors in Nuclear Power Plants*, Monterey, California, June 23-27 1985, Institute of Electronic and Electrical Engineers, New York (USA).
46. Hannaman GW, Spurgin AJ and Lukic Y. Human cognitive reliability model for PRA analysis. NUS-4531, EPRI Electric Power Research Institute, Palo Alto, California, 1984.
47. Moieni P, Spurgin AJ and Singh A. Advances in Human Reliability Analysis Methodology. Part I: Frameworks, Models and Data. *Reliab Eng Syst Saf* 1994, 44: 27–55.
48. Massaiu S and Holmgren L. Diagnosis and Decision-Making with Emergency Operating Procedures in Non-Typical Conditions: A HAMMLAB Study with U.S. Operators. HWR-1121. Halden, Norway: OECD Halden Reactor Project, 2014.
49. Massaiu S and Holmgren L. The 2013 Resilient Procedure Use Study with Swedish Operators: Final Results. HWR-1216. Halden, Norway: OECD Halden Reactor Project, 2017.
50. Mkrtchyan L, Podofillini L and Dang VN. Methods for building Conditional Probability Tables of Bayesian Belief Networks from limited judgment: An evaluation for Human Reliability Application. *Reliab Eng Syst Saf* 2016, 151: 93-112.
51. Gelman A, Carlin J, Stern H, et al. *Bayesian Data Analysis*, *Second Edition*. Chapman and Hall/CRC, 2003.

52. Kelly DL and Smith CL. *Bayesian Inference for Probabilistic Risk Assessment: A Practitioner's Guidebook*. London, UK: Springer-Verlag, 2011.
53. Venables WN. and Smith DM. *An Introduction to R. Notes on R: A Programming Environment for Data Analysis and Graphics*. *Version 3.6.1 (2019-07-05)*. 2019.

Chapter 3: Behavioral patterns to model crew performance variability

This chapter reproduces the author's article submitted to the Journal of Risk and Reliability (see "Publication details" below). Building on the variability model presented in the previous Chapter 2, this article proposes an alternative modelling approach, based on the concept of "behavioral patterns", to focus on the identification of crew performance drivers from simulator data and empirically incorporate their effects on the HEP variability distribution (research objective #2 in Section 1.2).

This chapter first introduces the use of behavioral patterns to categorize crew behavioral characteristics (e.g. in team decision-making, communication strategies, adherence to procedures) emerging from simulator observations, and represent the spectrum of performance variability over a finite ("discrete") set of crew behavioral groups, for a given constellation of task and PSF categories (research task 2.1, in Section 1.2). The chapter shows how the formulation with behavioral patterns is included in a new Bayesian hierarchical model, to quantitatively capture performance variability across the identified set of groups (research task 2.2). The chapter then presents the multi-step methodology to identify behavioral patterns from simulator data and use these to inform the crew behavioral groups of the Bayesian hierarchical model (research task 2.3). Finally, the chapter demonstrates the multi-step methodology on a case study from literature, involving crew behaviors observed in different emergency scenarios from recent simulator studies (research task 2.4). The chapter concludes discussing the benefits (as well as the limitations) of the proposed methodology with respect to HRA applications.

Publication details

This article is reproduced with permission from: **Greco SF**, Podofillini L and Dang VN. Crew performance variability in human error probability quantification: a methodology based on behavioral patterns from simulator data. *Proc I Mech E Part O: J Risk and Reliability* 2021, doi:10.1177/1748006X20986743.

Additional information relevant to this chapter

- The list of teamwork competences and metrics used to support the categorization of crew behaviors in the case study is presented in Appendix D.
- The results from the sensitivity analysis performed in the application to case study are reported in Appendix E.
- The code developed for the implementation of the Bayesian models compared in the numerical application is provided in Appendix F

Abstract

Current Human Reliability Analysis models express error probabilities as a function of task types and operational context, without explicitly modelling the influence of different crew behavioral characteristics on the error probability. The influence of such variability is treated only implicitly, by variability and uncertainty distributions with bounds primarily obtained by expert judgment. This chapter presents a methodology to empirically incorporate crew performance variability in error probability quantification, from simulator data. Crew behaviors are represented by a set of "behavioral patterns" that emerge in the observation of operating crews (e.g. in information sharing or in adhering to procedural guidance). The work demonstrates the use of a Bayesian hierarchical model to explicitly capture the performance variability emerging from data. The methodology is applied to a case study from literature. Numerical demonstrations are performed in order to compare the proposed approach to the existing quantification models used in HRA for treating simulator data.

3.1 Introduction

Human Reliability Analysis (HRA) assesses the contribution of human failures to the overall risk profile of industrial systems, e.g. nuclear power plants, chemical facilities and aerospace systems [1-2]. HRA methods support analysts to identify the safety-critical tasks performed by the personnel (e.g. operating crews in nuclear power plants), characterize the contextual factors influencing performance (the so-called Performance Shaping Factors, PSFs), and quantify the associated error probability (referred to as Human Error Probability, HEP). The HEPs are generally used in risk analysis for the quantification of the frequency of accident scenarios, typically in Probabilistic Safety Assessment (PSA).

HRA methods use quantitative models to produce HEP values depending on the task to be performed and the associated operational context [3], both represented by sets of categories (typically, task types and PSF levels/ratings). Through these categories, HRA models produce HEP values as a function of scenario-, task-, context-specific influences. HRA acknowledges that other aspects such as organizational factors as well as personal and team characteristics can have important influence on crew performance variability and, to some extent, addresses these in the qualitative analysis supporting HEP quantification [4-9]. However, their influence is typically not explicitly considered as input factors to quantitative HRA models (e.g. as PSFs) but implicitly, typically within the variability and uncertainty ranges associated to the HEP values [4, 10].

In recent years, the HRA Empirical Studies (the International [11] and the US [12]) highlighted the key importance of several crew behavioral aspects, such as "team dynamics, work processes, communication strategies, sense of urgency and willingness to take knowledge-based actions" [11], as main contributors to performance variability in operational tasks, especially in emergency situations where standard procedure following is challenged by a fast scenario progression and a limited procedural guidance. In such performance conditions, crew characteristics (e.g. in information sharing, task prioritization, adherence to procedural guidance) played a key role in determining not only the pace through the procedures, but also which procedural path to follow [11-12]. More recent studies in nuclear power plant control room simulators [13-14] further underscored that, for emergency scenarios characterized by a procedure-situation mismatch, "the crews that followed the procedures more strictly had lower performance than crews that engaged more in autonomous initiatives and extra-procedural activities". These works [11-14] acknowledged the benefits of using simulator studies to investigate the effects of crew behavioral characteristics on performance variability in operational tasks as well as the need to formally incorporate these in the HEP quantification, especially for those "scenarios that exceed the limits of the basic nuclear power plant design" and "include multiple equipment failures" [11]. Indeed, incorporation of some crew variability aspects in HRA is one of the distinctive characteristics of the emerging modern HRA methods, for example through the use of Crew Response Diagrams in the Integrated Human Event Analysis System (IDHEAS) method [15] or Crew Response Trees in [16].

In view of the increasing use of PSA and HRA results in licensing and operational decisions of nuclear power plants, HRA data collection from main control room simulators have gained new momentum [17-19]. Long-term, international simulator programs have been established, aiming at strengthening the empirical basis of future HEP estimates as well as at deriving insights for improving operating crew performance [20-21]. The exploratory approaches for the quantification of HEPs from the emerging data [22-24] maintained the traditional HEP formulation as a function of scenario-, task- and context-related factors, lumping together all other influences and performance variability aspects. These pioneering works focused on population-averaged HEP values, where the influence of other factors on the HEP values are thought of as a statistical population. These works demonstrated the advantages of using Bayesian methods (e.g. conjugate beta-binomial models [22-23]) in quantifying the HEP for sets of task and PSF categories of data collection taxonomies [20-21], but did not address the actual variability (e.g. organizational, plant, team and personal) within these sets of categories [25]. As the on-going data collection efforts will provide more evidence, it becomes important to strengthen the empirical basis of both the averaged HEP values, as well as of the HEP spectrum of variability and uncertainty, for the categories of HRA models.

Previous work by the same authors have addressed crew performance variability as a continuum, without distinguishing crew behavioral characteristics in HEP quantification from simulator data [25[1]-26]. In order to explicitly address these characteristics, this chapter puts

---

[1] In this thesis, the referred article is reproduced as Chapter 2: A Bayesian variability model for simulator data.

forward a new methodology based on the identification of “behavioral patterns” manifested during task performance (e.g. “collective” or “non-inclusive” information sharing, “proactive” or “reactive” interpretation of procedures). The analysis via behavioral patterns builds on literature works on models of crew response in emergency situations for simulation-based applications [27] and retrospective analysis of past event [28]. Similarly to the present work, both works interpret variability in crew behaviors as the result of the dynamic interaction between crew-specific and task-, context-related factors (modelled by “performance adjustment factors” in [27] and by “situation factors” in [28]). However, neither of these works had the objective of incorporating performance variability in HEP quantification.

The identified set of behavioral patterns is included in a variability model to capture the influence of different crew behavioral groups on the error probability, for a given combination of task type and PSF ratings (representing the given scenario-, task- and context-related influences). The underlying concept is that crews sharing similar patterns are aggregated in the same behavioral group and associated the same value of error probability. A Bayesian hierarchical model is then used as framework for the HEP quantification from simulator data. Bayesian hierarchical models have been widely adopted in probabilistic safety assessment to treat source-to-source variability [29-36], as well as in many other applications for inference of population-level quantities from group-level evidence and vice versa [37-42].

The chapter is structured as follows. Section 3.2 first introduces the concept of crew behavioral patterns to characterize behavioral aspects in nuclear power plant operations. Section 3.2 then presents how the patterns are quantitatively incorporated in the model for crew performance variability in HEP estimation. Section 3.3 presents the methodology as two blocks: the first block derives the behavioral categories emerging from the simulator data and the second block groups the crews based on patterns of behavioral categories and quantifies the associated HEP. Section 3.4 presents the application of the methodology to a case study from literature, involving diagnosis tasks performed in different emergency scenarios [12, 43]. Crew behavioral aspects empirically observed during task performance are systematically characterized using a taxonomy of teamwork competences for nuclear power plant operating crews [44]. The results from the numerical application are compared to alternative quantitative approaches for simulator data [22-23, 25] to demonstrate the effects of incorporating operating crew behavioral variability on HEP estimates. The application and the underlying model assumptions are further discussed in Section 3.5, along with recommendations on the feasibility and applicability of the proposed methodology to HRA problems. Conclusions are given at closure.

3.2 Concepts: behavioral patterns from simulator data and variability modelling

3.2.1 Behavioral patterns: definition and relationship with typical HRA quantification

Figure 3.1 shows the relationship between the scope of the factors typically considered by HRA models, with respect to the whole set of human and organizational factor influences (an overview of the whole set of influences can be found in Appendix A of [45]): the figure also compares the factor-HEP links in typical models and in the present work. The models used in HRA explicitly address factors characterizing the operator tasks, as well as the scenario and context in which the tasks are carried out (e.g. adequacy of procedural guidance, of time available, human-machine interface). Examples are the generic task types (e.g. "shift or restore system to a new or original state") and error producing conditions (e.g. "poor, ambiguous or ill-matched system feedback") in the Human Error Assessment and Reduction Technique (HEART, [5-6], newly issued in [46]); examples from newer methods are the crew macro-cognitive functions (e.g. "action", "detecting and noticing") and performance influencing factors (e.g. "high" or "low" workload, "poor" or "good" human-system interface) in IDHEAS [15]. Similar factor scope can be found in all other HRA methods, for example in the Technique for Human Error Rate Prediction (THERP, [4]), the Standardized Plant Analysis Risk Human Reliability Analysis (SPAR-H, [8]), and the Cognitive Reliability and Error Analysis Method (CREAM, [7]), to name a few.

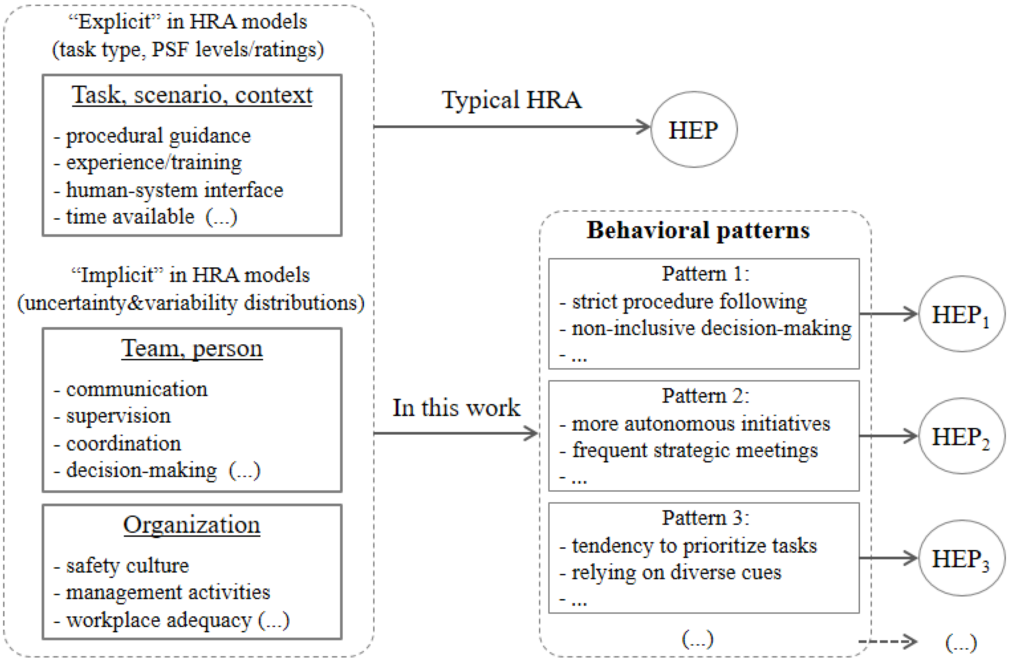


Figure 3.1. Relationship between performance influencing factors (taxonomy from IAEA [45]) and behavioral patterns used in this work to represent crew performance variability in HEP quantification.

The influence on human performance of the other human and organizational factors (e.g. team dynamics, work processes, communication strategies, as well as managerial and organizational factors) is generally considered in the variability and uncertainty distributions associated to the HEP, as shown in Figure 3.1 [4, 47]. The uncertainty and variability bounds account also for several other aspects of uncertainty in the HRA results, e.g. uncertainty on the assessment of the PSF ratings, epistemic uncertainty due to model limitation and scarcity of data [10]. The variability and uncertainty distributions and bounds are derived by expert judgment. The main source is represented by the values proposed in the THERP handbook [4], themselves based on THERP authors' judgment. One exception is the HEART method, in which the HEP uncertainty bounds are derived from human error data across different industries. The HEART bounds indeed reflect the empirical variability of the data, but their quantification does not explicitly address the source of the performance variability (the behavioral aspects that result in variability in performance and, consequently, in the HEP).

This work presents a first-of-a-kind attempt to empirically include crew performance variability in the HEP quantification, from simulator data. The concept blends elements from classical HRA methods as well as human factor studies, especially teamwork, decision-making and situation awareness studies in main control room simulators. In the proposed quantification model, the HEP is still expressed as a function of task-, scenario-, and context-based factors (task type and PSF levels/ratings in Figure 3.1), as in typical HRA models. On the other hand, human performance variability is captured by different "patterns" of crew behavioral categories (in teamwork, decision-making and situation awareness) emerging from simulator observations. As shown in Figure 3.1, "behavioral patterns" are interpreted as manifestations of the overall spectrum of influences: task, scenario, context, as well as person, team and organizational ones. Therefore, similar to typical HRA quantification models, the HEP is expressed explicitly as a function of task-, scenario-, and context-based factors. Differently, in the proposed concept, HEP variability is expressed via a model (based on behavioral differences across groups of crews) and estimated from empirical data, whereas in most other HRA models HEP variability is not incorporated and not informed by data.

The work addresses performance data from large-scale simulator programs (e.g. the HUman Reliability data EXtraction framework, HuREX [21]; the Scenario Authoring, Characterization, And Debriefing Application, SACADA [20]), an example of which is provided in Table 3.1. Data comes in the form of records of performance outcome (failure/success), behaviors gathered from different plants and operating crews, performing tasks in different simulated emergency scenarios (e.g. in Table 3.1, identification of the faulted steam generator in a SGTR scenario), under a given combination of PSF levels. The quantity of interest in this work is the HEP associated to a given set of task type / PSF levels (referred to as set $\boldsymbol{F}$) adopted by the specific data collection taxonomy: $HEP = HEP(\boldsymbol{F})$. For instance, in Table 3.1 (from SACADA taxonomy), ***F*** represents the task type "understanding the situation/problem" and PSF "information quality" with level "missing/masked" (the latter capturing the operational context "failure of secondary radiation indications"). Depending on the taxonomy, the PSF levels can be defined as a binary (e.g. low/high; adequate/not adequate) or multi-valued (e.g. rating) variable.

Besides information on tasks and PSFs, i.e. the set ***F***, the proposed methodology requires information on observed crew behaviors to populate the behavioral patterns, such as those in the last column of Table 3.1. Note that the current version of the HuREX taxonomy does not foresee the collection of such observed behaviors. For SACADA, such details on performance are foreseen only if failures or any performance issues are observed, but not for every simulator run as shown in the exemplification case in Table 3.1. This indeed has implications on the possibility to apply the proposed model to the currently available HRA data, as further discussed in Section 3.5.

In the present work, the crew behaviors collected for a given set ***F*** (Table 3.1, last column) are systematically analyzed adopting teamwork, decision-making and situation awareness taxonomies and classified into "behavioral categories" accordingly, for instance: concerning communication, the frequency with which strategic meetings are held (e.g. "frequent strategic meetings" in Figure 3.1); concerning work attitudes, the compliance to procedure indications (e.g. "strict procedure following" or "more autonomous initiatives" in Figure 3.1), and the like. Each crew performance is then represented by a specific combination (i.e. a specific pattern) of behavioral categories (see examples in Figure 3.1): according to this classification, crew performances can be clustered in "behavioral groups" (each group being identified by a specific behavioral pattern), representing the spectrum of performance variability empirically observed for the set ***F***. Each behavioral group is then associated an HEP value (Figure 3.1) in the variability model presented in the next subsection 3.2.2. This concept emphasizes the impact of crew behavioral characteristics on performance and, ultimately, on the resulting HEP value. For instance, Forester et al. [11] observed several crews performing a complex diagnosis tasks with masked indications (defining the set ***F***): seven crews "followed procedures too literally" with "no structured meeting for decision making", a pattern leading to five failures (five failures out of seven); two crews "investigated alternative causes to the increasing level" in the ruptured steam generator and overall were "well updated on the process" thanks to frequent meetings, a pattern resulting in task success (no failures out of 2). Similar situations can be found in [13].

The following list briefly restates the key terminology used in Section 3.2, in order to support the understanding of model development in the remainder of this subchapter, as well as the methodology presented in Section 3.3:

- "set ***F***": set of task and PSF categories, respectively representing the task characteristics and the operational context (e.g. "understanding the situation/problem", PSF "information quality" with level "missing/masked"). Category definitions vary with the given data collection taxonomy (e.g. SACADA [20], HuREX [21]);

- "crew behaviors": behaviors observed during crew performances in simulated scenarios, typically recorded in simulator logs (examples in Table 3.1, last column). Represent the "observable" of crew behavioral characteristics (in teamwork, decision-making and situation awareness) emerging from simulator observations;
- "behavioral categories": classification of the crew observed behaviors via categorical definitions (from Figure 3.1: "strict procedure following", "frequent strategic meetings"). In this work, behavioral categories are intended to represent the relevant aspects of teamwork, decision-making and situation awareness in crew performances. Definitions vary with the adopted taxonomy of metrics (e.g. [44]);
- "behavioral pattern": refers to a specific combination of the aforementioned categories (e.g. from Figure 3.1, pattern #1: "strict procedure following & non-inclusive decision making & [...]"). In this work, patterns are interpreted as the direct manifestation of the overall spectrum of influencing factors in Figure 3.1 (task, scenario, context, as well as person, team and organizational ones);
- "behavioral group": group of crews uniquely identified by a specific behavioral pattern (e.g. in Figure 3.1, the three patterns represent three different behavioral groups). All crew performances manifesting the same behavioral pattern are clustered in the same group and associated to a unique HEP value in the variability model (subsection 3.2.2). In this work, the set of behavioral groups emerging from data is used to model performance variability in the given $\boldsymbol{F}$.

Table 3.1. Grouping hypothetical data from different simulator contexts to inform the set of categories (***F***) of a generic HRA model. Operational contexts and crew observed behaviors are adapted from [13, 43].

Set ***F***: task type = "understanding the situation/problem", PSF "information quality" = "missing/masked" (taxonomy from SACADA, [20])						
Scenario	Operational context	Task realization	Plant	Crews	Failures	Observed behaviors
SGTR	Failure of secondary radiation indications	Identification of faulted SG	A	5	2	Crew 1 (failure): "shift supervisor makes most decisions", "did not try extra procedural isolations"… Crew 2 (success): "performed isolations that were not contained in the procedures", "shift supervisor is hesitant about what to do"…
SGTR	Radiation alarms already activated by early releases	Identification of faulted SG	B	6	1	Crew 3 (success): "reactor operator works alone and does not wait for answers from the assistant", "shift supervisor is very active in asking questions, and discussing the situation with the crew"… Crew 4 (success): "shift supervisor quickly orders important actions", "worked well with extensive three-way communication"…
SGTR	(…)	(…)	(…)	(…)	(…)	(…)
			Total	50	12	
ISLOCA	No indications on leaks' specific location	Identification and isolation of leaks	A	5	3	Crew 5 (failure): "shift supervisor leads communication without having structured meetings", "board operators more involved in decisions"… Crew 6 (failure): "shift supervisor gives orders without discussion", "waits for the expected result without questioning the situation"…
ISLOCA	No indications on leaks' specific location	Identification and isolation of leaks	B	6	2	Crew 7 (failure): "investigated an alternative cause to the increasing level in steam generator", "stuck in discussions" … Crew 8 (success): "shift supervisor is good at prioritizing", "good updates and briefings"…
ISLOCA	(…)	(…)	(…)	(…)	(…)	(…)
			Total	50	15	

3.2.2 Using behavioral patterns in a variability model for HEP

This subsection presents the variability model for $HEP(\boldsymbol{F})$ (shown in Figure 3.2, left) to capture HEP variability across behavioral groups (the identification of the groups will be presented in Section 3.3).

The model is based on the assumption that each "behavioral group" (pedix c in Figure 3.2, left) is characterized by a unique error probability, $p_{c|\boldsymbol{F}}$; therefore, $p_{c|\boldsymbol{F}}$ is intended as the failure probability associated to the crews of the c-th group in performing a task described by the task type and PSF levels in the set $\boldsymbol{F}$. In this formulation, $p_{c|\boldsymbol{F}}$ represents possible outcomes of $HEP(\boldsymbol{F})$: the HEP is intended as a variable quantity, discretized over the number of identified behavioral groups (C in Figure 3.2, left). The $p_{c|\boldsymbol{F}}$'s (the arrows in Figure 3.2, left) are interpreted as group-specific realizations of the HEP variability in $\boldsymbol{F}$.

The variability across the $p_{c|\boldsymbol{F}}$'s is captured assuming that the $p_{c|\boldsymbol{F}}$'s are continuously distributed according to a parametric variability distribution, represented by the following function:

$$p_{c|\boldsymbol{F}} \sim f_{\boldsymbol{F}}(p_{c|\boldsymbol{F}}|\boldsymbol{\theta}_{\boldsymbol{F}}) \tag{3.1}$$

where $\boldsymbol{\theta}_{\boldsymbol{F}}$ represents the vector of the unknown parameters of the variability distribution (e.g. for a lognormal, the mean and standard deviation). The parameters in $\boldsymbol{\theta}_{\boldsymbol{F}}$ are uncertain quantities and are inferred from simulator data, aggregated within the c-th group (from here, "aggregation by groups" in Figure 3.2, left) in the form of observed failures and crew observations (respectively k_c and N_c in Figure 3.2, left).

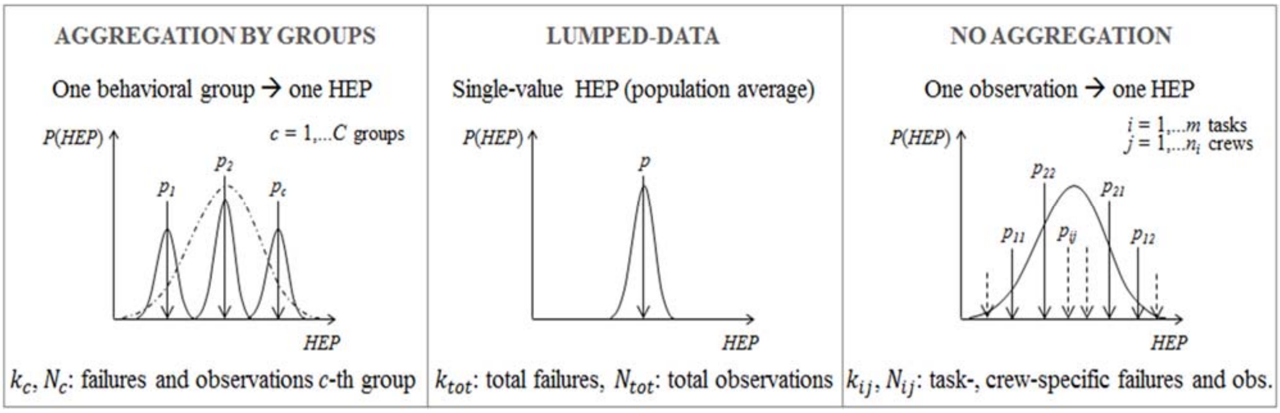

AGGREGATION BY GROUPS
One behavioral group → one HEP
P(HEP)
c = 1,...C groups
p_1
p_2
p_c
HEP
k_c, N_c: failures and observations c-th group
LUMPED-DATA
Single-value HEP (population average)
P(HEP)
p
HEP
k_{tot}: total failures, N_{tot}: total observations
NO AGGREGATION
One observation → one HEP
P(HEP)
i = 1,...m tasks
j = 1,...n_i crews
p_{22}
p_{21}
p_{11}
p_{ij}
p_{12}
HEP
k_{ij}, N_{ij}: task-, crew-specific failures and obs.

Figure 3.2. Comparison of the HEP formulations and the associated data aggregation adopted by the proposed variability model (left: "aggregation by groups") and the alternative approaches tested in the case study (center: "lumped-data", with HEP as population average; right: "no aggregation", with HEP as a "continuum" of task-, crew-specific error probabilities). All the p's are intended as conditional on the given set $\boldsymbol{F}$, e.g. $p_{c|\boldsymbol{F}}$, $p_{ij|\boldsymbol{F}}$.

In the numerical application of Section 3.4, the proposed variability model is tested and compared against two alternative modelling approaches for HEP quantification: a "lumped-data" model (as in [22-23]), and the "continuous" variability model presented in previous work by the same authors [25]. The lumped-data model (Figure 3.2, center) associates all the simulator records relevant to $\boldsymbol{F}$ (the rows in Table 3.1) to a single-value HEP, $p_{\boldsymbol{F}}$, i.e. the population average over the variability within $\boldsymbol{F}$: failures and crew observations relevant to $\boldsymbol{F}$ are lumped into a single piece of evidence (respectively k_{tot} and N_{tot} in Figure 3.2, center) to infer on the unique, unknown $p_{\boldsymbol{F}}$. The variability model proposed in Greco et al. [25] formulates performance variability in *HEP*($\boldsymbol{F}$) as a "continuum" of crew-, task-specific error probabilities, $p_{ij|\boldsymbol{F}}$'s (Figure 3.2, right). The variable $p_{ij|\boldsymbol{F}}$ models the failure probability of the j-th crew in performing the i-th task of a specific simulator scenario in data collection, i.e. one realization of the set $\boldsymbol{F}$ (e.g. from Table 3.1, identification and isolation of the leaks in a ISLOCA scenario). Similarly to the $p_{c|\boldsymbol{F}}$ variable in eq. 3.1, the $p_{ij|\boldsymbol{F}}$'s are assumed to be continuously distributed according to a known variability function, namely $f_{\boldsymbol{F}}(p_{ij|\boldsymbol{F}}|\boldsymbol{\theta}_{\boldsymbol{F}})$. Contrary to the formulation proposed in this work, the set of parameters $\boldsymbol{\theta}_{\boldsymbol{F}}$ of the continuous variability model is inferred from crew-, task-specific data in the form of couples k_{ij}/N_{ij}, respectively representing the k_{ij} failures observed for the i-th crew in N_{ij} repetitions of the i-th task (from here the term "no aggregation" in Figure 3.2, right). Simply put, Greco et al. [25] assumes different failure probabilities per each crew, while the present chapter per each behavioral group, aggregating different crews manifesting with similar behaviors.

The unknown parameters for the three mathematical formulations of *HEP*($\boldsymbol{F}$) described in this section (the single $p_{\boldsymbol{F}}$ in the lumped model, the sets $\boldsymbol{\theta}_{\boldsymbol{F}}$ in both the variability models) are derived from simulator data by Bayesian inference models and used to quantify the population-level HEP uncertainty distribution for the set $\boldsymbol{F}$, namely *P*(*HEP*). The development of the Bayesian models for the HEP quantification is discussed in details in subsection 3.3.3.

It is important to stress the conceptual differences in *HEP*($\boldsymbol{F}$) formulation between the three modelling approaches. Compared to the variability models, the lumped approach does not explicitly model performance variability across the crews but rather treat HEP as population average for the set of categories. The more data is collected for $\boldsymbol{F}$ (k_{tot} and N_{tot}), the more the epistemic uncertainty on the population average is reduced: ideally, with infinite data, the resulting *P*(*HEP*) will shrink to the single-value HEP ($p_{\boldsymbol{F}}$ in Figure 3.2, center). Compared to the quantitative approach proposed in this chapter, the continuous variability model captures performance variability in *HEP*($\boldsymbol{F}$) at a lower level with crew-, task-specific error probabilities. However, the continuous variability model does not formally consider the different behavioral characteristics manifested by the crews during task performance (in this chapter characterized by behavioral patterns): rather, it considers their behavioral differences in the realizations of the spectrum of *HEP*($\boldsymbol{F}$) variability (the $p_{ij|\boldsymbol{F}}$'s, i.e. the arrows in Figure 3.2, right). Compared to the continuous variability formulation, in this work performance variability in *HEP*($\boldsymbol{F}$) is modelled across behavioral groups, assuming the HEP population can be ideally represented by a finite number of group-specific error probabilities (the $p_{c|\boldsymbol{F}}$'s, i.e. the arrows in Figure 3.2, left). With increasing data available (k_{ij} and N_{ij} for the continuous variability model, k_c and N_c for the variability model with behavioral groups), epistemic uncertainty on the p's of both variability models is reduced and the *P*(*HEP*)'s estimated by the models tend to the actual *HEP*($\boldsymbol{F}$) variability distribution.

3.3 A methodology to incorporate crew behavioral patterns in HEP quantification

This section presents the multi-step methodology to identify the crew behavioral groups and account for them in the HEP quantification from simulator data (Figure 3.3). The methodology is presented for a generic combination of task type and PSF ratings ($\boldsymbol{F}$), e.g.: task type "understanding the situation/problem" and PSF "information quality" rated as "missing/masked", from SACADA taxonomy [20]; cognitive activity "response planning and instruction" and task type "transferring step in procedure", from HuREX [21]. In Section 3.4, it is applied to a specific $\boldsymbol{F}$ characterizing a case study from literature.

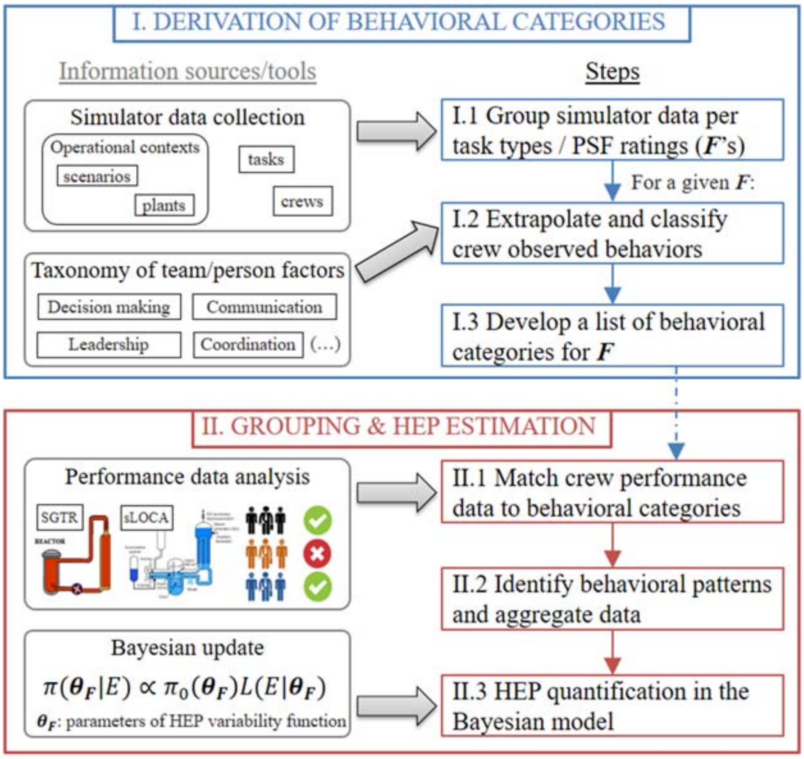

I. DERIVATION OF BEHAVIORAL CATEGORIES
Information sources/tools
Steps
Simulator data collection
Operational contexts
scenarios
plants
tasks
crews
I.1 Group simulator data per task types / PSF ratings (F's)
For a given F:
I.2 Extrapolate and classify crew observed behaviors
Taxonomy of team/person factors
Decision making
Communication
Leadership
Coordination
(...)
I.3 Develop a list of behavioral categories for F
II. GROUPING & HEP ESTIMATION
Performance data analysis
SGTR
sLOCA
REACTOR
II.1 Match crew performance data to behavioral categories
II.2 Identify behavioral patterns and aggregate data
Bayesian update
$\pi(\boldsymbol{\theta}_F|E) \propto \pi_0(\boldsymbol{\theta}_F)L(E|\boldsymbol{\theta}_F)$
$\boldsymbol{\theta}_F$: parameters of HEP variability function
II.3 HEP quantification in the Bayesian model

Figure 3.3. Overview of the multi-step methodology to derive and use behavioral patterns in HEP quantification for a generic set $\boldsymbol{F}$.

The methodology comprises two blocks (Figure 3.3). The first block (Figure 3.3, blue box) derives the behavioral categories emerging from the simulator data relevant to the combination ***F***. The second block (Figure 3.3, red box) groups the crews based on patterns of behavioral categories and quantifies the associated HEP. The set of behavioral categories can indeed be already available from other studies: in this case, the second block can be applied directly.

3.3.1 Derivation of behavioral categories from data collection

Steps I.1-I.3 in Figure 3.3 (blue box) address the derivation of the behavioral categories:

I.1. grouping simulator data per task type / PSF ratings,
I.2. extrapolation and classification of crew observed behaviors,
I.3. development of a list of behavioral categories.

In step I.1, the simulator records are grouped by different ***F***'s, where each ***F*** represents a combination of task types and PSF ratings for which data is available. The definition of representative sets ***F*** depends on the purpose of the application. For instance, if interested in deriving HEP estimates for the task categories of a data collection taxonomy (similarly to [23], for HuREX taxonomy), then the set ***F*** reduces to a single element, i.e. the specific task type of interest (e.g. from [23]: "transferring step in procedure"), grouping the observations from all the relevant task realizations in data collection. On the other hand, if interested in the effect of a specific combination of PSFs on task HEP (e.g. to inform an HRA model, as in [22] with the SPAR-H), then the set ***F*** comprehends both task type and PSF ratings (e.g. from [22], ***F***: {task type: "action"; PSF: "time available" with rating "barely adequate, PSF: "procedures" with rating "available but poor", etc.}).

The proposed methodology is intended to identify a manageable set of patterns for the given set ***F*** (e.g. in Table 3.1, task type "understanding the situation/problem" and PSF "information quality" rated as "missing/masked"), comprehensive enough, but not leading to a combinatorial explosion of possibilities. This requires a set of behavioral indicators ("metrics") in order to support the classification of crew behaviors (step I.2) across the respective team- and person-based performance influencing factors discussed in subsection 3.2.1 (e.g. in Figure 3.1, in communication, supervision, coordination etc.).

Different taxonomies of metrics in teamwork and individual aspects of nuclear power plant operations are available in literature [44, 48-49]. Amongst those examined, the taxonomy provided by Skjerve and Holmgren [44] was selected by the authors for the purposes of this work. This taxonomy accomplishes two important requirements. First, it comprehensively covers a broad range of team- and person-based factors: attitudes, communication, coordination, decision making, interpersonal competences, leadership, and situation awareness. Second, being the taxonomy originally derived to support the data collection protocol for

Halden simulator [50], the metrics provided per each dimension are compatible with what is "observable" in the context of a simulator study during different operational phases (normal operations, outage, emergency situations). This aspect eases the interpretation of crew behaviors in a given operational context and allows for a systematic classification of behaviors across the teamwork and individual dimensions. An example of the classification in step I.2 is provided in Table 3.2 (second column), with crew behaviors (first column) adapted from [13, 43]. The full list of factor-specific metrics can be found in [44].

In step I.3, the behavioral categories are derived from crew behaviors and assigned to labels reflecting the classification performed in step I.2. For example, in Table 3.2: for behaviors relevant to "team orientation in decision making", the categories "collective decision making" or "non-inclusive decision making" characterize crews within which all members were involved in the decision process or the supervisor took most of the decisions, respectively; for "progression in decision making", "prioritizing, fast decision maker" or "hesitating, slowly building up" refer to crews showing the tendency to prioritize goals and resources or a step-by-step progression during the scenario, respectively. A more detailed description of the categories shown in Table 3.2 is given in the application of steps I.1-I.3 to the case study (Section 3.4, Table 3.5).

Different modelling aspects should be considered when developing the list of behavioral categories for a given $\boldsymbol{F}$. First, the same category of behaviors can have different influences on task performance, based on the scenario progression. For example, in a complex diagnosis task (e.g. from Table 3.1, the identification of the ruptured steam generator in a STGR scenario masked by the failure of secondary radiation indications), a "collective decision making" can have positive effects on the diagnosis at an early stage of the scenario, when more time is available to the crew. On the other hand, the same category can have negative effects when the diagnosis is performed in the final phase of the scenario (e.g. due to a slow progression in previous tasks of the operational sequence). In the latter case, with limited time available for the diagnosis, a participatory approach in decision making can delay the diagnosis as opposed to a more authoritarian approach ("non-inclusive decision making"). Considering this aspect, the behavioral categories should be defined with "neutral" attributes (see definitions in Table 3.2) rather than being *a priori* characterized as "negative" or "positive".

Second, the number of categories identified for $\boldsymbol{F}$ is expected to grow with increasing available data: taking as reference behaviors relevant to "progression in decision making" in Table 3.2, a third category could emerge from simulator observations, e.g. "fast decision maker without prioritizing". This aspect can have practical implications on HEP estimation, considering that a larger number of categories potentially leads to a larger number of patterns identified across crews and consequently hinder data aggregation in crew groups (in step II.2 in Figure 3.3, red box). On the other hand, with limited data, a small number of categories may not adequately represent the performance variability observed across crews for the set ***F***. There is obviously not an "optimum" number of behavioral categories: being an empirically-driven process, the number will depend on the information available from simulator observations. Data analysis and statistical tests could be used in step I.3 to rank the most relevant categories for the given set ***F*** and inform the final list accordingly (e.g. ruling out the categories with no meaningful impact on task performance). On the other hand, when simulator observations are not sufficient to apply data analysis tools with statistically significant results, the set of categories preliminarily identified from available data could be refined by expert-based aggregation, consistently with the purposes of the application. As a general rule, the authors recommend avoiding partially-overlapping definitions and to aggregate, as reasonably as possible, affine behavioral aspects into the same category (e.g. in Table 3.2, the behaviors "crew worked well with extensive three-way communication" and "good updates and briefings" are enveloped as different realizations of the category "adhering" in "adherence to communication and meeting protocol").

3.3.2 Grouping crew performance data and HEP quantification

Steps II.1-II.3 in Figure 3.3 (red box) address the use of behavioral patterns to group performance data and estimate the HEP for the set ***F***:

II.1. matching crew performance data to behavioral categories,
II.2. identification of behavioral patterns and aggregation in crew groups,
II.3. HEP quantification in the Bayesian model.

In step II.1, for each simulator record associated to ***F***, crew behaviors reported in performance data are analyzed and matched to the relevant behavioral categories. Examples of matching are shown in Table 3.3, with reference to the categories reported in Table 3.2: for instance, a crew within which "the shift supervisor leads the communication without having structured meetings" and "board operators are more involved in decisions" is matched to the categories "diverging" in "adherence to communication and meeting protocol" and "collective" in "team orientation in decision making"; a second crew "investigated an alternative cause to the increasing level in steam generator" in a SGTR scenario during which members were often "stuck in discussions", both behaviors corresponding to the categories "beyond/proactive" in "adherence to/interpretation of procedures" and "hesitating, slowly building up" in "progression in decision making". Note that the crew factor-specific behavioral metrics in [44] can also be used to support the matching in step II.1, in case a list of behavioral categories is already available from external sources (e.g. from previous applications of steps I.1-I.3 to the same set ***F***).

Table 3.2. Derivation of behavioral categories from crew observed behaviors classified by team-, person-based factors: examples of application of steps I.2-I.3 in Figure 3.3, adapted from the case study in subchapter 3.4.

