This document is the accepted manuscript version of the following article:
Greco, S. F., Podofillini, L., & Dang, V. N. (2021). Crew performance variability in human error probability quantification: a methodology based on behavioral patterns from simulator data. Proceedings of the Institution of Mechanical Engineers, Part O: Journal or Risk and Reliability, 235(4), 637-659. https://doi.org/10.1177/1748006X20986743

Crew performance variability in human error probability quantification: a methodology based on behavioral patterns from simulator data

Salvatore F. Greco, Luca Podofillini, Vinh N. Dang

Risk and Human Reliability Group, Laboratory for Energy Systems Analysis
Paul Scherrer Institute, Villigen PSI, Switzerland

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 paper 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 paper 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.

Keywords: human reliability analysis, human performance, simulator data, uncertainty and variability, teamwork, nuclear power plant, HAMMLAB, SACADA, HuREX, Bayesian hierarchical models.

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 [25-26]. In order to explicitly address these characteristics, this paper puts 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 paper, 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 paper is structured as follows. Section 2 first introduces the concept of crew behavioral patterns to

characterize behavioral aspects in nuclear power plant operations. Section 2 then presents how the patterns are quantitatively incorporated in the model for crew performance variability in HEP estimation. Section 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 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 5, along with recommendations on the feasibility and applicability of the proposed methodology to HRA problems. Conclusions are given at closure.

2 Concepts: behavioral patterns from simulator data and variability modelling

2.1 *Behavioral patterns: definition and relationship with typical HRA quantification*

Fig. 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.

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 Fig. 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 paper 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 Fig. 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 Fig. 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 Tab. 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 Tab. 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 in this paper as set $\boldsymbol{F}$) adopted by the specific data collection taxonomy: $HEP = HEP(\boldsymbol{F})$. For instance, in Tab. 1 (from SACADA taxonomy), $\boldsymbol{F}$ 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 $\boldsymbol{F}$, the proposed methodology requires information on observed crew behaviors to populate the behavioral patterns, such as those in the last column of Tab. 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 Tab. 1. This indeed has implications on the possibility to apply the proposed model to the currently available HRA data, as further discussed in Section 5.

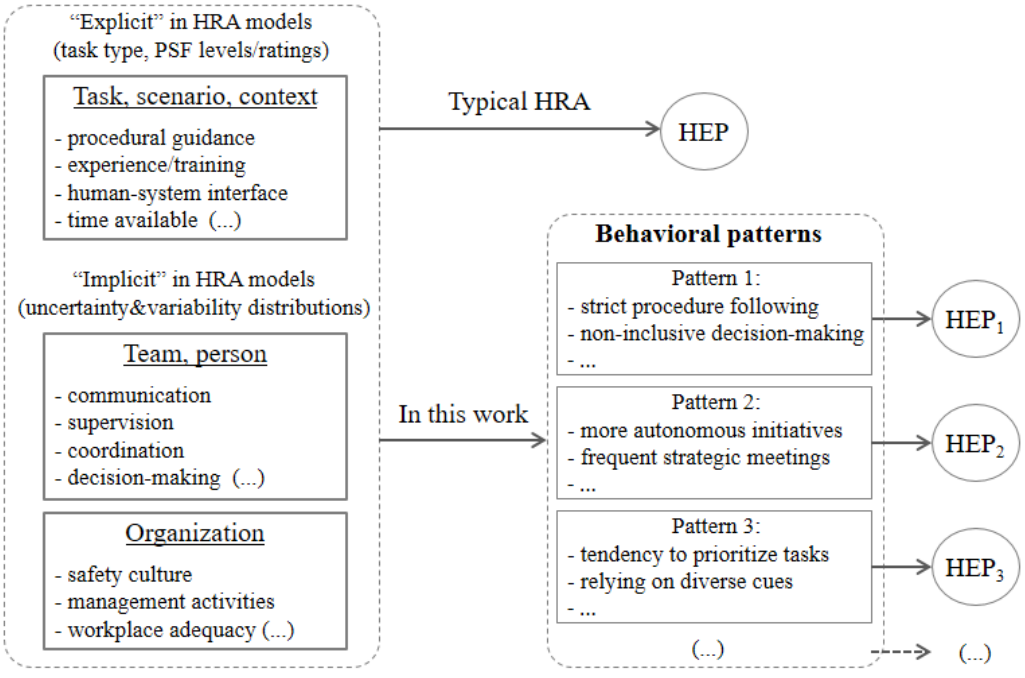

Fig. 1. Relationship between performance influencing factors (taxonomy from [45]) and behavioral patterns used in this work to represent crew performance variability in HEP quantification.

In the present work, the crew behaviors collected for a given set ***F*** (Tab. 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 Fig. 1); concerning work attitudes, the compliance to procedure indications (e.g. "strict procedure following" or "more autonomous initiatives" in Fig. 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 Fig. 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 (Fig. 1) in the variability model presented in the next Section 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 2, in order to support the understanding of model development in the remainder of this section, as well as the methodology presented in Section 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 Tab. 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 Fig. 1: "strict procedure following", "frequent strategic meetings"). In this paper, 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 Fig. 1, pattern #1: "strict procedure following & non-inclusive decision making & [...]"). In this paper, patterns are interpreted as the direct manifestation of the overall spectrum of influencing factors in Fig. 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 Fig. 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 (Section 2.2). In this paper, the set of behavioral groups emerging from data is used to model performance variability in the given ***F***.

Tab. 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	

2.2 *Using behavioral patterns in a variability model for HEP*

This subsection presents the variability model for *HEP*(***F***) (shown in Fig. 2, left) to capture HEP variability across behavioral groups (the identification of the groups will be presented in Section 3).

The model is based on the assumption that each "behavioral group" (pedix c in Fig. 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 ***F***. In this formulation, $p_{c|\boldsymbol{F}}$ represents possible outcomes of *HEP*(***F***): the HEP is intended as a variable quantity, discretized over the number of identified behavioral groups (C in Fig 2, left). The $p_{c|\boldsymbol{F}}$'s (the arrows in Fig. 2, left) are interpreted as group-specific realizations of the HEP variability in ***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:

$$p_{c|\boldsymbol{F}} \sim f_{\boldsymbol{F}}(p_{c|\boldsymbol{F}}|\boldsymbol{\theta}_{\boldsymbol{F}}) \tag{1}$$

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 Fig. 2, left) in the form of observed failures and crew observations (respectively k_c and N_c in Fig. 2, left).

In the numerical application of Section 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 (Fig. 2, center) associates all the simulator records relevant to ***F*** (the rows in Tab. 1) to a single-value HEP, $p_{\boldsymbol{F}}$, i.e. the population average over the variability within ***F***: failures and crew observations relevant to ***F*** are lumped into a single piece of evidence (respectively k_{tot} and N_{tot} in Fig. 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*(***F***) as a "continuum" of crew-, task-specific error probabilities, $p_{ij|\boldsymbol{F}}$'s (Fig. 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 ***F*** (e.g. from Tab. 1, identification and isolation of the leaks in a ISLOCA scenario). Similarly to the $p_{c|\boldsymbol{F}}$ variable in eq. 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 paper, 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 Fig. 2, right). Simply put, Greco et al. [25] assumes different failure probabilities per each crew, while the present paper per each behavioral group, aggregating different crews manifesting with similar behaviors.