Crew observed behaviors (from Table 3.1)	Classification by team-, person-based factors and associated metrics (taxonomy from [44]).	Behavioral categories
(a) “board operators more involved in decisions”, “shift supervisor is very active in asking questions, and discussing the situation with the crew”. (b) “shift supervisor makes most decisions”, “shift supervisor gives orders without much discussion”.	COMMUNICATION: upholding continuous communication during complex situations to promote collective sense-making. LEADERSHIP: developing strategies based on consultations with subordinates; mastering a more authoritarian leadership style during emergencies. ATTITUDE: team orientation.	In “Team orientation in decision making”: (a) Collective (b) Non-inclusive
(c) “shift supervisor is good at prioritizing”, “shift supervisor quickly orders important actions”. (d) “shift supervisor is hesitant about what to do”, “crew is stuck in discussions”.	LEADERSHIP: setting well-defined, realistic goals. DECISION MAKING: prioritize safety goals and concerns; Stop-Think-Act-Reflect when needed; develop a tactic/strategy for how to achieve performance goal.	In “Progression in decision making”: (c) Prioritizing, fast decision maker (d) Hesitating, slowly building up
(e) “crew worked well with extensive three-way communication”, “good updates and briefings”. (f) “shift supervisor leads the communication without having structured meetings”, “reactor operator works alone and does not wait for answers from the assistant”.	COMMUNICATION: three-way; active listening and follow up/verify/provide feedback. COORDINATION: carry out pre-job briefings when required/needed. SITUATION AWARENESS: informing colleagues when initiating important tasks.	In “Adherence to communication and meeting protocol”: (e) Adhering (f) Diverging
(g) “crew performed isolations that were not contained in the procedures”, “crew investigated an alternative cause to the increasing level in steam generator” (h) “crew did not try extra procedural isolations”, “crew waits for expected result, instead of questioning the situation”	COORDINATION: proactivity: think ahead possibilities for optimizing activities. DECISION MAKING: thinking outside the box. ATTITUDE: uphold a questioning attitude and willingness to consider a situation from multiple perspectives.	In “Adherence to / interpretation of procedures”: (g) Beyond / proactive (h) Close / reactive

Table 3.3. Matching crew behaviors to behavioral patterns: examples of application of steps II.1-II.2 in Figure 3.3. Note that a predefined list of behavioral categories has to be available prior to the matching, e.g. from the application of steps I.1-I.3 in Figure 3.3.

Performance data (Table 3.2)	Behavioral patterns and associated categories	Failures k_c, observations N_c
Crew 1 (failure): "shift supervisor makes most decisions", "did not try extra procedural isolations"… Crew 6 (failure): "shift supervisor gives orders without discussion". "waits for expected result, instead of questioning the situation"…	Pattern 1 Team orientation in decision making: "non-inclusive" + Adherence to/interpretation of procedures: "close/reactive" + ...	k_c= 5, N_c= 6
Crew 2 (success): "performed isolations that were not contained in the procedures", "shift supervisor is hesitant about what to do"… Crew 7 (failure): "investigated an alternative cause to the increasing level in steam generator", "stuck in discussions" …	Pattern 2 Adherence to/interpretation of procedures: "beyond/proactive" + Progression in decision making: "hesitating, slowly building up" + ...	k_c= 2, N_c= 5
Crew 3 (success): "reactor operator works alone and does not wait for answers from assistant", "shift supervisor is very active in asking questions, and discussing the situation with the crew"… Crew 5 (failure): "shift supervisor leads the communication without structured meetings", "board operators more involved in decisions"…	Pattern 3 Adherence to communication and meeting protocol: "diverging" + Team orientation in decision making: "collective" + ...	k_c= 2, N_c= 4
Crew 4 (success): "shift supervisor quickly orders important actions", "worked well with extensive three-way communication"… Crew 8 (success): "shift supervisor is good at prioritizing", "good updates and briefings"…	Pattern 4 Progression in decision making: "prioritizing, fast decision maker" + Adherence to communication and meeting protocol: "adhering" + ...	k_c= 1, N_c= 10
(...)	(...)	k_{tot}= 22, N_{tot}= 50

In step II.2, combinations of behavioral categories emerging across crew performances are identified and clustered as behavioral patterns. For instance, in Table 3.3, “pattern 1” refers to all crews manifesting a “non-inclusive decision making process” and a “close adherence to procedures” during the respective performances; “pattern 2” comprehends crews performing with a “proactive interpretation of procedures” and “slowly building-up in their decision making process”. The output of step II.2 is therefore an aggregated dataset populated by group-specific failures and crew observations (k_c and N_c in Table 3.3).

In the last step of the methodology, i.e. II.3, the aggregated dataset enters as input in the Bayesian model in order to infer on the group-specific error probabilities (the $p_{c|\boldsymbol{F}}$’s in eq. 3.1) and quantify the HEP uncertainty distribution, i.e. the $P(HEP)$, for the set $\boldsymbol{F}$. In the Bayesian framework [10], the initial degree of belief on the parameters of the HEP variability function ($\boldsymbol{\theta}_{\boldsymbol{F}}$ in eq. 3.1) is modelled by the so-called “prior distribution”, $\pi_0(\boldsymbol{\theta}_{\boldsymbol{F}})$ in Figure 3.3. The prior is updated by the group-specific simulator data (the “evidence” E in Figure 3.3) in the likelihood function, i.e. $L(E|\boldsymbol{\theta}_{\boldsymbol{F}})$ in Figure 3.3. The output of the Bayesian update (i.e. the “posterior distribution” $\pi(\boldsymbol{\theta}_{\boldsymbol{F}}|E)$ in Figure 3.3) represents the final state of knowledge on model parameters after the evidence. The $P(HEP)$ associated to the set $\boldsymbol{F}$ is eventually derived by averaging the variability function (i.e. $f_{\boldsymbol{F}}(p_{c|\boldsymbol{F}}|\boldsymbol{\theta}_{\boldsymbol{F}})$ in eq. 3.1) over the posterior $\pi(\boldsymbol{\theta}_{\boldsymbol{F}}|E)$:

$$P(HEP) = \int_{\boldsymbol{\theta}_{\boldsymbol{F}}} f_{\boldsymbol{F}}(p_{c|\boldsymbol{F}}|\boldsymbol{\theta}_{\boldsymbol{F}})\pi(\boldsymbol{\theta}_{\boldsymbol{F}}|E)d\boldsymbol{\theta}_{\boldsymbol{F}} \tag{3.2}$$

The development of the Bayesian model is discussed in details in the next subsection 3.3.3.

3.3.3 Development and implementation of the Bayesian model

In the numerical application (Section 3.4), the $f_{\boldsymbol{F}}(p_{c|\boldsymbol{F}}|\boldsymbol{\theta}_{\boldsymbol{F}})$ of eq. 3.1 is modelled with a beta distribution ($p_{c|\boldsymbol{F}} \sim \mathrm{Beta}(\alpha,\beta)$ in Figure 3.4, left) in a hierarchical beta-binomial model [51], to capture performance variability across the different behavioral groups. Accordingly, $\boldsymbol{\theta}_{\boldsymbol{F}} = \{\alpha,\beta\}$ become the parameters of the variability model to be inferred from data collection. The hierarchical structure reflects the mathematical formulation of HEP proposed in subsection 3.2.2. Indeed, the evidence (failures, k_c, and crew observations, N_c, for the c-th group, Figure 3.4, left) enters at group-level in the binomial likelihood function ($k_c \sim \mathrm{B}(N_c, p_{c|\boldsymbol{F}})$ in Figure 3.4, left) to inform the specific $p_{c|\boldsymbol{F}}$, i.e. the group-specific realization of $f_{\boldsymbol{F}}(p_{c|\boldsymbol{F}}|\boldsymbol{\theta}_{\boldsymbol{F}})$ associated to $\boldsymbol{F}$. The $p_{c|\boldsymbol{F}}$’s are then used to infer, at population-level, the unknown $\boldsymbol{\theta}_{\boldsymbol{F}}$ of $f_{\boldsymbol{F}}(p_{c|\boldsymbol{F}}|\boldsymbol{\theta}_{\boldsymbol{F}})$, i.e. the so-called “hyper-parameters” of the Bayesian model.

Beta distributions are commonly adopted in PSA domain for Bayesian hierarchical models where the group-level variable ($p_{c|\boldsymbol{F}}$ in this formulation) represents a probability value, as to constrain the outcomes of the latter between 0 and 1 [35]. Alternative choices for $f_{\boldsymbol{F}}(p_{c|\boldsymbol{F}}|\boldsymbol{\theta}_{\boldsymbol{F}})$ are discussed in subsection 3.4.2.2. Further information on Bayesian hierarchical models can be found in Bayesian literature [51-52].

In the numerical application, the lumped-data model is coupled to a simple Bayesian conjugate beta-binomial model (Figure 3.4, center) to derive the single $p_{\boldsymbol{F}}$ (Figure 3.4, center) from the lumped data (k_{tot} and N_{tot} in Figure 3.4, center). The continuous variability formulation (Figure 3.4, right) is coupled to a Bayesian model with a population variability curve (PVC in Figure 3.4, right) representing the variability in the crew-, task-specific $p_{ij|\boldsymbol{F}}$. To ensure a fair comparison between the models, the variability function of the continuous model, i.e. $f_{\boldsymbol{F}}(p_{ij|\boldsymbol{F}}|\boldsymbol{\theta}_{\boldsymbol{F}})$, is specialized with a beta PVC ($p_{ij|\boldsymbol{F}}\sim$Beta$(\alpha,\beta)$ in Figure 3.4, right), with $\boldsymbol{\theta}_{\boldsymbol{F}}=\{\alpha,\beta\}$ to be inferred from the crew-, task- specific data (k_{ij} and N_{ij} in Figure 3.4, right). Figure 3.4 shows an overview of the three Bayesian models tested in Section 3.4.

The Bayesian models are implemented in “Just Another Gibbs Sampler” (JAGS, [53]), a software using Markov Chain Monte Carlo (MCMC) simulation to approximate the solution of $\pi(\boldsymbol{\theta}_{\boldsymbol{F}}|E)$. The JAGS models are run in R programming environment via the “runjags” library [54].

In both the hierarchical beta-binomial and continuous variability models, the Beta(α,β) functions are reparametrized in terms of mean (μ) and a dispersion measure (i.e. the concentration, κ) as to improve the computational efficiency of MCMC simulations, as recommended in [35]. In the numerical application, non-informative priors are set on the hyper-parameters of the hierarchical model ($\pi_0(\boldsymbol{\theta}_{\boldsymbol{F}})$ in Figure 3.4, left) as common practice in lack of information [35], specifically: a diffuse $\pi_0(\mu)$, defined between 1e-5 and 1; a diffuse $\pi_0(\kappa)$, defined between 0 and 10. Similar priors are set on the parameters of both the conjugate beta-binomial and the continuous with beta PVC models, respectively: a diffuse $\pi_0(p)$ for the single-value HEP; diffuse $\pi_0(\mu)$ and $\pi_0(\kappa)$ for mean and concentration. Tests on the convergence of the MCMC simulations were performed using “diagMCMC”, a set of diagnostic tools provided by [52]. Further information on MCMC methods are given in Bayesian literature [51-52].

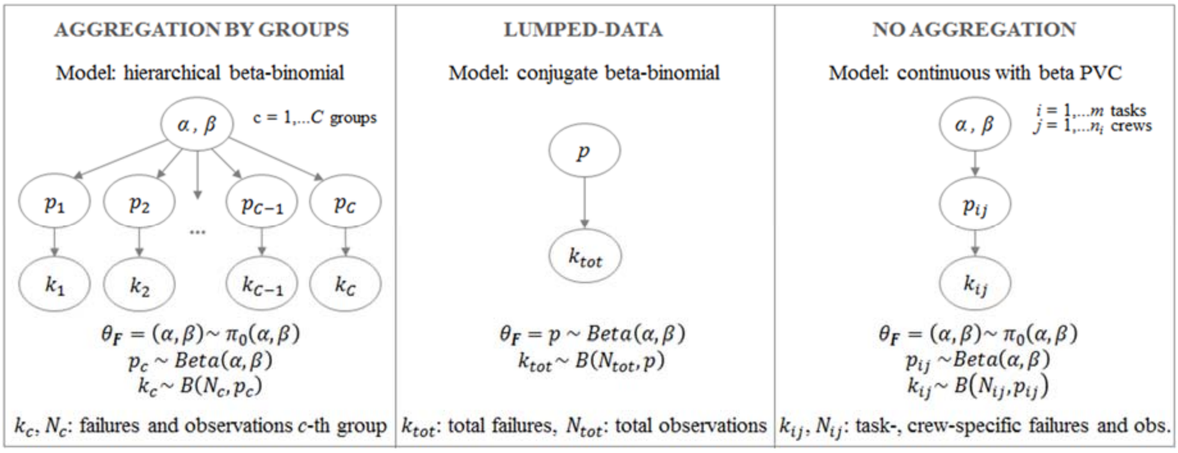

AGGREGATION BY GROUPS
Model: hierarchical beta-binomial
α,β
c = 1,...C groups
p_1
p_2
p_{C-1}
p_C
...
k_1
k_2
k_{C-1}
k_C
$\theta_F=(\alpha,\beta)\sim\pi_0(\alpha,\beta)$
$p_c\sim Beta(\alpha,\beta)$
$k_c\sim B(N_c,p_c)$
k_c, N_c: failures and observations c-th group
LUMPED-DATA
Model: conjugate beta-binomial
p
k_{tot}
$\theta_F=p\sim Beta(\alpha,\beta)$
$k_{tot}\sim B(N_{tot},p)$
k_{tot}: total failures, N_{tot}: total observations
NO AGGREGATION
Model: continuous with beta PVC
α,β
i = 1,...m tasks
j = 1,...n_i crews
p_{ij}
k_{ij}
$\theta_F=(\alpha,\beta)\sim\pi_0(\alpha,\beta)$
$p_{ij}\sim Beta(\alpha,\beta)$
$k_{ij}\sim B(N_{ij},p_{ij})$
k_{ij}, N_{ij}: task-, crew-specific failures and obs.

Figure 3.4. Bayesian models for HEP quantification coupled to the three modelling approaches of Figure 3.2. All the p’s are intended as conditional on the given set $\boldsymbol{F}$, e.g. $p_{c|\boldsymbol{F}}$, $p_{ij|\boldsymbol{F}}$.

3.4 Case study from literature data

This section presents the application of the multi-step methodology to the case study. Subsection 3.4.1 first describes the data source, then presents the set of behavioral categories identified and the HEP quantification considering behavioral groups via the hierarchical beta-binomial model (Figure 3.4, left). The results are compared to the alternative modelling approaches, i.e. the lumped-data and the continuous variability models (Figure 3.4, respectively center and right). Sensitivity analysis on model results is presented in subsection 3.4.2.

3.4.1. Case study

The case study processes data from two simulator experiments [13, 43], involving different emergency scenarios characterized by multiple concurrent malfunctions. As discussed in Section 2, due to procedural guidance-situation mismatches, crew behavioral characteristics played a key role in task performance.

3.4.1.1. Derivation of behavioral categories (methodology: block I)

From the application of step I.1 of the methodology (Figure 3.3), 27 crew observations were identified as belonging to the combination $\boldsymbol{F}$ of task type and PSF ratings reported in Table 3.4 (the SACADA taxonomy is used for illustration purposes). The selection of the PSF ratings was done by the authors of the present chapter, based on the information available in [13, 43]. In the simulated scenarios, the operating crews performed different diagnosis tasks (task type "understanding the situation" in Table 3.4), in all cases with masked indicators (PSF "information availability" with rating "missing/masked" in Table 3.4). All the involved crews experienced for the first time the operational situation replicated by the simulated scenario (PSF "familiarity" with rating "anomaly" in Table 3.4); moreover, the diagnosis had to be performed in absence of alarms directly pointing to the problem (PSF "information specificity" with rating "not specific" in Table 3.4) and with relatively high-tempo (PSF "time criticality" with rating "barely adequate" in Table 3.4).

Table 3.4 summarizes the failure data extrapolated from the simulator records (total observations N_{tot}= 27, with failures k_{tot}= 15). Note the high ratio of failures in the dataset (overall, ~0.56), justified by the complex nature of the tasks and the associated operational contexts under investigation.

Table 3.4. Simulator data used in the case study.

<table>
<tr><td colspan="6">Set $\boldsymbol{F}$ (taxonomy from SACADA, [20]):
{task type = understanding the situation, information quality = missing/masked, information specificity = not specific, familiarity = anomaly, time criticality = barely adequate[2]}</td></tr>
<tr><td>Source</td><td>Scenario</td><td>Realization of contextual factors</td><td>Task</td><td>Observations</td><td>Failures</td></tr>
<tr><td>[43][3]</td><td>SGTR</td><td>Failure of secondary radiation indications</td><td>Identification and isolation of faulted SG (“HFE1B”)</td><td>12</td><td>6</td></tr>
<tr><td>[13][4]</td><td>Multi SGTR</td><td>Radiation alarms already activated by early releases due to initiating event</td><td>Identification and isolation of faulted SG</td><td>5</td><td>4</td></tr>
<tr><td></td><td>ISLOCA</td><td>No indications on leaks’ specific location</td><td>Identification and isolation of leaks</td><td>5</td><td>2</td></tr>
<tr><td></td><td>LOFW+SGTR</td><td>Water level increase and absence of radiation indication mask faulted SG identification</td><td>Identification and isolation of faulted SG</td><td>5</td><td>3</td></tr>
<tr><td></td><td></td><td></td><td>Aggregated data (k_{tot},N_{tot})</td><td>27</td><td>15</td></tr>
</table>

In the application of steps I.2-I.3 of the methodology (Figure 3.3), for each of the simulator records in Table 3.4, the observed crew behaviors were analyzed using the team and person-specific metrics from [44] and classified in behavioral categories accordingly. Examples of the classification process are provided in Table 3.2. Table 3.5 (left) shows the list of the twenty behavioral categories preliminarily identified for the case study and organized by ten dimensions, together with a short description for each category. For instance, in Table 3.5 (left), crew behaviors relevant to the dimension “adherence to / interpretation of procedures” were classified in two categories, “beyond / proactive” and “close / reactive”, based on metrics from [44] concerning team coordination (e.g. “proactivity: think ahead possibilities for optimizing activities”), decision making (“thinking outside the box: regularly considering the situation at hand from different perspective”), and attitude (“uphold a questioning attitude and willingness to consider a situation from multiple perspectives”. The category “beyond / proactive” refers to crews that considered alternative causes and upheld a questioning attitude during the diagnosis, trying extra-procedural tasks not contained in procedures; on the other hand, “close / reactive” describes crews that waited for the procedures to provide explicit indications on how to perform

---

[2] In “LOFW+SGTR” scenario, the execution of the considered task (“identification and isolation of the faulted steam generator”) can overlap in time with the other main safety-critical operator actions (e.g. restore feed-water to the steam generators, control cooling system cool-down and pressurization to prevent “pressurized thermal shock” condition): therefore, the effective time available for the diagnosis can differ according to the scenario evolution experienced by each crew. For the purposes of the application, the authors assumed a “barely adequate” time for all the five crew observations.

[3] Performance outcome (failure or success) was considered according to the time criterion (25 minutes) set by the trainers for the task.

[4] Given that task-specific time criteria are not adopted, the outcome of each task was considered as a failure when the performance standards established by trainers were not met at the end of the scenario, e.g.: for the task in the ISLOCA scenario, failure when crews did not try to identify and isolate the leaks, success in the opposite case.

the diagnosis. The full set of metrics associated to each behavioral category is provided in Table D.1 (Appendix D).

As mentioned in subsection 3.3.1, the aggregation in behavioral groups (steps II.1-II.2 in Figure 3.3) can be problematic in presence of a large set of categories but only few data points at disposal. For the purposes of the present application, in order to avoid too much data dispersion over the categories due to the small number of observations available (N_{tot}= 27 in Table 3.4), the category list was further compacted by expert-based aggregation, and the respective metrics combined into ten categories as shown in Table 3.5 (right), with dimensions: "progress through procedures", "flexibility in dealing with procedures and cues", "role awareness", "prioritization of goals and resources", and "decision making and information sharing".

3.4.1.2. Grouping in behavioral groups and HEP quantification (methodology: block II)

The crew behaviors collected for each of the 27 observations were first matched to the categories of the compact list in Table 3.5, right (application of step II.1). The matching was based on the crew performance analysis available in the information sources [13, 43]. The metrics associated to each category (see Table D.1 in Appendix D) were used as basis for the category association. The behavioral patterns were identified by the combinations of categories (step II.2), similarly to the examples provided in Table 3.3. Table 3.6 shows the seven groups identified for the case study, together with the respective behavioral patterns and the group-specific failure data (number of failures k_c in N_c observations). The aggregated dataset in Table 3.6 highlights in qualitative terms what discussed in Section 3.2: different behavioral patterns can have different impacts on the task outcome (see the frequentist ratios k_c/N_c for each group) and determine crew performance variability within the set $\boldsymbol{F}$. For instance, the pattern associated to "group 2" in Table 3.6 overall exerts a positive impact on task outcome (zero failures out of six observations) compared to "group 5" (eight failures out of nine observations).

The group-specific failure data (k_c and N_c in Table 3.6) was used as input in the hierarchical beta-binomial model to infer the error probabilities for the seven behavioral groups and quantify the HEP uncertainty distribution, $P(HEP)$, for the set $\boldsymbol{F}$ (step II.3). Figure 3.5 shows the results, along with the comparison with the alternative models conjugate beta-binomial model with lumped-data (k_{tot}= 15 and N_{tot}= 27 in Table 3.6) and the beta-PVC variability model with crew-, task-specific data (k_{ij} and N_{ij}). For the latter, considering that each crew (index j) performed only one repetition of the same task (index i) in data collection, each N_{ij} was set to one, with the k_{ij} equal to one in case of failure (zero otherwise). Numerical results are given in Table 3.7.

Table 3.5. Left: preliminary list of behavioral categories emerging from the empirical data for the case study [13, 43]. The associated team-, person-based metrics [44] used for the preliminary categorization are provided in Table D.1 (Appendix D). Right: compact set after expert-based aggregation, for use in the numerical application (subsection 3.4.1.2).

Preliminary set of behavioral categories identified from empirical data (categorization supported by metrics in [44])			Aggregated set for the numerical application		
Dimensions	Behavioral categories		Dimensions	Behavioral categories	
Progress through procedures	"Sequential": systematic procedure reading (inc. foldout pages and warnings in appendix), transferring only when conditions are met.	"Adaptive": move forward and loop back through procedures, sometimes anticipating transferring conditions.	Progress through procedures	Thorough	Jumping
Adherence to / interpretation of procedures	"Beyond / proactive": address alternative causes with questioning attitude and willing to perform extra procedural tasks.	"Close / reactive": wait for explicit indications from procedures, performing tasks only if prescribed.	Flexibility in dealing with procedures and cues	Beyond	Close
Diversity of information sources	"Diverse cues": rely on diverse, redundant information, including outside-control room indications (local information).	"Prescribed cues": rely mostly on cues indicated in procedures.			
Monitoring indications when reacting to anomalies	"Follow-up trends": anomalies are immediately addressed and followed up over time.	"Focus only on initial deviations": indications are mostly monitored at the early stage of the anomaly.			
Role awareness	"Adhering": operators adhere to prescribed roles, with the supervisor maintaining a global overview.	"Diverging": some members perform tasks outside their responsibilities, with the supervisor more involved in details	Role awareness	Adhering	Diverging
Progression in decision making	"Prioritizing, fast decision maker": schedule tasks and goals to favor quick response.	"Hesitating, slowly building up": proceed step-by-step, upholding an explanation-building orientation.	Prioritization of goals and resources	Fast adaptation	Slow adaptation
Operator involvement	"All are involved": everyone is active during task execution.	"Some involved, some passive": some members are more active, some other more passive.			
Resource optimization during scenario	"Flexible redistribution": tend to optimize resources and flexibly adapt work redistribution according to task progression.	"Rigid": focus more on getting on with the work, keeping constant workload distribution during scenario (e.g. no parallel tasks).			
Team orientation in decision making	"Collective": supervisor develops strategies consulting the operators, taking into account their opinions and suggestions.	"Non-inclusive": supervisor takes most decisions alone, without much discussion with the rest of the team.	Decision making and information sharing	Collective	Non-inclusive
Adherence to communication and meeting protocol	"Adhering": meetings and briefings are held when necessary and structured according to protocols, with follow-up when needed.	"Diverging": meetings and briefings held with low frequency, when held: operators do not stick to form (e.g. not definitive endings).			

Table 3.6. Seven crew groups and associated behavioral patterns identified in the case study (note that each group corresponds to a specific behavioral pattern).

	Progress through procedures		Flexibility in dealing with procedures/cues		Role awareness		Prioritization of goals and resources		Decision making and information sharing		
Categories / Groups	Sequential	Adaptive	Beyond	Close	Adhere	Diverge	Fast adaptation	Slow adaptation	Collective	Non-inclusive	k_c/N_c
Group 1	X		X		X		X			X	0 / 1
Group 2[5]	X		X		X		X		X		0 / 6
Group 3	X		X		X			X	X		1 / 2
Group 4	X			X		X		X		X	2 / 4
Group 5		X		X		X		X		X	8 / 9
Group 6		X		X	X			X		X	3 / 3
Group 7		X		X	X			X	X		1 / 2

---

[5] For crew "N " in the SGTR scenario from [43], the available information was not sufficient to fully characterize crew performance in three out of five categories (i.e. "progress through procedures", "role awareness", and "prioritization of goals and resources"): in this case, for practical reasons, the categories in line with the crew behaviors "recommended" by the training standards (see Section 2.4 in [43]) were assigned (respectively: "sequential", "adhere", and "fast adaptation").

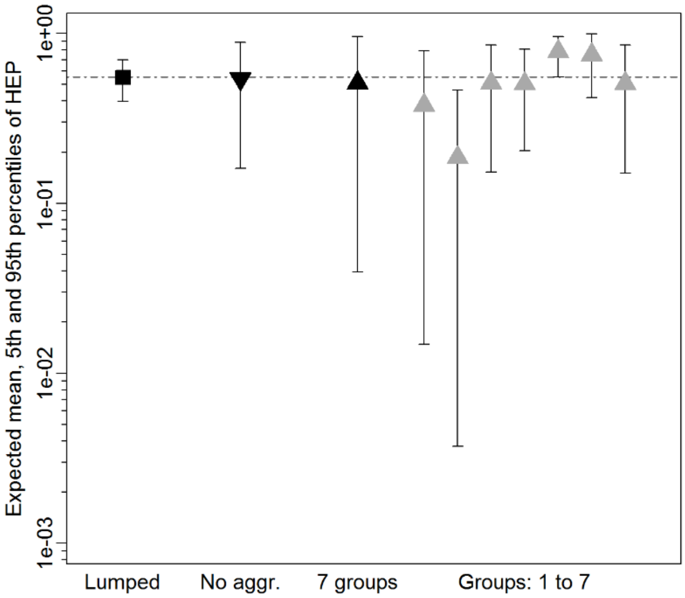


Figure 3.5. Results from the numerical application to the case study. In x axis, from left to right: conjugate beta-binomial with lumped data ("Lumped"); continuous variability model with crew-, task-specific data ("No aggregation"); hierarchical beta-binomial model with seven behavioral groups ("7 groups"); and group-specific $P(HEP)$'s ("Groups: 1 to 7"). In y axis (log scale): mean (symbols), 5th and 95th percentiles (whiskers) of the $P(HEP)$. Dotted line: overall frequentist failure ratio (k_{tot}/N_{tot}).

Table 3.7. Numerical results from Figure 3.5 (note that each group corresponds to a specific behavioral pattern).

Model (data aggregation in Figure 3.5)	Mean	5th	50th	95th	EF
Conjugate beta-binomial ("lumped")	5.5e-01	4.0e-01	5.5e-01	7.0e-01	1.3
Continuous with beta PVC ("no aggregation")	5.4e-01	1.6e-01	5.5e-01	8.9e-01	2.4
Hierarchical beta-binomial ("7 groups")	5.1e-01	3.9e-02	5.2e-01	9.6e-01	4.9
Group 1	3.8e-01	1.5e-02	3.7e-01	7.9e-01	7.3
Group 2	1.9e-01	3.7e-03	1.6e-01	4.6e-01	11.2
Group 3	5.1e-01	1.5e-01	5.1e-01	8.6e-01	2.4
Group 4	5.1e-01	2.0e-01	5.1e-01	8.1e-01	2.0
Group 5	7.8e-01	5.6e-01	8.0e-01	9.6e-01	1.3
Group 6	7.5e-01	4.2e-01	7.8e-01	9.9e-01	1.5
Group 7	5.1e-01	1.5e-01	5.1e-01	8.6e-01	2.4

The three models return similar values of the expected HEP ("lumped": 5.5e-01, "no aggregation": 5.4e-01, "7 groups": 5.1e-01) in line with the overall frequentist ratio (5.6e-01), but with very different expected variability (see Figure 3.5 and error factor, EF, values in Table 3.7). The differences in the variability distributions provided by the models can be interpreted according to the HEP formulations. In the lumped-data approach, variability in crew performances is averaged in the single piece of evidence (k_{tot}/N_{tot}): the $P(HEP)$ tends to shrink around the population average (with EF = 1.3). The continuous model with beta PVC "breaks down" HEP variability at crew-, task-level. Since the crews performed only one task repetition, the disaggregated data ($k_{ij}/1$) informs the $p_{ij|\boldsymbol{F}}$'s of the beta variability distribution only with 0's and 1's: consequently, variability across differently performing crews does not clearly emerge in the uncertainty distribution, resulting in a lower spread around the mean value (EF = 2.4) compared to the hierarchical beta binomial model. On the other hand, the latter "clusters" performance data in seven behavioral groups: the group-specific data (k_c/N_c) informs the seven $p_{c|\boldsymbol{F}}$'s of the beta variability distribution with less uncertainty compared to the continuous formulation (for example, 8/9 and 2/4 are more informative evidence compared to 0/1 and 1/1). In this case, the group-specific error probabilities capture variability in crew performances (see the $p_{c|\boldsymbol{F}}$ expected values in Table 3.7) and this reflects in a larger spread around the mean value (EF = 4.9) of the HEP uncertainty distribution.

3.4.2. Sensitivity analysis

This subsection discusses the influence of the number of and the degree of performance variability across the identified behavioral groups (subsection 3.4.2.1), and the choice of the variability function (subsection 3.4.2.2) on the estimated HEP uncertainty distribution. The artificial datasets used in the tests are adapted from the case study.

3.4.2.1. Number of and degree of performance variability across behavioral groups

As discussed in subsection 3.3.1, the number of identified crew groups is directly influenced by the amount of behavioral categories used to classify crew behaviors: this number depends on how many team- and person-based dimensions in Skjerve and Holmgren taxonomy [44] are considered by the analyst. As a general rule, the more behavioral categories are modelled, the higher the number of groups emerging from data. To investigate the extent to which this number can influence model results, the categorization in Table 3.6 is reinterpreted by not explicitly modelling behaviors related to "decision making and information sharing" and "role awareness": this specific case would be equivalent to considering crew as a "single entity", averaging the effects of interpersonal aspects (e.g. team coordination, communication strategies) over the remaining categories (i.e. in "progress through procedures", "flexibility in dealing with procedures/cues", "prioritization of goals and resources"). This corresponds to a higher level of data aggregation with only four behavioral groups, as shown in Table 3.8.

An additional aspect to consider is that the case study focused on a set $\boldsymbol{F}$ characterized by large performance variability. The influence of data aggregation (see Table 3.6 vs Table 3.8) on model results would need to be reconsidered in case of lower variability in performance data, e.g. as observed for those $\boldsymbol{F}$'s representing tasks/operational contexts for which the effect of crew behaviors plays a minor role in determining task failure (e.g. tasks in the base SGTR scenario in [11]). In order to include this aspect in the analysis, the group-specific failure data in Table 3.6 (seven groups) and Table 3.8 (four groups) was arbitrarily redistributed as to simulate conditions of lower performance variability across the behavioral groups, i.e. "equalizing" the frequentist ratios k_c/N_c towards the population average k_{tot}/N_{tot}. The resulting datasets are summarized in Table E.1 (Appendix E), together with the numerical results of the sensitivity analysis.

Figure 3.6 shows the $P(HEP)$'s provided by the hierarchical beta-binomial model informed with real data from Tables 10 and 12 ("large variability", respectively "7 groups" and "4 groups") and the artificial data from Table E.1 ("low variability", "7 groups" and "4 groups"). For "large variability" datasets, the aggregation from seven to four groups corresponds to a more heterogeneous failure data across the behavioral groups (e.g. in Table E.1: k_1/N_1= 0/7 and k_4/N_4= 12/14 for "4 groups" vs k_1/N_1= 0/6 and k_4/N_4= 8/9 for "7 groups"): consequently, the hierarchical beta-binomial model captures a larger variability in $p_{c|\boldsymbol{F}}$ values (see $E[p_{c|\boldsymbol{F}}]$'s in Table E.1) and returns a $P(HEP)$ with an increased spread around the population average (EF = 8.3 for "4 groups" vs EF = 4.9 for "7 groups"). For "low variability" cases, being failure data more homogeneously distributed across the groups, the $E[p_{c|\boldsymbol{F}}]$'s in Table E.1 get closer to the population average and the number of identified behavioral groups plays a minor influence on the estimated $P(HEP)$: mean = 5.3e-01 and EF = 2.2 for "7 group" case, mean = 5.4e-01 and EF = 2.4 for "4 group" case. Note that the results for "low variability" datasets are identical to the continuous variability formulation (mean = 5.4e-01 and EF = 2.4 in Table 3.7). The practical implications on HRA applications are discussed in the next section.

To summarize the results from the sensitivity analysis, the investigation showed that the proposed model is able to capture differences in performance variability compared to the alternative approaches. Also, the more heterogeneous is the group-specific failure data (see Table 3.8), the more the results diverge from the lumped and continuous variability formulations. On the other hand, the benefits of using a variability model based on behavioral patterns compared to simpler approaches (e.g. the continuous variability formulation) diminish with reduced performance variability underlying the dataset.

3.4.2.2. Choice of the variability function

Alternative variability functions (i.e. lognormal, logistic-normal) were tested for both the hierarchical and the continuous variability models: the numerical results are included in Table E.2 (Appendix E). In general, the considerations drawn from the sensitivity analysis still apply (see in Table E.2 the evolution of $P(HEP)$ statistics with varying number of groups and degree of performance variability). Concerning the case study, the hierarchical model set with lognormal and logistic-normal variability functions returns uncertainty distributions with higher

EFs (EF = 10.5 and EF = 16.2, respectively) compared to beta case (EF = 4.9). The reason is because the lognormal and logistic-normal PVCs converge more slowly with smaller datasets compared to the beta PVC (i.e. with few observations, the beta distribution peaks faster and returns less uncertain $p_{c|\boldsymbol{F}}$ estimates).

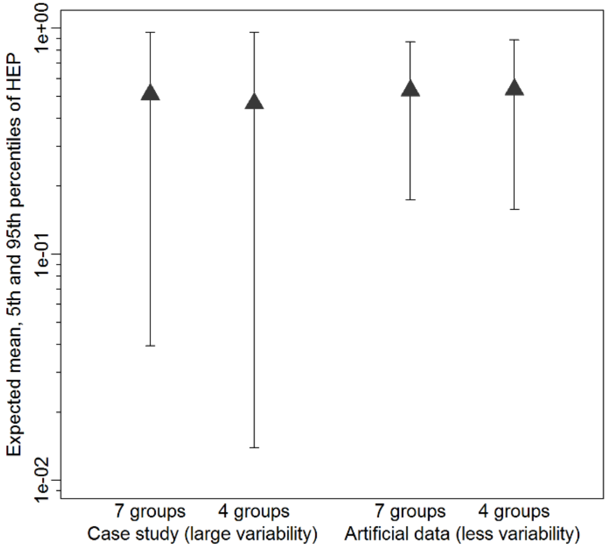
Note that the choice of an appropriate variability function should also take into consideration the expected HEP order of magnitude of the investigated $\boldsymbol{F}$. For instance, when treating higher HEP values as in the case study of this chapter (between 1e-1 and 1), the lognormal variability function tends to systematically underestimate the mean HEP (e.g. 3.4e-01 for the case study) compared to beta and logistic-normal functions (respectively, 5.1e-01 and 4.9e-01), as confirmed by Kelly and Smith [35]. In addition, the authors tested the sensitivity to different (reasonable) hyper-priors (e.g. constrained non-informative, Jeffreys, etc.) for the mean of the beta PVC, i.e. $\pi_0(\mu)$. The results did not highlight any significant dependence from the adopted $\pi_0(\mu)$, given the strong informative power of the particular dataset (15 failures over 27 observations, with very high frequentist ratio, i.e. 0.56). Indeed more in-depth analysis of possible choices for the prior function and the associated parameters would be required for less informative data sets.

Different techniques for model comparison (e.g. posterior predictive checks) are available in Bayesian literature to assist the analyst in selecting an appropriate variability function (for further details, the reader should refer to [51]).

Table 3.8. Higher level of data aggregation: an example with four behavioral groups (note that each group corresponds to a specific behavioral pattern).

	Progress through procedures		Flexibility in dealing with procedures/cues		Role awareness		Prioritization of goals and resources		Decision making and information sharing		
Categories / Groups	Sequential	Adaptive	Beyond	Close	Adhere	Diverge	Fast adaptation	Slow adaptation	Collective	Non-inclusive	k_c/N_c
Group 1	X		X				X				0 / 7
Group 2	X		X					X			2 / 4
Group 3	X			X				X			1 / 2
Group 4		X		X				X			12 / 14

Figure 3.6. Influence of the number of identified behavioral groups ("7 groups" vs "4 groups") on the $P(HEP)$ estimated by the hierarchical beta-binomial model for both the case study (Tables 10 and 12) and the artificially-generated dataset with less performance variability (Table E.1).



3.5 Discussion

Concerning data requirements, the methodology presented in this paper requires the availability of crew behavior records to classify the crews in groups. As anticipated in Section 3.2, this information goes beyond what would be collected if strictly adopting currently available protocols for large-scale data collection programs (SACADA [20], HuREX [21]). To some extent, the SACADA taxonomy could be extended relatively easily, given that some piece of information on crew behaviors is already collected in the SACADA framework, although only for crews for which performance issues are observed. Indeed, for the methodology to be applicable, information on crew behaviors should be available for all sessions, independently on the crew performance outcome. It has to be mentioned that SACADA has been developed to collect data within the operator training sessions: therefore any extension of the amount of data collected would have to be evaluated in terms of overload on trainers and operators.

On the other hand, records of crew behaviors are available from other human factor studies, not necessarily intended for HRA applications. Indeed, this has been the case for the application presented in this chapter. Therefore, the collection of crew behaviors does not necessarily have to be integrated in HRA data collection protocols such as SACADA and HuREX. An alternative could be to decouple data collection on crew variability from those on the mean HEP values. Specific data collection studies could be directed only to subsets of tasks type and PSF combinations to identify dominant crew behavioral groups and their associated variability, while maintaining the available taxonomies for estimation of population-averaged HEPs. Indeed, although the aim of the methodology presented in this paper is to estimate HEP distributions conditional on the set $\boldsymbol{F}$ of task types and PSF levels, the methodology is not intended for application to all possible combinations. Besides being unrealistic for the amount of data required, this would also be unnecessary for those sets $\boldsymbol{F}$ expected to trigger a similar spectrum of crew behaviors. These studies may give empirical indications of the actual HEP spread, which could be then assigned to the estimates from the population-averaging data collection protocols.

The crew behavioral categories identified in this work emerged from very challenging scenarios, characterized by high failure probabilities. The scenarios were characterized by masked indications and symptoms-procedural mismatches, with stringent requirements on which behaviors would lead to successful performance: ability to adapt, fast decisions, questioning attitude were all crew characteristics necessary to success. The result is a large variability in crew performance: those crews manifesting these characteristics were much more likely to succeed compared to other crews (compare performance results of group 2 and group 5 in Table 3.6). This also converts in the large variability for the resulting HEP distribution, EF of about 5 in Table 3.7). The characteristics of the scenarios analyzed in the present chapter were imposed by the available data; for future analysis, with larger amount of data available, it would be beneficial to address diverse scenarios, as well as less challenging situations to investigate more comprehensively the effect of crew behavioral variability on the HEP variability.

The proposed methodology acknowledges that crew behaviors are neither merely "situation-driven" (i.e. "task, scenario, context" factors predominantly determine mechanisms and pace of performance, independently on crew characteristics) nor "crew-driven" (i.e. each crew has an "inherent" problem solving style and communication strategy, independently of task/context). For instance, in the empirical observations analyzed in the case study [13, 43], on the one hand, in the same situation (scenario) significant differences in crew behaviors were observed. On the other hand, the same crew did not always adopted the same problem solving style (e.g. fixation-prone or prioritization-oriented) or communication strategy (e.g. frequent meetings/briefings or few strategic discussions) across different simulated scenarios [13]. Indeed as shown in Figure 3.1, all factors (situation- as well as crew-driven) interplay in the determination of the crew behaviors. The proposed methodology acknowledges this and generalizes both the situation- and the crew-driven interpretations: indeed, the analysis of the behavioral characteristics is made conditional on the "situation-driven" set $\boldsymbol{F}$, but the actual set of characteristics is made emerging from the actual observations, which are a result of the interactions of all factors. Besides the specific analyses of the present chapter, the proposed framework offers a tool for future works to study the interplay across these influences.

For the purposes of the present work, the set of behavioral categories has been defined based on the analysis of the crew performances and the expertise of the authors, aiming at a tradeoff between coverage of characterization and the number of categories. Since the behavioral groups are identified based on the category combination, the number of categories has to be maintained low enough to avoid combinatorial explosion. Note that, while the set of categories adopted in this work is indeed subjective, the authors linked the definition of each category to an established set of teamwork competences (see Table 3.2 and Table D.1 in Appendix D), which in turn can be associated to observable crew behaviors [44]. When processing a simulator record, the categorization is based on the observed crew behaviors (step II.1) and not directly on the categories: this has been done to make the behavior categorization more objective and traceable. Additionally, this opens to the test of different category sets: as long as the crew behaviors are recorded and a link to these behaviors is established as shown in Table 3.2 and Table D.1. In the long term, as more data on crew behaviors may be available, consolidated sets of behavioral categories may be identified and reused across studies to investigate their relative importance and impact on crew performance. As mentioned, this "library of categories" would identify the categories relevant for groups of $\boldsymbol{F}$, ideally defined to group situations by type, e.g. "fast-pacing", "standard procedure-following", "conflicting goals" in a similar way as proposed in [28]. Also, with more data available, data analysis and statistical tests could be used to derive the groups (e.g. via cluster analysis), identify dominant categories, and rule out or aggregate categories with limited impact on task performance and support accordingly the library of categories, reducing the subjective component in category definitions. Besides more established sets of categories and groups, the accumulated data can be used to provide information on the frequency of each group, per given set $\boldsymbol{F}$. This information (possibly complemented with expert judgment on the plant crew specificity) can be used to inform HRA prospective analyses for which many crew observations are not possible.

The methodology presented in this chapter could be used to support the development of future, advanced crew performance models, representing the complex relationships among the performance influencing factors (task-, context-, team-, and person-based, see Figure 1) and the HEP. In this direction, modern approaches based on Bayesian Belief Networks (BBN) [55-56] resort to a flexible framework to represent crew performance variability, either implicitly (into the BBN conditional probability distribution), as well as explicitly (as dedicated input nodes). Concerning the former (implicit incorporation), the variability model presented in this work can enhance the empirical basis of the BBN distributions, e.g. producing anchoring distributions to populate the BBN relationships via filling algorithms (such as those in [57]). Concerning the latter (explicit representation), the proposed methodology could inform crew-to-crew variability nodes with behavioral patterns that are relevant for a given status of the task and PSF nodes (i.e. for a given $\boldsymbol{F}$).

3.6 Conclusions

As acknowledged by recent simulator studies, crew performance variability plays an important role in nuclear power plant operational tasks and requires explicit consideration in the estimation of the HEP (and the associated uncertainty). Characterizing the performance drivers for different task types and operational contexts is not straightforward, given the complexity of both human behaviors and emergency scenarios typically addressed in PSA applications.

As a first-of-a-kind attempt in this direction, the present work shows how to formally incorporate crew behavioral characteristics observed in simulator experiments in a variability model for HEP quantification. Crew behaviors are here categorized by behavioral patterns, modelling the dynamic influence of crew-specific (e.g. communication strategies, attitude, decision making and leadership styles) and task-, scenario-specific factors (e.g. task complexity, procedural guidance, information availability) on crew performance. This approach allows aggregating crews sharing a similar behavioral profile in a unique behavioral group, and associate each group to a specific error probability value in the HEP variability model.

This chapter presented a multi-step methodology that can be generally applied to multiple sets of HRA method categorical elements (task type, PSF ratings) to systematically process the information on crew behaviors in simulator data collection, identify behavioral groups and finally use group-specific failure data to inform a Bayesian hierarchical model for HEP estimation. A case study demonstrates the feasibility of the proposed methodology to a practical HRA application, focusing on data from complex emergency scenarios where diagnosis tasks are challenged by masked indicators. The numerical application showed that, compared to existing approaches in treating simulator data, the Bayesian hierarchical model with behavioral groups is able to capture variability across different-performing crews, representing a versatile solution for estimating HEP uncertainty and variability distributions to feed HRA methods with empirically-based reference data.

Besides enabling data aggregation from different crews on the basis of their behavioral commonalities, this new formulation allows identifying the crew characteristics that determine performance variability in the failure probability. From this perspective, the proposed methodology can be also used to highlight those crew behavioral patterns that favor lower failure probability values for a given task and operational context, therefore supporting training of operators accordingly.

Acknowledgments

The work was funded by the Swiss Federal Nuclear Safety Inspectorate (ENSI), under contract Nr. 101163. The views expressed in this work are solely those of the authors.

References

1. Kirwan B. *A guide to practical Human Reliability Assessment*. CRC press: Boca Raton, FL, USA, 1994.
2. Podofillini L. Human Reliability Analysis. In: Moller N, Hansson SO, Holmberg JE, and Rollenhagen C. (eds) *Handbook of Safety Principles*. Wiley, 2017, pp.565-592.
3. Spurgin AJ. *Human Reliability Assessment – theory and practice*. CRC press: Boca Raton, FL, USA, 2010.
4. Swain AD and Guttman HE. *Handbook of human reliability analysis with emphasis on nuclear power plant applications*. NUREG/CR-1278, U.S. Nuclear Regulatory Commission, Washington DC, USA, 1983.
5. Williams JC. HEART – A Proposed Method for Assessing and Reducing Human Error. In: *9th Advance in Reliability Technology Symposium*, University of Bradford, 1986.
6. Williams JC. A data-based method for assessing and reducing human error to improve operational performance. In: *Proceedings of the IEEE Fourth Conference on Human Factors and Power Plants*, Monterey, California, 5–9 June, pp. 436–450, 1988.
7. Hollnagel E. *Cognitive Reliability and Error Analysis Method (CREAM)*. Oxford: Elsevier Science Ltd, 1998.
8. Gertman DI, Blackman HS, Marble JL, et al. *The SPAR-H Human Reliability Analysis Method*. NUREG/CR-6883, U.S. Nuclear Regulatory Commission, Washington DC, USA, 2005.
9. Whaley AM, Kelly DL, Boring RL, et al. SPAR-H step-by-step guidance. INL/EXT-10-18533, Idaho National Labs, Idaho Falls, Idaho 83415, 2011.
10. Mosleh A and Smith C. *The Feasibility Of Employing Bayesian Techniques And Other Mathematical Formalisms In Human Reliability Analysis, in The Employment of Empirical Data and Bayesian Methods in Human Reliability Analysis: A Feasibility Study*, NUREG/CR-6949, pp. 5-15, INL/EXT-06-11670, Washington, D.C.: U.S. Nuclear Regulatory Commission, 2007.
11. Forester J, Dang VN, Bye A, et al. *The International HRA Empirical Study Lessons Learned from Comparing HRA Methods Predictions to HAMMLAB Simulator Data*. NUREG-2127, US Nuclear Regulatory Commission, Washington DC, USA, 2014.
12. Forester J, Liao H, Dang VN, et al. *The US HRA Empirical Study - Assessment of HRA Method Predictions against Operating Crew Performance on a US Nuclear Power Plant Simulator*. NUREG-2156, US Nuclear Regulatory Commission, Washington DC, USA, 2016.
13. Massaiu S and Holmgren L. *Diagnosis and Decision-Making with Emergency Operating Procedures in Non-Typical Conditions: A HAMMLAB Study with U.S. Operators*. HWR-1121. Halden, Norway: OECD Halden Reactor Project, 2014.
14. Massaiu S and Holmgren L. *The 2013 Resilient Procedure Use Study with Swedish Operators: Final Results*. HWR-1216. Halden, Norway: OECD Halden Reactor Project, 2017.
15. Xing J, Parry G, Presley M, et al. *An Integrated Human Event Analysis System (IDHEAS) for Nuclear Power Plant Internal Events At-Power Application*. NUREG-2199 Vol.1, U.S. Nuclear Regulatory Commission, Washington DC and Electric Power Research Institute,

Palo Alto CA, USA, 2017

16. Ekanem NJ, Mosleh A, and Shen SH. Phoenix – A model-based Human reliability analysis methodology: Qualitative analysis procedure. *Reliab Eng Syst Saf* 2015, 145: 301-315.
17. Mosleh A and Chang YH. Model-based human reliability analysis: prospects and requirements. *Reliab Eng Syst Saf* 2004, 83: 241–253.
18. Hallbert B and Kolaczkowski A. *The Employment of Empirical Data and Bayesian Methods in Human Reliability Analysis: A Feasibility Study*. NUREG/CR-6949, pp. 1-4, INL/EXT-06-11670, Washington, D.C.: U.S. Nuclear Regulatory Commission, 2007.
19. Liao H, Forester J, Dang VN, et al. Assessment of HRA method predictions against operating crew performance: Part III: Conclusions and achievements. *Reliab Eng Syst Saf* 2019, 191: 106511.
20. Chang JY, Bley D, Criscione L, et al. The SACADA database for human reliability and human performance. *Reliab Eng Syst Saf* 2014, 125: 117-133.
21. Park J, Jung W, Kim S, et al. *A guideline to collect HRA data in the simulator of nuclear power plants*. KAERI/TR-5206, Korea Atomic Energy Research Institute, Republic of Korea, 2013.
22. Groth KM, Smith CL, and Swiler LP. A Bayesian method for using simulator data to enhance human error probabilities assigned by existing HRA methods. *Reliab Eng Syst Saf* 2014, 128 (Supplement C): 32-40.
23. Jung W, Park J, Kim Y, et al. HuREX – A framework of HRA data collection from simulators in nuclear power plants. *Reliab Eng Syst Saf* 2020, 194: 106235.
24. Azarm MA, Kim IS, Marks C, et al. Analyses methods and pilot applications of SACADA database. In: *14th Probabilistic Safety Assessment and Management*, PSAM 14 2018: UCLA Meyer & Renee Luskin Conference Center, Los Angeles, California.
25. Greco SF, Podofillini L, and Dang VN. A Bayesian model to treat within-category and crew-to-crew variability in simulator data for Human Reliability Analysis. *Reliab Eng Syst Saf* 2021; 206: 107309.
26. Greco SF, Podofillini L, and Dang VN. Crew performance variability in simulator data for Human Reliability Analysis: investigation of modelling options. In: *Proceedings of the 29th European Safety and Reliability Conference*, ESREL 2019. ISBN: 981-973-0000-00-0.
27. Woods DD and Roth EM. *Cognitive environment simulation: an artificial intelligence system for human performance assessment*. NUREG/CR-4862-Vol. 3, Westinghouse Research and Development Center, Pittsburgh, PA, USA. Technical report, May 1986-June 1987.
28. Mosneron-Dupin F, Reer B, Heslinga G, et al. Human-centered modeling in human reliability analysis: some trends based on case studies. *Reliab Eng Syst Saf* 1997, 58(3): 249-274.
29. Apostolakis G, Kaplan S, Garrick BJ, et al. Data specialization for plant specific risk studies. *Nucl Eng Des* 1980, 56(2): 321-329.
30. Kaplan S. On a two-stage Bayesian procedure for determining failure rates. *IEEE Trans Power Apparatus Syst* 1983, 102(1):195–262.
31. Mosleh A. Bayesian modeling of expert-to-expert variability and dependence in estimating rare event frequencies. *Reliab Eng Syst Saf* 1992, 38(1): 47-57.
32. Siu NO and Kelly DL. Bayesian parameter estimation in probabilistic risk assessment.

Reliab Eng Syst Saf 1998, 62(1): 89-116.

33. Droguett EL, Groen F, and Mosleh A. Bayesian assessment of the variability of reliability measures. *Pesq Oper* 2006, 26: 109-127.
34. Yue M and Chu TL. Estimation of failure rates of digital components using a hierarchical Bayesian method. In: *8th International conference on probabilistic safety assessment and management (PSAM8)*, New Orleans, LA, 14–19 May, 2006, ISBN-10: 0791802442.
35. Kelly DL and Smith CL. *Bayesian Inference for Probabilistic Risk Assessment: A Practitioner's Guidebook*. London, UK: Springer-Verlag, 2011.
36. Podofillini L and Dang VN. A Bayesian Approach to Treat Expert-Elicited Probabilities in Human Reliability Analysis Model Construction. *Reliab Eng Syst Saf* 2013, 117: 52-64.
37. Lee MD. How cognitive modeling can benefit from hierarchical Bayesian models. *J Math Psychol* 2011, 55(1): 1-7.
38. Bartlema A, Lee M, Wetzels R, et al. A Bayesian hierarchical mixture approach to individual differences: Case studies in selective attention and representation in category learning. *J Math Psychol* 2014, 59: 132-150.
39. Krauss M, Tappe K, Schuppert A, et al. Bayesian Population Physiologically-Based Pharmacokinetic (PBPK) Approach for a Physiologically Realistic Characterization of Interindividual Variability in Clinically Relevant Populations. *PLoS ONE* 2015, 10(10): e0139423.
40. Moura MC, Azevedo RV, Droguett EL, et al. Estimation of expected number of accidents and workforce unavailability through Bayesian population variability analysis and Markov-based model. *Reliab Eng Syst Saf* 2016, 150: 136-146.
41. Chiu W, Wright F, and Rusyn I. A tiered, Bayesian approach to estimating population variability for regulatory decision-making. *ALTEX – Altern Anim Ex* 2017, 34(3), pp. 377-388.
42. Shao K, Allen BC, and Wheeler MW. Bayesian Hierarchical Structure for Quantifying Population Variability to Inform Probabilistic Health Risk Assessments. *Risk Anal* 2017, 37(10): 1865-1878.
43. Lois E, Dang V, Forester J, et al. *International HRA Empirical Study - Phase 1 Report: Description of Overall Approach and Pilot Phase Results from Comparing HRA Methods to Simulator Performance Data*. NUREG/IA-0216 Vol. 1, US Nuclear Regulatory Commission, Washington DC, USA, 2009.
44. Skjerve AB, and Holmgren L. *An investigation of Teamwork Competence Requirements in Nuclear Power Plant Control-Room Crews across Operational States – a Field Study*. HWR-1107. Halden, Norway: OECD Halden Reactor Project, 2016.
45. IAEA-TECDOC-1846. *Regulatory Oversight of Human and Organizational Factors for Safety of Nuclear Installations*, https://www-pub.iaea.org/MTCD/Publications/PDF/TE-1846web.pdf (2018, accessed June 2020).
46. Williams JC. HEART – a proposed method for achieving high reliability in process operation by means of human factors engineering technology. *Saf Reliab* 2015; 35 (3).
47. Bye A, Lois E, Dang VN, et al. *International HRA Empirical Study – Phase 2 Report: Results from Comparing HRA Method Predictions to Simulator Data from SGTR Scenarios*. NUREG/IA-0216 Vol. 2, US Nuclear Regulatory Commission, Washington DC, USA, 2011.

48. Crichton M and Flin R. Identifying and training non-technical skills of nuclear emergency response teams. *Ann Nucl Energy* 2004. 31: 1317-1330.
49. O'Connor P, O'Dea A, Flin R, et al. Identifying the team skills required by nuclear power plant operations personnel. *Int J Ind Ergon* 2008, 28: 1028-1037.
50. Holmgren L and Skjerve AB. *Team Self-Assessment Tool (TESA)*. HWR-1082 Rev. 2. Halden, Norway: OECD Halden Reactor Project, 2016.
51. Gelman A, Carlin J, Stern H, et al. *Bayesian Data Analysis*. 2nd edition. Chapman and Hall/CRC, 2003.
52. Kruschke JK. *Doing Bayesian Data Analysis*. 2nd edition. Academic Press, 2015.
53. Plummer M. JAGS: A Program for Analysis of Bayesian Graphical Models using Gibbs Sampling. In: *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, DSC 2003, March 20-22, Vienna, Austria.
54. Denwood MJ. Runjags: An R Package Providing Interface Utilities, Model Templates, Parallel Computing Methods and Additional Distributions for MCMC Models in JAGS. *J Stat Softw* 2016, 71(9): 1–25.
55. Groth K.M., Mosleh A. Deriving causal Bayesian networks from human reliability analysis data: a methodology and example mode. *Proc Inst Mech Eng, Pt O: J Risk Reliab* 2012; 226(4) p. 361–79.
56. Groth KM. A framework for using SACADA to enhance the qualitative and quantitative basis of HRA. In: *14th Reliability Safety Assessment and Management*, PSAM 14 2018, UCLA Meyer & Renee Luskin Conference Center, Los Angeles, California.
57. Mkrtchyan L, Podofillini L and Dang VN. Methods for building Conditional Probability Tables of Bayesian Belief Networks from limited judgment: An evaluation for Human Reliability Application. *Reliab Eng Syst Saf* 2016, 151: 93-112.

Chapter 4: Traceable integration of data and judgment in HEP estimation

This chapter contains the last of the three author's articles reproduced by this thesis (see "Publication details" below). As discussed in subsection 1.2.3, a traceable and transparent incorporation of expert judgment is required whenever the latter is used to complement simulator data in the estimation of human error probability values and bounds. Building on the Bayesian variability model for simulator data discussed in Chapter 2, this chapter addresses how formally integrate data with judgment in the HEP estimation process (research objective #3 in Section 1.2).

The chapter first shows how judgment (in the form of expert estimates on task failure probability) is mathematically combined with simulator data in an upgraded formulation of the HEP variability model presented in Chapter 2 (research task 3.1, in Section 1.2). The chapter then introduces the two-stage Bayesian model (research task 3.2), developed to systematically and traceably integrate data and judgment in the derivation of average HEP values and variability bounds for various constellations of task and PSF categories (first stage), as well as in the estimation of failure probabilities for plant-specific tasks (second stage). The developed two-stage Bayesian model was first verified with artificially-generated evidence (research task 3.3). Then, the model was applied to a collection of human failure events from the recent HRA Empirical Studies (research task 3.4). The insights from the application are further discussed in the closing section of this chapter.

Publication details

This chapter reproduces the content of the following paper: **Greco SF**, Podofillini L and Dang VN. A Bayesian two-stage approach to integrate simulator data and expert judgment in human error probability estimation. Currently under internal review, expected submission date: June 2021.

Additional information relevant to this chapter

- Numerical results from both sensitivity analysis and model application to case study are reported in Appendix G.
- The code developed for the implementation of the two-stage Bayesian model is provided in Appendix H.

Abstract

With the ongoing efforts to collect new data for Human Reliability Analysis (HRA) (in particular, from nuclear power plant control room simulators), it becomes important that the coming data will be processed traceably, addressing its underlying variability, eventually in combination with expert judgment. In this direction, this work presents a two-stage Bayesian model to integrate expert-elicited probability estimates and empirical evidence from simulator data collection programs in the quantification of HEP values and of the associated variability distributions. The general aim is to provide a data aggregation framework able to mathematically combine diverse information sources throughout the HEP estimation process, in a systematic and reproducible way, contributing to strengthening the empirical basis of future HRA methods. The Bayesian model can be used to produce reference values and bounds for various HRA methods' task and factor categories, as well as to improve the quality of plant-specific HEP estimates for use in Probabilistic Safety Assessment applications. The model is first verified with artificial data and then applied to quantify the HEP of human failure events from literature, to demonstrate its applicability to practical HRA problems. Model sensitivity to biases in expert estimates is also investigated.

Nomenclature

$\boldsymbol{F}$:	combination of taxonomy categories (e.g. task type and PSF levels/ratings), referred as "constellation".
$f_{\boldsymbol{F}}(p_{t\|\boldsymbol{F}}\|\boldsymbol{\theta}_{\boldsymbol{F}})$:	parametric variability function, modelling HEP population variability across the task/context realizations within the given $\boldsymbol{F}$.
$\boldsymbol{\theta}_{\boldsymbol{F}}$:	set of (unknown) parameters of the variability function $f_{\boldsymbol{F}}(p_{t\|\boldsymbol{F}}\|\boldsymbol{\theta}_{\boldsymbol{F}})$.
$p_{t\|\boldsymbol{F}}$:	task-, context-specific HEP variable.
$t\|\boldsymbol{F}$:	index for the task/context realization within the given $\boldsymbol{F}$.
p_i:	specific numerical value of $p_{t\|\boldsymbol{F}}$ associated to the i-th realization of $\boldsymbol{F}$.
$\{k_i, N_i\}$:	number of k_i failures on N_i crew observations of the i-th task/context realization associated to $\boldsymbol{F}$ (i.e. the piece of evidence $E_{S,i}$). $i = \{1, 2 \ldots, m\}$, where m: total number of realizations in the dataset.
$\{k_{tot}, N_{tot}\}$:	total number of failures and observations for $\boldsymbol{F}$ ("lumped data").
$\tilde{p}_{t\|\boldsymbol{F}}$:	point estimate of the task-, context-specific HEP value provided by the domain expert ($\tilde{p}_i$: point estimate for the i-th task/context realization).
$g_{\boldsymbol{F}}(\tilde{p}_{t\|\boldsymbol{F}}\|p_{t\|\boldsymbol{F}})$:	probability function modelling the analyst's belief in the expert's ability to provide a correct estimate of $p_{t\|\boldsymbol{F}}$ (represented by the lognormal error model from references [38-39]).
σ_i:	logarithmic standard deviation of $g_{\boldsymbol{F}}(\tilde{p}_{t\|\boldsymbol{F}}\|p_{t\|\boldsymbol{F}})$ reflecting analyst's confidence on the specific estimate $\tilde{p}_i$ provided by the expert. Alternatively expressed as error factor EF_i, forms the piece of evidence $E_{J,i}$: $\{\tilde{p}_i, EF_i\}$.

$LN(\ldots)$:	lognormal distribution.
$\{E_{S\mid\boldsymbol{F}}, E_{J\mid\boldsymbol{F}}\}$:	evidence for the constellation $\boldsymbol{F}$, entering Stage I of the Bayesian model.
$\pi_0(\boldsymbol{\theta}_{\boldsymbol{F}}\mid E_0)$:	prior distribution of Stage I, representing the knowledge of the set of parameters $\boldsymbol{\theta}_{\boldsymbol{F}}$ before collecting the evidence $\{E_{S\mid\boldsymbol{F}}, E_{J\mid\boldsymbol{F}}\}$. E_0 : prior knowledge of $\boldsymbol{\theta}_{\boldsymbol{F}}$.
$\pi\left(\boldsymbol{\theta}_{\boldsymbol{F}}\mid E_{S\mid\boldsymbol{F}}, E_{J\mid\boldsymbol{F}}\right)$:	posterior distribution of Stage I, representing the knowledge of the set of parameters $\boldsymbol{\theta}_{\boldsymbol{F}}$ after collecting the evidence $\{E_{S\mid\boldsymbol{F}}, E_{J\mid\boldsymbol{F}}\}$.
$L\left(E_{S\mid\boldsymbol{F}}, E_{J\mid\boldsymbol{F}}\mid\boldsymbol{\theta}_{\boldsymbol{F}}\right)$:	likelihood function of Stage I, i.e. the probability density that evidence $\{E_{S\mid\boldsymbol{F}}, E_{J\mid\boldsymbol{F}}\}$ is observed.
$P_{\boldsymbol{F}}(p_{t\mid\boldsymbol{F}})$:	estimated HEP population variability distribution for the constellation $\boldsymbol{F}$ (used as prior distribution in Stage II).
$\{k_{HFE}, N_{HFE}\}$:	the piece of evidence $E_{S\mid HFE}$, where k_{HFE} is the number of failures observed over N_{HFE} crew observations collected for the given human failure event (HFE) in the specific plant-simulator.
$\{\tilde{p}_{HFE}, EF_{HFE}\}$:	the piece of evidence $E_{J\mid HFE}$, where $\tilde{p}_{HFE}$ is the expert estimate on the HEP value of the given HFE, and EF_{HFE} the associated confidence level.
$\{E_{S\mid HFE}, E_{J\mid HFE}\}$:	evidence for the given HFE, entering Stage II of the Bayesian model.
$\pi(p_{t\mid\boldsymbol{F}}\mid E_{S\mid HFE}, E_{J\mid HFE})$:	posterior distribution of Stage 2, representing the knowledge of the HEP value of the given HFE after collecting the evidence $\{E_{S\mid HFE}, E_{J\mid HFE}\}$. For simplicity, the posterior is referred as HEP uncertainty distribution $P_{HFE}(HEP)$.
$L_{HFE}(E_{S\mid HFE}, E_{J\mid HFE}\mid p_{t\mid\boldsymbol{F}})$	likelihood function of Stage II, i.e. the probability density that evidence $\{E_{S\mid HFE}, E_{J\mid HFE}\}$ is observed.
$\{\mu_{\boldsymbol{F}}, \sigma_{\boldsymbol{F}}\}$:	parameters of the lognormal variability distribution (mean and standard deviation) used in the numerical application of Stage I.
$\{HEP_5, HEP_{95}\}$:	recommended HEP bounds for $\boldsymbol{F}$ from HRA literature, used as prior information to construct the lognormal informative prior for $\mu_{\boldsymbol{F}}$, i.e. $\pi_0(\mu_{\boldsymbol{F}})$.
$\{\mu_{\mu_{\boldsymbol{F}}}, \sigma_{\mu_{\boldsymbol{F}}}\}$:	parameters (mean and standard deviation) of $\pi_0(\mu_{\boldsymbol{F}})$.
$\{\alpha, \beta\}$:	shape parameters of the beta prior distributions in the conjugate beta-binomial model with lumped data.
b:	bias factor.
$E^r_{S\mid\boldsymbol{F}}$:	r-th replicated dataset $\{E^r_{S\mid\boldsymbol{F}} = (k^r_i, N_i), i = 1, \ldots m; r = 1, \ldots R\}$, where R is the total number of replicates.
$\{T(E^r_{S\mid\boldsymbol{F}}), T(E_{S\mid\boldsymbol{F}})\}$:	generic test quantities associated to $E^r_{S\mid\boldsymbol{F}}$ and $E_{S\mid\boldsymbol{F}}$, respectively.
$\{\bar{k}^r, \bar{k}\}$:	mean values of the replicated k^r_i and the observed k_i, respectively.
p_B:	Bayesian p-value, i.e. the probability $P(T(E^r_{S\mid\boldsymbol{F}}) \geq T(E_{S\mid\boldsymbol{F}}))$.

4.1 Introduction

As part of Probabilistic Safety Assessment (PSA), Human Reliability Analysis (HRA) addresses the contribution of human failures to risk in complex technical systems, e.g. nuclear power plants, chemical and aerospace systems [1-3]. HRA methods support the identification of the safety-critical tasks performed by the personnel, the characterization of the contextual factors influencing crew performances, and finally the assessment of the task failure probabilities (referred as Human Error Probabilities, HEPs).

The estimation of HEP values is supported by quantitative models that represent both the operational tasks and the context-related influencing factors via categories (typically, of task types and of Performance Shaping Factors, PSFs), and relate these categories to values of failure probability. The HRA models (e.g. [4-11]) are parametrized on reference HEP values: these provide baseline HEP values, e.g. the HEP corresponding to tasks performed under optimal/nominal performance conditions, as well as the PSF's effect, typically as multipliers to the baseline. Advanced HRA models, such as based on Bayesian Belief Networks (BBNs) [12-16], require reference HEP values too, for example to inform the BBN Conditional Probability Distributions (CPDs).

The data underlying the reference HEP values is generally obtained by combining empirical evidence and expert judgment [1-3]. Empirical data has been traditionally gathered from a variety of information sources: licensee event reports, retrospective analyses of accidents and operational events, human factors and behavioral science experiments [17]. Similarly, judgment has typically assumed different forms, e.g. quantitative probability estimates and/or qualitative statement on the importance of influencing factors [17]. In addition, an overarching contribution of expert judgment exists in the evaluation of the suitability of the different data sources to the specific HRA model development.

Due to the general lack of data, its diversity and its often uncertain quality, there has been very limited traceability in the aggregation of the various data sources, as well as in their combination with expert judgment. For instance, quoting the THERP Handbook [4], "the data underlying THERP's model is mostly coming from human factor experiments and field studies (…). The probability values are generally derived data, in the sense that they contain much extrapolation and judgment". The result is that it is difficult to determine to what extent the HEP values produced by HRA models are empirically-based. Also, in absence of a traceable process, it is not clear how to incorporate new data eventually becoming available to feed HRA models. Despite the data collection challenges, important attempts to validate HRA methods have been performed in the past, e.g. by [18-19], and more recently from the International [20] and US [21] HRA Empirical Studies. These studies demonstrate the usefulness of HRA methods and the trustworthiness of the produced HEPs (at least for some types of operator tasks), beyond the limitations of the HRA data processing. However, with the recent emphasis to improve the empirical basis of HRA methods [20-22] and the modern ongoing data collection programs [23-24], it becomes important that methods for data integration are ready to accommodate sustained data generation: on the long term, these programs are expected to produce a significant amount of new observations, that can be used to empirically-inform reference HEP values and bounds of HRA models. Some preliminary studies [25-31] have been performed to assess the feasibility of estimating HEP values; however, these studies have not addressed the incorporation of data and expert judgment in the HEP calculation.

The current amount of data collected by the on-going initiatives [23-24] is still not sufficient to derive statistically-significant HEP estimates for the entire spectrum of task and PSF categories of HRA models. Hence, the incorporation of expert judgment will still play an important role for future HRA, also in consideration that current simulators are not suited for data collection for all scenarios of interest for PSA applications (e.g. tasks in response to events external to the plant such as seismic events and tasks in severe accident situations) [32].

The general goal of the present chapter is to contribute to the development of HRA data aggregation models to traceably process diverse data sources, including expert judgment [33-34]. In particular, this chapter presents a Bayesian two-stage model to integrate judgment (in the form of expert-elicited failure probability estimates) and simulator data, in the quantification of HEP values. The two stages of the developed Bayesian model address different purposes. The first stage aims at producing reference HEP values and bounds to feed HRA methods (e.g. to parametrize task and PSF categories or to inform BBN's CPDs). The product of the first stage is a population distribution, to represent the diversity of data types and sources, plant-to-plant and crew-to-crew variability, as well as variability within the HRA methods categories (e.g. of task types and PSFs). The second stage utilizes plant-specific information (again data and judgement) to produce plant-specific HEPs, updating the generic distributions from the first stage, eventually to be used in plant-specific PSAs.

Bayesian inference methods [35] represent a natural framework to formally treat expert-elicited estimates and combine these with empirical data in the estimation of PSA-related quantities (e.g. a single probability value, or the parameters of a population distribution) [36-38]; their use has been well-established over the years by different PSA applications [39-43]. Similarly to the present work, references [41-42] adopt Bayesian two-stage approaches for the aggregation of expert opinions and reliability data (possibly sparse and from diverse plants) to derive generic population variability distributions of reliability parameters (e.g. core melt frequency of nuclear power reactors [41]; pump failure rate [42]), and support parameter estimation for plant-specific components. In this regard, the application performed in this work represents a first-of-a-kind attempt to explore the use of Bayesian two-stage models in a practical HRA problem.

The chapter is structured as follows. Section 4.2 presents the concepts and the general formulation of the combined use of expert estimates and simulator data to represent HEP population variability. Section 4.3 describes the structure of the developed Bayesian two-stage model and its implementation with lognormal distributions for use in the numerical application. The model is verified with artificially-generated data, to analyze the effects of judgment incorporation on the HEP estimates and investigate model sensitivity to biases in expert judgment. Then, Section 4.4 presents its practical application to a case study, involving a collection of human failure events from the recent HRA Empirical Studies [20-21]. The results are further discussed in Section 4.5, together with recommendations and insights on the applicability of the developed model. Conclusions are given at closure.

4.2 Integrated use of data and expert estimates in modelling HEP population variability: concepts and mathematical formulation

4.2.1. Concept: data and expert opinions as "mixed evidence" for HEP estimation

Figure 4.1 gives an overview of how the work presented in this chapter fits in the HRA method development and application processes. The figure focuses on the data used to develop HRA models: other aspects of model development (e.g. development of task and PSF taxonomies) are not shown. The upper part of the figure addresses the HRA method development. Data is typically diverse by type (e.g. simulator data and expert judgment) and by subject (collections on different accident scenarios, crews, personnel tasks). This data is used to determine reference HEP values for combinations ("constellations") of task and PSF categories, specific for the different HRA methods (shaded box in Figure 4.1). The generic constellation $\boldsymbol{F} = \{F_1, F_2, \ldots\}$ of categories F_1, F_2,…, may represent a diagnosis task under nominal conditions, or the constellation required to assess PSFs multipliers, or the factor combination for CPDs in a Bayesian Belief Network [12-16].

In the derivation of reference values for the given $\boldsymbol{F}$, it becomes important to represent the spectrum of "population variability" (e.g. plant-to-plant, task-to-task, crew-to-crew) intrinsic to the data sources and subjects, to avoid overconfidence in HRA model results [44]. In the present work, this is done via population variability distributions obtained in the first stage, building on the Bayesian approach presented by the authors in recent work [44[1]]. The modelling approach in [44] applies to simulator data, while the present work addresses both data and expert judgment, integrated as "mixed evidence" in the Bayesian model (subsection 4.2.2).

HRA applications (lower part of Figure 4.1) typically address specific Human Failure Events (HFEs), analyzed in plant-, scenario-, and task-specific contexts. The context analysis determines the representative task and PSF categories. In this sense, the conventional HRA

---

[1] In this thesis, the referred article is reproduced as Chapter 2: A Bayesian variability model for simulator data.

analyses produce context-specific HEPs, but based on generic data (because of the diverse data feeding HRA models). The present work considers the case that plant-specific data is also available to inform the HEP estimate, for example from simulator training sessions and expert judgment addressing the specific HFE of interest. This plant-specific evidence can be incorporated in the HEP estimate (and incorporated in the plant PSA) via the second stage of the Bayesian model presented in this chapter, thus obtaining context-specific HEPs, based on both generic and plant-specific data (shown as "Long-term concept" in Figure 4.1). Since the focus of the present work is on the Bayesian process, this chapter will apply the second stage to the same constellations $\boldsymbol{F}$ used for the demonstration of the first stage of the Bayesian model ("This work (demonstration)" in Figure 4.1). In practice, the present chapter does not address the intermediate step of HRA method development and application.

Within the diversity of the information sources potentially available (upper left box in Figure 4.1), this chapter addresses the type of data shown in Table 4.1, i.e. from main control room simulators and a specific fashion of expert judgment, i.e. direct HEP elicitation (eventually via the application of an existing HRA method). As an example, Table 4.1 reports data pertaining to the constellation ($\boldsymbol{F}$) of the following taxonomy categories: task type "understanding the situation/problem" (F_1), with PSF "information quality" rated as "missing/masked" (F_2), taking the example naming from the Scenario Authoring Characterization And Debriefing Application (SACADA [24]) taxonomy. The simulator records (i.e. the rows in Table 4.1) are collected from different task realizations (e.g. in Table 4.1, "identification of faulted steam generator") performed by operating crews in different scenarios (e.g. in Table 4.1, a Steam Generator Tube Rupture, SGTR, and a Small break Loss Of Coolant Accident, SLOCA) that are simulated in different plants, hence under different design of human-machine interfaces, training programs, and procedural guidance (representing the subject diversity from Figure 4.1). For each record, the simulator data is in the form of number of task failures and observations (respectively, k_i and N_i in Table 4.1). For each record (Table 4.1), in addition to these failure counts, it is assumed to have HEP estimates derived from domain experts, e.g. via direct elicitation or through the application of an HRA method [38]. Such estimates (the $\tilde{p}_i$'s in Table 4.1, with associated level of confidence EF_i as presented in the next Section 4.2.2) reflect the experts' state of knowledge on the HEP for the specific task realization (e.g. "identification of faulted steam generator") in the specific operational context (e.g. "failure of secondary radiation indications") [35, 43]. Note that how to elicit HEP estimates from experts is not within the scope of the present work: formal approaches for judgment elicitation can be found in literature, for HRA [45] as well as outside HRA field [46-47].

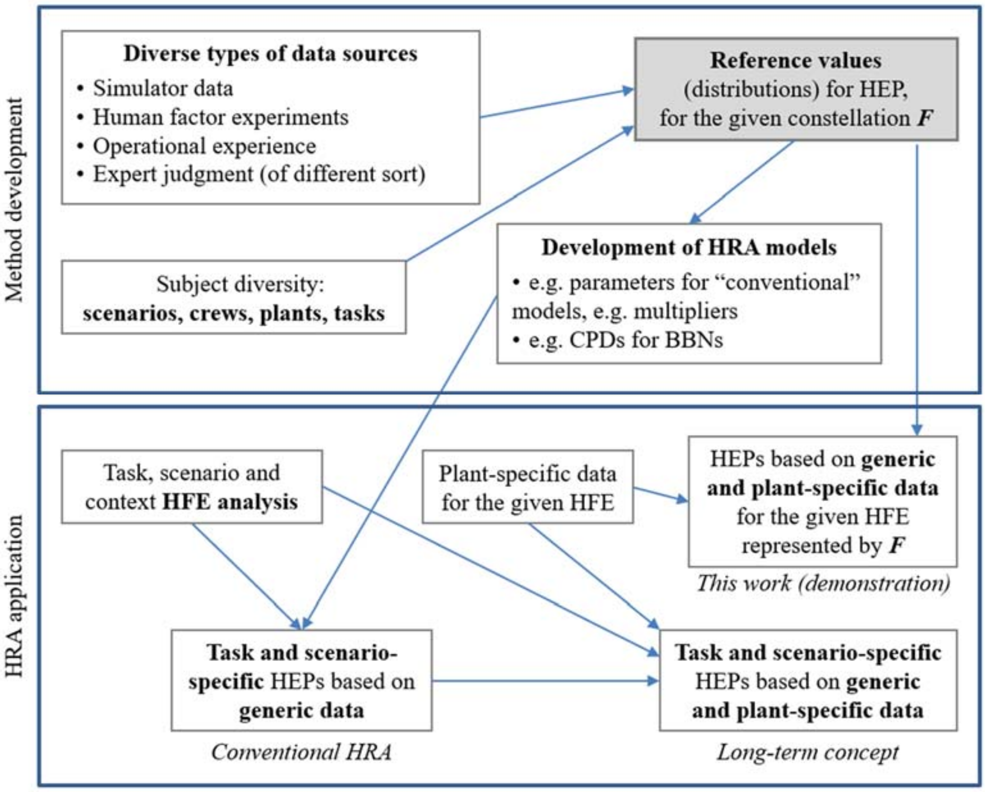

Method development
Diverse types of data sources
• Simulator data
• Human factor experiments
• Operational experience
• Expert judgment (of different sort)
Reference values (distributions) for HEP, for the given constellation F
Subject diversity: scenarios, crews, plants, tasks
Development of HRA models
• e.g. parameters for "conventional" models, e.g. multipliers
• e.g. CPDs for BBNs
HRA application
Task, scenario and context HFE analysis
Plant-specific data for the given HFE
HEPs based on generic and plant-specific data for the given HFE represented by F
This work (demonstration)
Task and scenario-specific HEPs based on generic data
Conventional HRA
Task and scenario-specific HEPs based on generic and plant-specific data
Long-term concept

Figure 4.1. Concept for use of reference HEP values and its application in the present work.

Table 4.1. Hypothetical simulator and expert-elicited data for different tasks and operational contexts, for constellation (***F***) (the examples of task and PSF categories are taken from SACADA taxonomy [24] for illustration purpose).

F: task type = “understanding the situation/problem”, PSF “information quality” = “missing/masked” (taxonomy from SACADA [24])							
Scenario	Operational context	Task	Plant simulator	Failures (k_i)	Observations (N_i)	Expert estimates ($\tilde{p}_i$)	Confidence (EF_i)
SGTR	Failure of secondary radiation indications	Identification of faulted steam generator	A	0	4	5.46e-02	3
SGTR	Radiation alarms already activated by early releases	Identification of faulted steam generator	B	0	3	3.20e-02	5
SGTR	Failure of secondary radiation indications	Identification of faulted steam generator	C	1	5	1.15e-01	5
SGTR	(…)	(…)	(…)	(…)	(…)	(…)	
SLOCA	No indications on leaks’ specific location	Identification of leak source	A	0	1	2.10e-02	7
SLOCA	No indications on leaks’ specific location	Identification of leak source	D	0	2	7.20e-03	5
SLOCA	(…)	(…)	(…)	(…)	(…)	(…)	
				$k_{tot} = 1$	$N_{tot} = 25$		

4.2.2. HEP population variability within $\boldsymbol{F}$: mathematical formulation and modelling assumptions

The quantity of interest is the HEP associated to the given constellation $\boldsymbol{F}$ of task and PSF categories adopted by the specific data collection taxonomy, namely: *HEP* = *HEP*($\boldsymbol{F}$). Figure 4.2 provides a sketch of the mathematical formulation of *HEP*($\boldsymbol{F}$); each task/context realization (i.e. each row of Table 4.1) associated to the given $\boldsymbol{F}$ is characterized by a unique HEP, modelled in Figure 4.2 by the variable $p_{t|\boldsymbol{F}}$ (with "t" indicating the dependence on the specific task realization). Accordingly, the six realizations of $\boldsymbol{F}$ contained in Table 4.1 are associated to six different values of the variable $p_{t|\boldsymbol{F}}$. The population variability across the different task-, context-specific $p_{t|\boldsymbol{F}}$ values is (Figure 4.2, top):

$$p_{t|\boldsymbol{F}} \sim f_{\boldsymbol{F}}(p_{t|\boldsymbol{F}}|\boldsymbol{\theta_F}) \tag{4.1}$$

where $\boldsymbol{\theta_F}$ represents the vector of unknown parameters (e.g. in the numerical application in Section 4.4, the mean and standard deviation of a lognormal variability distribution). The population parameters $\boldsymbol{\theta_F}$ are inferred in the first stage of the developed Bayesian model (subsection 4.3.2).