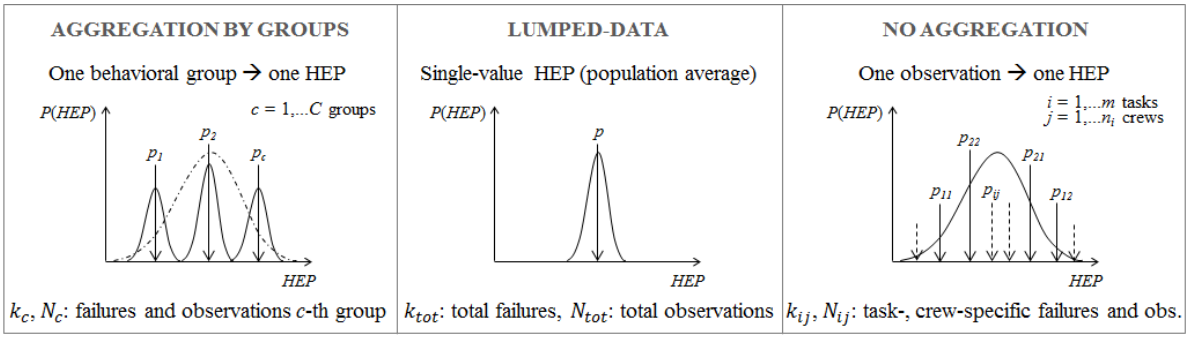

Fig. 2. Comparison of the HEP formulations and the associated data aggregation adopted by the proposed variability model (left: "aggregation by groups") and the alternative approaches tested in the case study (center: "lumped-data", with HEP as population average; right: "no aggregation", with HEP as a "continuum" of task-, crew-specific error probabilities). All the p's are intended as conditional on the given set ***F***, e.g. $p_{c|\boldsymbol{F}}$, $p_{ij|\boldsymbol{F}}$.

The unknown parameters for the three mathematical formulations of $HEP(\boldsymbol{F})$ described in this subsection (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.

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 Fig. 2, center). Compared to the quantitative approach proposed in this paper, 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 paper 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 Fig. 2, right). Compared to the continuous variability formulation, in this paper 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 Fig. 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 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 (Fig. 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 4, it is applied to a specific $\boldsymbol{F}$ characterizing a case study from literature.

The methodology comprises two blocks (Fig. 3). The first block (Fig. 3, blue box) derives the behavioral categories emerging from the simulator data relevant to the combination $\boldsymbol{F}$. The second block (Fig. 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.

3.1 *Derivation of behavioral categories from data collection*

Steps I.1-I.3 in Fig. 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 $\boldsymbol{F}$'s, where each $\boldsymbol{F}$ represents a combination of task types and PSF ratings for which data is available. The definition of representative sets $\boldsymbol{F}$ 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 $\boldsymbol{F}$ 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 Tab. 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 2.1 (e.g. in Fig. 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 paper. 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 Tab. 2 (second column), with crew behaviors (first column) adapted from [13,43]. The full list of factor-specific metrics can be found in [44].

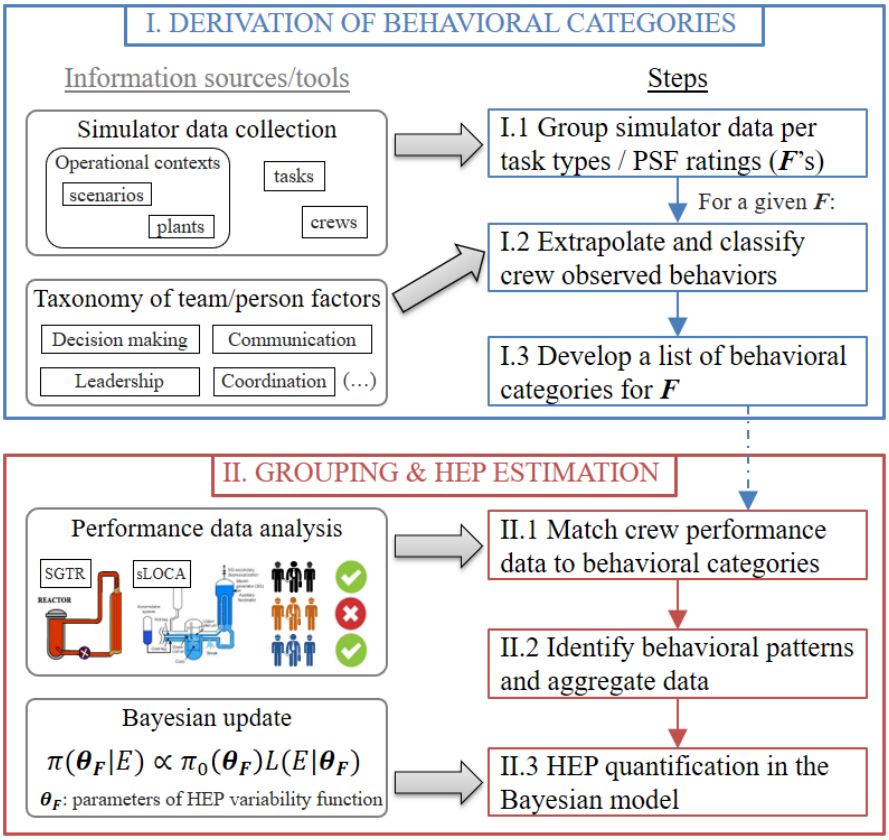

Fig. 3. Overview of the multi-step methodology to derive and use behavioral patterns in HEP quantification for a generic set ***F***.