For the i-th realization, the specific value p_i is informed by combining the empirical data (Figure 4.2, bottom left: the count of failures k_i over N_i crew observations) with the corresponding expert-elicited failure probability estimate (Figure 4.2, bottom right: the point estimate $\tilde{p}_i$). For ease of discussion, we assume that a single estimate $\tilde{p}_i$ is available, either from a single expert or aggregated across multiple experts (the methods in [43, 47] can be used to aggregate estimates). The probabilistic relationship between $p_{t|\boldsymbol{F}}$ and the associated failure data - e.g. for the i-th realization, between p_i and the pair (k_i, N_i) in Figure 4.2 (bottom left) - is modelled as a binomial aleatory process. Whereas the probabilistic relationship between $p_{t|\boldsymbol{F}}$ and the corresponding expert estimate $\tilde{p}_{t|\boldsymbol{F}}$ - e.g. for the i-th realization, between p_i and $\tilde{p}_i$ in Figure 4.2 (bottom right) - is captured by the function (as in [43]):

$$\tilde{p}_{t|\boldsymbol{F}} \sim g_{\boldsymbol{F}}(\tilde{p}_{t|\boldsymbol{F}}|p_{t|\boldsymbol{F}}) \tag{4.2}$$

expressing the probability density that the given expert provides the value $\tilde{p}_{t|\boldsymbol{F}}$ as point estimate, given that the true failure probability value is $p_{t|\boldsymbol{F}}$. The function $g_{\boldsymbol{F}}(\tilde{p}_{t|\boldsymbol{F}}|p_{t|\boldsymbol{F}})$ models the analyst's belief in the expert's ability to provide a correct estimate of $p_{t|\boldsymbol{F}}$: different options for $g_{\boldsymbol{F}}(\tilde{p}_{t|\boldsymbol{F}}|p_{t|\boldsymbol{F}})$ are available in the PSA literature [39-40], to account for expert's level of experience, known biases, as well as any dependence to other experts [40]. For the purposes of the present work, we assume the experts are independent (e.g. the $\tilde{p}_i$'s in Table 4.1 are independent estimates) and provide unbiased failure probability estimates for the assessed task/context realizations. Under these hypotheses, the lognormal error model from [39-40] is chosen to represent $g_{\boldsymbol{F}}(\tilde{p}_{t|\boldsymbol{F}}|p_{t|\boldsymbol{F}})$ in eq. 4.2, formally:

$$\tilde{p}_{t|\boldsymbol{F}} \sim g_{\boldsymbol{F}}(\tilde{p}_{t|\boldsymbol{F}}|p_{t|\boldsymbol{F}}) = LN(\tilde{p}_{t|\boldsymbol{F}}|p_{t|\boldsymbol{f}},\sigma_i) = LN(\tilde{p}_{t|\boldsymbol{F}}|p_{t|\boldsymbol{f}}, 1.645^{-1}\log(EF_i)) \qquad (4.2\text{bis})$$

where *LN* is used to denote a lognormal distribution. According to eq. 4.2bis, $\tilde{p}_i$ (i.e. the specific $\tilde{p}_{t|\boldsymbol{F}}$ value for the i-th realization of $\boldsymbol{F}$ in Table 4.1) is expected to be lognormally distributed around the true value p_i (Figure 4.2, bottom right), with a logarithmic standard deviation σ_i that reflects the analyst's confidence on the given expert (note σ_i in eq. 4.2bis can be alternatively expressed in terms of error factor[2] EF_i, as in Table 4.1). For instance, in case σ_i is taken equal to zero (i.e. EF_i is equal to one), the function $g_{\boldsymbol{F}}(\tilde{p}_{t|\boldsymbol{F}}|p_{t|\boldsymbol{F}})$ becomes a delta function: the elicited $\tilde{p}_i$ matches the real value p_i ("perfect expert", [39-40]); in the general case of "imperfect experts", the larger the σ_i (and correspondingly the EF_i), the lower the confidence in the expert's ability to estimate p_i.

Note that, compared to the HEP variability formulation in [44[3]], the function $f_{\boldsymbol{F}}(p_{t|\boldsymbol{F}}|\boldsymbol{\theta}_{\boldsymbol{F}})$ does not distinguish the crew-to-crew variability component from the other sources of population variability (e.g. task-to-task, plant-to-plant) within the constellation $\boldsymbol{F}$. The variable $p_{t|\boldsymbol{F}}$ in eq. 4.1 represents indeed the failure probability of an "average crew" in performing the given task-, context-specific realization of $\boldsymbol{F}$: for instance, with reference to Table 4.1, the four crew observations available for the first realization (i.e. the first row of Table 4.1) are associated to the same $p_{t|\boldsymbol{F}}$ value, independently of the specific crew involved. Simply put, the approach in [44] assumes different failure probabilities per each crew, while the present chapter per each task/context realization, aggregating different behavioral characteristics across crews (e.g. in teamwork, decision-making, communication strategies) in the crew-generic $p_{t|\boldsymbol{F}}$ variable. The explicit treatment of crew-to-crew variability within $\boldsymbol{F}$ is outside the scope of the present work (this aspect is further discussed in Section 4.5).

---

[2] In PSA/HRA, the error factor is a commonly-adopted measure of dispersion for characterizing the spread of a lognormal distribution. Typically, the EF is expressed by the square root of the ratio 95th/5th percentiles.

[3] In the present thesis, the referred HEP variability model from the authors' article [44] is presented in Section 2.3.

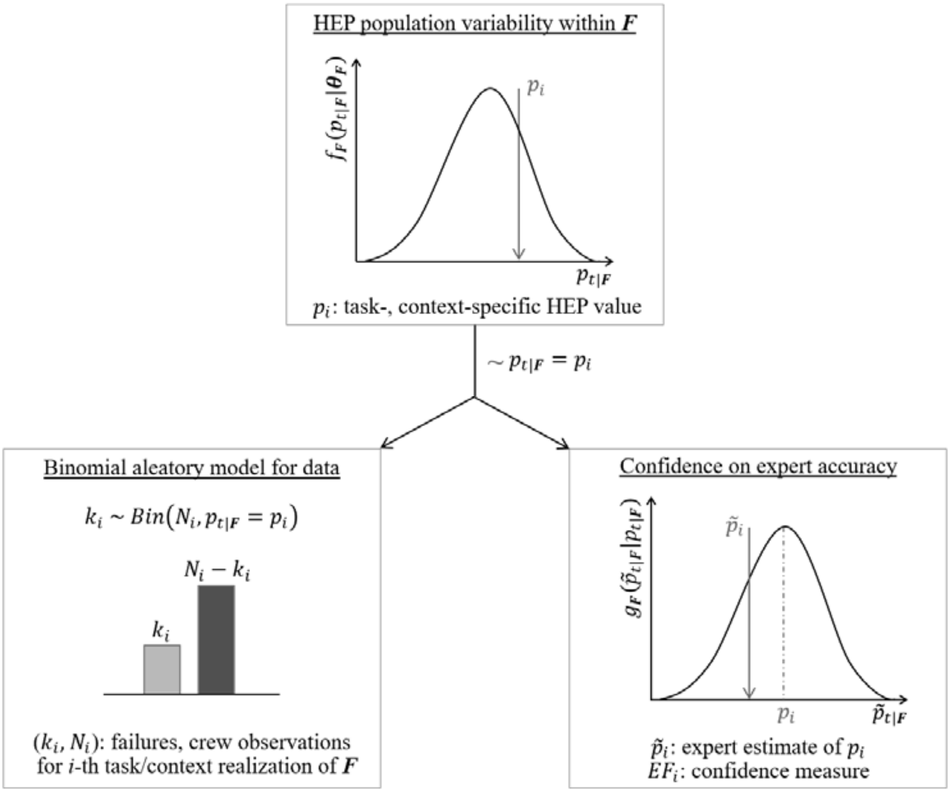

HEP population variability within F
$f_F(p_{t|F}|\theta_F)$
p_i
$p_{t|F}$
p_i: task-, context-specific HEP value
$\sim p_{t|F} = p_i$
Binomial aleatory model for data
$k_i \sim Bin(N_i, p_{t|F} = p_i)$
$N_i - k_i$
k_i
(k_i, N_i): failures, crew observations for i-th task/context realization of F
Confidence on expert accuracy
$g_F(\tilde{p}_{t|F}|p_{t|F})$
$\tilde{p}_i$
p_i
$\tilde{p}_{t|F}$
$\tilde{p}_i$: expert estimate of p_i
EF_i: confidence measure

Figure 4.2. Mathematical formulation of HEP(***F***) adopted in this work (adapted from [44] and extended to incorporate expert estimates). Top: continuous parametric distribution for $p_{t|\boldsymbol{F}}$, with unknown parameters $\boldsymbol{\theta_F}$. Bottom left: binomial aleatory process (failures k_i and observations N_i) for the i-th task, governed by the associated HEP value (p_i). Bottom right: probability distribution representing the expert estimate $\tilde{p}_i$ and associate accuracy.

4.3 The Bayesian two-stage model for HEP quantification

This section presents the Bayesian two-stage model for HEP quantification. First, the section introduces the general equations underlying each stage (subsection 4.3.1). The specific formulation with lognormal variability distributions adopted in the numerical application to case study (Section 4.4) is provided in subsection 4.3.2 and then tested with artificially-generated data in subsection 4.3.3.

4.3.1. General formulation

The flowchart in Figure 4.3 gives an overview of the inference steps (“Bayesian update” in Figure 4.3) throughout the two-stage model. The structure of both stages is based on the general formulation of the Bayesian update [38]:

$$\pi(x|E) = A^{-1}L(E|x)\pi_0(x|E_0) \tag{4.3}$$

where:

- x is the unknown quantity (e.g. a probability value, or a set of parameters) of the inference problem;
- π_0 and π are the prior and posterior probability functions for x, modelling the state of knowledge of the analyst on the investigated quantity, before and after the evidence E is collected, respectively (with E_0 expressing the prior evidence of x, if available);
- $L(E|x)$ is the likelihood function, representing the probability density that the evidence E is observed;
- E is the available evidence for the quantity x ("Information sources" in Figure 4.3);
- $A^{-1} = \int L(E|x)\pi_0(x|E_0)dx$, the denominator of eq. 4.3, normalizes the (updated) posterior distribution π to a probability density function.

In the first stage of the model ("Stage I" in Figure 4.3, blue box), the quantity x in eq. 4.3 is represented by the set of parameters $\boldsymbol{\theta_F}$ of $f_{\boldsymbol{F}}(p_{t|\boldsymbol{F}}|\boldsymbol{\theta_F})$, of unknown values. The evidence (E in eq. 4.3) for Stage I (Figure 4.3, blue box) comes from the following information sources:

- human failure data from plant simulators ($E_{S|\boldsymbol{F}}$ in Figure 4.3, blue box), from different task/context realizations of $\boldsymbol{F}$. Considering m realizations (m different records in Table 4.1), the evidence $E_{S|\boldsymbol{F}}$ is expressed as the set of pairs $\{E_{S,i} = (k_i, N_i)$, $i = 1, 2, \ldots m\}$ of k_i failures on N_i observations for the i-th realization. Each pair $E_{S,i} = (k_i, N_i)$ informs the specific $p_{t|\boldsymbol{F}}$value associated to the i-th realization (i.e. p_i in Figure 4.2).
- judgment-based probability estimates by domain experts ($E_{J|\boldsymbol{F}}$ in Figure 4.3, blue box), in the form of point estimates (i.e. $\tilde{p}_{t|\boldsymbol{F}}$ in eq. 4.2) of the task-, context-specific $p_{t|\boldsymbol{F}}$ values. The evidence $E_{J|\boldsymbol{F}}$ is then expressed as the set of pairs $\{E_{J,i} = (\tilde{p}_i, EF_i)$, $i = 1, 2, \ldots m\}$, where $EF_i = e^{1.645\sigma_i}$ represents the analyst's confidence on the accuracy of $\tilde{p}_i$, see eq. 4.2bis [39-40].

From eq. 4.3, the Bayesian update for Stage I can be written as:

$$\pi\left(\boldsymbol{\theta_F}|E_{S|\boldsymbol{F}}, E_{J|\boldsymbol{F}}\right) = \frac{L\left(E_{S|\boldsymbol{F}}, E_{J|\boldsymbol{F}}|\boldsymbol{\theta_F}\right)\pi_0(\boldsymbol{\theta_F}|E_0)}{\int_{\boldsymbol{\theta_F}} L\left(E_{S|\boldsymbol{F}}, E_{J|\boldsymbol{F}}|\boldsymbol{\theta_F}\right)\pi_0(\boldsymbol{\theta_F}|E_0)d\boldsymbol{\theta_F}} \tag{4.4}$$

The core element of eq. 4.4 is the likelihood term, i.e. $L\left(E_{S|\boldsymbol{F}}, E_{J|\boldsymbol{F}}|\boldsymbol{\theta_F}\right)$: here, simulator data $E_{S|\boldsymbol{F}}$ and expert estimates $E_{J|\boldsymbol{F}}$ (i.e. the mixed evidence for the task-, context-specific $p_{t|\boldsymbol{F}}$ values) update the prior probability distribution $\pi_0(\boldsymbol{\theta_F}|E_0)$ to the posterior $\pi\left(\boldsymbol{\theta_F}|E_{S|\boldsymbol{F}}, E_{J|\boldsymbol{F}}\right)$.

For the i-th task/context realization, the likelihood of observing the evidence $E_{S,i}, E_{J,i}$ is:

$$L_i\left(E_{S,i}, E_{J,i}|\boldsymbol{\theta_F}\right) = \int_{p_{t|\boldsymbol{F}}} Bin(k = k_i|p_{t|\boldsymbol{F}}, N_i)g_{\boldsymbol{F}}(\tilde{p}_{t|\boldsymbol{F}}|p_{t|\boldsymbol{F}})f_{\boldsymbol{F}}(p_{t|\boldsymbol{F}}|\boldsymbol{\theta_F})dp_{t|\boldsymbol{F}} =$$

$$\int_{p_{t|\boldsymbol{F}}} Bin(k = k_i|p_{t|\boldsymbol{F}}, N_i)LN(\tilde{p}_i|p_{t|\boldsymbol{f}}, \sigma_i)f_{\boldsymbol{F}}(p_{t|\boldsymbol{F}}|\boldsymbol{\theta_F})dp_{t|\boldsymbol{F}} \tag{4.5}$$

where:

- $f_{\boldsymbol{F}}(p_{t|\boldsymbol{F}}|\boldsymbol{\theta}_{\boldsymbol{F}})$ is the probability density that the failure probability of the i-th specific realization is $p_{t|\boldsymbol{F}}$;
- the binomial distribution $Bin(k = k_i|p_{t|\boldsymbol{F}}, N_i)$ expresses the probability of observing k_i failures over N_i trials of the specific i-th realization (Figure 4.2, bottom left);
- the probability distribution $g_{\boldsymbol{F}}(\tilde{p}_{t|\boldsymbol{F}}|p_{t|\boldsymbol{F}})$ (specifically $LN(\tilde{p}_i|p_{t|\boldsymbol{f}}, \sigma_i)$) expresses the probability (density) that the expert's estimate is $\tilde{p}_{t|\boldsymbol{F}} = \tilde{p}_i$, given that the true failure probability value of the i-th realization is $p_{t|\boldsymbol{F}}$ (Figure 4.2, bottom right);
- the likelihood of the evidence is then obtained by averaging the expression $Bin(k = k_i|p_{t|\boldsymbol{F}}, N_i)LN(\tilde{p}_i|p_{t|\boldsymbol{f}}, \sigma_i)$ over the variability function $f_{\boldsymbol{F}}(p_{t|\boldsymbol{F}}|\boldsymbol{\theta}_{\boldsymbol{F}})$.

In particular cases where only one type of evidence is available for the i-th realization, i.e. $E_{S,i}$ or $E_{J,i}$, eq. 4.5 reduces to $\int_{p_{t|\boldsymbol{F}}} Bin(k = k_i|p_{t|\boldsymbol{F}}, N_i) f_{\boldsymbol{F}}(p_{t|\boldsymbol{F}}|\boldsymbol{\theta}_{\boldsymbol{F}}) dp_{t|\boldsymbol{F}}$ or $\int_{p_{t|\boldsymbol{F}}} LN(\tilde{p}_i|p_{t|\boldsymbol{f}}, \sigma_i) f_{\boldsymbol{F}}(p_{t|\boldsymbol{F}}|\boldsymbol{\theta}_{\boldsymbol{F}}) dp_{t|\boldsymbol{F}}$, respectively.

Extending eq. 4.5 to the set of m task/context realizations of the constellation $\boldsymbol{F}$, the likelihood term in eq. 4.4 becomes:

$$L\left(E_{S|\boldsymbol{F}}, E_{J|\boldsymbol{F}}|\boldsymbol{\theta}_{\boldsymbol{F}}\right) = \prod_{i=1}^{m} L_i\left(E_{S,i}, E_{J,i}|\boldsymbol{\theta}_{\boldsymbol{F}}\right) =$$

$$\prod_{i=1}^{m} \int_{p_{t|\boldsymbol{F}}} Bin(k = k_i|p_{t|\boldsymbol{F}}, N_i) LN(\tilde{p}_i|p_{t|\boldsymbol{f}}, \sigma_i) f_{\boldsymbol{F}}(p_{t|\boldsymbol{F}}|\boldsymbol{\theta}_{\boldsymbol{F}}) dp_{t|\boldsymbol{F}} \tag{4.6}$$

The posterior degree of belief on the parameters of the HEP population variability distribution for the generic constellation $\boldsymbol{F}$, i.e. $\pi\left(\boldsymbol{\theta}_{\boldsymbol{F}}|E_{S|\boldsymbol{F}}, E_{J|\boldsymbol{F}}\right)$ in Figure 4.3 (blue box), is then obtained by substituting eq. 4.6 into the likelihood term of eq. 4.4. We consider that prior knowledge (E_0) relevant to $\boldsymbol{F}$ is available from HRA methods (e.g. in Figure 4.3, blue box: upper/lower bounds of *HEP*(***F***) from Technique for Human Error Rate Prediction, THERP, database [4]) and can be used to construct the prior distribution $\pi_0(\boldsymbol{\theta}_{\boldsymbol{F}}|E_0)$ in eq. 4.4: a practical example is given in the numerical application of the model in Section 4.4.

The posterior $\pi\left(\boldsymbol{\theta}_{\boldsymbol{F}}|E_{S|\boldsymbol{F}}, E_{J|\boldsymbol{F}}\right)$ computed by the Bayesian update in eq. 4.4 is finally used to derive the expected HEP population variability distribution for the constellation $\boldsymbol{F}$, namely $P_{\boldsymbol{F}}(p_{t|\boldsymbol{F}})$ in Figure 4.3 (blue box), by weighting the parametric variability distribution $f_{\boldsymbol{F}}(p_{t|\boldsymbol{F}}|\boldsymbol{\theta}_{\boldsymbol{F}})$ by $\pi\left(\boldsymbol{\theta}_{\boldsymbol{F}}|E_{S|\boldsymbol{F}}, E_{J|\boldsymbol{F}}\right)$; formally:

$$P_{\boldsymbol{F}}(p_{t|\boldsymbol{F}}) = \int_{\boldsymbol{\theta}_{\boldsymbol{F}}} f_{\boldsymbol{F}}(p_{t|\boldsymbol{F}}|\boldsymbol{\theta}_{\boldsymbol{F}})\, \pi\left(\boldsymbol{\theta}_{\boldsymbol{F}}|E_{S|\boldsymbol{F}}, E_{J|\boldsymbol{F}}\right) d\boldsymbol{\theta}_{\boldsymbol{F}} \tag{4.7}$$

The produced $P_{\boldsymbol{F}}(p_{t|\boldsymbol{F}})$'s are used in the second stage of the model ("Stage II" in Figure 4.3, red box) as prior state of knowledge on the failure probability of plant-specific Human Failure Events (HFEs), as typically referred to in plant-specific PSA studies. The quantity of interest (i.e. the x in eq. 4.3) in Stage II is the unknown HEP value associated to the specific HFE. The concept underlying the Bayesian update in Stage II (Figure 4.3, red box) is that, in lack of plant-specific evidence, the HEP is represented by the Stage I variability distribution associated to the applicable context (represented by the given $\boldsymbol{F}$). If evidence becomes available for the specific HFE, this is then used in Stage II to update the generic prior $P_{\boldsymbol{F}}(p_{t|\boldsymbol{F}})$ and incorporate the plant-specific evidence (Figure 4.3, red box):

- $E_{S|HFE}$: human performance data, expressed as $E_{S|HFE} = (k_{HFE}, N_{HFE})$, where k_{HFE} is the number of failures observed over N_{HFE} crew observations, collected in the plant simulator for the specific HFE (hence the pedix $S|HFE$);
- $E_{J|HFE}$: judgment-based evidence, in the form $E_{J|HFE} = (\tilde{p}_{HFE}, EF_{HFE})$, where $\tilde{p}_{HFE}$ is the point estimate of the HEP value provided by the consulted expert and $EF_{HFE} = e^{1.645\sigma_{HFE}}$ the associated error factor, representing the HRA analyst's confidence on expert accuracy according to eq. 4.2bis.

The Bayesian update (eq. 4.3) for Stage II is written as follows:

$$P_{HFE}(HEP) \equiv \pi\left(p_{t|\boldsymbol{F}} | E_{S|HFE}, E_{J|HFE}\right) = \frac{L_{HFE}\left(E_{S|HFE}, E_{J|HFE} | p_{t|\boldsymbol{F}}\right) P_{\boldsymbol{F}}(p_{t|\boldsymbol{F}})}{\int_{p_{t|\boldsymbol{F}}} L_{HFE}\left(E_{S|HFE}, E_{J|HFE} | p_{t|\boldsymbol{F}}\right) P_{\boldsymbol{F}}(p_{t|\boldsymbol{F}}) dp_{t|\boldsymbol{F}}} =$$
$$\frac{L_{HFE}\left(E_{S|HFE}, E_{J|HFE} | p_{t|\boldsymbol{F}}\right) \int_{\boldsymbol{\theta}_{\boldsymbol{F}}} f_{\boldsymbol{F}}(p_{t|\boldsymbol{F}} | \boldsymbol{\theta}_{\boldsymbol{F}})\, \pi\left(\boldsymbol{\theta}_{\boldsymbol{F}} | E_{S|\boldsymbol{F}}, E_{J|\boldsymbol{F}}\right) d\boldsymbol{\theta}_{\boldsymbol{F}}}{\int_{p_{t|\boldsymbol{F}}} L_{HFE}\left(E_{S|HFE}, E_{J|HFE} | p_{t|\boldsymbol{F}}\right) \int_{\boldsymbol{\theta}_{\boldsymbol{F}}} f_{\boldsymbol{F}}(p_{t|\boldsymbol{F}} | \boldsymbol{\theta}_{\boldsymbol{F}})\, \pi\left(\boldsymbol{\theta}_{\boldsymbol{F}} | E_{S|\boldsymbol{F}}, E_{J|\boldsymbol{F}}\right) d\boldsymbol{\theta}_{\boldsymbol{F}} dp_{t|\boldsymbol{F}}} \tag{4.8}$$

where the posterior $\pi(p_{t|\boldsymbol{F}} | E_{S|HFE}, E_{J|HFE})$, referred as $P_{HFE}(HEP)$ in the remainder of this chapter to ease the notation, formally represents the HEP uncertainty distribution for the given HFE (updated after the evidence $E_{S|HFE}$ and $E_{J|HFE}$). Similarly to eq. 4.5, the likelihood term $L_{HFE}(E_{S|HFE}, E_{J|HFE} | p_{t|\boldsymbol{F}})$ in eq. 4.8 is expressed as:

$$L_{HFE}\left(E_{S|HFE}, E_{J|HFE} | p_{t|\boldsymbol{F}}\right) = L_{HFE}(k_{HFE}, \tilde{p}_{HFE} | \boldsymbol{\theta}_{\boldsymbol{F}}, N_{HFE}, \sigma_{HFE}) =$$
$$\int_{p_{t|\boldsymbol{F}}} Bin(k = k_{HFE} | p_{t|\boldsymbol{F}}, N_{HFE}) LN(\tilde{p}_{HFE} | p_{t|\boldsymbol{f}}, \sigma_{HFE}) f_{\boldsymbol{F}}(p_{t|\boldsymbol{F}} | \boldsymbol{\theta}_{\boldsymbol{F}}) dp_{t|\boldsymbol{F}} \tag{4.9}$$

The final expression for $P_{HFE}(HEP)$ can be obtained by substituting eq. 4.9 into eq. 4.8.

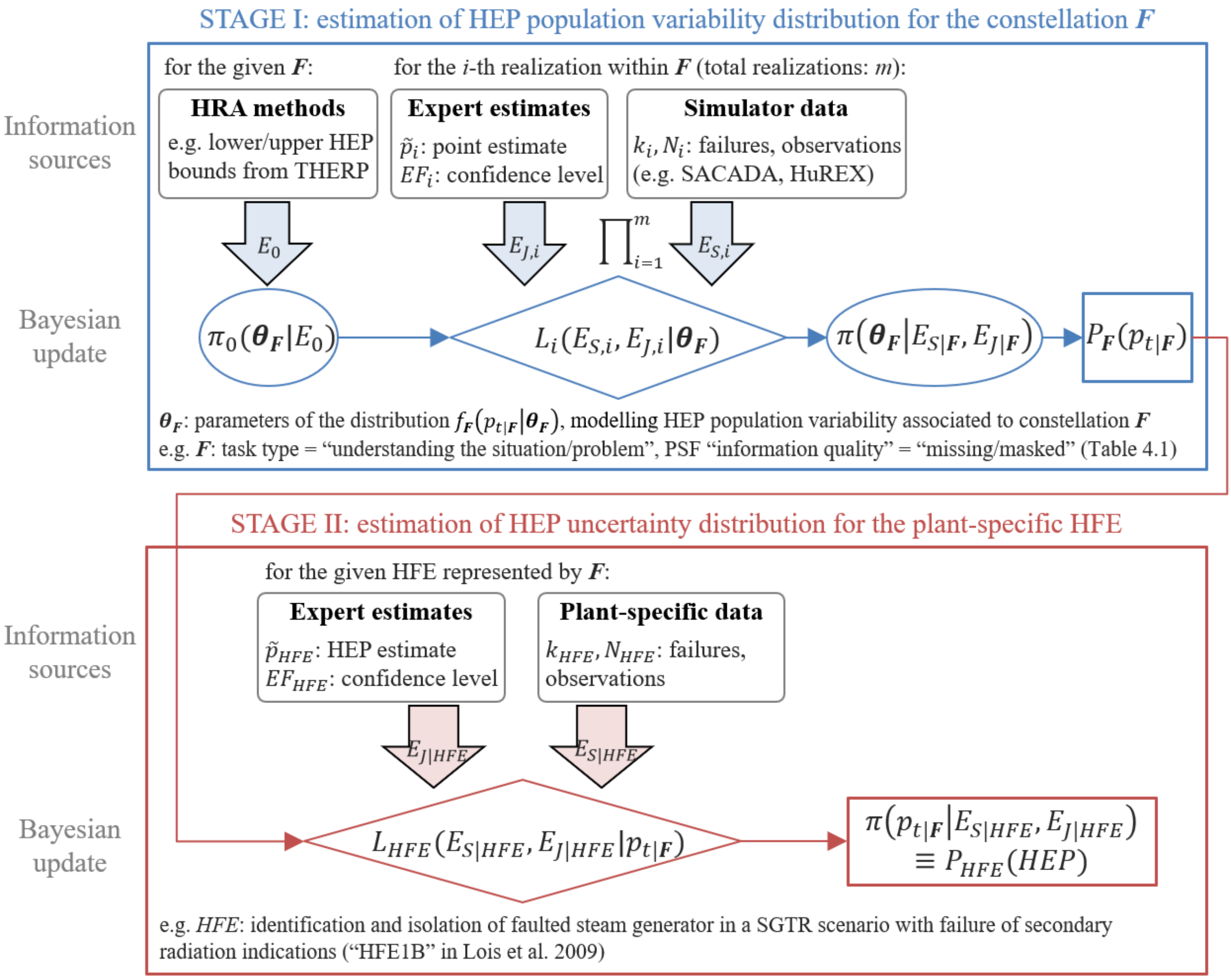

STAGE I: estimation of HEP population variability distribution for the constellation $\boldsymbol{F}$
Information sources
Bayesian update
for the given $\boldsymbol{F}$:
for the i-th realization within $\boldsymbol{F}$ (total realizations: m):
HRA methods
e.g. lower/upper HEP bounds from THERP
Expert estimates
$\tilde{p}_i$: point estimate
EF_i: confidence level
Simulator data
k_i, N_i: failures, observations
(e.g. SACADA, HuREX)
E_0
$E_{J,i}$
$\prod_{i=1}^{m}$
$E_{S,i}$
$\pi_0(\boldsymbol{\theta_F}|E_0)$
$L_i(E_{S,i}, E_{J,i}|\boldsymbol{\theta_F})$
$\pi(\boldsymbol{\theta_F}|E_{S|\boldsymbol{F}}, E_{J|\boldsymbol{F}})$
$P_{\boldsymbol{F}}(p_{t|\boldsymbol{F}})$
$\boldsymbol{\theta_F}$: parameters of the distribution $f_{\boldsymbol{F}}(p_{t|\boldsymbol{F}}|\boldsymbol{\theta_F})$, modelling HEP population variability associated to constellation $\boldsymbol{F}$
e.g. $\boldsymbol{F}$: task type = "understanding the situation/problem", PSF "information quality" = "missing/masked" (Table 4.1)
STAGE II: estimation of HEP uncertainty distribution for the plant-specific HFE
Information sources
Bayesian update
for the given HFE represented by $\boldsymbol{F}$:
Expert estimates
$\tilde{p}_{HFE}$: HEP estimate
EF_{HFE}: confidence level
Plant-specific data
k_{HFE}, N_{HFE}: failures, observations
$E_{J|HFE}$
$E_{S|HFE}$
$L_{HFE}(E_{S|HFE}, E_{J|HFE}|p_{t|\boldsymbol{F}})$
$\pi(p_{t|\boldsymbol{F}}|E_{S|HFE}, E_{J|HFE})$
$\equiv P_{HFE}(HEP)$
e.g. HFE: identification and isolation of faulted steam generator in a SGTR scenario with failure of secondary radiation indications ("HFE1B" in Lois et al. 2009)

Figure 4.3. Flowchart of the inference process in the developed Bayesian two-stage model. Stage I (blue box): estimation of the HEP population variability distribution associated to the given constellation $\boldsymbol{F}$ of task type / PSF ratings, i.e. $P_{\boldsymbol{F}}(p_{t|\boldsymbol{F}})$. Stage II (red box): estimation of the HEP uncertainty distribution for the plant-specific human failure event, i.e. $P_{HFE}(HEP)$.

4.3.2. Configuration with lognormal variability distribution and implementation

This subsection presents the Stage I configuration specifically adopted for the numerical application (Section 4.4), where the parametric variability distribution $f_{\boldsymbol{F}}(p_{t|\boldsymbol{F}}|\boldsymbol{\theta}_{\boldsymbol{F}})$ is a lognormal probability density function ($p_{t|\boldsymbol{F}} \sim \mathrm{LN}(\mu_{\boldsymbol{F}}, \sigma_{\boldsymbol{F}})$ in Figure 4.4, left). Therefore, $\boldsymbol{\theta}_{\boldsymbol{F}} = \{\mu_{\boldsymbol{F}}, \sigma_{\boldsymbol{F}}\}$ - respectively the mean and standard deviation of the HEP population variability distribution in the logarithmic space - become the Stage I parameters to be inferred from the evidence ($E_{S|\boldsymbol{F}}$, $E_{J|\boldsymbol{F}}$), eq. 4.4. The use of lognormal probability density functions as population variability distributions is a common practice in population variability analysis with Bayesian hierarchical or two-stage models for PSA applications [35, 40-44]. Alternative options for $f_{\boldsymbol{F}}(p_{t|\boldsymbol{F}}|\boldsymbol{\theta}_{\boldsymbol{F}})$, e.g. beta or logistic-normal probability density functions, can be found in the Bayesian literature [35].

In the next subsection 4.3.3, the configuration is compared against two alternative approaches for $P_{\boldsymbol{F}}(p_{t|\boldsymbol{F}})$ estimation: the Bayesian variability model presented by the authors in [44] (Figure 4.4, center), and a lumped-data model (Figure 4.4, right) as the one adopted in [25-27]. The lognormal variability model from [44] differs from the proposed Stage 1 configuration in the type of evidence processed: both data ($E_{S|\boldsymbol{F}}$) and expert estimates ($E_{J|\boldsymbol{F}}$) in the latter (referred as "Stage 1: data & estimates" in Figure 4.4, left); only data ($E_{S|\boldsymbol{F}}$) in the former (referred as "Stage 1: only data" in Figure 4.4, center). The lumped-data approach from [25-27] (referred as "Stage 1: lumped-data" in Figure 4.4, right) consists of a simple Bayesian conjugate beta-binomial model where the aggregated failure data (i.e. the total failures $k_{tot} = \sum_i k_i$ and observations $N_{tot} = \sum_i N_i$ in Table 4.1 and Figure 4.4, right) is used to infer the population-average HEP (p in Figure 4.4, right).

The Bayesian two-stage model - as well as the alternative modelling approaches for Stage 1 compared in Figure 4.4 (center and right) - are implemented in "Just Another Gibbs Sampler" (JAGS, [48]), a software using Markov Chain Monte Carlo (MCMC) simulation to approximate the solution of the posterior probability distributions for both Stage I (eq. 4.4) and Stage II (eq. 4.8). The JAGS models are run in R programming environment via the "runjags" library [49]; the R code is provided in Appendix H. Convergence of the MCMC simulations was tested using "diagMCMC", a set of diagnostic tools provided by [50]. Further information on MCMC methods can be found in the Bayesian literature [50-51].

Concerning the prior probability distributions, in all numerical cases, recommended HEP bounds for $\boldsymbol{F}$ (e.g. from existing HRA methods) are used (i.e. the E_0 in eq. 4.4 and Figure 4.3, blue box) to construct a lognormal informative prior for the logarithmic mean, i.e. $\pi_0(\mu_{\boldsymbol{F}})$. In particular, the parameters of $\pi_0(\mu_{\boldsymbol{F}})$ are obtained by fitting the 5th and 95th percentiles of the lognormal prior distribution to the lower and upper HEP bounds,. according to the following formulas:

$$\mu_{\mu_F} = \log\sqrt{HEP_{95}HEP_5}\ ;\ \sigma_{\mu_F} = 1.645^{-1}\log\sqrt{HEP_{95}/HEP_5} \tag{4.10}$$

where: *HEP*5 and *HEP*95 respectively refer to the lower and upper HEP bounds; μ_{μ_F} and σ_{μ_F} represent the parameters of the lognormal prior $\pi_0(\mu_F)$. The $\pi_0(\mu_F)$ is defined between natural log(1E-5) and 0 (corresponding to the upper limit HEP = 1), as to cover HEP values of interest for HRA applications. A diffuse prior $\pi_0(\sigma_F)$ is then set on the logarithmic standard deviation σ_F, defined between 0.01 and 5 (corresponding to error factors of 1.02 and 3733, respectively). Further information on the development of non-conjugate lognormal prior distributions from available information can be found in PSA literature [52]. For the numerical demonstrations in the next subsection 4.3.3, the same prior distributions are set on the hyperparameters of the lognormal variability model from [44] ("Stage 1: only data" in Figure 4.4, center). Concerning the conjugate beta-binomial model ("Stage 1: lumped-data" in Figure 4.4, right), a constrained non-informative (CNI) prior is used for the single-value HEP ($\pi_0(p)$ in Figure 4.4, right). The shape parameters of the CNI prior (α_0 and β_0) are derived as in [25], consistently with the prior information (i.e. *HEP*5 and *HEP*95) used to build the $\pi_0(\mu_F)$ in the variability models (Figure 4.4, left and center).

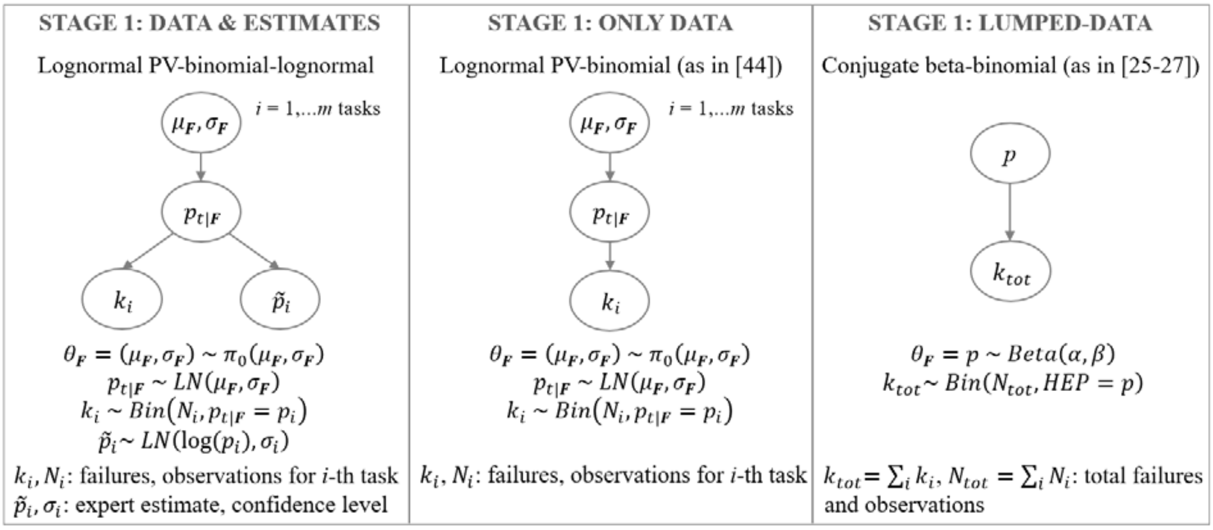

STAGE 1: DATA & ESTIMATES
Lognormal PV-binomial-lognormal
i = 1,...m tasks
μ_F, σ_F
$p_{t|F}$
k_i
$\tilde{p}_i$
$\theta_F = (\mu_F, \sigma_F) \sim \pi_0(\mu_F, \sigma_F)$
$p_{t|F} \sim LN(\mu_F, \sigma_F)$
$k_i \sim Bin(N_i, p_{t|F} = p_i)$
$\tilde{p}_i \sim LN(\log(p_i), \sigma_i)$
k_i, N_i: failures, observations for i-th task
$\tilde{p}_i, \sigma_i$: expert estimate, confidence level
STAGE 1: ONLY DATA
Lognormal PV-binomial (as in [44])
i = 1,...m tasks
μ_F, σ_F
$p_{t|F}$
k_i
$\theta_F = (\mu_F, \sigma_F) \sim \pi_0(\mu_F, \sigma_F)$
$p_{t|F} \sim LN(\mu_F, \sigma_F)$
$k_i \sim Bin(N_i, p_{t|F} = p_i)$
k_i, N_i: failures, observations for i-th task
STAGE 1: LUMPED-DATA
Conjugate beta-binomial (as in [25-27])
p
k_{tot}
$\theta_F = p \sim Beta(\alpha, \beta)$
$k_{tot} \sim Bin(N_{tot}, HEP = p)$
$k_{tot} = \sum_i k_i$, $N_{tot} = \sum_i N_i$: total failures and observations

Figure 4.4. Left: Stage 1 configuration with lognormal population variability distribution (PV: population variability). Centre and right: alternative modelling approaches for Stage 1 tested in subsection 4.3.3, respectively based on the lognormal variability model for simulator data presented in [44] and the lumped-data approach as in [25-27].

4.3.3. Stage 1: verification with artificial data and sensitivity analysis

With reference to a generic constellation $\boldsymbol{F}$, both data ($E_{S|\boldsymbol{F}}$) and expert estimates ($E_{J|\boldsymbol{F}}$) are artificially generated with known characteristics (e.g. median and percentiles) of the underlying HEP population variability distribution, in order to verify model behavior against a known distribution. In particular, the target HEP population variability distribution for all numerical tests (subsections 4.3.3.1-4.3.3.2) is lognormal, with median = 5e-02, mean = 5.46e-02, and error factor = 2. The considered case represents a failure probability range of interest for practical HRA applications, with moderately high HEP values; cases with lower HEP values (e.g. with median in the range 1e-03÷1e-04) are not addressed in this chapter.

The evidence $E_{S|\boldsymbol{F}}$ and $E_{J|\boldsymbol{F}}$ is generated by first sampling possible $p_{t|\boldsymbol{F}}$ values from the target distribution, each representing the specific failure probability value of an hypothetical task/context realization relevant to the considered $\boldsymbol{F}$. Then, for the i-th realization, the sampled $p_{t|\boldsymbol{F}}$ value (i.e. p_i) is used to generate the number of observed failures k_i (sampling from a Binomial distribution on N_i trials) as well as the corresponding expert estimate $\tilde{p}_i$ (sampling from the lognormal error model of eq. 4.2bis with the given confidence level, i.e. EF_i). An example of artificially-generated dataset is given in Table 4.2, where the different pairs $E_{S,i} = (k_i, N_i)$ and $E_{J,i} = (\tilde{p}_i, EF_i)$ - i.e. the rows of Table 4.2 - are obtained from different $p_{t|\boldsymbol{F}}$ values, according to the total number of task/context realizations of $\boldsymbol{F}$ for which evidence is assumed to be collected (e.g. ten realizations in Table 4.2). The amount of trials pertaining to each realization (i.e. the N_i's in Table 4.2) is randomly assigned as to reflect realistic sets of crew observations gathered from each plant simulator. Note that, for each realization, we consider that both data (k_i, N_i) and expert estimate $(\tilde{p}_i)$ are available and assign a relatively low confidence level (EF_i = 5) to each $\tilde{p}_i$.

Table 4.2. Example of artificially-generated dataset, containing evidence hypothetically collected for ten task/context realizations (index i) relevant to the generic constellation $\boldsymbol{F}$. Sampled from a lognormal HEP population variability distribution with median = 5e-02, mean = 5.46e-02, and error factor = 2.

Simulator data ($E_{S\|\boldsymbol{F}}$)		Expert estimates ($E_{J\|\boldsymbol{F}}$)	
k_i	N_i	$\tilde{p}_i$	EF_i
0	5	2.70e-02	5
0	7	6.80e-02	5
0	4	8.60e-03	5
0	2	8.50e-03	5
0	3	4.90e-02	5
0	8	4.60e-02	5
0	4	7.30e-02	5
0	10	4.90e-03	5
1	3	8.90e-03	5
2	8	7.00e-01	5

For the purposes of the numerical demonstration, the parameters of the lognormal prior $\pi_0(\mu_{\boldsymbol{F}})$ are obtained from eq. 4.10 assuming HEP_5 = 5e-03 and HEP_{95} = 5e-01 as recommended HEP bounds for $\boldsymbol{F}$, resulting in a $\pi_0(\mu_{\boldsymbol{F}})$ informed around the target median HEP value (i.e. 5e-02). Note that model sensitivity to alternative choices of priors for the hyper-parameters of the HEP variability model, i.e. $\pi_0(\mu_{\boldsymbol{F}})$ and $\pi_0(\sigma_{\boldsymbol{F}})$, has already been investigated by the authors in [44] and therefore is not further discussed in the present chapter. Consistently with the specific $\pi_0(\mu_{\boldsymbol{F}})$ adopted for both population variability formulations (Figure 4.4, left and center), the conjugate beta-binomial model with lumped data (Figure 4.4, right) is set with a CNI prior $\pi_0(p)$ with shape parameters α_0 = 0.50 and β_0 = 3.25.