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 Tab. 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 Tab. 2 is given in the application of steps I.1-I.3 to the case study (Section 4, Tab. 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 Tab. 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 Tab. 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 Tab. 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 Fig. 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 $\boldsymbol{F}$. 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 $\boldsymbol{F}$ 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 Tab. 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").

3.2 *Grouping crew performance data and HEP quantification*

Steps II.1-II.3 in Fig. 3 (red box) address the use of behavioral patterns to group performance data and estimate the HEP for the set $\boldsymbol{F}$:

- 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 $\boldsymbol{F}$, crew behaviors reported in performance data are analyzed and matched to the relevant behavioral categories. Examples of matching are shown in Tab. 3, with reference to the categories reported in Tab. 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 $\boldsymbol{F}$).

In step II.2, combinations of behavioral categories emerging across crew performances are identified and clustered as behavioral patterns. For instance, in Tab. 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 Tab. 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. 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 ($\boldsymbol{\theta_F}$ in eq. 1) is modelled by the so-called "prior distribution", $\pi_0(\boldsymbol{\theta_F})$ in Fig. 3. The prior is updated by the group-specific simulator data (the "evidence" E in Fig. 3) in the likelihood function, i.e. $L(E|\boldsymbol{\theta_F})$ in Fig. 3. The output of the Bayesian update (i.e. the "posterior distribution" $\pi(\boldsymbol{\theta_F}|E)$ in Fig. 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. $f_{\boldsymbol{F}}(p_{c|\boldsymbol{F}}|\boldsymbol{\theta_F})$ in eq. 1) over the posterior $\pi(\boldsymbol{\theta_F}|E)$:

$$P(HEP) = \int_{\boldsymbol{\theta_F}} f_{\boldsymbol{F}}(p_{c|\boldsymbol{F}}|\boldsymbol{\theta_F})\pi(\boldsymbol{\theta_F}|E)d\boldsymbol{\theta_F} \quad (2)$$

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

Tab. 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 Fig. 3, adapted from the case study in Section 4.

Crew observed behaviors (from Tab. 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

Tab. 3. Matching crew behaviors to behavioral patterns: examples of application of steps II.1-II.2 in Fig. 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 Fig. 3.

Performance data (Tab. 2)	Behavioral patterns and associated categories	Failures k_i, observations N_i
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" + ...	$k_i = 5$, $N_i = 6$
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" + ...	$k_i = 2$, $N_i = 5$
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" + ...	$k_i = 2$, $N_i = 4$
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" + ...	$k_i = 1$, $N_i = 10$
(...)	(...)	$k_{tot} = 22$, $N_{tot} = 50$

3.3 *Development and implementation of the Bayesian model*

In the numerical application (Section 4), the $f_{\boldsymbol{F}}(p_{c|\boldsymbol{F}}|\boldsymbol{\theta_F})$ of eq. 1 is modelled with a beta distribution ($p_{c|\boldsymbol{F}} \sim$ Beta(α, β) in Fig. 4, left) in a hierarchical beta-binomial model [51], to capture performance variability across the different behavioral groups. Accordingly, $\boldsymbol{\theta_F} = \{\alpha, \beta\}$ 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 2.2. Indeed, the evidence (failures, k_c, and crew observations, N_c, for the c-th group, Fig. 4, left) enters at group-level in the binomial likelihood function ($k_c \sim$ B(N_c, $p_{c|\boldsymbol{F}}$) in Fig.4, left) to inform the specific $p_{c|\boldsymbol{F}}$, i.e. the group-specific realization of $f_{\boldsymbol{F}}(p_{c|\boldsymbol{F}}|\boldsymbol{\theta_F})$ associated to $\boldsymbol{F}$. The $p_{c|\boldsymbol{F}}$'s are then used to infer, at population-level, the unknown $\boldsymbol{\theta_F}$ of $f_{\boldsymbol{F}}(p_{c|\boldsymbol{F}}|\boldsymbol{\theta_F})$, 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 $f_{\boldsymbol{F}}(p_{c|\boldsymbol{F}}|\boldsymbol{\theta_F})$ are discussed in subsection 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 (Fig. 4, center) to derive the single $p_{\boldsymbol{F}}$ (Fig. 2, center) from the lumped data (k_{tot} and N_{tot} in Fig. 4, center). The continuous variability formulation (Fig. 2, right) is coupled to a Bayesian model with a population variability curve (PVC in Fig. 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. $f_{\boldsymbol{F}}(p_{ij|\boldsymbol{F}}|\boldsymbol{\theta_F})$, is specialized with a beta PVC ($p_{ij|\boldsymbol{F}} \sim$ Beta(α, β) in Fig. 4, right), with $\boldsymbol{\theta_F} = \{\alpha, \beta\}$ to be inferred from the crew-, task- specific data (k_{ij} and N_{ij} in Fig. 4, right). Fig. 4 shows an overview of the three Bayesian models tested in Section 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 $\pi(\boldsymbol{\theta_F}|E)$. 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 ($\pi_0(\boldsymbol{\theta_F})$ in Fig. 4, right) 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.

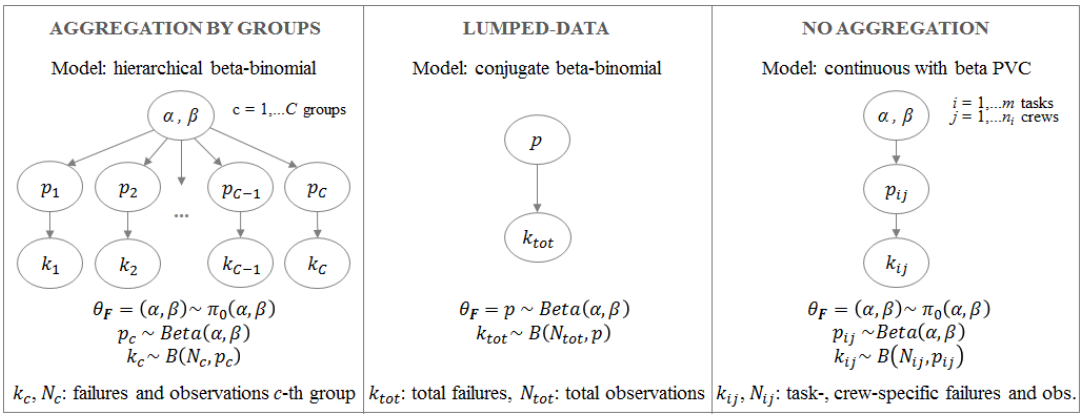

Fig. 4. Bayesian models for HEP quantification coupled to the three modelling approaches of Fig. 2. All the p's are intended as conditional on the given set $\boldsymbol{F}$, e.g. $p_{c|\boldsymbol{F}}$, $p_{ij|\boldsymbol{F}}$.

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].

4 Case study from literature data

This section presents the application of the multi-step methodology to the case study. Subsection 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 (Fig. 4, left). The results are compared to the alternative modelling approaches, i.e. the lumped-data and the continuous variability models (Fig. 4, respectively center and right). Sensitivity analysis on model results is presented in subsection 4.2.

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.

4.1.1 Derivation of behavioral categories (methodology: block I)

From the application of step I.1 of the methodology (Fig. 3), 27 crew observations were identified as belonging to the combination ***F*** of task type and PSF ratings reported in Tab. 4 (the SACADA taxonomy is used for illustration purposes). The selection of the PSF ratings was done by the authors of the present paper, 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 Tab. 4), in all cases with masked indicators (PSF "information availability" with rating "missing/masked" in Tab. 4). All the involved crews experienced for the first time the operational situation replicated by the simulated scenario (PSF "familiarity" with rating "anomaly" in Tab. 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 Tab. 4) and with relatively high-tempo (PSF "time criticality" with rating "barely adequate" in Tab. 4).

Tab. 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.