In the remainder of this section, subsection 4.3.3.1 numerically demonstrates the effects of expert judgment incorporation on the expected HEP population variability distribution, i.e. $P_{\boldsymbol{F}}(p_{t|\boldsymbol{F}})$. Then, subsection 4.3.3.2 analyses model sensitivity to sample size, with the goal to investigate the influence of judgment incorporation on data requirements. Lastly, subsection 4.3.3.3 discusses model sensitivity to biased expert estimates, showing the use of Bayesian p-values [50-51] to spot potential biases in the $\tilde{p}_i$'s provided by the experts and support the HRA analyst in selecting an appropriate confidence level.

4.3.3.1. Effects of judgment incorporation on the estimated $P_{\boldsymbol{F}}(p_{t|\boldsymbol{F}})$

This subsection discusses the numerical differences in the expected $P_{\boldsymbol{F}}(p_{t|\boldsymbol{F}})$ among the proposed HEP population variability model ("Stage 1: data and judgment" in Figure 4.4, left) and the alternative modelling approaches presented in Figure 4.4: the lognormal variability model from [44] ("Stage 1: only data" in Figure 4.4, center) and a beta-binomial model representative of the lumped-data approach ("Stage 1: lumped data" in Figure 4.4, right). Figure 4.5 and Table 4.3 compare the expected mean, 5th and 95th percentiles of the corresponding $P_{\boldsymbol{F}}(p_{t|\boldsymbol{F}})$. Different sample sizes are considered (Figure 4.5, x-axis), to test on progressively increasing numbers of task/context realizations of $\boldsymbol{F}$ (referred as "tasks" in Figure 4.5 and Table 4.3, for simplicity): from relatively small datasets (10÷20 tasks) to larger datasets hypothetically accumulated in the long-term (50÷100 tasks).

Comparing the results from the two variability models ("Stage 1: only data" and "Stage 1: data and judgment"), the incorporation of expert estimates in Stage 1 allows to obtain a better approximation of the target HEP population variability distribution (i.e. lognormal with mean HEP = 5.46e-02 and error factor = 2) in presence of limited datasets (e.g. in Table 4.3, datasets: 10, 20 tasks). Already at 20 task realizations, the model combining data and expert estimates returns a value of expected error factor close to seven, about seven times smaller than the case when only data is used (i.e. 6.9 vs 46 in Table 4.3). The smaller error factors are not surprising, given the different informative power of the evidence between the models. Indeed, expert estimates bring additional information on the unknown $p_{t|\boldsymbol{F}}$ values of the corresponding task/context realizations, compensating the scarce empirical data available for each realization (e.g. in Table 4.2: few observations per task, very few observed failures). Since large values of error factor are not of practical use for PSA applications, the incorporation of expert estimates has therefore important implications on data requirements to inform HEP population variability associated to a given constellation $\boldsymbol{F}$: this aspect is the focus of the next subsection 4.3.3.2. Note that the confidence level (EF_i) assigned to the experts determines the informative power of their estimates ($\tilde{p}_i$'s). Recalling from subsection 4.2.2, values of EF_i greater than one imply considering experts as "imperfect", according to which the provided $\tilde{p}_i$'s are treated as uncertain evidence in the Bayesian update process. The effects of this additional source of uncertainty on model results emerge with increasing sample size: this explains why, for the case with both data and judgment (with relatively low confidence level, i.e. $EF_i = 5$), the numerical value of expected error factor does not significantly decrease between 50 and 100 task realizations (Figure 4.5 and Table 4.3). This aspect is further discussed in the sensitivity analysis in subsection 4.3.3.2.

Compared to the variability model ("Stage 1: only data"), the $P_{\boldsymbol{F}}(p_{t|\boldsymbol{F}})$'s estimated by the lumped-data model ("Stage 1: lumped data") tend to shrink around the population-average HEP (see the smaller values of error factor in Table 4.3). As discussed in [44], in the lumped-data approach, population variability across the different task/context realizations of $\boldsymbol{F}$ is averaged in the single piece of evidence (e.g. $k_{tot}/N_{tot} = 33/541$, for 100 realizations in Table 4.3). This reflects, with increasing evidence (e.g. 100 realizations in Figure 4.5 and Table 4.3), in an overly-narrow $P_{\boldsymbol{F}}(p_{t|\boldsymbol{F}})$ with respect to the target HEP population variability distribution, with values of expected error factor significantly smaller than the target one (e.g. 1.3 vs 2 at 100 realizations, Table 4.3). The tendency of the lumped-data approach to return overconfident $P_{\boldsymbol{F}}(p_{t|\boldsymbol{F}})$ estimates is further analyzed in [44].

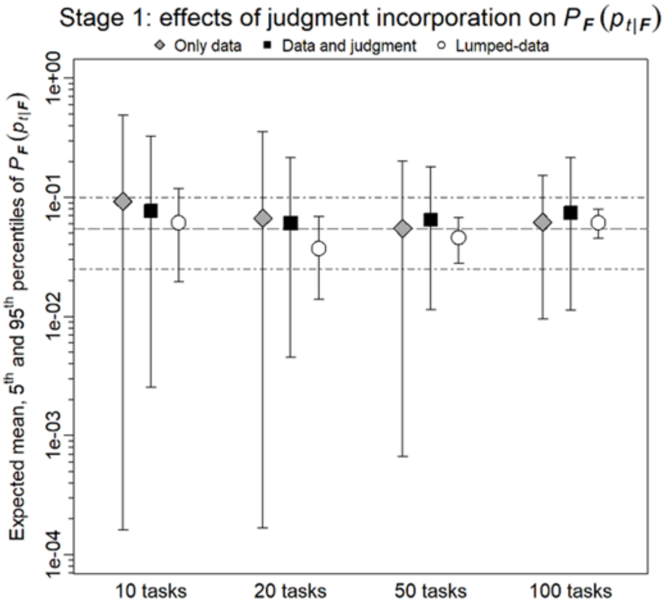


Figure 4.5. In y axis (logarithmic scale): expected mean (filled symbols), 5^{th} and 95^{th} percentiles (whiskers) of $P_F(p_{t|F})$ returned by three Stage 1 formulations in Figure 4.4, tested against the same datasets (in x axis, number of task realizations: 10, 20, 50, 100). Datasets generated from a known lognormal variability distribution with mean = 5.46e-02 (dashed line) and error factor = 2 (dot-dashed lines at 5th percentile = 2.5e-02 and 95th percentile = 1.0e-01).

Table 4.3. Numerical results from Figure 4.5.

Dataset	Modelling approach for Stage 1	Mean	5th	50th	95th	EF
10 tasks:	Only data (Figure 4.4, center)	9.30e-02	1.62e-04	2.53e-02	4.91e-01	55.1
3 failures,	Data and judgment (Figure 4.4, left)	7.70e-02	2.55e-03	3.26e-02	3.25e-01	11.3
54 observations	Lumped-data (Figure 4.4, right)	6.10e-02	1.95e-02	5.57e-02	1.19e-01	2.5
20 tasks:	Only data (Figure 4.4, center)	6.70e-02	1.68e-04	1.88e-02	3.55e-01	46.0
4 failures,	Data and judgment (Figure 4.4, left)	6.10e-02	4.53e-03	3.20e-02	2.15e-01	6.9
117 observations	Lumped-data (Figure 4.4, right)	3.70e-02	1.39e-02	3.47e-02	6.92e-02	2.2
50 tasks:	Only data (Figure 4.4, center)	5.50e-02	6.65e-04	2.78e-02	2.01e-01	17.4
13 failures,	Data and judgment (Figure 4.4, left)	6.50e-02	1.14e-02	4.59e-02	1.80e-01	4.0
290 observations	Lumped-data (Figure 4.4, right)	4.60e-02	2.78e-02	4.50e-02	6.76e-02	1.6
100 tasks:	Only data (Figure 4.4, center)	6.20e-02	9.52e-03	4.98e-02	1.52e-01	4.0
33 failures,	Data and judgment (Figure 4.4, left)	7.40e-02	1.13e-02	4.99e-02	2.16e-01	4.4
541 observations	Lumped-data (Figure 4.4, right)	6.10e-02	4.55e-02	6.10e-02	7.91e-02	1.3

4.3.3.2. Sensitivity to available evidence

The goal of this sensitivity analysis is to investigate how effectively the incorporation of expert estimates ($E_{J|F}$) can reduce the amount of data ($E_{S|F}$) required in Stage I such that the estimated $P_F(p_{t|F})$'s are of practical use for PSA (i.e. the uncertainty on $P_F(p_{t|F})$ is not too large). This is an important aspect: since simulator data collection is a resource-intensive, long-term process, for some constellations of task/PSF categories the amount of evidence $E_{S|F}$ currently available can be indeed limited. To this end, the subsection compares the $P_F(p_{t|F})$'s yielded by the lognormal population variability models ("Stage 1: data and judgment" and "Stage 1:

only data”, in Figure 4.4 left and center) with increasing sample sizes, following the convergence of the expected statistics of $P_{\boldsymbol{F}}(p_{t|\boldsymbol{F}})$ to the corresponding statistics of the target distribution (i.e. median HEP = 5.0e-02, error factor = 2). In particular, the focus is on the sample size required by the models to obtain (on average) values of error factor compatible with typical HRA applications, e.g. values close to 5.

Figure 4.6 shows the expected error factor (left) and median (right) of $P_{\boldsymbol{F}}(p_{t|\boldsymbol{F}})$ returned by the variability models, as a function of the sample size (from 5 to 100 task/context realizations of $\boldsymbol{F}$). For each sample size, 100 datasets are generated via Monte Carlo sampling so that the quantitative indications obtained by the analysis are as independent as possible from the specific dataset. Error bars are used in Figure 4.6 to represent the spread of the estimates across the sampled datasets, with boxes and whiskers respectively corresponding to the 50% (25th - 75th percentiles) and the 90% (5th - 95th percentile) confidence intervals.

The results in Figure 4.6 indicate that, for both variability models, the expected error factor and median of $P_{\boldsymbol{F}}(p_{t|\boldsymbol{F}})$ across the 100 datasets tend to the target statistics with increasing sample size. Concerning the expected median (Figure 4.6, right), its value gets close to the target one at about 60 task realizations when only data is used, with an average value of 4.5e-02 across the datasets (50% interval: 3.4e-02, 5.4e-02). On the other hand, when data is combined with expert estimates, five realizations are already sufficient to reach an average expected median of 4.8e-02 (with 50% interval: 2.9e-02, 6.0e-02). Concerning the expected error factor (Figure 4.6, left), 60 task realizations are still required by the variability model using only data to obtain an average value close to 5, i.e. 5.6 (50% interval: 2.8, 5.8). When combining both data and expert estimates, the expected error factor gets close to 5 at approximately 20 task realizations, with an average value of 6.3 (50% interval: 3.5, 8.1).

To conclude, the analysis demonstrates that, for constellations $\boldsymbol{F}$ characterized by moderately high HEP values (i.e. around 0.01), the incorporation of expert estimates in Stage 1 allows to produce HEP population variability distributions (i.e. $P_{\boldsymbol{F}}(p_{t|\boldsymbol{F}})$) of practical use for PSA applications with significantly lower requirements of empirical data. In particular, assigning a relatively low confidence level (EF_i= 5) to expert estimates, the variability model observed a reduction of data requirements of approximately a factor of three (in Figure 4.6, left: 60 vs 20 task realizations to reach error factors close to 5). Tests performed with lower HEP values (e.g. in the range of HEP ~ 0.001 and below) confirmed a similar trend, with expectedly more pronounced benefits on data requirements. As discussed in [44], when lower HEP values are involved, empirical data becomes indeed less informative (i.e. fewer failures are observed): in such case, Stage I without expert estimates would require very large datasets (i.e. with few hundred data points [44]).

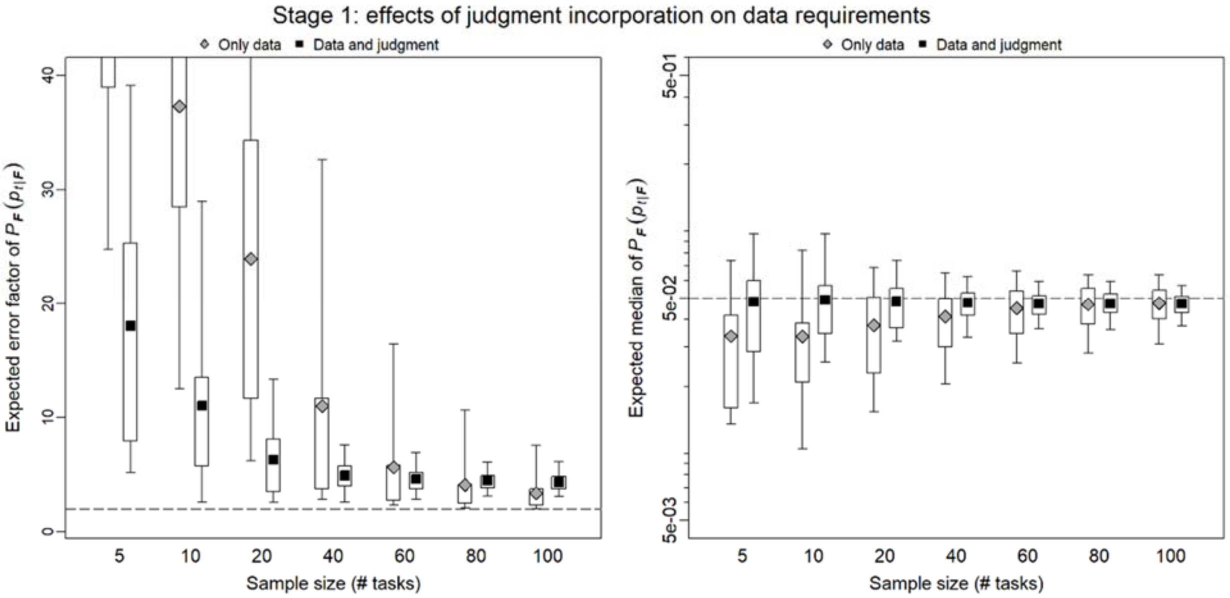


Figure 4.6. Data requirements of the lognormal population variability models ("Stage 1: only data" and "Stage 1: data and judgment", in Figure 4.4 left and center). For each sample size in x axis, 100 datasets are Monte Carlo-sampled from the target HEP variability distribution with median = 5e-02 and error factor = 2 (dashed lines). From left to right: expected error factor and median (log-scale) of $P_F(p_{t|F})$ returned by the models, in the form of 50% and 90% confidence intervals (filled symbols: average value of the 100 datasets).

Lastly, Figure 4.6 (left) shows also the influence of the relatively low confidence level (i.e. EF_i = 5) on the convergence of the expected error factor with increasing availability of evidence. If, on the one hand, the more data ($E_{S|F}$) is collected, the more the epistemic uncertainty of HEP population variability (i.e. the target error factor) is reduced, on the other hand the uncertainty associated to expert estimates ($E_{J|F}$) still remains. The effects of this residual uncertainty source on the expected error factor numerically emerge with larger sample sizes (see for instance the datasets with 80÷100 task realizations in Figure 4.6, left). While for small sample sizes the expected EF for the case of data only is larger than for the case of data and judgment, for larger sample sizes the situation is inverted, with the data and judgment EFs levelled to values larger than for the data only case. Note however that, in presence of such large datasets, the comparison between the variability models is not realistic: indeed, if empirical data is available for e.g. 80÷100 task realizations of F, this should be already sufficient to derive statistically-significant estimates of $P_F(p_{t|F})$, hence the incorporation of expert estimates in Stage 1 would not be required.

4.3.3.3. Sensitivity to biases in expert estimates and to the confidence level

The numerical tests in subsections 4.3.3.1-4.3.3.2 consider the estimates $\tilde{p}_i$'s are provided from unbiased experts. In real applications, it becomes important to investigate the effects of potential biases in expert estimates on the $P_{\boldsymbol{F}}(p_{t|\boldsymbol{F}})$ returned by Stage 1: this is the goal of the sensitivity analysis presented in this subsection.

Biased estimates are generated by adding a multiplicative factor (the bias factor b) to the sampling process of $\tilde{p}_i$ (eq. 4.2bis), as follows: $\tilde{p}_i \sim LN(\log(b \cdot p_{t|\boldsymbol{f}}), 1.645^{-1}\log(EF_i))$. The factor b is intended to simulate a conservative (for $b > 1$) or optimistic (for $b < 1$) bias in expert assessments, with respect to the actual PSF effects on task failure probability (i.e. on the specific $p_{t|\boldsymbol{F}}$ values). The following three cases are considered:

- "Unbiased" experts ($b = 1$): the provided $\tilde{p}_i$'s are not affected by any bias (same as for subsections 4.3.3.1-4.3.3.2);
- "Conservative" experts ($b = 10$): the provided $\tilde{p}_i$'s are overall shifted towards HEP values one order of magnitude above the actual $p_{t|\boldsymbol{f}}$'s;
- "Optimistic" experts ($b = 0.1$): the provided $\tilde{p}_i$'s are overall shifted towards HEP values one order of magnitude below the actual $p_{t|\boldsymbol{f}}$'s.

Figure 4.7 compares the expected mean, 5th and 95th percentiles of the $P_{\boldsymbol{F}}(p_{t|\boldsymbol{F}})$ returned by Stage 1 in the three cases, for 10 and 50 task realizations. As in subsections 4.3.3.1-4.3.3.2, a relatively low confidence level ($EF_i = 5$) is assigned to each $\tilde{p}_i$ (hence, the results for "unbiased" in Figure 4.7 corresponds to those in Figure 4.5, at 10 and 50 tasks). Numerical values are summarized in Table G.1 (Appendix G).

Compared to the base case ("Unbiased"), the use of biased expert estimates results in a sensible overestimation ("Conservative") or underestimation ("Optimistic") of the expected mean, with the bias effects tending to increase with the amount of evidence $E_{J|\boldsymbol{F}}$. At 50 tasks, the case with "conservative" experts returned an expected mean value of 1.7e-01, i.e. approximately 3 times higher than the target value (i.e. 5.46e-02); with "optimistic" experts, the expected mean value is 1.4e-02, i.e. approximately 4 times lower than the target one. Biased $\tilde{p}_i$'s also affect the expected error factor returned by the model. As more $E_{J|\boldsymbol{F}}$ becomes available, the more the $P_{\boldsymbol{F}}(p_{t|\boldsymbol{F}})$'s in Figure 4.7 narrows towards their conservative or optimistic mean value, with an error factor that depends on the extent to which the type of bias is compatible with the characteristics of the empirical data ($E_{S|\boldsymbol{F}}$) at hand. This explains why, for the datasets considered in this analysis (e.g. from Table 4.2: 0 failures on 7 observations; 0 on 5; and the like), the $P_{\boldsymbol{F}}(p_{t|\boldsymbol{F}})$'s informed by "conservative" experts present smaller error factors (e.g. at 50 tasks: 1.4), compared to the case with "optimistic" experts (e.g. at 50 tasks: 7.4).

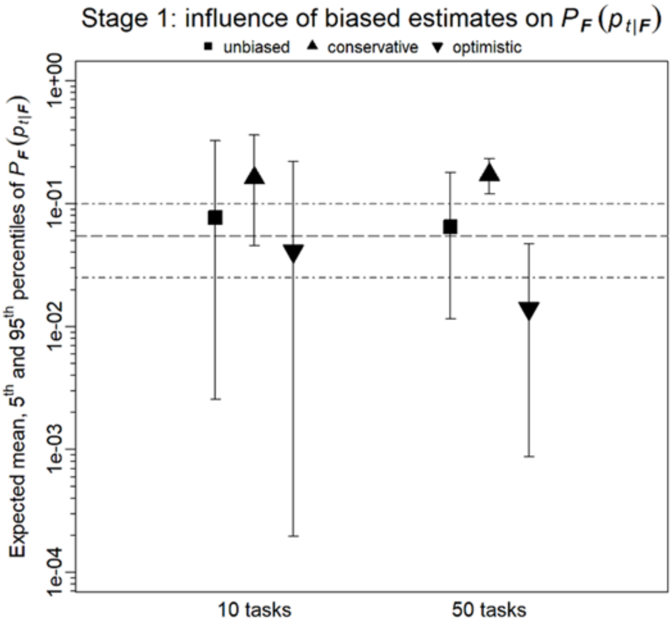


Figure 4.7. Sensitivity of the lognormal PV-binomial-lognormal model (Figure 4.4, left) to biases in expert estimates (results for "unbiased" case are taken from Figure 4.5). In y axis (logarithmic scale): expected mean (filled symbols), 5th and 95th percentiles (whiskers) of $P_F(p_{t|F})$ returned by the model for three different bias cases ("unbiased", "conservative", "optimistic"), at 10 and 50 task realizations (x axis). Target HEP variability distribution: lognormal with mean = 5.46e-02 (dashed line) and error factor = 2 (dot-dashed lines at 5th percentile = 2.5e-02 and 95th percentile = 1.0e-01).

An important aspect to consider is that the specific confidence level (EF_i) chosen by the analyst can influence model sensitivity to expert biases, amplifying or mitigating their effects on the expected $P_F(p_{t|F})$. To complement this sensitivity analysis, additional tests with different confidence levels (e.g. moderate confidence: EF_i = 3; low confidence: EF_i = 7) have been performed: the numerical results can be found in Table G.1. As general rule, the higher the confidence on the experts (i.e. the smaller the values of EF_i assigned to their $\tilde{p}_i$'s), the more relevant is the weight of the evidence $E_{J|F}$ in the Bayesian update process: consequently, the effects of expert biases on $P_F(p_{t|F})$ tend to be amplified (e.g. in Table G.1, compare the results for "conservative" and "optimistic" experts with EF_i = 5 vs EF_i = 3). On the contrary, the lower the confidence on the experts (i.e. the larger the values of EF_i), the larger the uncertainty associated to $E_{J|F}$: in such case, the effects of expert biases on $P_F(p_{t|F})$ tend to be mitigated (e.g. in Table G.1, compare the results with EF_i = 5 versus EF_i = 7). Naturally, the downside of selecting larger values of EF_i is that, with a weakly-informative $E_{J|F}$, Stage 1 requires more empirical data (i.e. more evidence $E_{S|F}$) to estimate $P_F(p_{t|F})$ with error factors of practical use (see subsection 4.3.3.2).

Bayesian model checking techniques [50-51] could be adopted to identify *a priori* potential biases in the provided $\tilde{p}_i$'s, in order to support the HRA analyst in selecting appropriate confidence levels for the application at hand. In the following, we show how a basic predictive check with Bayesian p-values [50-51] allows to verify the extent to which the set of expert estimates ($E_{J|\boldsymbol{F}}$) is consistent with the empirical data ($E_{S|\boldsymbol{F}}$) collected for the given constellation $\boldsymbol{F}$. The verification consists of the following steps:

I. Run Stage I with only $E_{J|\boldsymbol{F}}$ as evidence, assigning EF_i = 1 to each estimate $\tilde{p}_i$. In such configuration, the model returns a $P_{\boldsymbol{F}}(p_{t|\boldsymbol{F}})$ exclusively informed by experts, with the maximum confidence level possible;

II. From the expert-informed $P_{\boldsymbol{F}}(p_{t|\boldsymbol{F}})$, draw R replicated datasets $\{E^r_{S|\boldsymbol{F}} = (k^r_i, N_i), i = 1, \dots m; r = 1, \dots R\}$ with the same characteristics as the actual data $E_{S|\boldsymbol{F}}$ (i.e. same number m of task/context realizations; same number of observations N_i per each realization). The notation r denotes the index of the replicated dataset;

III. For each $E^r_{S|\boldsymbol{F}}$, calculate test quantities - namely, $T(E^r_{S|\boldsymbol{F}})$ - to measure the discrepancy between the replicated datasets and $E_{S|\boldsymbol{F}}$. In this test, we use the mean value of the replicated failures as test quantity (namely, $\bar{k}^r$), obtained by averaging the k^r_i's over the m task realizations, i.e. for the r-th dataset: $T(E^r_{S|\boldsymbol{F}}) = \bar{k}^r = (\sum_{i=1}^{m} k^r_i)/m$;

IV. Finally, compute the Bayesian p-value (namely, p_B) for the given test quantity, as the probability: $p_B = P(T(E^r_{S|\boldsymbol{F}}) \geq T(E_{S|\boldsymbol{F}}))$. In practical terms, p_B expresses the proportion of replicated $E^r_{S|\boldsymbol{F}}$'s for which the test quantity (i.e. $T(E^r_{S|\boldsymbol{F}})$) equals or exceeds the corresponding quantity of $E_{S|\boldsymbol{F}}$ (i.e. $T(E_{S|\boldsymbol{F}})$). When the test quantity is $\bar{k}^r$, the expression becomes: $p_B = P(\bar{k}^r \geq \bar{k})$, where $\bar{k}$ is the mean value of the observed failures, i.e. $T(E_{S|\boldsymbol{F}}) = \bar{k} = (\sum_{i=1}^{m} k_i)/m$.

The concept behind is that, if the estimates $\tilde{p}_i$'s are overall unbiased with respect to the actual population of task failure probability values (i.e. the $p_{t|\boldsymbol{F}}$'s) within $\boldsymbol{F}$, the replicated datasets $E^r_{S|\boldsymbol{F}}$'s should then look similar to the observed data $E_{S|\boldsymbol{F}}$. According to this interpretation, Bayesian p-values around 0.5 indicate an overall consistency between expert estimates and empirical data, hence providing no clear evidence of biases. On the contrary, the closer the p-values get to 0 or 1, the more likely is the presence of biases amongst the experts. As an example, for the case "10 tasks" in Figure 4.7, the multi-step verification returns p_B = 0.32 when the model is informed by "unbiased experts", and p_B = 1 and p_B = 0 when informed respectively by "conservative" and "optimistic" experts. Such extreme Bayesian p-values for both "conservative" and "optimistic" cases are due to the strong bias factors (i.e. b = 10 and b = 0.1, respectively) assumed for the sensitivity analysis in Figure 4.7. Nevertheless, the verification effectively provides indications of bias also with smaller bias factors, e.g.: with b = 2, p_B = 0.82; with b = 0.5, p_B = 0.08. The Bayesian p-values in the examples are computed drawing $R = 10^5$ replicated datasets from the posterior predictive distribution of Stage I: the code developed for the predictive check is provided in Appendix H. Further information on the use of Bayesian p-values can be found in [50-51]. Formal methods for the explicit treatment of expert bias in population variability analysis are available in PSA literature [39-40].

To sum up, biases in expert estimates can lead to a significant overestimation or underestimation of the expected value of $P_{\boldsymbol{F}}(p_{t|\boldsymbol{F}})$, i.e. of the population-average HEP associated to the given constellation $\boldsymbol{F}$. Predictive checks with Bayesian p-values proved to be effective in diagnosing possible biases amongst the experts; also, the computed p-values can provide recommendations on which confidence levels to assign in order to mitigate bias effects on the estimated $P_{\boldsymbol{F}}(p_{t|\boldsymbol{F}})$. For instance, with p-values below 0.2 or above 0.8, low confidence (e.g. EF_i = 5÷7) may be advisable.

4.4 Application to case study

The developed Bayesian two-stage model is applied to literature data to demonstrate its feasibility for the quantification of HEPs for plant-specific human failure events. Subsection 4.4.1 describes the set of HFEs selected for the case study, as well as the literature sources [53-57]. Then, subsection 4.4.2 presents the numerical results.

4.4.1. Case study: set of HFEs and evidence from literature

The authors selected 16 HFEs from the recent HRA Empirical Studies (the US [54] and the International [55-57]), involving operating crew tasks at nuclear power plant simulators. The selected HFEs (listed in Table G.2 left, Appendix G) are representative of different task types and operational contexts, spanning from routine tasks in normally-trained scenarios (e.g. standard SGTR) to more challenging tasks in scenarios characterized by conflicting or masked cues (e.g. variants of a SGTR with multiple, concurrent system malfunctions). Task types and PSF ratings from SACADA taxonomy [24] were adopted to categorize task and context characteristics of each HFE. Accordingly, the 16 HFEs were identified as belonging to 13 different combinations (i.e. constellations $\boldsymbol{F}$'s) of task type and PSF ratings: the associated $\boldsymbol{F}$'s are reported in Table G.2. The selection of task type and PSF ratings for each HFE was performed by the authors of the present work, based on the information available in [53-57].

The evidence entering the two stages of the Bayesian model is reported in Table 4.4, and consists of empirical data and expert estimates processed from the following literature sources:

- $E_{S|\boldsymbol{F}}$ (Stage 1): failure data relevant to the identified constellations $\boldsymbol{F}$'s, extrapolated from the SACADA database[4] [58] (Table 4.4, second column). Note that no $E_{J|\boldsymbol{F}}$ (Stage

---

[4] As at September 2018 [28], the SACADA database counts more than 25000 data points distributed across few

1) was available for the present application;

- $E_{S|HFE}$ (Stage 2): failure data for the set of HFEs, gathered from crew performances at the HAMMLAB plant simulator [53-57] (Table 4.4, fourth column);
- $E_{J|HFE}$ (Stage 2): HFE probability estimates derived from the expert-based HFE difficulty rankings in [53-57] (Table 4.4, last column).

According to the data aggregation framework provided in Figure 4.3, $E_{S|\boldsymbol{F}}$ is used in Stage I to construct HEP population variability distributions, i.e. $P_{\boldsymbol{F}}(p_{t|\boldsymbol{F}})$'s, for the identified constellations $\boldsymbol{F}$'s (Table G.2, right). The estimated $P_{\boldsymbol{F}}(p_{t|\boldsymbol{F}})$'s are then updated by the plant-specific evidence ($E_{S|HFE}$ and $E_{J|HFE}$) in Stage II to quantify HEP uncertainty distributions, i.e. $P_{HFE}(HEP)$'s, for the associated HFEs (Table G.2, left).

Concerning $E_{S|\boldsymbol{F}}$, at the time of this analysis, simulator records relevant to 5 out of 13 constellations (i.e. $\boldsymbol{F}_3$, $\boldsymbol{F}_5$, $\boldsymbol{F}_9$, $\boldsymbol{F}_{12}$, and $\boldsymbol{F}_{13}$ in Table G.2, right) were not available in the SACADA database; therefore, the corresponding HFEs (i.e. US-HFE2A, INT-SGTR-HFE1B, INT-SGTR-HFE5B1, INT-LOFW-HFE1B, and INT-LOFW-HFE2B in Table G.2, left) have been excluded from the HEP quantification in subsection 4.4.2 (data availability aspects are further discussed in Section 4.5). For each of the remaining constellations (i.e. $\boldsymbol{F}_1$, $\boldsymbol{F}_2$, $\boldsymbol{F}_4$, $\boldsymbol{F}_6$, $\boldsymbol{F}_7$, $\boldsymbol{F}_8$, $\boldsymbol{F}_{10}$, and $\boldsymbol{F}_{11}$ in Table G.2, right), the SACADA database provided only aggregated data, i.e. in the form of total number of failures and crew observations (i.e. k_{tot} and N_{tot} in Table 4.4, left) collected over different tasks and plants. Similarly to example provided in Table 4.1, the aggregated pairs (k_{tot}, N_{tot}) were arbitrarily distributed across hypothetical task/context realizations (Table 4.4, second column), as to replicate realistic data collection conditions for each $\boldsymbol{F}$: e.g. in Table 4.4, four realizations are assumed for constellation $\boldsymbol{F}_1$, with k_1/N_1 = 0/5, k_2/N_2 = 0/5, k_3/N_3 = 0/5, and k_4/N_4 = 0/1. Obviously, the specific sets of k_i/N_i assumed for each $\boldsymbol{F}$ influence the $P_{\boldsymbol{F}}(p_{t|\boldsymbol{F}})$'s returned by Stage I. However, it is important to highlight that the focus of the present application is not on the specific numeric results, rather on providing a practical demonstration of the use of HEP population variability distributions to support HEP estimation of plant-specific HFEs.

Concerning $E_{J|HFE}$, the qualitative HFE difficulty rankings provided by domain experts in the HRA Empirical Studies [53-57] (e.g. "easy", "somewhat difficult" in Table 4.4, last column) were converted into HFE probability estimates according to the scaling guidance reported in Table 4.5 (adapted from the qualitative likelihood scale suggested in [59] for HEP elicitation). Similarly to the numerical tests in subsection 4.3.3, a relatively low confidence level (i.e. EF_i = 5) was assigned to each expert estimate in Table 4.4.

---

hundred constellations of task and PSF categories (a portion of the database is publicly available at the US Nuclear Regulatory Commission website [58]).

Table 4.4. Datasets for Stage 1 (left) and Stage 2 (right) used in the case study. For $E_{J|HFE}$, EF_i = 5 is assigned to each expert estimate.

Stage 1: estimation of HEP population variability distribution, $P_{\boldsymbol{F}}(p_{t|\boldsymbol{F}})$		Stage 2: estimation of HEP uncertainty distribution for the plant-specific HFE, $P_{HFE}(HEP)$		
Constellation	$E_{S|\boldsymbol{F}}$: failure data (k_i/N_i) extrapolated from [58]	HFE	$E_{S|HFE}$: failure data (k_{HFE}/N_{HFE}) from [54-57])	$E_{J|HFE}$: expert estimates derived from the HFE difficulty scale in [54-57]
$\boldsymbol{F_1}$	0/5, 0/5, 0/5, 0/1 (k_{tot} = 0, N_{tot} = 16)	US-HFE1A	0/4	3.20e-02 (“Fairly difficult/difficult”)
$\boldsymbol{F_2}$	2/5, 2/5, 1/5, 1/5, 1/5, 1/5, 0/1 (k_{tot} = 8, N_{tot} = 31)	US-HFE1C	1/4	1.00e-01 (“Difficult”)
$\boldsymbol{F_4}$	0/5, 0/5, 0/5, 0/5, 0/5, 0/5, 0/5, 0/5, 0/5, 0/5, 0/5, 0/4 (k_{tot} = 0, N_{tot} = 59)	US-HFE3A INT-SGTR-HFE1A	0/3 1/14	1.00e-03 (“Easy”) 3.20e-03 (“Easy/somewhat difficult”)
$\boldsymbol{F_6}$	0/5, 0/5, 0/4 (k_{tot} = 0, N_{tot} = 14)	INT-SGTR-HFE2A INT-SGTR-HFE3A INT-SGTR-HFE3B	1/14 1/14 2/14	3.20e-03 (“Easy/somewhat difficult”) 1.00e-02 (“Somewhat difficult”) 1.00e-02 (“Somewhat difficult”)
$\boldsymbol{F_7}$	1/5, 0/5, 0/5, 0/5, 0/3 (k_{tot} = 1, N_{tot} = 23)	INT-SGTR-HFE2B	0/14	3.20e-03 (“Easy/somewhat difficult”)
$\boldsymbol{F_8}$	0/5, 0/5, 0/5, 0/5, 0/5, 0/5, 0/5, 0/5, 0/5, 0/5, 0/5, 0/5, 0/2 (k_{tot} = 0, N_{tot} = 62)	INT-SGTR-HFE4A	0/14	3.20e-04 (“Very easy”)
$\boldsymbol{F_{10}}$	0/5, 0/5, 0/5, 0/5, 0/5, 0/5, 0/5, 0/5, 0/5, 0/5, 0/5, 0/4 (k_{tot} = 0, N_{tot} = 59)	INT-SGTR-HFE5B2	0/7	1.00e-03 (“Easy”)
$\boldsymbol{F_{11}}$	0/5, 0/5, 0/5 (k_{tot} = 0, N_{tot} = 15)	INT-LOFW-HFE1A	0/10	3.20e-03 (“Easy/somewhat difficult”)

Table 4.5. Scaling guidance used to convert the expert-based HFE difficulty rankings from [54-57] into HFE probability estimates (Table 4.4, last column). Adapted from the qualitative likelihood scale provided in Table 3.8-2 of NUREG-1880 (ATHEANA User's Guide [59]).

HFE difficulty ranking	Failure probability estimate
Extremely difficult	1.00e-00
Very difficult	3.20e-01
Difficult	1.00e-01
Fairly difficult / difficult	3.20e-02
Somewhat difficult	1.00e-02
Easy / somewhat difficult	3.20e-03
Easy	1.00e-03
Very easy	3.20e-04
Extremely easy	1.00e-04

4.4.2. HEP quantification

To distinguish the effects of expert judgment incorporation on the HEP quantification, the following two cases are considered:

- Case 1: Stage 2 is informed only by data $E_{S|HFE}$ (Table 4.4, fourth column);
- Case 2: Stage 2 is informed by combining data $E_{S|HFE}$ with expert estimates $E_{J|HFE}$ (Table 4.4, fourth and last column).

The statistics (mean, median, 5th-95th percentiles, and error factor) of the $P_{HFE}(HEP)$'s returned by the two-stage model for the full set of HFEs are summarized in Table G.3 (Appendix G). Figure 4.8 shows the expected mean and the 5th-95th percentiles for a representative subset of HFEs, i.e. in x-axis: US-1A, US-3A, and US-1C from [54]; INT-SGTR-FB2 and INT-SGTR-3B from [55-57]. Figure 4.8 also includes the $P_{\boldsymbol{F}}(p_{t|\boldsymbol{F}})$'s returned by Stage I for the associated $\boldsymbol{F}$'s (in x-axis, $\boldsymbol{F}_1$, $\boldsymbol{F}_4$, $\boldsymbol{F}_2$, $\boldsymbol{F}_{10}$, and $\boldsymbol{F}_6$, respectively), as well as the HEP uncertainty distributions estimated in the HRA Empirical Studies [53-57]. In the US Study [54], a conjugate beta-binomial model is set up with a Jeffreys prior distribution, i.e. a non-informative beta distribution with both shape parameters (i.e. α_0 and β_0 in subsection 4.3.2) equal to 0.5. In the International Study [55-57], a lognormal-binomial model is set up with a weakly-informative lognormal prior, with 5th and 95th percentiles respectively equal to 1.2e-04 and 3.0e-01. According to [55], such percentiles "represent some of the lowest and highest values expected for the HEPs of operator actions and correspond to an error factor of 50". For convenience, the same percentiles are assigned to HEP_5 and HEP_{95} in eq. 4.10 to derive the parameters of the lognormal prior $\pi_0(\mu_{\boldsymbol{F}})$ in Stage I (resulting in $\mu_{\mu_{\boldsymbol{F}}}$ = log(6.0e-3) and $\sigma_{\mu_{\boldsymbol{F}}}$ = 2.4).

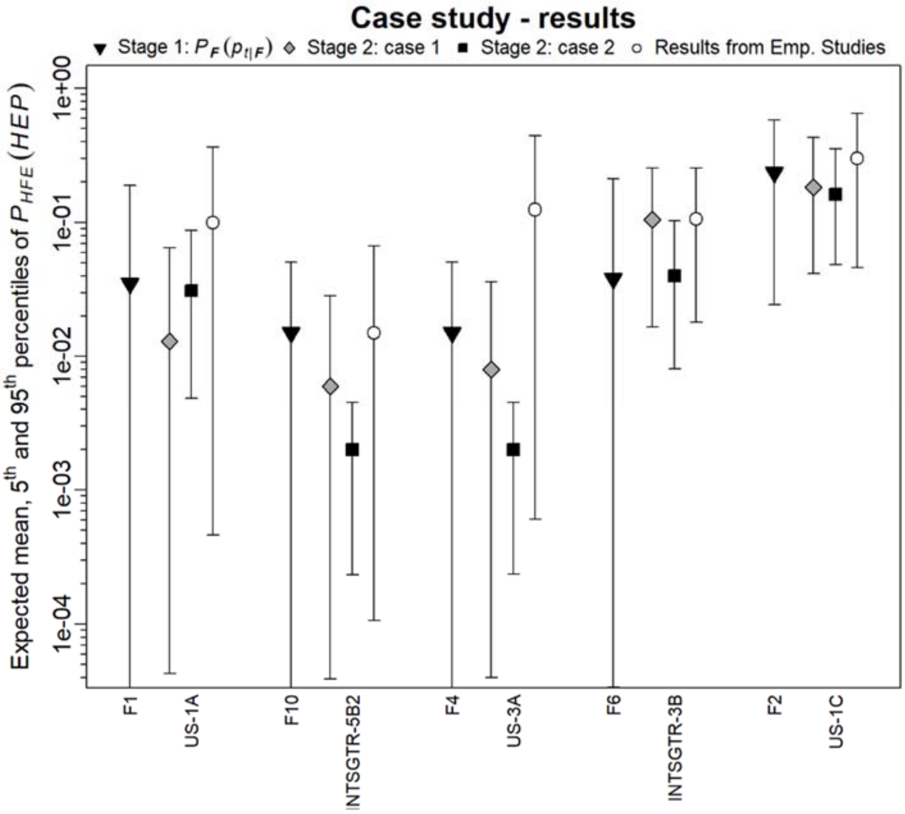


Figure 4.8. Results from the application of the two-stage model to a subset of HFEs (in x-axis) from the case study (the numeric results for the complete set of HFEs are provided in Table G.3, Appendix G). On y-axis (in log-scale): expected mean (filled symbols), 5th and 95th percentiles (whiskers) of the $P_{HFE}(HEP)$'s returned by the two-stage model (for both Case 1 and Case 2) and the lumped-data approaches from literature sources [53-57].

Comparing the results for Case 1 and Case 2 in Figure 4.8, a general tendency can be observed: the incorporation of expert estimates ($E_{J|HFE}$) in Case 2 tends to reduce the uncertainty on the estimated HEP values compared to Case 1 where only failure data ($E_{J|HFE}$) is used. Overall, this tendency replicates across all the HFEs analyzed in the case study, with the effects of $E_{J|HFE}$ becoming more evident for HFEs characterized by scarce failure data. For instance, for US-HFE1A (k_{HFE}/N_{HFE} = 0/4 in Table 4.4, right), the uncertainty on the expected HEP (Case 1: 1.3e-02; Case 2: 3.10e-02) is reduced by about a factor of nine, with values of expected error factor equal to 39 and 4.2 for Case 1 and Case 2, respectively. Similarly, for INT-SGTR-HFE5B2 (k_{HFE}/N_{HFE} = 0/7 in Table 4.4, right), the uncertainty on the expected HEP (Case 1: 6.0e-03; Case 2: 2.0e-02) is reduced by approximately a factor of six (expected error factors: 26.9 versus 4.4). On the contrary, for HFEs with at least one observed failure, the differences between Case 1 and Case 2 are small: see, for instance, US-HFE1C and INT-SGTR-HFE3B in Figure 4.8 (respectively with k_{HFE}/N_{HFE} = 1/4 and k_{HFE}/N_{HFE} = 2/14).