In the application of steps I.2-I.3 of the methodology (Fig. 3), for each of the simulator records in Tab. 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 Tab. 2. Tab. 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 Tab. 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 the diagnosis. The full set of metrics associated to each behavioral category is provided in Tab. A1 (Appendix).

As mentioned in subsection 3.1, the aggregation in behavioral groups (steps II.1-II.2 in Fig. 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 Tab. 2), the category list was further compacted by expert-based aggregation, and the respective metrics combined into ten categories as shown in Tab. 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".

Tab. 4. Simulator data used in the case study.

Set ***F*** (taxonomy from SACADA, [20]): {task type = understanding the situation, information quality = missing/masked, information specificity = not specific, familiarity = anomaly, time criticality = barely adequate[1]}					
Source	Scenario	Realization of contextual factors	Task	Observations	Failures
[43][2]	SGTR	Failure of secondary radiation indications	Identification and isolation of faulted SG ("HFE1B")	12	6
[13][3]	Multi SGTR	Radiation alarms already activated by early releases due to initiating event	Identification and isolation of faulted SG	5	4
	ISLOCA	No indications on leaks' specific location	Identification and isolation of leaks	5	2
	LOFW+SGTR	Water level increase and absence of radiation indication mask faulted SG identification	Identification and isolation of faulted SG	5	3
			Aggregated data (k_{tot}, N_{tot})	27	15

.

[1] 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.

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

[3] 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.

Tab. 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 Tab. A1 (Appendix). Right: compact set after expert-based aggregation, for use in the numerical application (subsection 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).			

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 Tab. 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 Tab. A1 in Appendix) 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 Tab. 3. Tab. 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 Tab. 6 highlights in qualitative terms what discussed in Section 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 Tab. 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 Tab. 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). Fig. 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 Tab. 5) 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 Tab. 7.

Tab. 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 adapt.	Slow adapt.	Collective	Non-inclusive	k_c / N_c
Group 1	X		X		X		X			X	0 / 1
Group 2[4]	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

[4] 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 Chapter 2.4 in [43]) were assigned (respectively: "sequential", "adhere", and "fast adaptation").

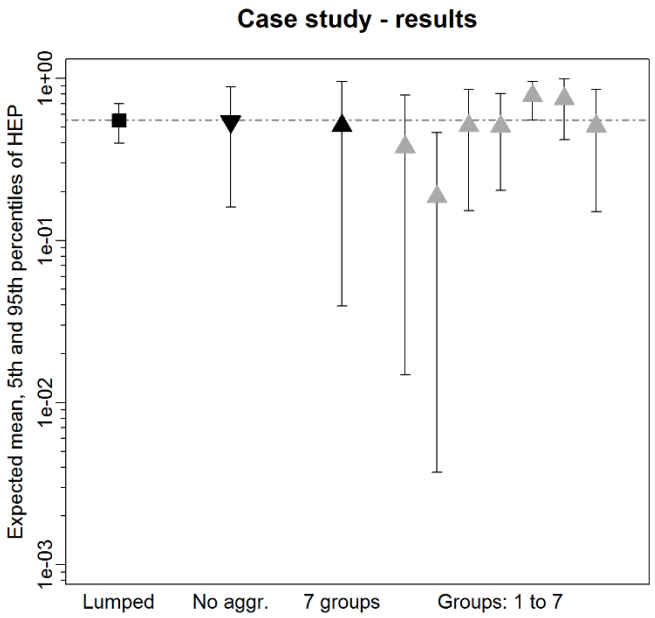

Fig. 5. Results from the numerical application to the case study. In x axis, from left to right: conjugate beta-binomial with lumped data ("Lumped"); continuous variability model with crew-, task-specific data ("No aggregation"); hierarchical beta-binomial model with seven behavioral groups ("7 groups"); 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}).

Tab. 7. Numerical results from Fig. 5 (note that each group corresponds to a specific behavioral pattern).

Model (data aggregation in Fig. 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 Fig. 5 and error factor, EF, values in Tab. 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 $p_{ij|\mathbf{F}}$'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 Tab. 7) and this reflects in a larger spread around the mean value (EF = 4.9) of the HEP uncertainty distribution.

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 4.2.1), and the choice of the variability function (subsection 4.2.2) on the estimated HEP uncertainty distribution. The artificial datasets used in the tests are adapted from the case study.