Similar considerations apply when comparing Case 2 with the lumped-data approaches used in the Empirical Studies [53-57]. On the one hand, both the proposed two-stage model and the lumped-data model return similar $P_{HFE}(HEP)$'s for HFEs with $k_{HFE} \geq 1$. On the other hand, with poor failure data, the lumped-data model provides uncertain HEP estimates characterized by unpractical error factors (e.g. for US-HFE1A and INT-SGTR-HFE5B2, the expected error factors are respectively 28 and 25.1).

4.5 Discussion

The construction of HEP population variability distributions in Stage I and their use as generic priors for plant-specific HFEs require the availability of evidence (empirical data $E_{S|\boldsymbol{F}}$ and/or expert estimates $E_{J|\boldsymbol{F}}$) relevant to the representative constellations of task/PSF categories. Concerning $E_{S|\boldsymbol{F}}$, such data requirements are generally met by the current availability of simulator data for many constellations $\boldsymbol{F}$ that are normally trained in large-scale programs, as highlighted by the application to case study (Section 4.4) with data extrapolated from the SACADA public database [58] (Table 4.4). The same does not apply when dealing with constellations that are not-frequently trained in simulators: an example of such constellations are $\boldsymbol{F}_5$ and $\boldsymbol{F}_9$ in Table G.2 (respectively representing the HFEs INT-SGTR-HFE1B and INT-SGTR-HFE5B1 from the complex SGTR variant in [55]), for which no data was found available in the SACADA database at the time of the application. In the latter case, to increase data usability, the $P_{\boldsymbol{F}}(p_{t|\boldsymbol{F}})$'s in Stage I could be alternatively informed by evidence collected for "similar" constellations, e.g. that share a subset of PSFs with the constellation of interest: for instance, in Table G.2, $E_{S|\boldsymbol{F}}$ relevant to $\boldsymbol{F}_2$ could be adapted to both $\boldsymbol{F}_5$ and $\boldsymbol{F}_9$. Note however that the compatibility between constellations must be carefully evaluated in order to avoid underrepresentation (or overrepresentation) of the actual performance influencing factors characterizing the plant-specific HFE at hand.

Given the demonstration purposes of the case study (Section 4.4), each of the HFEs in Table G.2 is associated to a unique task type, representing the predominant macro-cognitive function from the SACADA taxonomy [24]. It is important to note that, whilst such modelling choice would be more appropriate for operator tasks defined at a more microscopic granularity level (e.g. monitoring a specific alarm, or operating a specific a valve), it may however oversimplify the representation of those HFEs whose task characteristics are defined at a more macroscopic level (e.g. in Table G.2, "failure to identify and isolate the ruptured steam generator"), for which more cognitive functions are expected to play a role in operator performances. Guidelines on the use of SACADA taxonomy to inform HEP quantification models can be found in [24].

The HEP population variability formulation adopted for Stage I interprets the HEP as a crew-generic quantity (i.e. the variable $p_{t|\boldsymbol{F}}$), without explicitly considering crew-to-crew variability aspects stemming from different crew behavioral characteristics or operating styles (e.g. in team decision-making or communication strategies). As mentioned in subsection 4.2.2, the focus of the present work was indeed on modelling source-to-source variability (i.e. plant-to-plant, scenario-to-scenario, task-to-task) within the categories of task type and performance factors of the given data collection taxonomy. Previous work from the authors [60] developed a Bayesian hierarchical model based on the concept of crew behavioral patterns to explicitly treat crew performance variability aspects in simulator data. In this regard, the mathematical formulation of Stage I can be extended by future works to integrate crew behavioral patterns [60] in the HEP quantification process. Note however that the use of behavioral patterns to model crew-to-crew variability would require the availability of records of crew behaviors from simulator experiments or human factor studies [60], in order to be applicable.

As stated earlier in Section 4.1, besides their use as priors for plant-specific HEP estimation, the HEP population variability distributions produced in Stage I can inform reference HEP values and variability bounds to parametrize HRA methods, with general applicability to different constellations of categories (e.g. generic task types; PSF levels or ratings) of the given method taxonomy. Similarly, the estimated HEP distributions can be used as anchoring information (i.e. the CPDs) to parametrize the node categories of the emerging BBN-based models [12-16]. In this regard, the proposed Bayesian model allows for a formal and traceable incorporation of the judgment-based evidence (i.e. the expert-elicited estimates $E_{J|F}$) in the reference HEP values and bounds of HRA models: this feature is of key importance especially for those constellations of task/PSF categories for which current data availability from data collection programs is not sufficient to derive statistically significant information. Lastly, it is important to note that this work considered expert judgment only in the form of task failure probability estimates. However, expert judgment can be available also in other fashions, e.g. as likelihood rankings or qualitative statements on the importance of influencing factors. Also, besides HEP quantification, expert judgment is involved in the construction of HRA models (e.g. in the selection of the nodes or to inform causal relationships in the BBN-based models [12-16]), as well as in the definition of protocols for HRA data collection. Future studies should investigate more in detail how integrate the results from the developed HEP quantification framework into HRA model parameters (e.g. into the BBN model relationships). Work by the authors is ongoing along this direction.

4.6 Conclusions

The increasing use of HRA results to support safety-relevant decision-making of nuclear power plants licensees and regulators requires that the HEPs estimated by the models, as well as the associated bounds, be to the extent possible empirically grounded. Therefore, a traceable incorporation of expert judgment is required whenever the latter is combined with empirical data in the HEP estimation process, to distinguish the empirical basis of HEP estimates from the judgment-based component.

This chapter presents a Bayesian two-stage model to mathematically integrate the new batches of simulator data produced by the currently-ongoing data collection campaigns with expert-elicited probability estimates, in the derivation of HEP population variability distributions for various constellations of task types and PSF levels (Stage I) as well as in the estimation of HEP values for HFEs in plant-specific PSA analyses (Stage II). The possibility to systematically combine diverse information sources in a traceable and reproducible way makes the proposed Bayesian model a versatile, ready-to-use data aggregation framework for HEP quantification. Traceability is a feature of key importance, since it allows continuous updates of the HEP estimates as new empirical evidence becomes available (i.e. from the long-running data collection programs, or from the specific plant), progressively replacing the judgment-based information in the reference HEP values and bounds underlying HRA models, as well as in the plant-specific estimates. Also, traceability in judgment incorporation is expected to increase the acceptability of HRA results for use in safety-relevant applications.

The application to case study demonstrates that the combined use of data and expert estimates in the two-stage model can significantly improve the quality of the quantified HEP values for those HFEs characterized by scarce plant-specific data. This is an important aspect considering that, with poor data available, the HEP estimates returned by the commonly-adopted lumped-data approaches are not of practical use for PSA applications (i.e. the uncertainty on the expected values is too large).

The sensitivity analysis performed on Stage I has shown that, for constellations of task/PSF categories characterized by moderately high HEP values (i.e. around 0.01), the integration of data and expert judgment yields practical estimates of the associated HEP population variability distributions already with few dozen data points. Overall, such data requirements are already achievable for most of the constellations of task and performance factors addressed by current simulator programs. Numerical tests with artificial data have also shown that a simple predictive checks with Bayesian p-values may effectively spot the presence of biases in the probability estimates provided by the experts. Such checks can support the HRA analyst in assigning appropriate confidence levels to the consulted experts, in order to mitigate the effects of their biases on the HEP estimates.

Acknowledgments

The work was funded by the Swiss Federal Nuclear Safety Inspectorate (ENSI), under contract Nr. 101163. The views expressed in this work are solely those of the authors.

References

1. Spurgin AJ. *Human Reliability Assessment – theory and practice*. CRC press: Boca Raton, FL, USA, 2010.
2. Kirwan B. *A guide to practical Human Reliability Assessment*. CRC press: Boca Raton, FL, USA, 1994.
3. Podofillini L. Human Reliability Analysis. In: Moller N, Hansson SO, Holmberg JE, and Rollenhagen C. (eds) *Handbook of Safety Principles*. Wiley, 2017, pp.565-592.
4. Swain AD and Guttman HE. Handbook of human reliability analysis with emphasis on nuclear power plant applications. NUREG/CR-1278, U.S. Nuclear Regulatory Commission, Washington DC, USA, 1983.
5. Williams JC. HEART – A Proposed Method for Assessing and Reducing Human Error. In: *9th Advance in Reliability Technology Symposium*, University of Bradford, 1986.
6. Williams JC. A data-based method for assessing and reducing human error to improve operational performance. In: *Proceedings of the IEEE Fourth Conference on Human Factors and Power Plants*, Monterey, California, 5–9 June, pp. 436–450, 1988.
7. Williams JC. HEART – a proposed method for achieving high reliability in process operation by means of human factors engineering technology. *Saf. Reliab*. 2015, 35(3): 5–25.
8. Gertman DI, Blackman HS, Marble JL, et al. The SPAR-H Human Reliability Analysis Method. NUREG/CR-6883, U.S. Nuclear Regulatory Commission, Washington DC, USA, 2005.
9. Whaley AM, Kelly DL, Boring RL, et al. *SPAR-H step-by-step guidance*. INL/EXT-10-18533, Idaho National Labs, Idaho Falls, Idaho 83415, 2011.
10. Hollnagel E. *Cognitive Reliability and Error Analysis Method* (*CREAM*). Oxford: Elsevier Science Ltd, 1998.
11. Xing J, Parry G, Presley M, et al. An Integrated Human Event Analysis System (IDHEAS) for Nuclear Power Plant Internal Events At-Power Application. NUREG-2199 Vol.1, U.S. Nuclear Regulatory Commission, Washington DC and Electric Power Research Institute, Palo Alto CA, USA, 2017.
12. Groth KM and Mosleh A. Deriving causal Bayesian networks from human reliability analysis data: a methodology and example mode. *Proc Inst Mech Eng*, *Pt O: J Risk Reliab* 2012, 226(4): 361-379.
13. Mkrtchyan L, Podofillini L and Dang VN. Bayesian belief networks for human reliability analysis: A review of applications and gaps. *Reliab Eng Syst Saf* 2015, 139:1-16.
14. Sundarmurthi R and Smidts C. Human reliability modelling for Next Generation System Code. *Ann Nucl Energy* 2013, 137-156.
15. Zhao Y and Smidts C. A method for systematically developing the knowledge base of reactor operators in nuclear power plants to support cognitive modeling of operator performance. *Reliab Eng Syst Saf* 2019, 186:64-77.
16. Shirley RB, Smidts C and Zhao Y. Development of a quantitative Bayesian network mapping objective factors to subjective performance shaping factor evaluations: An example using student operators in a digital nuclear power plant simulator. *Reliab Eng Syst Saf* 2020, 194:106416.

17. Hallbert B and Kolaczkowski A. The Employment of Empirical Data and Bayesian Methods in Human Reliability Analysis: A Feasibility Study. NUREG/CR-6949, pp. 1-4, INL/EXT-06-11670, Washington, D.C.: U.S. Nuclear Regulatory Commission, 2007.
18. Kirwan B. Validation of human reliability assessment techniques: Part 1 — Validation issues. *Saf Sci* 1997a, 27(1):25-41.
19. Kirwan B. Validation of human reliability assessment techniques: Part 2 — Validation results. *Saf Sci* 1997b, 27(1):43-75.
20. Forester J, Dang VN, Bye A, et al. The International HRA Empirical Study Lessons Learned from Comparing HRA Methods Predictions to HAMMLAB Simulator Data. NUREG-2127, US Nuclear Regulatory Commission, Washington DC, USA, 2014.
21. Forester J, Liao H, Dang VN, et al. The US HRA Empirical Study - Assessment of HRA Method Predictions against Operating Crew Performance on a US Nuclear Power Plant Simulator. NUREG-2156, US Nuclear Regulatory Commission, Washington DC, USA, 2016.
22. Hallbert B, Morgan T, Hugo J, et al. A Formalized Approach for the Collection of HRA Data from Nuclear Power Plant Simulators. NUREG/CR-7163; INL/EXT-12-26327, US Nuclear Regulatory Commission, Washington DC, USA and Idaho National Laboratories, Idaho, USA, 2013.
23. Park J, Jung W, Kim S, et al. A guideline to collect HRA data in the simulator of nuclear power plants. KAERI/TR-5206, Korea Atomic Energy Research Institute, Republic of Korea, 2013.
24. Chang JY, Bley D, Criscione L, et al. The SACADA database for human reliability and human performance. *Reliab. Eng. Syst. Saf.* 2014, 125: 117-133.
25. Groth KM, Smith CL, and Swiler LP. A Bayesian method for using simulator data to enhance human error probabilities assigned by existing HRA methods. *Reliab Eng Syst Saf* 2014, 128 (Supplement C): 32-40.
26. Azarm MA, Kim IS, Marks C, et al. Analyses methods and pilot applications of SACADA database. In: *14th Probabilistic Safety Assessment and Management, PSAM 14 2018*, UCLA Meyer & Renee Luskin Conference Center, Los Angeles, California.
27. Jung W, Park J, Kim Y, et al. HuREX – A framework of HRA data collection from simulators in nuclear power plants. *Reliab Eng Syst Saf* 2020, 194: 106235.
28. Chang JY and Franklin C. SACADA Data for HEP Estimates. In: *14th Probabilistic Safety Assessment and Management, PSAM 14 2018*, UCLA Meyer & Renee Luskin Conference Center, Los Angeles, California.
29. Nelson PF and Grantom CR. Methodology for Supporting the Determination of Human Error Probabilities from Simulator Sourced Data. In: *14th Reliability Safety Assessment and Management, PSAM 14 2018*, UCLA Meyer & Renee Luskin Conference Center, Los Angeles, California.
30. Groth KM. A framework for using SACADA to enhance the qualitative and quantitative basis of HRA. In: *14th Reliability Safety Assessment and Management, PSAM 14 2018*, UCLA Meyer & Renee Luskin Conference Center, Los Angeles, California.
31. Kim Y, Park J, Jung W, et al. Estimating the quantitative relation between PSFs and HEPs from full-scope simulator data. *Reliab Eng Syst Saf* 2018, 173: 12-22.
32. Apostolakis G. On the use of judgment in probabilistic risk analysis. *Nucl. Eng. Des*. 1986,

93(2–3): 161-166.
33. Mosleh A and Chang YH. Model-based human reliability analysis: prospects and requirements. *Reliab Eng Syst Saf* 2004, 83: 241–253.
34. Groth KM, Smith R and Moradi R. A hybrid algorithm for developing third generation HRA methods using simulator data, causal models, and cognitive science. *Reliab Eng Syst Saf* 2019, 191:106507.
35. Siu NO and Kelly DL. Bayesian parameter estimation in probabilistic risk assessment. *Reliab Eng Syst Saf* 1998, 62(1): 89-116.
36. Apostolakis G. The concept of probability in safety assessments of technological systems. *Science* 1990, 250(4986): 1359-1364.
37. Hallbert B and Kolaczkowski A. The Employment of Empirical Data and Bayesian Methods in Human Reliability Analysis: A Feasibility Study. NUREG/CR-6949, pp. 1-4, INL/EXT-06-11670, Washington, D.C.: U.S. Nuclear Regulatory Commission, 2007.
38. Mosleh A and Smith C. The Feasibility Of Employing Bayesian Techniques And Other Mathematical Formalisms In Human Reliability Analysis, in The Employment of Empirical Data and Bayesian Methods in Human Reliability Analysis: A Feasibility Study, NUREG/CR-6949, pp. 5-15, INL/EXT-06-11670, Washington, D.C.: U.S. Nuclear Regulatory Commission, 2007.
39. Mosleh A and Apostolakis G. The assessment of probability distributions from expert opinions with an application to seismic fragility curves. *Risk Anal* 1986, 6(4): 447-461.
40. Mosleh A. Bayesian modeling of expert-to-expert variability and dependence in estimating rare event frequencies. *Reliab Eng Syst Saf* 1992, 38(1-2).
41. Apostolakis G and Mosleh A. Expert Opinion and Statistical Evidence: An Application to Reactor Core Melt Frequency. *Nucl Sci Eng* 1979, 70(2):135-149.
42. Droguett EL, Groen F and Mosleh A. The combined use of data and expert estimates in population variability analysis. *Reliab Eng Syst Saf* 2004, 83(3): 311-321.
43. Podofillini L and Dang VN. A Bayesian Approach to Treat Expert-Elicited Probabilities in Human Reliability Analysis Model Construction. *Reliab Eng Syst Saf* 2013, 117: 52-64.
44. Greco SF, Podofillini L, and Dang VN. A Bayesian model to treat within-category and crew-to-crew variability in simulator data for Human Reliability Analysis. *Reliab Eng Syst Safe* 2021, 206:107309, ISSN 0951-8320.
45. Forester J, Bley D, Cooper S, et al. Expert elicitation approach for performing ATHEANA quantification. *Reliab Eng Syst Saf* 2004, 83(2):207–20.
46. O'Hagan A, Buck CH, Daneshkhah A, et al. Uncertain judgments: eliciting experts' probabilities. Chichester, England: John Wiley & Sons, 2006.
47. Cooke RM. *Experts in uncertainty*. New York: Oxford University Press, 1991.
48. Plummer M. JAGS: A Program for Analysis of Bayesian Graphical Models using Gibbs Sampling. In: *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, DSC 2003, March 20-22, Vienna, Austria.
49. Denwood MJ. Runjags: An R Package Providing Interface Utilities, Model Templates, Parallel Computing Methods and Additional Distributions for MCMC Models in JAGS. *J Stat Softw* 2016, 71(9): 1–25.
50. Kruschke JK. *Doing Bayesian Data Analysis*. 2nd edition. Academic Press, 2015.
51. Gelman A, Carlin J, Stern H, et al. *Bayesian Data Analysis*. 2nd edition. Chapman and

Hall/CRC, 2003.

52. Kelly DL and Smith CL. *Bayesian Inference for Probabilistic Risk Assessment: A Practitioner's Guidebook*. London, UK: Springer-Verlag, 2011.
53. Forester J, Dang VN, Bye A, et al. The International HRA Empirical Study Lessons Learned from Comparing HRA Methods Predictions to HAMMLAB Simulator Data. NUREG-2127, US Nuclear Regulatory Commission, Washington DC, USA, 2014.
54. Forester J, Liao H, Dang VN, et al. The US HRA Empirical Study - Assessment of HRA Method Predictions against Operating Crew Performance on a US Nuclear Power Plant Simulator. NUREG-2156, US Nuclear Regulatory Commission, Washington DC, USA, 2016.
55. Lois E, Dang V, Forester J, et al. International HRA Empirical Study - Phase 1 Report: Description of Overall Approach and Pilot Phase Results from Comparing HRA Methods to Simulator Performance Data. NUREG/IA-0216 Vol. 1, U.S. Nuclear Regulatory Commission, Washington DC, USA, 2009.
56. Bye A, Lois E, Dang VN, et al. International HRA Empirical Study – Phase 2 Report: Results from Comparing HRA Method Predictions to Simulator Data from SGTR Scenarios. NUREG/IA-0216 Vol. 2, US Nuclear Regulatory Commission, Washington DC, USA, 2011.
57. Dang VN, Forester J, Boring R, et al. International HRA Empirical Study – Phase 3 Report: Results from Comparing HRA Methods Predictions to HAMMLAB Simulator Data on LOFW Scenarios. NUREG/IA-0216 Vol 3, US Nuclear Regulatory Commission, Washington DC, USA, 2014.
58. US Nuclear Regulatory Commission (USNRC). NRC's High-Value Datasets: Human Reliability Analysis, https://www.nrc.gov/data (2019, accessed 21 September 2020).
59. Forester J, Kolaczkowski A, Cooper S, et al. ATHEANA User's Guide Final Report. NUREG-1880, US Nuclear Regulatory Commission, Washington DC, USA, 2017.
60. Greco SF, Podofillini L, and Dang VN. Crew performance variability in human error probability quantification: a methodology based on behavioral patterns from simulator data. *Proc I Mech E Part O: J Risk and Reliability* 2021, doi:10.1177/1748006X20986743.

Chapter 5: Conclusions and future work

The closing chapter of this dissertation is structured as follows. Section 5.1 summarizes the research objectives and tasks of the Ph.D. work presented in Chapter 1. Section 5.2 first provides a general overview of the main contributions delivered by this thesis; then, subsections 5.2.1-5.2.3 address more in detail the key findings and achievements from Chapters 2-4, providing a link to the stated research objectives. Section 5.3 collects ideas for future studies, whilst Section 5.4 provides at closure the list of publications produced by the Ph.D. work.

5.1 Overview of research objectives and tasks

The overall motivation of this Ph.D. work was to develop new quantitative models, based on Bayesian statistical methods, integrating simulator data and expert judgment in the estimation of human error probabilities. The developed models are intended to improve the traceability in the aggregation of HRA data sources, as well as in the use of expert judgment, in the production of reference HEP values and bounds for task and PSF categories of HRA models. Focusing on three specific research gaps (see subsections 1.1.1-1.1.3), the Ph.D. work accomplished the following research objectives. First, the formal treatment of variability sources (crew-to-crew, within-category) for statistical inference of HEP estimates from simulator data collection (e.g. from the HuREX [1], SACADA [2] taxonomies) (research objective #1). Second, the identification (from simulator data) of crew behavioral characteristics (e.g. in team decision-making, communication strategies, etc.) that determine performance variability in given scenarios, and the incorporation of their effects in the estimated HEP variability distributions (research objective #2). Third, the systematic and transparent integration of expert judgment in the HEP quantification process, to allow for updates as new empirical evidence becomes available and strengthen the technical basis of HEP estimates (research objective #3). The three above-stated research objectives foresaw also that each of the developed Bayesian models be demonstrated to a case study of interest for practical HRA applications.

The development and application of the Bayesian models were broken into three blocks of research tasks, i.e. tasks 1.1-1.4, tasks 2.1-2.4, and tasks 3.1-3.4, respectively accomplishing research objectives #1, #2, and #3. The following list summarizes the research tasks and links them to the corresponding chapters of this thesis:

- **Chapter 2**: **Bayesian variability model for simulator data** (**research objective #1**):
  1.1. Characterization of variability aspects (crew-to-crew, within-category) in the constellations of task/PSF categories of data collection taxonomies (Section 2.2).
  1.2. Mathematical formulation of HEP variability model with continuous parametric

  distributions to represent data variability for a given constellation, and development of a Bayesian model to empirically estimate the parameters of the variability distribution from simulator data (Section 2.3).
  1.3. Model verification and sensitivity analysis with artificial data, to investigate data requirements to inform HEP variability in presence of different types of prior information on model parameters (subsection 2.4.2).
  1.4. Application to simulator datasets from literature [3-4], to demonstrate the effects of modelling variability on HEP estimates (subsection 2.4.3).

- **Chapter 3: Behavioral patterns to model crew performance variability (<u>research objective #2</u>)**:
  2.1. Concept of behavioral patterns to explicitly represent the influence of crew behavioral characteristics observed in simulator studies on performance variability: discrete formulation of HEP variability (Section 3.2).
  2.2. Development of a Bayesian hierarchical model to capture (from data) performance variability across crew behavioral patterns/groups, and incorporate their effects on the HEP estimate for the given constellation of task/PSF categories (subsection 3.3.3).
  2.3. Multi-step methodology to support the identification of crew behavioral patterns from simulator data and their use in HEP quantification (subsections 3.3.1-3.3.2).
  2.4. Application to crew behaviors collected from different emergency scenarios in recent simulator studies [5-6], to demonstrate the effects of empirically incorporating crew behavioral characteristics in the HEP estimates (Section 3.4).

- **Chapter 4: Traceable integration of data and judgment in HEP estimation (<u>research objective #3</u>)**:
  3.1. Extension of the HEP variability formulation with continuous parametric distributions (research task 1.2) to mathematically incorporate judgment (Section 4.2).
  3.2. Development of a two-stage Bayesian model to formally combine data and judgment in the estimation of HEP values and bounds for constellations of task/PSF categories (first stage), and plant-specific task failure probabilities (second stage) (subsection 4.3.2).
  3.3. Numerical test with artificially-generated data and judgment, to analyze the effects of judgment incorporation on HEP estimates and investigate model sensitivity to

biases in expert judgment (subsection 4.3.4).

3.4. Application to a collection of human failure events from the recent HRA Empirical Studies [6-9] (Section 4.4).

5.2 Conclusions

Through the accomplishment of the aforementioned research tasks, the Ph.D. work successfully produced a versatile set of modelling solutions to the research gaps discussed in Section 1.1 (an overview is given in Table 5.1).

First, the characterization of the variability sources in simulator data collection (research task 1.1, in Section 2.2) led to the development of a Bayesian variability model to formally treat crew-to-crew and within-category variability aspects in the estimation of human error probabilities from simulator data (research task 1.2, in Section 2.3). The developed model is generally applicable to different constellations of task and PSF categories of existing data collection taxonomies (e.g. HuREX [1], SACADA [2]). For the given constellation, the variability model mathematically represents HEP variability stemming from differences in tasks, scenarios, plants and crew characteristics via continuous parametric distributions: in this formulation, the model can be flexibly adapted to address specific variability aspects (e.g. plant-to-plant, scenario-to-scenario, crew-to-crew) according to data availability and the scope of the application. Contrarily to lumped-data approaches [3-4, 10], the developed model uses simulator data not just to inform the average HEP values of the constellations of task/PSF categories, but also the associated variability bounds. Indeed, the Bayesian variability model can produce empirically-based reference HEP values and bounds to inform HRA methods' task type and PSF categories, as well as anchoring distributions to parametrize advanced HRA models (such as the modern BBN-based models [11-15]). The variability model was first verified on artificially-generated data (research task 1.3, in subsections 2.4.1-2.4.2) and then applied to a case study involving simulator datasets from literature [3-4] (research task 1.4, in subsection 2.4.3). The numerical demonstration showed a significant overconfidence in the HEP estimates if variability within the constellations of task/PSF categories is not considered, e.g. if all data is lumped to inform the population average HEP as in the existing approaches with beta-binomial models [3-4, 10]. Also, not considering variability can result in significant biases for plant-specific human error probabilities.

Second, the Ph.D. work delivered a new modelling approach to empirically incorporate crew behavioral characteristics emerging from simulator studies in the estimation of HEP variability distributions. To this end, the thesis introduced the use of behavioral patterns to categorize the spectrum of crew behavioral characteristics (e.g. in team decision-making, communication strategies, adherence to procedures) for a given constellation of task and PSF categories, and represent performance variability over a finite (“discrete”) set of crew behavioral groups (research task 2.1, in Section 3.2). The formulation with behavioral patterns was included in a new Bayesian hierarchical model, to quantitatively capture performance variability across crew behavioral groups (research task 2.2, in subsection 3.3.3), and provided with a multi-step methodology, to support the identification of behavioral patterns from data and inform the behavioral groups of the Bayesian hierarchical model (research task 2.3, in subsections 3.3.1-3.3.2). Both the multi-step methodology and the Bayesian hierarchical model were applied to a case study from literature, involving different emergency scenarios from recent simulator studies [5-6] (research task 2.4, in Section 3.4). The application successfully demonstrated the capabilities of the proposed methodology in identifying relevant crew performance drivers determining performance variability from data, and efficiently incorporating this information in the HEP variability distributions estimated by the Bayesian hierarchical model. The application also highlighted the potential of the proposed methodology in detecting those crew behavioral patterns favoring lower failure probability values, per given task and operational context: in this regard, the methodology could be used in future HRA applications to suggest safety-enhancing measures to nuclear power plant managers (e.g. informing crew training; implementations of new procedural steps).

Lastly, the Ph.D. work addressed how mathematically incorporate expert judgment (in the form of expert estimates on task failure probability) in an upgraded formulation of the Bayesian variability model for simulator data (research task 3.1, in Section 4.2). This new formulation of the variability model was used as the basis for the development of a two-stage Bayesian model, with the goal to improve the estimation of plant-specific task failure probabilities in presence of limited empirical data (research task 3.2, in Section 4.3). The developed two-stage Bayesian model was first verified with artificially-generated evidence (research task 3.3, in subsection 3.4.3), to analyze the effects of judgment incorporation on HEP estimates and investigate model sensitivity to biases in expert judgment. Then, the model was applied to a collection of human failure events from the recent HRA Empirical Studies [6-9] (research task 3.4, in Section 4.4). The application demonstrated that the combined use of data and judgment in the two stages of the model effectively reduces the uncertainty on the estimated HEP values for those human failure events characterized by scarce empirical observations (i.e. with only few data points). In this regard, the numerical demonstration on the case study showed the potential of the two-stage Bayesian model for use in plant-specific PSA applications, to improve the quality of HEP estimates in presence of limited data availability.

As discussed in Section 1.1, the increasing use of HRA results to support safety-relevant decision-making of nuclear power plants licensees and regulators requires that the HEPs estimated by the models, as well as the associated bounds, be to the extent possible empirically grounded. In this direction, the modelling solutions presented in this thesis are expected to contribute to the advancement in the empirical foundation of future HRA models, representing versatile, ready-to-use statistical tools for deriving HEP values and variability bounds from the new batches of simulator data produced by the currently-ongoing data collection campaigns. In this regard, the systematic, traceable aggregation of simulator data and judgment will allow feeding HRA models with new data as it becomes available, progressively replacing judgment and older evidence that may become outdated because of new advances in plant operation and design. Also, a more transparent incorporation of judgment in the HEP estimation process is expected to increase the acceptability of HRA results for use in safety-relevant applications.

Table 5.1. General overview of the deliverables of this thesis and their uses in HRA/PSA.

Research objectives	Deliverables	Use in HRA/PSA
#1: Formal treatment of variability aspects (crew-to-crew, within-category) in the constellations of task and PSF categories of simulator data collection taxonomies (e.g. HuREX, SACADA), with general applicability to different constellations	Bayesian variability model to mathematically treat crew-to-crew and within-category variability in the estimation of HEP values and variability distributions from simulator data (Chapter 2)	- Production of empirically-grounded reference HEP values and bounds to inform HRA methods' task type and PSF categories, as well as anchoring distributions to parametrize advanced HRA models (e.g. the modern BBN-based models) - Enhance capabilities of future HRA models in treating specific data variability aspects (e.g. plant-to-plant, scenario-to-scenario) in simulator data, according to data availability and the scope of the application
#2: Identification of relevant crew behavioral characteristics (e.g. in team decision-making, communication strategies, adherence to procedures) emerging from simulator data that determine performance variability for a given constellation and incorporation of their effects on the HEP, jointly with the influence of the set of PSFs	Behavioral patterns to model crew performance variability: a multi-step methodology to support the identification of patterns from observed crew behaviours and their use in a Bayesian hierarchical model for HEP quantification (Chapter 3)	- Incorporation of crew-to-crew variability aspects in advanced crew performance models, representing the relationships among the spectrum of performance influencing factors (task-, context-, team-, and person-based) and the HEP - Highlight crew behavioral patterns that favor larger failure probability values in a given accidental scenario and, accordingly, suggest safety-enhancing measures to nuclear power plant managers (e.g. support training of operators, implementation of new steps in procedural guidance)
#3: Systematic and traceable incorporation of expert judgment in the HEP estimation process, to allow for updates as new empirical evidence becomes available and strengthen the technical basis of HEP estimate	Bayesian two-stage model integrating data and expert estimates in the quantification of HEP population variability distributions and plant-specific HEP values (Chapter 4)	- Data aggregation framework for HEP estimation combining diverse information sources in a reproducible way: expected to improve traceability in the use of expert judgment in future HRA models - Support the derivation of HEP variability distributions to parametrize HRA models, in particular those constellations of task/PSF categories for which current availability of simulator data is still not sufficient to derive statistically significant information - Support HEP estimation of human failure events in plant-specific PSA studies characterized by limited availability of plant-specific observations

The remainder of this section addresses more in detail the achievements and contributions from Chapters 2-4, with respect to the three research objectives reported in Section 5.1.

5.2.1. Bayesian variability model for simulator data (Chapter 2)

Chapter 2 presented the development of the Bayesian variability model to treat data variability in the estimation of HEP values and variability bounds from simulator data (research objective #1). The chapter accomplished the following main achievements:

- the mathematical formulation of HEP variability via continuous parametric distributions represents a first-of-a-kind attempt to formally represent crew-to-crew and within-category variability in the estimation of error probabilities from simulator data;
- the numerical application demonstrated the implications of neglecting variability within the constellations, notably: overconfidence in the estimated HEP uncertainty distributions; significant biases in plant-specific HEPs;
- the sensitivity analysis provided quantitative indications on the amount of simulator data required to empirically inform the HEP variability distributions, as well as insights on the range of practical applicability of the proposed model.

The formulation of HEP variability provided in Section 2.2 considers the HEP as a quantity that is specific for the given crew and for the given task/operational context; correspondingly, the developed Bayesian variability model (Section 2.3) adopts continuous variability distributions to capture (from simulator data) variability stemming from different operating crews (e.g. due to different behavioral characteristics), as well as from different tasks and operational conditions corresponding to different realizations of the associated task and PSF categories (see examples in Table 2.1). This interpretation of HEP as inherently variable quantity is opposite to the interpretation provided by the existing lumped-data approaches with conjugate beta-binomial models [3-4, 10]: for the given constellation of task/PSF categories, these approaches considers the HEP as a unique quantity and aggregates data from different crews, tasks, plants, and scenarios to inform a population-average HEP value. As discussed in Section 2.5, determining which of the two HEP interpretations should be adopted depends on the application at hand. For instance, when investigating PSF influences across different constellations, then the effect on the HEP of changes in one (or more) PSF state(s) may be studied by focusing on the aggregated effect, i.e. on the population average, therefore adopting the typical beta-binomial model. On the contrary, when the HEP estimates are used as reference values and bounds to inform the constellations of categories of an HRA model, adopting a variability model becomes important to capture the variability sources in data and ideally allow for plant-specific HEP estimates.

The second achievement results from the numerical demonstration of the Bayesian variability model on artificially-generated datasets (model verification in subsection 2.4.1). For the purpose of the application, the model was implemented with lognormal distributions (with parameters: mean and standard deviation) to represent both crew-to-crew and within-category variability terms. The numerical application demonstrated the “significant overconfidence in the estimated HEP uncertainty distributions if variability within the constellations is not considered, e.g. if all data is lumped to feed a beta-binomial Bayesian model” [16]. In addition, the application showed that overconfident HEP distributions can lead to significant biases in task failure probabilities when used as prior information for plant-specific applications [16].

Compared to the lumped-data approaches, empirically informing variability intuitively requires a larger amount of simulator observations. Given that the collection of simulator data is resource-intensive and requires important time and money investments, the sensitivity analysis presented in subsection 2.4.2 aimed at investigating the amount of data required such that the HEP estimates produced by the model are of practical use for PSA/HRA applications (i.e. the associated uncertainties are not too large). The results from the sensitivity analysis demonstrated that, for moderately high HEP values (in the range of 1e-2) and when setting diffuse hyper-priors on model parameters, HEP estimates with error factors of practical use (i.e. around 5) “can be obtained with few hundred, say below 500, data points (i.e. simulator runs)” [16]. The achieved results proved that data requirements of the Bayesian variability model are overall compatible with data availability from current simulator programs (though depending on the specific data collection taxonomy: see Section 2.5 for further details). In addition, the sensitivity analysis showed that setting informative priors on model parameters (e.g. from information available in failure database) can effectively reduce data requirements of the variability model. Notably, for HEP values in the range of 1e-2, “about 50 data points are sufficient to know HEP with acceptable error factors” [16]. For lower HEP values, in the range of 1e-3, “estimates of practical use become achievable with few hundred data points” [16]. An important aspect to consider when using informative priors is that biases in the prior distributions may result in biases in the posterior estimates. In this regard, “a simple check of the change between the prior and posterior estimates may reveal the presence of the initial bias” [16], as numerically demonstrated in subsection 2.4.2.

5.2.2. Behavioral patterns to model crew performance variability (Chapter 3)

Chapter 3 presented the multi-step methodology based on crew behavioral patterns and the associated Bayesian hierarchical model to capture performance variability across crew behavioral groups. The chapter accomplished the following main achievements:

- the concept of behavioral patterns as novel (model-based) approach to represent the effects of crew behavioral characteristics on performance variability, and empirically incorporate them (from simulator data) in the HEP variability distributions;

- the multi-step methodology is effective in processing information on crew observed behaviors to identify relevant crew performance drivers for the given constellation of task/PSF categories, as demonstrated in the application to case study;
- the numerical application showed that the Bayesian hierarchical model with crew behavioral groups is more sensitive in capturing performance variability from data, compared to the alternative quantitative approaches (the lumped-data models [3-4, 10] and the continuous variability model [16] presented in Chapter 2).

Contrary to the continuous formulation (Chapter 2), where performance variability within the constellations of task/PSF categories is mathematically treated as a "continuum" of different crew-, task-specific error probabilities, the modelling approach with behavioral patterns represent the HEP variability spectrum over a finite ("discrete") set of crew behavioral groups (Section 3.2). In this discrete formulation, crews sharing similar behavioral characteristics (e.g. in team decision-making, communication strategies, role awareness) during task performance, i.e. sharing the same behavioral pattern, are aggregated in the same behavioral group and associated the same value of error probability in the Bayesian hierarchical model presented in subsection 3.3.3. The hierarchical structure of the Bayesian model reflects indeed the discrete formulation of HEP variability: failure data (task failures and crew observations for each of the identified behavioral groups) enters at group level to inform the discrete set of group-specific error probabilities, with the latter then used to infer the HEP variability distribution for the constellation of task/PSF categories (hence, the "discretization" of the continuous variability formulation of Chapter 2). The element of newness in the proposed approach with behavioral patterns is that, whereas in existing HRA models crew behavioral characteristics are not incorporated in the HEP estimates and not informed by data (rather confounded in the provided, judgment-based uncertainty bounds, as in THERP method [17]), crew-to-crew variability is here expressed via a model (based on behavioral differences across groups of crews) and estimated from empirical data (i.e. the simulator observations).

The methodology for the identification of behavioral patterns from data and their use in HEP quantification (Section 3.3) was successfully applied in the case study (Section 3.4) to a collection of crew behaviors observed in emergency scenarios from recent simulator studies [5-6]. The considered tasks (Table 3.4) were all representative of the same task type (diagnosis), and characterized by masked indications (PSF "information quality" with level "masked") and symptoms-procedural mismatches (PSF "familiarity" with level "anomaly"), with stringent requirements on which behaviors would lead to successful performance. The methodology was structured as a multi-step process, consisting of two blocks. In the first block, crew behaviors were systematically processed using a taxonomy of teamwork competences from literature [18], comprehensively covering a broad range of teamwork and individual metrics relevant to nuclear power plant operations, and then classified according to the following behavioral categories: "progress through procedures", "flexibility in dealing with procedures and cues", "role awareness", "prioritization of goals and resources", and "decision making and information sharing" (Table 3.5). In the second block, crew performance data was matched to the corresponding combination of behavioral categories to identify seven behavioral patterns (Table 3.6), according to which data was then aggregated across seven behavioral groups to inform the Bayesian hierarchical model. The methodology was also provided with step-by-step guidance and recommendations to support the HRA analyst in the definition of behavioral categories and in the identification of patterns from the available data (subsections 3.3.1-3.3.2). The numerical results from the application highlighted a large performance variability across the seven behavioral groups: in particular, "ability to adapt, fast decisions, questioning attitude were all crew characteristics necessary to success" [19]. Besides enabling data aggregation from different crews on the basis of their behavioral commonalities, the proposed methodology efficiently spotted those crew behavioral patterns that favored lower failure probability values in the scenarios analyzed in the case study (see for instance "group 2" in Table 3.6). This opens to different applications: from the above-mentioned use to inform crew training, to the definition of "library" of behavioral profiles in different operational contexts (i.e. different constellations of task/PSF categories) to inform future crew performance models (see "Future works and recommendations" in Section 5.3).

The numerical application to case study (Section 3.4) also provided a comparison with two alternative models, i.e.: the existing conjugate beta-binomial models with lumped-data [3-4, 10], and the Bayesian variability model with continuous attributes [16] presented in Chapter 2. In addition, the application was complemented by a sensitivity analysis (with artificial data) on model results (subsection 3.4.2), to investigate the influence of the following aspects on the estimated HEP variability distribution: the number of identified behavioral groups; the degree of performance variability across the groups; and the choice of parametric variability distribution in the Bayesian hierarchical model. Concerning the first two aspects, the results from the sensitivity analysis highlighted that the model with behavioral patterns is more sensitive to performance variability compared to the alternative approaches. Notably, "the more heterogeneous is the group-specific failure data, the more the results diverge from the lumped and continuous variability formulations" [19]. On the other hand, the results showed that "the benefits of using the Bayesian hierarchical model […] compared to simpler approaches (e.g. the lumped-data models) diminish with reduced performance variability underlying the dataset" [19].

5.2.3. Traceable integration of data and judgment (Chapter 4)

Chapter 4 developed a two-stage Bayesian model to formally integrate empirical data and expert judgment in the estimation of task failure probabilities. The chapter accomplished the following main achievements:

- The developed two-stage model allows for a systematic and traceable aggregation of diverse information sources (simulator data, expert-elicited probability estimates, and plant-specific failure data) throughout the HEP quantification process;
- The numerical application to case study successfully demonstrated that the combined use of data and judgment in both stages of the model overall improves the quality of HEP estimates for data-poor human failure events;
- The sensitivity analysis on the first stage quantitatively proved how judgment incorporation can reduce data requirements to inform HEP variability, and suggested a technique to effectively spot potential biases in expert estimates.

Chapter 4 first built on the Bayesian variability model with continuous parametric distributions proposed in Chapter 2, and extended its formulation to incorporate judgment in the estimation of HEP variability distributions for the constellations of task/PSF categories of simulator data collection taxonomies (Section 4.2). The key element of the new formulation is that judgment (in the form of task failure probability estimates provided by domain experts, via direct elicitation or through the application of an existing HRA method) is mathematically combined to simulator data in the likelihood function of the Bayesian model, to inform the HEP realizations (i.e. the different task failure probabilities) of the HEP variability distribution. To this end, the lognormal error model proposed by reference [20] was adopted here to represent expert accuracy on the provided failure probability estimates, according to the associated uncertainty measures (e.g. an error factor on the point estimate). The new formulation of the HEP variability model was then included in the first stage of the two-stage Bayesian model (Section 4.3), to integrate simulator data (e.g. from HuREX [1], or SACADA [2]) with expert-elicited probability estimates in the derivation of HEP variability distributions. The two-stage configuration can flexibly address different purposes. On the one hand, the first stage of the model can be used to produce reference HEP values and variability bounds to feed HRA methods' task type and PSF categories (as well as to produce anchoring information, i.e. the CPDs, for the modern BBN-based models [11-15]), especially for those constellations of task/PSF categories for which current data availability from data collection programs is still not sufficient to derive statistically significant information. On the other hand, the output of the first stage (i.e. the estimated HEP variability distributions) can be also used as prior information in the second stage, where plant-specific failure data and expert estimates update the HEP variability distributions to quantify plant-specific task failure probabilities: in this regard, the developed two-stage Bayesian model can be used to estimate the HEPs (and the associated uncertainty) of human failure events in plant-specific PSA applications (similarly to the application to case study presented in Section 4.4). As stated earlier in this section, an important contribution of the developed two-stage Bayesian model is the possibility to combine diverse information sources throughout the HEP estimation process, in a systematic and reproducible way. This feature is of key importance, since enables for continuous updates of the HEP estimates as new empirical evidence becomes available (i.e. from the long-running data collection programs [1-2], or from the specific plant), progressively replacing the judgment-based information in the reference HEP values and bounds underlying HRA models, as well as in the plant-specific estimates. A similar use of Bayesian two-stage approaches can be found in references [21-22] for the aggregation of expert opinions and reliability data (possibly sparse and from diverse plants) to derive generic population variability distributions of reliability parameters (e.g. core melt frequency of nuclear power reactors [21]; pump failure rate [22]), and support parameter estimation for plant-specific components. In this regard, the work presented in Chapter 4 represents a first-of-a-kind attempt to explore the use of Bayesian two-stage models in a practical HRA problem.

In the numerical application to case study presented in Section 4.4, the developed two-stage Bayesian model was used to quantify the HEPs of a collection of human failure events simulated in the recent HRA Empirical Studies [6-9]. The selected HFEs were representative of different constellations of task and PSF categories, spanning from routine tasks in normally-trained scenarios (e.g. standard SGTR) to more challenging tasks in scenarios characterized by conflicting or masked cues (e.g. variants of a SGTR with multiple, concurrent system malfunctions). The dataset for the case study was informed by multiple sources: in the first stage of the model, simulator observations from the SACADA database [23] were used to derive the HEP variability distributions for the corresponding constellations; in the second stage, crew failure data and expert estimates from the Empirical Studies [6-9] were combined to update the variability distributions and estimate the HEP of the failure events. The results from the application demonstrated that the combined use of data and expert estimates in the two stages of the model overall improved the quality of the HEP estimates, reducing the uncertainty on the estimated values for those HFEs characterized by scarce empirical data. For instance, for the HFE "US-HFE1A" [7] (four task observations, with no failures), the uncertainty on the expected HEP was reduced by a factor of nine, compared to when only empirical data was used (error factors: 4.2 vs 39). Similarly, for "INT-SGTR-HFE5B2" [8] (seven task observations, with no failures), the uncertainty was reduced by a factor of six (error factors: 4.4 vs 26.9).

Finally, the sensitivity analysis performed on the first stage of the two-stage Bayesian model (subsection 4.3.4) quantitatively investigated the effectiveness of judgment incorporation in reducing data requirements to inform the HEP variability distributions. For the purpose of the analysis, simulator data and expert estimates were artificially-generated by sampling from a known HEP variability distribution. The sensitivity analysis demonstrated that, for moderately high HEP values (in the range of 1e-2), about 20 simulator tasks seem to be already sufficient to approximate the target variability distribution with acceptable error factors (i.e. around 5). Without judgment incorporation, the same approximation would require on average more than 60 simulated tasks, i.e. three times higher requirements of simulator data. The numerical tests with artificial evidence also proved the efficiency of model checking techniques, based on Bayesian p-values, in spotting potential biases (e.g. overestimation or underestimation) in the probability estimates provided by experts. As recommended in subsection 4.3.4, such techniques could support the analyst in assigning appropriate confidence levels to the experts, in order to reduce the effects of biased experts on the HEP estimates.

5.3 Future works and recommendations

The Ph.D. work presented in this thesis represents a first step towards the next generation of quantitative approaches for the treatment of simulator observations emerging from the ongoing data collection campaigns, and their use to inform future HRA models. The outcomes of this thesis raised further research directions that require explicit consideration in future works, a selection of which is listed in the following.

- The Bayesian models presented in this thesis were developed for general applicability to different constellations of categories (task types, PSF levels/ratings) of HRA models. It has to be noted however that existing HRA models significantly differ in the task types and PSFs considered, as well as in the granularity of their definition. It can be expected that both aspects are strongly connected with the variability terms that the model shall be able to represent, e.g.: the coarser the granularity of task type and PSFs definitions, the larger the expected within-category variability; the more decision-making and communication at crew level is involved, the more crew-to-crew variability will be relevant. With the current interest by the HRA community on empirically estimated HEPs, future studies should address the extent to which variability shall be modelled according to the characteristics and scope of the considered HRA model, as well as develop guidelines to do it.
- Concerning data requirements, the methodology based on behavioral patterns (Chapter 3) specifically requires information on crew behaviors in order to identify patterns from simulator observations. For the methodology to be applicable, information on crew

behaviors should be available for all sessions, independently on the crew performance outcome. If, on the one hand, this goes beyond the data collection intentions of the ongoing large-scale simulator programs (see for instance the current taxonomy of HuREX [1] and SACADA [2], more focused on population-averaged information), on the other hand records of crew behaviors can be provided by other human factor studies, not necessarily intended for HRA applications (similarly to the case study addressed in Section 3.4). In future, these studies may be used to derive empirical indications of the actual HEP spread for subsets of tasks and PSF categories, to complement the average HEP values estimated from large scale data collection protocols.

- Coherently with the empirical observations in simulator studies, the modelling approach with behavioral patterns (Chapter 3) acknowledges that crew behaviors are neither merely “situation-driven” nor “crew-driven”: rather, the approach generalizes both interpretations and makes the analysis of behavioral characteristics conditional on the constellation of task/PSF categories, with the actual set of characteristics emerging from the actual observations, as a result of the interactions of all factors (situation- as well as crew-driven) interplaying in the determination of crew behaviors (Section 3.5). In this regard, the proposed approach with behavioral patterns offers a tool for future works to study the interplay across these influences.
- Concerning the spectrum of crew behavioral characteristics determining crew-to-crew variability, the application to case study in Section 3.4 focused only on those characteristics relevant to the teamwork dimensions addressed by reference [18], in particular: in team coordination, team decision making, communication, leadership, situation awareness and working attitude (see Table D1 in Appendix D). The application did not consider behavioral characteristics stemming from intra-personal factors (such as self-management of fatigue or stress, personal distractions or concerns), as well as cultural aspects. Although such aspects play a role in determining crew-to-crew variability, their observation in a simulator study is not straightforward. Future works should address what type of information needs to be collected and establish metrics for interpretation of such aspects, to allow for their incorporation HRA models.
- The crew behavioral patterns identified in the case study of Section 3.4 emerged from very challenging scenarios, characterized by masked indications and symptoms-procedural mismatches that led to large variability in crew performance. The considered case study was imposed by the available data. For future analysis, with larger amount of data available, it would be beneficial to address diverse scenarios, as well as less challenging situations, to investigate more comprehensively the effect of crew behavioral characteristics on the HEP variability. In the long term, as more data on crew

behaviors is collected, a consolidated “library” of behavioral categories may be identified and reused across studies to investigate their relative importance and impact on crew performance. Also, with more data available, data analysis and statistical tests could be used to identify dominant behavioral categories (e.g. via cluster analysis) and accordingly rule out (or aggregate) categories with limited impact on task performance, to support the identification of behavioral patterns from data and at the same time reduce the subjective component in category definitions. Besides defining more established sets of categories, such analyses can be used to provide information on the frequency of each behavioral pattern, per given constellation of task and PSF categories. This information (possibly complemented with expert judgment on the plant crew specificity) can be used to inform HRA prospective analyses for which many crew observations are not possible.

- The development of the two-stage Bayesian model (Chapter 4) did not explicitly consider crew behavioral characteristics in the mathematical formulation of the HEP variability model used in the first stage. As mentioned in Section 4.5, this specific configuration was driven by the data availability for the case study addressed in Section 4.4. With more information available, future works could reformulate the first stage of the Bayesian model as to explicitly model crew performance variability (e.g. via behavioral patterns, as in Chapter 3) in the HEP quantification process.
- The Bayesian models delivered by this thesis can be used to produce anchoring information (i.e. the CPDs) to parametrize the node categories of the emerging BBN-based models [11-15], empirically incorporating data variability in the CPDs. Also, the methodology presented in Chapter 3 could be used to inform crew-to-crew variability nodes in the BBN, via behavioral patterns relevant to the specific combination of task and PSF nodes. Future studies should investigate more in detail how integrate the results from the developed Bayesian models into the BBN relationships, compatibly with the scope of the application at hand.

5.4 Publications

The Ph.D. work delivered a total of five articles, out of which three journal articles and two conference papers. The three journal articles represent the backbone of the present thesis and were reproduced (with permission from the authors) as Chapters 2-4. Out of the three journal articles, two have been peer-reviewed and accepted by the editors, and one is currently under internal review. The list of publications and the corresponding chapters of this thesis are presented in the following.

Journal papers:

- Greco SF, Podofillini L, and Dang VN. A Bayesian model to treat within-category and crew-to-crew variability in simulator data for Human Reliability Analysis. *Reliab Eng Syst Safe* 2021, 206:107309, ISSN 0951-8320 (Chapter 2).
- Greco SF, Podofillini L, and Dang VN. Crew performance variability in human error probability quantification: a methodology based on behavioral patterns from simulator data. *Proc I Mech E Part O: J Risk and Reliability* 2021, doi:10.1177/1748006X20986743 (Chapter 3).
- Greco SF, Podofillini L, and Dang VN. A Bayesian two-stage approach to integrate simulator data and expert judgment in human error probability estimation. Currently under internal review, expected submission date: June 2021 (Chapter 4).

Conference papers:

- Greco SF, Podofillini L and Dang VN. Modelling crew performance variability in emergency situations from simulator data for human reliability analysis. In: *Proceedings of the 30th European Safety and Reliability Conference and 15th Probabilistic Safety Assessment and Management Conference*, ESREL2020 PSAM15, 1-5 November 2020, Venice, Italy. ISBN: 978-981-14-8593-0.
- Greco SF, Podofillini L and Dang VN. Crew performance variability in simulator data for Human Reliability Analysis: investigation of modelling options. In: *Proceedings of the 29th European Safety and Reliability Conference*, ESREL 2019, 22-26 September, Hannover, Germany. ISBN: 981-973-0000-00-0.

.

References

1. Park J, Jung W, Kim S, et al. A guideline to collect HRA data in the simulator of nuclear power plants. KAERI/TR-5206, Korea Atomic Energy Research Institute, Republic of Korea, 2013.
2. Chang JY, Bley D, Criscione L, et al. The SACADA database for human reliability and human performance. *Reliab Eng Syst Saf* 2014, 125: 117-133.
3. Groth KM, Smith CL, and Swiler LP. A Bayesian method for using simulator data to enhance human error probabilities assigned by existing HRA methods. *Reliab Eng Syst Saf* 2014, 128 (Supplement C): 32-40.
4. Jung W, Park J, Kim Y, et al. HuREX – A framework of HRA data collection from simulators in nuclear power plants. *Reliab Eng Syst Saf* 2020, 194: 106235.
5. Massaiu S and Holmgren L. Diagnosis and Decision-Making with Emergency Operating Procedures in Non-Typical Conditions: A HAMMLAB Study with U.S. Operators. HWR-1121. Halden, Norway: OECD Halden Reactor Project, 2014.
6. Lois E, Dang V, Forester J, et al. International HRA Empirical Study - Phase 1 Report: Description of Overall Approach and Pilot Phase Results from Comparing HRA Methods to Simulator Performance Data. NUREG/IA-0216 Vol. 1, U.S. Nuclear Regulatory Commission, Washington DC, USA, 2009.
7. Forester J, Liao H, Dang VN, et al. The US HRA Empirical Study - Assessment of HRA Method Predictions against Operating Crew Performance on a US Nuclear Power Plant Simulator. NUREG-2156, US Nuclear Regulatory Commission, Washington DC, USA, 2016.
8. Bye A, Lois E, Dang VN, et al. International HRA Empirical Study – Phase 2 Report: Results from Comparing HRA Method Predictions to Simulator Data from SGTR Scenarios. NUREG/IA-0216 Vol. 2, US Nuclear Regulatory Commission, Washington DC, USA, 2011.
9. Dang VN, Forester J, Boring R, et al. International HRA Empirical Study – Phase 3 Report: Results from Comparing HRA Methods Predictions to HAMMLAB Simulator Data on LOFW Scenarios. NUREG/IA-0216 Vol 3, US Nuclear Regulatory Commission, Washington DC, USA, 2014.
10. Azarm MA, Kim IS, Marks C, et al. Analyses methods and pilot applications of SACADA database. In: 14th Probabilistic Safety Assessment and Management, PSAM 14 2018, UCLA Meyer & Renee Luskin Conference Center, Los Angeles, California.
11. Shirley RB, Smidts C and Zhao Y. Development of a quantitative Bayesian network mapping objective factors to subjective performance shaping factor evaluations: An example using student operators in a digital nuclear power plant simulator. *Reliab Eng Syst Saf* 2020, 194:106416.
12. Groth KM, Smith R and Moradi R. A hybrid algorithm for developing third generation HRA methods using simulator data, causal models, and cognitive science. *Reliab Eng Syst Saf* 2019, 191:106507.
13. Nelson PF and Grantom CR. Methodology for Supporting the Determination of Human Error Probabilities from Simulator Sourced Data. In: *14th Reliability Safety Assessment and Management, PSAM 14 2018*, UCLA Meyer & Renee Luskin Conference Center, Los

Angeles, California.

14. Groth KM. A framework for using SACADA to enhance the qualitative and quantitative basis of HRA. In: *14th Reliability Safety Assessment and Management, PSAM 14 2018*, UCLA Meyer & Renee Luskin Conference Center, Los Angeles, California.
15. Mkrtchyan L, Podofillini L and Dang VN. Methods for building Conditional Probability Tables of Bayesian Belief Networks from limited judgment: An evaluation for Human Reliability Application. *Reliab Eng Syst Saf* 2016, 151: 93-112.
16. Greco SF, Podofillini L, and Dang VN. A Bayesian model to treat within-category and crew-to-crew variability in simulator data for Human Reliability Analysis. *Reliab Eng Syst Safe* 2021, 206:107309, ISSN 0951-8320.
17. Swain AD and Guttman HE. Handbook of human reliability analysis with emphasis on nuclear power plant applications. NUREG/CR-1278, U.S. Nuclear Regulatory Commission, Washington DC, USA, 1983.
18. Skjerve AB and Holmgren L. An investigation of Teamwork Competence Requirements in Nuclear Power Plant Control-Room Crews across Operational States – a Field Study. HWR-1107. Halden, Norway: OECD Halden Reactor Project, 2016.
19. Greco SF, Podofillini L, and Dang VN. Crew performance variability in human error probability quantification: a methodology based on behavioral patterns from simulator data. *Proc I Mech E Part O: J Risk and Reliability* 2021, doi:10.1177/1748006X20986743.
20. Mosleh A. Bayesian modeling of expert-to-expert variability and dependence in estimating rare event frequencies. *Reliab Eng Syst Saf* 1992, 38(1): 47-57.
21. Apostolakis G and Mosleh A. Expert Opinion and Statistical Evidence: An Application to Reactor Core Melt Frequency. *Nucl Sci Eng* 1979, 70(2):135-149.
22. Droguett EL, Groen F and Mosleh A. The combined use of data and expert estimates in population variability analysis. *Reliab Eng Syst Saf* 2004, 83(3): 311-321.
23. US Nuclear Regulatory Commission (USNRC). NRC's High-Value Datasets: Human Reliability Analysis, https://www.nrc.gov/data (2019, accessed 21 September 2020).

Appendix A

This appendix complements the information in Chapter 1, providing further background on HRA within PSA of nuclear power plant operations, as well as an overview on the international status of HRA research field.

PSA, also called Probabilistic Risk Assessment (PRA), is a comprehensive and structured discipline aimed at analyzing risks associated to operations in complex, socio-technical systems [1]. In the context of nuclear power plant operations, PSA methods and quantitative tools have been widely adopted over the past four decades to identify initiating events leading to severe accidents, quantify their probability of occurrence and evaluate their impact on both plant and population [2]. PSA results are used to support safety-related decision making of nuclear power plants licensees and regulators, complementing the traditional deterministic safety studies (e.g. design-basis accident, defense-in-depth concept, single-failure criterion).

Operational experience in nuclear industry has shown that human error is responsible for a significant proportion (60% to 80% according to [3]) of safety-relevant accidents. The term "human error", as originally defined by [4], addresses "any member of a set of human actions or activities that exceeds some limit of acceptability, i.e. an out of tolerance action [or failure to act] where the limits of performance are defined by the system". As discussed in reference [5], recent large-scale accidents in nuclear industry (e.g. Three Mile Island, 1979; Chernobyl, 1986), demonstrated the effects of human error on plant performance, suggesting the need for properly assessing the associated risks and reducing their impacts on system vulnerability.

As part of PSA, HRA addresses the human contribution to the overall risk profile of nuclear power plant operations [5]. In this capacity, an HRA study aims at answering the following key questions [6]:

- how may operating crews fail the performance of a task?
- what are the factors (context-, task-, scenario-, team-, and person-based) that influence crew performance?
- what is the likelihood of human error?

In order to address these questions, a general HRA application is structured according to three main blocks:

- task analysis and human error identification, to represent how tasks are performed by the operating crews, characterize possible failure modes and error mechanisms, and identify safety-critical tasks for different operational phases (normal operations, outage, emergency situations);
- qualitative analysis of performance conditions, according to which the operational

context is characterized in terms of performance influencing factors (e.g. generally via a set of PSF levels or ratings);

- estimation of task failure probability (i.e. the HEP), in quantitative models capturing the relationships between the HEP and the set of PSFs representative of the given scenario-, task-, context-specific influences, taking into consideration also the interactions among PSF effects..

These three elements can overlap or be part of an iterative process, according to the HRA method adopted [6].

Since early stages of HRA history, various HRA methods have been developed and applied in the context of nuclear power plant PSA. These methods differ in their characteristics and in the underlying models for HEP quantification, according to which reference [6] proposed the following classification: decomposition-based methods (e.g. the Technique for Human Error Rate Prediction, THERP [4]); error mechanisms-based methods (e.g. the Cause-Based Decision Tree, CBDT [7]); factor-based methods (e.g. the Human Error Assessment and Reduction Technique, HEART [8-9]; the Cognitive Reliability and Error Analysis Method, CREAM [10]; the Standardized Plant Analysis Risk–Human reliability, SPAR-H [11-12]); and narrative-based methods (e.g. A Technique for Human Error Analysis, ATHEANA [13-14]; the Methode d'Evaluation de la Realisation des Missions Operateur pour la Surete, MERMOS [15]). Amongst the new generation methods, it is worth mentioning the emerging Integrated Human Event Analysis System (IDHEAS) method [16], as well as the Phoenix method [17].

HRA method results are typically integrated into PSA quantitative tools (e.g. event trees and fault trees) to inform scenario-specific human failure events, with the goal to quantify the overall frequency of accidental scenarios [2]. Figure A.1 shows an example adapted from literature [18], where the SPAR-H method [11-12] is applied to estimate the error probability of the HFE "failure of the crew to cooldown the Reactor Coolant System (RCS) expeditiously", in the event tree of a Steam Generator Tube Rupture (SGTR) accidental scenario. Besides the quantitative aspects, HRA results are effectively used to evaluate improvements for reducing risk associated to plant operations (typically referred in HRA literature as "error reduction measures" [5]). The proposed improvements can relate to the performance conditions of operating crews (e.g. implementation of new steps in procedural guidance, enhancement of the human-machine interface design), as well as to modifications to the response strategy in order to increase the time available for task performance [3, 5-6].

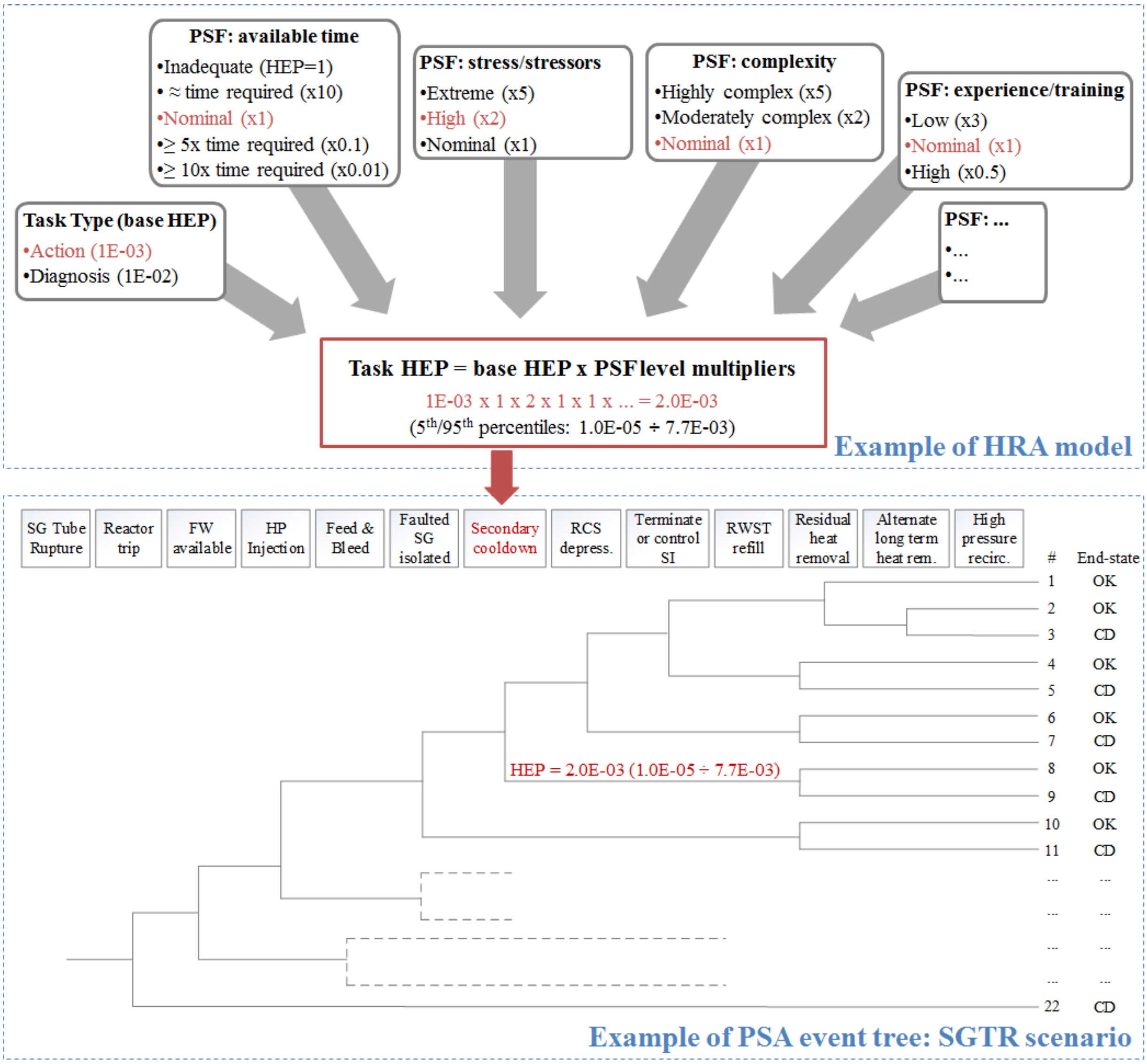


Figure A.1. Integration of HRA model results in a PSA event tree analysis: examples adapted from literature [17] (CD = Core Damage).

Current efforts in HRA research are motivated by the increasing use of PSA to support regulatory and operational decisions, in the nuclear power as well as other safety-relevant industrial sectors [6, 19]. As a result, a number of initiatives have been recently undertaken to understand strengths and weaknesses of the available HRA methods, as well as to improve guidance for their application. Landmark studies in this direction are the International Empirical HRA Study and the US Empirical Study [20-21], in which the strengths and weaknesses are investigated against empirical data from nuclear power plant control room simulator. The insights obtained in these studies were then consolidated in the large collaborative effort leading to the development of the IDHEAS method, coordinated by the US Nuclear Regulatory Commission [16]. Furthermore, the Nordic PSA Group[1] has sponsored a number of projects to

---

[1] http://www.npsag.org/home

improve practical application guidance and reduce variability in the HRA method application and results [22-23]. Fostered by the Empirical Studies, a revived impetus can be observed for the collection of data from nuclear power plant simulators (for an overview of the current data collection initiatives, as well as the open issues in the use of simulator data for HRA models, see the literature review in Chapter 2). The extensive use of PSA for risk-informed decision-making has also led to an increase in its application scope; in the nuclear power domain, this resulted in an enlarged set of performance conditions to be addressed by the methods, e.g. accidents initiated by fires in the plant [24] or seismic events [25] and human performance to mitigate severe accidents [26]. Further, newer HRA methods emphasize the treatment of decision failures (e.g. ATHEANA [13-14], MERMOS [15], IDHEAS [16], and the Commission Errors Search and Assessment method, CESA [27]), for example related to inappropriate strategies followed in response to an accident: indeed, investigations of accident reports and near misses emphasize the important contribution of decision failures leading to inappropriate actions (also known as Errors of Commission, EOCs [13-14, 27], along with the non-performance of the required actions, the latter being typically addressed in state-of-the-art PSA. As methods to deal with these failures mature, their treatment in modern PSA is increasingly covered [28]. New methods are being developed as well for application to industrial domains other than nuclear power where HRA has been traditionally mostly applied, e.g. [29-30].

As mentioned in the introduction of this thesis (Chapter 1), HRA methods are characterized by underlying models; these provide HEP values based on the method-specific characterization of the type of personnel tasks and factors believed to influence performance (e.g. the time available to respond, the quality of the human-machine interface, the salience of the main indicators in the main control room). Regarding the development of these HRA models, recent research efforts have predominantly addressed improvements in the link between HRA models and cognitive psychology [16, 31] and the development of new quantification frameworks, mostly adopting the Bayesian Belief Network framework, more suitable to reproduce the complex relationships among the factors influencing human performance and incorporate the diverse and sparse data available to inform them [32-35].

More detailed overviews of HRA process, methods, and state-of-art can be found in HRA literature [3, 5-6].

References

1. Verma AK, Srividya A and Karanki DR. Probabilistic Safety Assessment. In: Pham (ed) *Reliability and Safety Engineering*. 2nd ed. Springer-Verlag London, 2015, pp. 323-370.
2. International Atomic Energy Agency. Applications of Probabilistic Safety Assessment (PSA) for Nuclear Power Plants. IAEA-TECDOC-1200, IAEA, Wien, 2001.
3. Spurgin AJ. *Human Reliability Assessment – theory and practice*. CRC press: Boca Raton, FL, USA, 2010
4. Swain AD and Guttman HE. Handbook of human reliability analysis with emphasis on nuclear power plant applications. NUREG/CR-1278, U.S. Nuclear Regulatory Commission, Washington DC, USA, 1983.
5. Kirwan B. *A guide to practical Human Reliability Assessment*. CRC press: Boca Raton, FL, USA, 1994.
6. Podofillini L. Human Reliability Analysis. In: Moller N, Hansson SO, Holmberg JE, and Rollenhagen C. (eds) *Handbook of Safety Principles*. Wiley, 2017, pp.565-592.
7. Moieni P, Spurgin J and Singh A. Advances in Human Reliability Analysis Methodology. Part I: Frameworks, Models and Data. *Reliab Eng Syst Saf* 1994, 444:27–55.
8. Williams JC. HEART – A Proposed Method for Assessing and Reducing Human Error. In: *9th Advance in Reliability Technology Symposium*, University of Bradford, 1986.
9. Williams JC. A data-based method for assessing and reducing human error to improve operational performance. In: *Proceedings of the IEEE Fourth Conference on Human Factors and Power Plants*, Monterey, California, 5–9 June, pp. 436–450, 1988.
10. Hollnagel E. *Cognitive Reliability and Error Analysis Method* (*CREAM*). Oxford: Elsevier Science Ltd, 1998.
11. Gertman DI, Blackman HS, Marble JL, et al. The SPAR-H Human Reliability Analysis Method. NUREG/CR-6883, U.S. Nuclear Regulatory Commission, Washington DC, USA, 2005.
12. Whaley AM, Kelly DL, Boring RL, et al. *SPAR-H step-by-step guidance*. INL/EXT-10-18533, Idaho National Labs, Idaho Falls, Idaho 83415, 2011.
13. U.S. Nuclear Regulatory Commission. Technical basis and implementation guidelines for A Technique for Human Event Analysis (*ATHEANA*). NUREG-1624, U.S. Nuclear Regulatory Commission, Washington, DC, 2000.
14. Forester J, Kolaczkowski A, Cooper S, et al. ATHEANA User's Guide. NUREG-1880, U.S. Nuclear Regulatory Commission, Washington DC, USA, 2007.
15. Le Bot P. Human Reliability data, human error and accident models – illustration through Three Mile Island accident analysis. *Reliab Eng Syst Saf* 2004, 83: 153–167.
16. Xing J, Parry G, Presley M, et al. An Integrated Human Event Analysis System (IDHEAS) for Nuclear Power Plant Internal Events At-Power Application. NUREG-2199 Vol.1, U.S. Nuclear Regulatory Commission, Washington DC and Electric Power Research Institute, Palo Alto CA, USA, 2017.
17. Ekanem NJ, Mosleh A, and Shen SH. Phoenix – A model-based Human reliability analysis methodology: Qualitative analysis procedure. *Reliab Eng Syst Saf* 2015, 145: 301-315.
18. Lois E, Dang V, Forester J, et al. International HRA Empirical Study - Phase 1 Report: Description of Overall Approach and Pilot Phase Results from Comparing HRA Methods

to Simulator Performance Data. NUREG/IA-0216 Vol. 1, U.S. Nuclear Regulatory Commission, Washington DC, USA, 2009.

19. Mosleh A. PSA: A perspective on strengths, current limitations, and possible improvements. *Nucl Eng Technol* 2014, 46:1–10.
20. Forester J, Dang VN, Bye A, et al. The International HRA Empirical Study Lessons Learned from Comparing HRA Methods Predictions to HAMMLAB Simulator Data. NUREG-2127, US Nuclear Regulatory Commission, Washington DC, USA, 2014.
21. Forester J, Liao H, Dang VN, et al. The US HRA Empirical Study - Assessment of HRA Method Predictions against Operating Crew Performance on a US Nuclear Power Plant Simulator. NUREG-2156, US Nuclear Regulatory Commission, Washington DC, USA, 2016.
22. Johanson G, Jonsson S, Bladh K, et al. Exam-HRA Summary. Nordic PSA Group (NPSAG), Report 11-004-01, 2015.
23. He X. Dependences in HRA. Nordic PSA Group (NPSAG), Report 41-001-01, 2016.
24. EPRI/USNRC. Fire Human Reliability Analysis Guidelines – Final Report. NUREG-1921, Electric Power Research Institute (EPRI), Palo Alto, CA and U.S. Nuclear Regulatory Commission Washington DC, USA, 2012.
25. Julius JA. A Preliminary Approach to Human Reliability Analysis for External Events with a Focus on Seismic. EPRI 1025294, Palo Alto, CA, 2012.
26. Löffler H and Raimond E. Technical report ASAMPSA_E/ WP40 / D40.7/ 2017-39 volume 1. Rapport IRSN/PSN-RES-SAG/2017-00026, 2017.
27. Reer B, Dang VN and Hirschberg S. The CESA method and its application in a plant-specific pilot study on errors of commission. *Reliab Eng Syst Saf* 2004, 83(2): 187-205.
28. Kolaczkowski A, Forester J, Lois E, et al. Good practices for implementing Human Reliability Analysis (HRA). NUREG-1792, U.S. Nuclear Regulatory Commission, Washington DC, USA, 2005.
29. .Bye A, Laumann K, Taylor C, et al. The Petro-HRA Guideline. Version 1, Institute for Energy Technology, ISBN: 978-82-7017-901-5 (printed), 978-82-7017-902-02 (electronic), 2017.
30. Pandya D, Podofillini L, Emert F, et al. Developing the foundations of a cognition-based human reliability analysis model via mapping task types and performance-influencing factors: Application to radiotherapy. *P I Mech Eng O-J Ris* 2018, 232(1):3-37.
31. Whaley AM, Xing J, Boring RL, et al. Cognitive basis for Human Reliability Analysis. NUREG-2114, US Nuclear Regulatory Commission, Washington DC, USA, 2016.
32. Mkrtchyan L, Podofillini L and Dang VN. Bayesian belief Networks for Human reliability analysis: a review of applications and gaps. *Reliab Eng Syst Saf* 2015, 139:1-16.
33. Trucco P, Cagno E, Ruggeri F, et al. A Bayesian Belief Network modelling of organisational factors in risk analysis: A case study in maritime transportation. *Reliab Eng Syst Saf* 2008, 93(6):845-856.
34. Ale BJ, Bellamy LJ, Van der Boom R, et al. Further development of a causal model for air transport safety (CATS): building the mathematical heart. *Reliab Eng Syst Saf* 2009, 94(9): 1433-1441.
35. Groth KM and Mosleh A. Deriving causal Bayesian networks from human reliability analysis data: A methodology and example mode. *P I Mech Eng O-J Ris* 2012, 226(4):361-

379.

Appendix B

This appendix contains the numerical results from the sensitivity analysis on the choice of priors for the Bayesian variability model (with lognormal distributions) presented in Chapter 2.

Table B.1. Numeric results from sensitivity analysis on choice of priors for the lognormal variability model as shown in Figure 2.7 (Case 1, target HEP variability distribution with median = 5e-2, mean = 5.46e-2, and error factor = 2).

	Prior distribution	**Mean**	**Median**	**5th perc**	**95th perc**	**EF**
No evidence (marginal priors)	Diffuse	7.44e-02	3.35e-03	2.01e-05	4.98e-01	157.39
	Low mean	4.06e-02	4.75e-03	8.11e-05	2.21e-01	52.14
	High mean	2.23e-01	1.23e-01	2.10e-03	7.92e-01	19.40
	Good mean	1.20e-01	3.43e-02	5.21e-04	5.59e-01	32.75
	Low mean, with sigma	1.61e-02	4.75e-03	4.13e-04	6.14e-02	12.19
	High mean, with sigma	2.71e-01	1.96e-01	3.43e-02	7.92e-01	4.81
	Good mean, with sigma	1.11e-01	4.86e-02	3.76e-03	4.43e-01	10.85
N_F=10, 1 failure	Diffuse	7.47e-02	2.15e-02	1.63e-04	3.51e-01	46.42
	Low mean	4.97e-02	9.55e-03	1.29e-04	2.48e-01	43.79
	High mean	1.55e-01	9.77e-02	7.56e-03	4.98e-01	8.11
	Good mean	9.06e-02	3.85e-02	8.30e-04	3.94e-01	21.80
	Low mean, with sigma	3.67e-02	1.92e-02	1.87e-03	1.23e-01	8.11
	High mean, with sigma	1.56e-01	1.23e-01	2.72e-02	4.43e-01	4.04
	Good mean, with sigma	8.91e-02	5.46e-02	8.50e-03	2.78e-01	5.72
N_F =50, 2 failures	Diffuse	3.73e-02	1.52e-02	4.13e-04	1.38e-01	18.31
	Low mean	3.12e-02	1.07e-02	3.27e-04	1.23e-01	19.40
	High mean	7.15e-02	5.46e-02	9.55e-03	1.96e-01	4.53
	Good mean	4.53e-02	2.42e-02	1.48e-03	1.38e-01	9.66
	Low mean, with sigma	2.75e-02	1.71e-02	2.98e-03	8.70e-02	5.40
	High mean, with sigma	7.08e-02	5.46e-02	1.52e-02	1.75e-01	3.39
	Good mean, with sigma	4.42e-02	3.05e-02	5.99e-03	1.23e-01	4.53
N_F =200, 11 failures	Diffuse	5.24e-02	3.85e-02	3.35e-03	1.38e-01	6.43
	Low mean	4.74e-02	3.05e-02	1.67e-03	1.56e-01	9.66
	High mean	6.27e-02	5.46e-02	1.20e-02	1.38e-01	3.39
	Good mean	5.38e-02	3.85e-02	4.75e-03	1.38e-01	5.40
	Low mean, with sigma	4.97e-02	3.85e-02	1.07e-02	1.23e-01	3.39
	High mean, with sigma	6.30e-02	5.46e-02	1.71e-02	1.38e-01	2.85
	Good mean, with sigma	5.50e-02	4.33e-02	1.20e-02	1.38e-01	3.39
N_F =1000, 58 failures	Diffuse	5.75e-02	4.86e-02	1.07e-02	1.38e-01	3.59
	Low mean	5.62e-02	4.33e-02	8.50e-03	1.38e-01	4.04
	High mean	5.95e-02	5.46e-02	1.52e-02	1.23e-01	2.85
	Good mean	5.76e-02	4.86e-02	1.20e-02	1.38e-01	3.39
	Low mean, with sigma	5.67e-02	4.86e-02	1.35e-02	1.38e-01	3.20
	High mean, with sigma	5.96e-02	4.86e-02	1.71e-02	1.23e-01	2.69
	Good mean, with sigma	5.79e-02	4.86e-02	1.52e-02	1.23e-01	2.85

Table B.2. Numeric results from sensitivity analysis on choice of priors for the lognormal variability model as shown in Figure 2.9 (Case 2, target HEP variability distribution with median = 5e-3, mean = 6.25e-3, and error factor = 3).

	Prior distribution	**Mean**	**Median**	**5th perc**	**95th perc**	**EF**
No evidence (marginal priors)	Diffuse	7.44e-02	3.35e-03	2.01e-05	4.98e-01	157.39
	Low mean	1.25e-02	5.86e-04	2.26e-05	4.33e-02	43.79
	High mean	1.20e-01	3.43e-02	5.21e-04	5.59e-01	32.75
	Good mean	4.06e-02	4.75e-03	8.11e-05	2.21e-01	52.14
	Low mean, with sigma	1.68e-03	5.21e-04	4.04e-05	6.73e-03	12.92
	High mean, with sigma	1.11e-01	4.86e-02	3.76e-03	4.43e-01	10.85
	Good mean, with sigma	1.61e-02	4.75e-03	4.13e-04	6.14e-02	12.19
N_F =10, 0 failures	Diffuse	1.91e-02	1.32e-03	2.85e-05	8.70e-02	55.26
	Low mean	7.11e-03	6.58e-04	3.59e-05	1.92e-02	23.10
	High mean	5.23e-02	1.92e-02	6.58e-04	2.21e-01	18.31
	Good mean	2.18e-02	3.76e-03	1.29e-04	8.70e-02	25.95
	Low mean, with sigma	1.57e-03	5.21e-04	5.09e-05	5.99e-03	10.85
	High mean, with sigma	4.11e-02	2.15e-02	2.36e-03	1.38e-01	7.65
	Good mean, with sigma	1.06e-02	4.23e-03	3.68e-04	3.85e-02	10.24
N_F =50, 0 failures	Diffuse	5.22e-03	7.39e-04	3.59e-05	2.15e-02	24.48
	Low mean	2.75e-03	5.86e-04	5.09e-05	8.50e-03	12.92
	High mean	1.85e-02	8.50e-03	7.39e-04	6.14e-02	9.11
	Good mean	8.46e-03	2.66e-03	1.83e-04	3.05e-02	12.92
	Low mean, with sigma	1.32e-03	5.21e-04	5.72e-05	4.75e-03	9.11
	High mean, with sigma	1.65e-02	9.54e-03	1.32e-03	5.46e-02	6.43
	Good mean, with sigma	6.05e-03	2.98e-03	2.92e-04	2.15e-02	8.60
N_F =200, 2 failures	Diffuse	1.05e-02	5.34e-03	3.27e-04	3.43e-02	10.24
	Low mean	7.17e-03	2.66e-03	1.83e-04	2.42e-02	11.50
	High mean	1.45e-02	9.55e-03	1.05e-03	4.33e-02	6.43
	Good mean	1.05e-02	5.34e-03	4.13e-04	3.43e-02	9.11
	Low mean, with sigma	5.36e-03	3.35e-03	4.64e-04	1.71e-02	6.06
	High mean, with sigma	1.40e-02	1.07e-02	2.36e-03	3.85e-02	4.04
	Good mean, with sigma	9.44e-03	6.73e-03	1.18e-03	2.72e-02	4.81
N_F =1000, 8 failures	Diffuse	8.10e-03	5.34e-03	5.86e-04	2.15e-02	6.06
	Low mean	7.23e-03	4.23e-03	3.68e-04	2.15e-02	7.65
	High mean	9.07e-03	6.73e-03	1.05e-03	2.15e-02	4.53
	Good mean	8.10e-03	5.34e-03	5.86e-04	2.15e-02	6.06
	Low mean, with sigma	6.83e-03	5.34e-03	1.32e-03	1.71e-02	3.59
	High mean, with sigma	8.99e-03	7.56e-03	2.10e-03	2.15e-02	3.20
	Good mean, with sigma	7.90e-03	5.99e-03	1.67e-03	1.92e-02	3.39

Appendix C

This appendix shows the numerical results from the application of the Bayesian variability model (with lognormal distributions) to case study in Chapter 2.

Table C.1. Numeric results from the application of the lognormal variability model on real simulator data taken from Groth et al. 2014 [1] shown in Figure 2.11.

	Model	**Mean**	**Median**	**5th perc.**	**95th perc.**	**EF**
Context A	Beta-binomial	4.46e-03	3.57e-04	1.70e-06	1.74e-02	101.16
N_F =4, 0 failures	Variability model	9.13e-03	3.87e-04	2.35e-06	2.61e-02	105.34
Context B	Beta-binomial	2.00e-01	1.68e-01	2.05e-02	5.23e-01	5.05
N_F =4, 1 failure	Variability model	1.67e-01	8.80e-02	8.03e-04	6.15e-01	7.68
Context B bis	Beta-binomial	5.64e-01	5.67e-01	2.33e-01	8.50e-01	1.91
N_F =4, 3 failures	Variability model	3.54e-01	3.22e-01	1.74e-02	8.50e-01	6.99
Context C	Beta-binomial	8.60e-01	9.22e-01	5.67e-01	1.00e-00	1.33
N_F =4, 4 failures	Variability model	7.36e-01	7.84e-01	2.33e-01	1.00e-00	2.07
Context D	Beta-binomial	1.41e-03	4.35e-05	4.66e-07	3.18e-03	82.62
N_F =3, 0 failure	Variability model	2.62e-03	5.11e-05	5.94e-07	3.74e-03	79.34

Table C.2. Numeric results from the application of the lognormal variability model on real simulator data taken from Jung et al. 2018 [2] shown in Figure 2.11.

	Model	**Mean**	**Median**	**5th perc.**	**95th perc.**	**EF**
RP-manipulation	Beta-binomial	4.87e-02	4.99e-02	3.61e-02	6.37e-02	1.33
N_F =830, 40 failures	Variability model	4.83e-02	3.92e-02	8.41e-03	1.12e-01	3.65
RP-procedure	Beta-binomial	6.09e-03	4.77e-03	8.03e-04	1.61e-02	4.47
N_F =253, 1 failure	Variability model	5.94e-03	2.49e-03	7.07e-05	1.89e-02	16.35
RP-step	Beta-binomial	6.21e-02	5.87e-02	2.41e-02	1.12e-01	2.16
N_F =71, 4 failures	Variability model	5.80e-02	3.61e-02	1.41e-03	1.68e-01	10.91
SI-diagnosis	Beta-binomial	1.73e-02	9.12e-03	1.87e-04	6.37e-02	18.46
N_F =30, 0 failures	Variability model	1.68e-02	3.18e-03	1.78e-05	6.91e-02	62.23
SI-diagnosis	Beta-binomial	4.80e-02	3.92e-02	6.08e-03	1.22e-01	4.47
N_F =30, 1 failures	Variability model	4.46e-02	1.89e-02	2.03e-04	1.68e-01	28.83

References

1. Groth KM, Smith CL and Swiler LP. A Bayesian method for using simulator data to enhance human error probabilities assigned by existing HRA methods. *Reliab Eng Syst Saf* 2014, 128(Supplement C):32-40.
2. Jung W, Park J, Kim Y, et al. HuREX – A framework of HRA data collection from simulators in nuclear power plants. *Reliab Eng Syst Saf* 2020, 194:106235.

Appendix D

The table in this appendix complements the case study performed in Chapter 3 (Section 3.4).

Table D.1. List of teamwork competences and the associated metrics (taxonomy from Skjerve and Holmgren (2016) [1] used to categorize crew behaviors emerging from the case study (descriptions for each behavioral category are provided in Table 3.5).

Behavioral categories	Associated teamwork competences (dimensions and metrics)	
Progress through procedures: • “Sequential” • “Adaptive”	LEADERSHIP	- Analytical competence
		- Enforcing adherence to standards for plant and personnel safety (e.g. operational plans, documents)
		- Behaving as a good example for subordinates
	ATTITUDE	- Conscientious and commitment to quality
Adherence to / interpretation of procedures: • “Beyond / Proactive” • “Close / Reactive”	COORDINATION	- Proactivity: think ahead possibilities for optimizing activities
	LEADERSHIP	- Encourage out-of-the-box thinking if needed
	DECISION MAKING	- Thinking outside the box: regularly considering the situation at hand from different perspective
		- Understanding the overall goal and which decision(s) should aim at achieving
	ATTITUDE	- Uphold a questioning attitude and willingness to consider a situation from multiple perspectives
Diversity of information sources: • “Diverse cues” • “Prescribed cues”	LEADERSHIP	- Ensuring that preconditions exist for successful task execution
	COORDINATION	- Proactivity: collecting information that may be useful at later stages
	DECISION MAKING	- Proactively determining how to verify the consequences/adequacy of a decision
		- Acknowledging and proactively addressing uncertainties
	SITUATION AWARENESS	- Managing periods with incomplete/insufficient/uncertain information: distinguish facts from interpretations
Monitoring indications when reacting to anomalies: • “Follow-up trends” • “Focus only on initial deviations”	COORDINATION	- Timely updating on progress and deviations
	SITUATION AWARENESS	- Attending to details to identify unexpected states/occurrences and follow up on these
		- Monitoring control-board indications frequently
		- Addressing process deviations immediately, as well as important indications and trends
Role awareness: • “Adhering” • “Diverging”	LEADERSHIP	- Maintaining a global, stand-back, overview
		- Monitoring sub-ordinates and colleagues
	INTERPERS. COMPETENCE	- Built trust, treat colleagues with respect
		- Familiarity with the work organization, roles & responsibilities, as well as with individuals
		- Acknowledging that different roles have different authority associated (leadership, followership)
		- Mastering negotiation and conflict resolution
	SITUATION AWARENESS	- Ensuring (or helping to ensure) that someone on the shift always uphold a global overview
Progression in decision making: • “Prioritizing, fast decision maker”	COORDINATION	- Clarifying operational goals and the associated tasks, incl. addressing inter-dependencies
		- Summarizing and documenting plans, goals, tasks, and deviations on a joint surface
	LEADERSHIP	- Setting well-defined, realistic goals
	DECISION MAKING	- Prioritize safety goals and concerns

• “Hesitating, slowly building up”		- Stop-Think-Act-Reflect when needed, develop a tactic/strategy for how to achieve the performance goal
		- Develop a tactic/strategy for how to achieve performance goal.
	SITUATION AWARENESS	- Making sense of the situation based on a working mental model of the process system
		- Ability to make sense of the operational situation “on-the-fly”
Operator involvement: • “All are involved” • “Some involved, some passive”	COORDINATION	- Mutual performance monitoring and provision of needed support, to the extent possible
	DECISION MAKING	- Ensuring that crew members are adequately involved
	INTEPERS. COMPETENCE	- Assess if colleagues need assistance
		- Follow up on colleagues in situations where they do not provide any information
		- Contributing to ensure that the crew keeps functioning as a team, even under trying conditions
	ATTITUDE	- Engaging constructively in task performance
Resource optimization during scenario: • “Flexible redistribution” • “Rigid”	LEADERSHIP	- Delegating tasks
	COORDINATION	- Being ready for adapting performance on-the-fly, engaging back-up behavior
		- Thinking ahead for extra resources
	ATTITUDE	- Conservative attitude: safety concerns pervade all thinking and decision making processes
		- Mental preparedness for the unforeseen/unexpected: willingness to adapt performance
	SITUATION AWARENESS	- Demonstrating readiness to re-interpret information in light of new insights/events
Team orientation in decision making: • “Collective” • “Non-inclusive”	COMMUNICATION	- Upholding continuous communication during complex situations to promote collective sense-making
	LEADERSHIP	- Developing strategies based on consultations with subordinates
		- During emergencies: mastering a more authoritarian leadership style
	DECISION MAKING	- Less participatory approach when information is limited/incomplete and time pressure higher
	INTERPERS. COMPETENCE	- Recognizing the achievements of colleagues
	ATTITUDE	- Team orientation
	SITUATION AWARENESS	- Ability as a team to pool and assess information to make sense of the occurrences
		- Ensuring that updates, briefings and problem solving meetings are held when necessary
Adherence to communication and meeting protocol: • “Adhering” • “Diverging”	COMMUNICATION	- Communicating in an assertive way: concise, clear and calm manner
		- Communicating using required standards when giving orders and sharing safety-critical information
		- Three-way communication
		- Phonetic alphabet and tag numbers, especially when communicating over the phone
		- Communicating in such a way that there is never doubt
		- Adapting communication to the receiver(s)'s competencies
		- Active listening and follow up/verify/provide feedback
		- Using robust, “stress-resistant”, communication practices (e.g. more information channels)
	LEADERSHIP	- Announcing strategies and goals clearly
		- Giving orders clearly and follow-up on ask execution continuously
	COORDINATION	- Carry out pre-job briefings when required/needed
	SITUATION AWARENESS	- Informing colleagues when initiating important tasks

References

1. Skjerve AB and Holmgren L. An investigation of Teamwork Competence Requirements in Nuclear Power Plant Control-Room Crews across Operational States – a Field Study. HWR-1107. Halden, Norway: OECD Halden Reactor Project, 2016.

Appendix E

This appendix complements the sensitivity analysis on the application of the Bayesian hierarchical model presented in Chapter 3.

Table E.1 shows both datasets and numerical results relevant to the plot in Figure 3.6 (subsection 3.4.2), where the hierarchical beta-binomial model with crew behavioral groups (Figure 3.4, left, in subsection 3.3.3) was tested with varying number of groups and degrees of variability across groups.

Table E.2 presents the results from both variability models (i.e. the hierarchical with crew behavioral groups and the continuous variability formulation, respectively left and right in Figure 3.4) tested with alternative choices of parametric distribution (i.e. $\boldsymbol{f_F}(p_{c|\boldsymbol{F}}|\boldsymbol{\theta_F})$ in the hierarchical with groups, $\boldsymbol{f_F}(p_{ij|\boldsymbol{F}}|\boldsymbol{\theta_F})$ in the continuous formulation, see subsection 3.2), which are listed in the following :

- lognormal distribution: $p\sim\mathrm{LN}(\mu,\sigma^2)$, with hyper-priors: diffuse $\pi_0(\mu)$ between log(1e-5) and log(1), diffuse $\pi_0(\sigma)$ between 0 and 5;
- logistic-normal distribution: $p\sim\mathrm{P}(\mathrm{N}(\mu,\sigma^2))$, with hyper-priors: diffuse $\pi_0(\mu)$ between logit(1e-5) and logit(1), diffuse $\pi_0(\sigma)$ between 0 and 5.

Table E.1. Numerical results from the sensitivity analysis on the hierarchical model with varying number of and degrees of variability across behavioral groups (Figure 3.6).

Case	Failure data (group-specific $E[p_c]$)	Mean	5th	50th	95th	EF
Case study (large variability)	“7 groups”: k_1/N_1=0/1, k_2/N_2=0/6, k_3/N_3=1/2, k_4/N_4=2/4, k_5/N_5=8/9, k_6/N_6=3/3, k_7/N_7=1/2 ($E[p_1]$: 3.8e-01, $E[p_2]$: 1.9e-01, $E[p_3]$: 5.1e-01, $E[p_4]$: 5.1e-01, $E[p_5]$: 7.8e-01, $E[p_6]$: 7.5e-01, $E[p_7]$: 5.1e-01)	5.1e-01	3.9e-02	5.2e-01	9.6e-01	4.9
	“4 groups”: k_1/N_1=0/7, k_2/N_2=2/4, k_3/N_3=1/2, k_4/N_4=12/14 ($E[p_1]$: 1.5e-01, $E[p_2]$: 4.9e-01, $E[p_3]$: 4.8e-01, $E[p_4]$: 7.8e-01)	4.7e-01	1.4e-02	4.6e-01	9.6e-01	8.3
Artificial data (less variability)	“7 groups”: k_1/N_1=0/1, k_2/N_2=4/6, k_3/N_3=1/2, k_4/N_4=2/4, k_5/N_5=5/9, k_6/N_6=2/3, k_7/N_7=1/2 ($E[p_1]$: 4.6e-01, $E[p_2]$: 5.9e-01, $E[p_3]$: 5.2e-01, $E[p_4]$: 5.2e-01, $E[p_5]$: 5.5e-01, $E[p_6]$: 5.8e-01, $E[p_7]$: 5.2e-01)	5.3e-01	1.7e-01	5.3e-01	8.7e-01	2.2
	“4 groups”: k_1/N_1=4/7, k_2/N_2=2/4, k_3/N_3=1/2, k_4/N_4=8/14 ($E[p_1]$: 5.6e-01, $E[p_2]$: 5.2e-01, $E[p_3]$: 5.3e-01, $E[p_4]$: 5.6e-01)	5.4e-01	1.6e-01	5.4e-01	8.9e-01	2.4

Table E.2. Alternative variability functions tested for the variability models: lognormal, $p \sim LN(\mu,\sigma^2)$; logistic-normal, $p \sim P(N(\mu,\sigma^2))$. Prior distributions: diffuse $\pi_0(\mu)$ between log(1e-5) and log(1) for the lognormal formulations, between logit(1e-5) and logit(1) for logistic-normal formulations; diffuse $\pi_0(\sigma)$ between 0 and 5 (as recommended in Kelly and Smith, 2011 [1]).

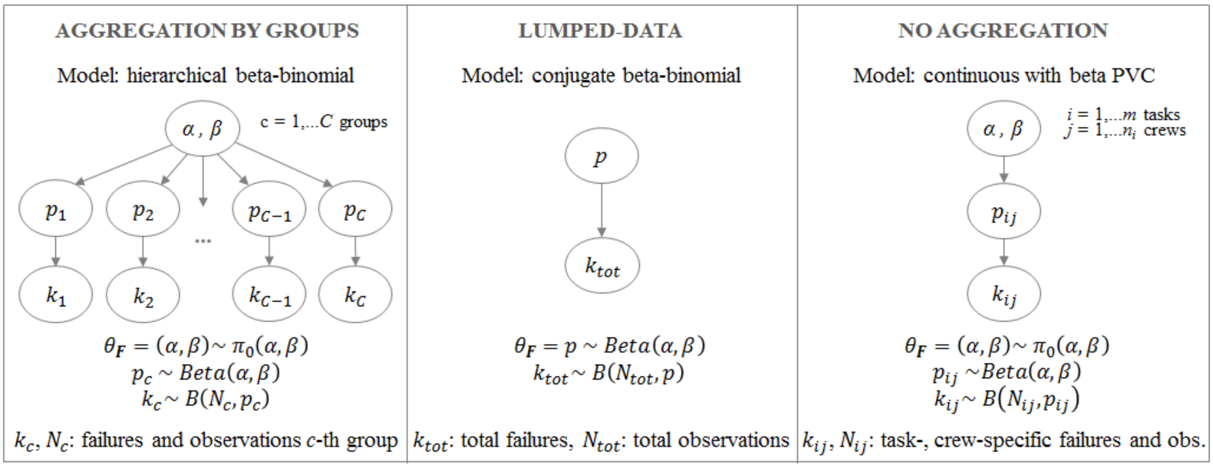
Model (variability function)	Dataset (group-specific $E[p_c]$ for hierarchical model)	Mean	5th	50th	95th	EF
Continuous (lognormal PVC)	Case study (Table 3.6 in Section 3.4)	4.9e-01	1.2e-01	4.9e-01	9.1e-01	2.8
Continuous (logistic-normal PVC)	Case study (Table 3.6 in Section 3.4)	5.5e-01	5.7e-03	6.0e-01	1.0e-00	13.2
Hierarchical with groups (lognormal-binomial)	Case study (large variability) – “7 groups” (Table E.1) ($E[p_1]$: 2.5e-01, $E[p_2]$: 1.3e-01, $E[p_3]$: 4.0e-01, $E[p_4]$: 4.3e-01, $E[p_5]$: 7.9e-01, $E[p_6]$: 7.4e-01, $E[p_7]$: 4.0e-01)	3.4e-01	8.0e-03	2.9e-01	8.7e-01	10.5
	Case study (large variability) – “4 groups” (Table E.1) ($E[p_1]$: 7.8e-02, $E[p_2]$: 4.1e-01, $E[p_3]$: 3.6e-01, $E[p_4]$: 8.0e-01)	2.5e-01	1.1e-03	1.6e-01	8.2e-01	28.0
	Artificial data (less variability) – “7 groups” (Table E.1) ($E[p_1]$: 4.3e-01, $E[p_2]$: 5.6e-01, $E[p_3]$: 4.9e-01, $E[p_4]$: 4.9e-01, $E[p_5]$: 5.2e-01, $E[p_6]$: 5.4e-01, $E[p_7]$: 4.9e-01)	4.9e-01	1.2e-01	4.9e-01	8.5e-01	2.6
	Artificial data (less variability) – “4 groups” (Table E.1) ($E[p_1]$: 5.3e-01, $E[p_2]$: 4.9e-01, $E[p_3]$: 4.8e-01, $E[p_4]$: 5.4e-01)	4.8e-01	6.8e-02	4.8e-01	8.7e-01	3.6
Hierarchical with groups (logistic-normal-binomial)	Case study (large variability) – “7 groups” (Table E.1) ($E[p_1]$: 2.6e-01, $E[p_2]$: 9.3e-02, $E[p_3]$: 5.0e-01, $E[p_4]$: 5.0e-01, $E[p_5]$: 8.5e-01, $E[p_6]$: 8.6e-01, $E[p_7]$: 5.0e-01)	4.9e-01	3.8e-03	4.9e-01	1.0e-00	16.2
	Case study (large variability) – “4 groups” (Table E.1) ($E[p_1]$: 7.1e-02, $E[p_2]$: 4.9e-01, $E[p_3]$: 4.8e-01, $E[p_4]$: 8.3e-01)	4.4e-01	1.3e-03	3.9e-01	1.0e-00	27.8
	Artificial data (less variability) – “7 groups” (Table E.1) ($E[p_1]$: 5.0e-01, $E[p_2]$: 5.8e-01, $E[p_3]$: 5.4e-01, $E[p_4]$: 5.4e-01, $E[p_5]$: 5.5e-01, $E[p_6]$: 5.7e-01, $E[p_7]$: 5.4e-01)	5.4e-01	2.2e-01	5.5e-01	8.3e-01	2.0
	Artificial data (less variability) – “4 groups” (Table E.1) ($E[p_1]$: 5.6e-01, $E[p_2]$: 5.4e-01, $E[p_3]$: 5.4e-01, $E[p_4]$: 5.6e-01)	5.5e-01	1.4e-01	5.6e-01	9.0e-01	2.6

References

1. Kelly DL and Smith CL. Bayesian Inference for Probabilistic Risk Assessment: A Practitioner’s Guidebook. London, UK: Springer-Verlag, 2011.

Appendix F

This appendix provides the “Just Another Gibbs Sampler” (JAGS, [1]) code developed for the implementation of the three Bayesian models compared in Figure 3.4, Section 3.3 (the figure is reported here to support code interpretation). JAGS is a software based on Markov Chain Monte Carlo (MCMC) simulation to approximate the solution of the posterior distributions of the Bayesian models. The JAGS models are run in R programming environment via the “runjags” library [2]. The beta distribution, i.e. Beta(α,β), of the hierarchical models (i.e. the beta-binomial with behavioral groups, left, and the continuous variability formulation, right) was reparametrized in terms of mean (μ) and a dispersion measure (i.e. the “concentration”, κ, or the “sample size”, V) in order to improve the computational efficiency, as recommended in the Bayesian literature [3-5].



<u>Hierarchical beta-binomial model (Figure 3.4, left)</u>

```
model = "
        model {
                for (i in 1:Ntotal) {  # Ntotal: total crew observations, x.obs: failures
                        x.obs[i] ~ dbin(HEP[s[i]], n[i])  # s: group index, i: crew index
                }
                for (s in 1:Nsubj) {  # Nsubj: total number of groups/patterns
                        HEP[s] ~ dbeta(alpha,beta) T(1e-5,0.99999) # population variability
                }
                HEP.pred ~ dbeta(alpha,beta) T(1e-5,0.99999) # HEP predictive posterior
                alpha = U*V  # reparametrization: U=A/(A+B)=mean of Beta pdf
                beta = (1-U)*V  # reparametrization: V = A+B = sample size
                U ~ dbeta(1, 1) T(1e-5,0.99999)  # diffuse hyperprior (mean)
                V ~ dunif(0.01, 10)   # diffuse hyperprior (sample size)
        }"
```

Conjugate beta-binomial model (Figure 3.4, center)

```
model = "
        model {
                x.obs ~ dbin(HEP, n) # x.obs: total failures, n: total observations (lumped data)
                HEP ~ dbeta(1,1) T(1e-5,0.99999) # diffuse prior
        }"
```

Hierarchical with beta population variability curve (PVC, Figure 3.4, right)

```
model = "
        model {
                for (i in 1:Ntotal) {  # Ntotal: total crew observations, x.obs: crew failures
                        x.obs[i] ~ dbin(HEP[s[i]], n[i])  # i: crew index
                        HEP[i] ~ dbeta(alpha,beta) T(1e-5,0.99999) # population variability
                }
                HEP.pred ~ dbeta(alpha,beta) T(1e-5,0.99999) # HEP predictive posterior
                alpha = ω*(κ-2)+1  # reparametrization: mean and concentration of Beta pdf
                beta = (1- ω)*(κ-2)+1
                ω ~ dbeta(1, 1) T(1e-5,0.99999)  # diffuse hyperprior (mean)
                κ = κminus2 +2  # dummy variable (avoid zero values in shape parameters)
                κminus2 ~ dunif(0.01, 10)  # diffuse hyperprior (concentration)
        }"
```

The three models were run using the following set of JAGS parameters (for further details on JAGS parameters, see references [4-5]):

- MCMCseed = 222
- n_burnin = 5000 # *number of steps to "burn-in"*
- n_adapt = 5000 # *number of steps to "adapt"*
- n_chains = 3 # *number of chains to run*
- n_iter = 100000 # *total number of steps across chains to save*
- thinSteps = 1 # *number of steps to "thin" (1=keep every step)*
- nPerChain = ceiling(n_iter/n_chains) # *steps per chain*

References

1. Plummer M. JAGS: A Program for Analysis of Bayesian Graphical Models using Gibbs Sampling. In: *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, DSC 2003, March 20-22, Vienna, Austria.
2. Denwood MJ. Runjags: An R Package Providing Interface Utilities, Model Templates, Parallel Computing Methods and Additional Distributions for MCMC Models in JAGS. J. Stat. Softw. 2016, 71(9): 1–25.
3. Kelly DL and Smith CL. Bayesian Inference for Probabilistic Risk Assessment: A Practitioner's Guidebook. London, UK: Springer-Verlag, 2011.
4. Gelman A, Carlin J, Stern H, et al. *Bayesian Data Analysis*. 2nd edition. Chapman and Hall/CRC, 2003.
5. Kruschke JK. *Doing Bayesian Data Analysis*. 2nd edition. Academic Press, 2015.

Appendix G

This appendix complements the numerical demonstration of the Bayesian two-stage model presented in Chapter 4.

Table G.1. Numerical results from the sensitivity analysis to expert biases (subsection 4.3.3.3).

Case	Confidence level	Type of bias (bias factor)	Mean	5th	50th	95th	EF
10 tasks	EF_i= 1	Unbiased (b=1)	5.10e-02	1.71e-02	4.31e-02	1.09e-01	2.5
		Conservative (b=10)	4.51e-01	1.43e-01	4.18e-01	8.66e-01	2.5
		Optimistic (b=0.1)	5.00e-03	1.79e-03	4.45e-03	1.12e-02	2.5
	EF_i= 3	Unbiased (b=1)	7.20e-02	4.13e-03	3.49e-02	2.71e-01	8.1
		Conservative (b=10)	2.08e-01	7.12e-02	1.86e-01	4.25e-01	2.4
		Optimistic (b=0.1)	2.10e-02	4.17e-04	4.90e-03	7.70e-02	13.6
	EF_i= 5 (Figure 4.7)	Unbiased (b=1)	7.70e-02	2.55e-03	3.26e-02	3.25e-01	11.3
		Conservative (b=10)	1.62e-01	4.57e-02	1.38e-01	3.61e-01	2.8
		Optimistic (b=0.1)	4.10e-02	1.95e-04	5.92e-03	2.19e-01	33.5
	EF_i= 7	Unbiased (b=1)	7.80e-02	1.79e-03	3.05e-02	3.46e-01	13.9
		Conservative (b=10)	1.43e-01	3.25e-02	1.18e-01	3.41e-01	3.2
		Optimistic (b=0.1)	5.30e-02	1.34e-04	6.72e-03	3.00e-01	47.3
50 tasks	EF_i= 1	Unbiased (b=1)	5.50e-02	2.47e-02	5.04e-02	1.03e-01	2.0
		Conservative (b=10)	5.32e-01	2.37e-01	5.10e-01	8.99e-01	1.9
		Optimistic (b=0.1)	6.00e-03	2.47e-03	5.05e-03	1.04e-02	2.1
	EF_i= 3	Unbiased (b=1)	6.70e-02	1.26e-02	4.82e-02	1.82e-01	3.8
		Conservative (b=10)	2.30e-01	1.64e-01	2.26e-01	3.08e-01	1.4
		Optimistic (b=0.1)	9.00e-03	1.16e-03	5.73e-03	2.84e-02	5.0
	EF_i= 5 (Figure 4.7)	Unbiased (b=1)	6.50e-02	1.15e-02	4.61e-02	1.79e-01	3.9
		Conservative (b=10)	1.72e-01	1.20e-01	1.70e-01	2.33e-01	1.4
		Optimistic (b=0.1)	1.40e-02	8.72e-04	6.41e-03	4.74e-02	7.4
	EF_i= 7	Unbiased (b=1)	6.40e-02	1.08e-02	4.49e-02	1.77e-01	4.0
		Conservative (b=10)	1.48e-01	1.01e-01	1.45e-01	2.02e-01	1.4
		Optimistic (b=0.1)	1.80e-02	7.27e-04	6.97e-03	6.65e-02	9.6

Table G.2. Set of HFEs from US [54] and International [55-57] Empirical Studies analysed in the case study (Section 4.4), and associated constellation of task/PSF categories ($\boldsymbol{F}$) from SACADA taxonomy [24].

ID: human failure event [54-57]	Constellation $\boldsymbol{F}$ [24]
US-HFE1A: “failure to establish feed and bleed within 45 minutes of the reactor trip, given that the crews initiate a manual reactor trip before an automatic reactor trip”	$\boldsymbol{F_1}$: task type = manipulation, PSFs: {time criticality = normal time available, extent of communication = extensive onsite, type of action = order, guidance = S.T.A.R, location = main/auxiliary control board, miscellaneous = non-standard conditions}
US-HFE1C: “failure to isolate the ruptured steam generator and control pressure below the SG PORV set-point to avoid SG PORV opening”	$\boldsymbol{F_2}$: task type = understanding the situation/problem, PSFs: {time criticality = normal time available, workload = concurrent demand, extent of communication = extensive within control room, information quality = missing, information specificity = not specific, familiarity = anomaly, diagnosis base = procedure, information integration = ambiguous}
US-HFE2A: “failure to trip the RCPs and start the PDP to prevent RCP seal loss of coolant accident (LOCA)”	$\boldsymbol{F_3}$: task type = detecting an alarm, PSFs: {time criticality = barely adequate time, workload = multiple concurrent demand, extent of communication = extensive within control room, detection mode = aware/inspection, status of alarm board = overloaded, expectation of alarm change = not expected}
US-HFE3A: “failure of crew to isolate the ruptured steam generator and control pressure below the SG PORV set point before SG PORV opening” INT-SGTR-HFE1A: “Failure to identify and isolate the ruptured steam generator”	$\boldsymbol{F_4}$: task type = detecting an alarm, PSFs: {time criticality = barely adequate time, workload = multiple concurrent demand, extent of communication = extensive within control room, detection mode = aware/inspection, status of alarm board = overloaded, expectation of alarm change = not expected}
INT-SGTR-HFE1B: “failure to identify and isolate the ruptured steam generator”	$\boldsymbol{F_5}$: task type = understanding the situation/problem, PSFs: {time criticality = barely adequate time, workload = concurrent demand, extent of communication = extensive within control room, information quality = conflicting, information specificity = not specific, familiarity = anomaly, diagnosis base = knowledge, information integration = ambiguous}
INT-SGTR-HFE2A: “failure to cool down the RCS expeditiously” INT-SGTR-HFE3A: “failure to depressurize the RCS expeditiously” INT-SGTR-HFE3B: “failure to depressurize the RCS expeditiously”	$\boldsymbol{F_6}$: task type = manipulation, PSFs: {time criticality = normal time available, workload = concurrent demand, extent of communication = extensive within control room, type of action = monitoring, guidance = procedure, location = main/auxiliary control board}
INT-SGTR-HFE2B: “failure to cool down the RCS expeditiously”	$\boldsymbol{F_7}$: task type = manipulation, PSFs: {time criticality = normal time available, workload = normal, extent of communication = normal, type of action = monitoring, guidance = procedure, location = main/auxiliary control board}
INT-SGTR-HFE4A: “failure to stop the safety injection”	$\boldsymbol{F_8}$: task type = manipulation, PSFs: {time criticality = normal time available, workload = normal, extent of communication = normal, type of action = order, guidance = procedure, location = main/auxiliary control board}
INT-SGTR-HFE5B1: “failure to close PORV block valve if it remains partially open and indications show ‘closed’”	$\boldsymbol{F_9}$: task type = understanding the situation/problem, PSFs: {time criticality = barely adequate time, workload = concurrent demand, extent of communication = normal, information quality = misleading, outcome = procedure, familiarity = anomaly, diagnosis base = knowledge, information integration = timing}
INT-SGTR-HFE5B2: “failure to close PORV block valve if it remains partially open and indications	$\boldsymbol{F_{10}}$: task type = detecting status change of indicator/alarm, PSFs: {time criticality = normal time available, extent of communication = normal, detection mode = procedure-directed check, degree of change = distinct}

show ‘open”	
INT-LOFW-HFE1A: “failure to establish Bleed and Feed before SG dryout”	**F_{11}**: task type = manipulation, PSFs: {time criticality = normal time available, workload = concurrent demand, extent of communication = normal, type of action = monitoring, guidance = procedure, location = main/auxiliary control board}
INT-LOFW-HFE1B: “failure to establish Bleed and Feed before SG dryout and depressurization of steam generator”	**F_{12}**: task type = manipulation, PSFs: {time criticality = barely adequate time, workload = multiple concurrent demand, extent of communication = extensive within control room, type of action = monitoring, guidance = S.T.A.R., location = main/auxiliary control board, miscellaneous = additional mental effort, unintuitive plant response}
INT-LOFW-HFE2B: “failure to establish Bleed and Feed within 25 minutes after SG dryout and depressurization of steam generator”	**F_{13}**: task type = manipulation, PSFs: {time criticality = normal time available, workload = normal, extent of communication = extensive within control room, type of action = monitoring, guidance = S.T.A.R., location = main/auxiliary control board, recoverability = unrecoverable}

Table G.3. Numerical results from the application to case study (Section 4.4).

Stage 1: estimation of $P_{\boldsymbol{F}}(p_{t\|\boldsymbol{F}})$						Stage 2: estimation of $P_{HFE}(HEP)$						
F	Mean	5th	50th	95th	EF	HFE	Case	Mean	5th	50th	95th	EF
F_1	3.50e-02	3.26e-05	2.47e-03	1.88e-01	76.0	US-HFE1A (Figure 4.8)	Case 1: simulator data	1.30e-02	4.29e-05	1.95e-03	6.52e-02	39.0
							Case 2: simulator data and estimate	3.10e-02	4.87e-03	2.12e-02	8.77e-02	4.2
							Results from Empirical Studies	1.00e-01	4.62e-04	5.20e-02	3.62e-01	28.0
F_2	2.34e-01	2.44e-02	2.02e-01	5.81e-01	4.9	US-HFE1C (Figure 4.8)	Case 1: simulator data	1.83e-01	4.16e-02	1.52e-01	4.30e-01	3.2
							Case 2: simulator data and estimate	1.61e-01	4.85e-02	1.37e-01	3.55e-01	2.7
							Results from Empirical Studies	3.00e-01	4.60e-02	2.72e-01	6.51e-01	3.8
F_4	1.50e-02	3.06e-05	1.37e-03	5.06e-02	40.7	US-HFE3A (Figure 4.8)	Case 1: simulator data	8.00e-03	4.00e-05	1.19e-03	3.62e-02	30.1
							Case 2: simulator data and estimate	2.00e-03	2.35e-04	1.03e-03	4.53e-03	4.4
							Results from Empirical Studies	1.25e-01	6.03e-04	6.74e-02	4.44e-01	27.1
						INT-SGTR-HFE1A	Case 1: simulator data	3.40e-02	1.38e-03	1.97e-02	1.18e-01	9.2
							Case 2: simulator data and estimate	8.00e-03	1.33e-03	5.68e-03	2.32e-02	4.2
							Results from Empirical Studies	4.90e-02	3.17e-03	3.33e-02	1.51e-01	6.9
F_6	3.80e-02	3.37e-05	2.53e-03	2.11e-01	79.1	INT-SGTR-HFE2A	Case 1: simulator data	4.70e-02	2.36e-03	3.03e-02	1.48e-01	7.9
							Case 2: simulator data and estimate	1.00e-02	1.52e-03	6.59e-03	2.74e-02	4.2
							Results from Empirical Studies	4.90e-02	3.17e-03	3.33e-02	1.51e-01	6.9
						INT-SGTR-HFE3A	Case 1: simulator data	4.70e-02	2.36e-03	3.03e-02	1.48e-01	7.9
							Case 2: simulator data and estimate	2.20e-02	3.84e-03	1.59e-02	5.86e-02	3.9
							Results from Empirical Studies	4.90e-02	3.17e-03	3.33e-02	1.51e-01	6.9
						INT-SGTR-HFE3B (Figure 4.8)	Case 1: simulator data	1.05e-01	1.66e-02	8.80e-02	2.54e-01	3.9
							Case 2: simulator data and estimate	4.00e-02	8.00e-03	3.01e-02	1.03e-01	3.6
							Results from Empirical Studies	1.07e-01	1.80e-02	8.97e-02	2.53e-01	3.8
F_7	6.10e-02	5.81e-05	1.00e-02	3.35e-01	75.9	INT-SGTR-HFE2B	Case 1: simulator data	9.00e-03	5.71e-05	2.23e-03	4.17e-02	27.0
							Case 2: simulator data and estimate	5.00e-03	7.09e-04	3.15e-03	1.36e-02	4.4
							Results from Empirical Studies	1.00e-02	9.61e-05	3.03e-03	4.56e-02	21.8
F_8	1.40e-02	2.68e-05	1.08e-03	4.76e-02	42.1	INT-SGTR-HFE4A	Case 1: simulator data	5.00e-03	3.55e-05	9.40e-04	2.13e-02	24.5
							Case 2: simulator data and estimate	1.00e-03	9.12e-04	3.85e-03	1.69e-02	4.4
							Results from Empirical Studies	1.00e-02	9.61e-05	3.03e-03	4.56e-02	21.8
F_{10}	1.50e-02	3.06e-05	1.37e-03	5.06e-02	40.7	INT-SGTR-HFE5B2 (Figure 4.8)	Case 1: simulator data	6.00e-03	3.90e-05	1.11e-03	2.83e-02	26.9
							Case 2: simulator data and estimate	2.00e-03	2.33e-04	1.02e-03	4.51e-03	4.4
							Results from Empirical Studies	1.50e-02	1.06e-04	3.73e-03	6.67e-02	25.1
F_{11}	3.80e-02	3.61e-05	2.91e-03	2.04e-01	75.1	INT-LOFW-HFE1A	Case 1: simulator data	9.00e-03	4.37e-05	1.75e-03	4.32e-02	31.4
							Case 2: simulator data and estimate	5.00e-03	6.69e-04	3.01e-03	1.34e-02	4.5
							Results from Empirical Studies	1.20e-02	1.00e-04	3.37e-03	5.51e-02	23.4

Appendix H

This appendix provides the JAGS code for the implementation of the Bayesian two-stage model presented in Chapter 4.

The code for Stage I is relevant to the configuration with lognormal population variability distribution (“lognormal PV-binomial-lognormal” in Figure 4.4 left, Section 4.3), and has been used in both the numerical tests with artificial data (subsections 4.3.3.1-4.3.3.3) and the application to case study (Section 4.4). The specific values of “mean_mu” and “mean_sigma” (i.e. in eq. 4.10, the parameters of the lognormal hyperprior $\pi_0(\mu_F)$) used in the numerical applications are reported in Sections 4.3-4.4. The code for Stage I configuration without expert estimates (Figure 4.4, center) can be derived from the code below by simply excluding the strings relevant to the incorporation of expert estimates. The code for the lumped-data model (Figure 4.4, right) can be found in Appendix F.

Concerning the code for Stage II, note that the symbol “[…]” in the last string represents a replacement for the numerical output of Stage I, i.e. the estimated HEP population variability distribution $P_{\boldsymbol{F}}(p_{t|\boldsymbol{F}})$. Note also that the HEP support is truncated below 1.0e-05 and above 1, as discussed in subsection 4.3.2. For further details on JAGS software, see Appendix F.

Stage I

```
model = "
        model {  # i= task index, Ntotal: number of task realizations across plants
                for (i in 1:Ntotal) {  # x.obs: task-specific failures
                     x.obs[i] ~ dbin(HEP[i], n[i])  # x.exp: task-specific expert estimate
                     x.exp[i] ~ dlnorm(log.HEP[i],tau.exp[i])  # lognormal error model
                     tau.exp[i] = pow((1/(1.645))*log(EF.exp[i]),-2)  # convert EFi into sigma
                     HEP[i] = exp(log.HEP[i]) # transformation log to real scale
                     log.HEP[i] ~ dnorm(mu,tau)  # population variability function
                     x.sim[i] ~ dbin(HEP[i], n[i]) # replicated datasets (for model checking)
                }
                HEP.pred = exp(log.HEP.pred)  # HEP posterior predictive
                log.HEP.pred ~ dnorm(mu,tau)
                mu ~ dnorm(mean_mu,mean_tau)  # hyperprior on mean (with ext. parameters)
                mean_tau = pow(mean_sigma,-2)
                sigma ~ dunif(0.01,5)  # hyperprior on sigma (diffuse)
                tau = pow(sigma,-2)
        }"
```

Stage II

```
model = "
        model {
                x.obs ~ dbin(HEP, n)  # n, x.exp: plant-specific failures and expert estimate
                x.exp ~ dlnorm(log.HEP,tau.exp)  # lognormal error model
        tau.exp = pow((1/(1.645))*log(EF.exp),-2)  # convert EF into sigma
        HEP = exp(log.HEP)  # transformation log to real space
        log.HEP ~ […] T(log(1E-5),log(1))  # marginal prior (output Stage I)
}"
```

Curriculum Vitae

Jan 2017 – till date	Eidgenössische Technische Hochschule Zürich (ETH Zürich, Switzerland). Doctoral Studies at the Department of Mechanical and Process Engineering (D-MAVT).
Nov 2016 – Oct 2020	Paul Scherrer Institut, Switzerland. Ph.D. Candidate at the Risk and Human Reliability Group, Laboratory of Energy Systems Analysis (LEA).
Sept 2015 – Mar 2016	University of Palermo, Italy. Post-graduate research assistant at the Dipartimento dell’Energia, Ingegneria dell'Informazione e Modelli Matematici (DEIM).
Sept 2014 – Dec 2014	Paul Scherrer Institut, Switzerland. Internship at the Risk and Human Reliability Group, Laboratory of Energy Systems Analysis (LEA).
Mar 2012 – Oct 2016	University of Palermo, Italy. Master of Science in Nuclear and Energy Engineering.
Feb 2007 – Feb 2012	University of Palermo, Italy. Bachelor of Science in Energy Engineering.
Sept 2001 – July 2006	Liceo Classico “F. Scaduto”, Bagheria, Italy. High School leaving qualification in classical studies.

List of Publications

Journal publications:

- Greco SF, Podofillini L, and Dang VN. A Bayesian two-stage approach to integrate simulator data and expert judgment in human error probability estimation. *Safety Sci* (expected submission date: July 2021).
- Greco SF, Podofillini L, and Dang VN. Crew performance variability in human error probability quantification: a methodology based on behavioral patterns from simulator data. *Proc I Mech E Part O: J Risk and Reliability* 2021, doi:10.1177/1748006X20986743.
- Greco SF, Podofillini L, and Dang VN. A Bayesian model to treat within-category and crew-to-crew variability in simulator data for Human Reliability Analysis. *Reliab Eng Syst Safe* 2021, 206:107309, ISSN 0951-8320.
- Giardina M, Greco SF, Buffa P, Dang VN, Podofillini L, and Prete G. Early-design improvement of human reliability in an experimental facility: A combined approach and application on SPES. *Safety Sci* 2019, 119:300-314, ISSN 0925-7535 (https://doi.org/10.1016/j.ssci.2018.08.008).

Conference papers:

- Greco SF, Podofillini L, and Dang VN. Modelling Crew Performance Variability in Emergency Situations from Simulator Data for Human Reliability Analysis. In: *e-Proceedings of the 30th European Safety and Reliability Conference and 15th Probabilistic Safety Assessment and Management Conference* (ESREL2020 PSAM15). ISBN: 978-981-14-8593-0.
- Greco SF, Podofillini L, and Dang VN. Crew performance variability in simulator data for Human Reliability Analysis: investigation of modelling options. In: *Proceedings of the 29th European Safety and Reliability Conference* (ESREL 2019). ISBN: 981-973-0000-00-0.
- Greco SF, Buffa P, Giardina M, Palermo G, Podofillini L, Dang VN, Esposito J, Prete G. Human Reliability Analysis to support operational planning of an experimental facility. In: *Proceedings of the 25th European Safety and Reliability Conference* (ESREL 2015), Zürich, Switzerland, September 7-10, Safety and Reliability of Complex Engineered Systems, Podofillini et al. (Eds), Taylor and Francis Group, London, ISBN 978-1-138-02879-1.
- Bruschetta S, Greco SF, Richiusa ML, Tomarchio E. An ALARA approach for designing an electron accelerator plant for industrial and research applications. SARA

2013, Prague, Czech Republic, February 17-March 2 (http://sara.fjfi.cvut.cz/pub/SARA2013).

- Buffa P, Castiglia C, Giardina M, Greco SF, Morana G. Progettazione e sviluppo del software RAD per analisi FMECA. VGR 2012, Pisa, Italia, October 3-5 (http://conference.ing.unipi.it/vgr2012).