4.2.1 Number of and degree of performance variability across behavioral groups

As discussed in subsection 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 Tab. 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 Tab. 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 Tab. 6 vs Tab. 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 Tab. 6 (seven groups) and Tab. 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 Tab. A2 (Appendix), together with the numerical results of the sensitivity analysis.

Tab. 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 adapt.	Slow adapt.	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

Fig. 6 shows the $P(HEP)$'s provided by the hierarchical beta-binomial model informed with real data from Tabs. 6 and 8 ("large variability", respectively "7 groups" and "4 groups") and the artificial data from Tab. A2 ("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 Tab. A2: 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 Tab. A2) 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 Tab. A2 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 Tab. 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 Tab. 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.

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 Tab. A3 (Appendix). In general, the considerations drawn from the sensitivity analysis still apply (see in Tab. A3 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 paper (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]).

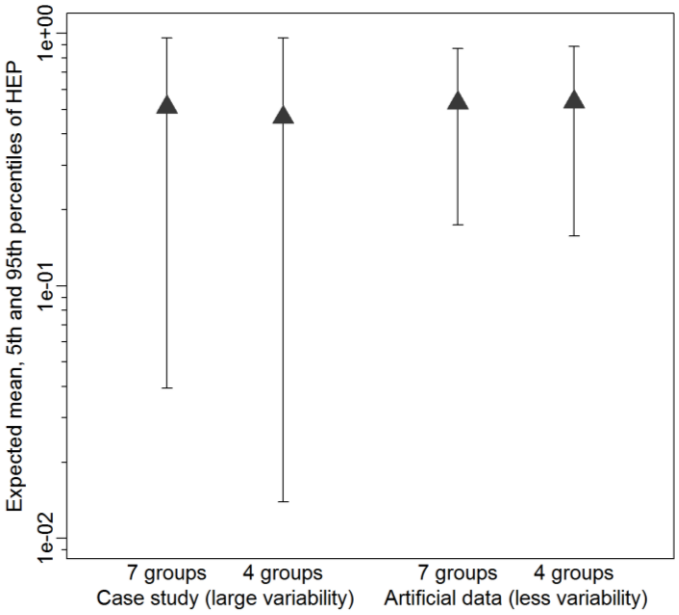

Fig. 6. Influence of the number of identified behavioral groups ("7 groups" vs "4 groups") on the $P(HEP)$ estimated by the hierarchical beta-binomial model for both the case study (Tabs. 6-8) and the artificially-generated dataset with less performance variability (Tab. A2, Appendix).

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 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 paper. 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 Tab. 6). This also coverts in the large variability for the resulting HEP distribution, EF of about 5 in Tab. 7). The characteristics of the scenarios analyzed in the present paper 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 Fig. 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 ***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 paper, 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 Tab. 2 and Appendix A), 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 Tab. 2 and Appendix A. 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 ***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 ***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 paper 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 Fig. 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 $\boldsymbol{F}$).

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.

The paper presents 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. 2020: in review.
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: *International Conference on Probabilistic Safety*, PSAM8 2006, New Orleans, Louisiana, May 14- 19.
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.
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.

Appendix A

Tab. A1: List of teamwork competences and the associated metrics (taxonomy from [44]) used to categorize crew behaviors emerging from the case study (descriptions for each behavioral category are provided in Tab. 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 LEADERSHIP DECISION MAKING	- Proactivity: think ahead possibilities for optimizing activities - Encourage out-of-the-box thinking if needed - 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 COORDINATION DECISION MAKING	- Ensuring that preconditions exist for successful task execution - Proactivity: collecting information that may be useful at later stages - 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 SITUATION AWARENESS	- Timely updating on progress and deviations - 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" • "Hesitating, slowly building up"	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 DECISION MAKING	- Setting well-defined, realistic goals - Prioritize safety goals and concerns - 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 DECISION MAKING INTEPERS. COMPETENCE	- Mutual performance monitoring and provision of needed support, to the extent possible - Ensuring that crew members are adequately involved - 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

Tab A2. Numerical results from the sensitivity analysis on the hierarchical model with varying number of and degrees of variability across behavioral groups (Fig. 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

Tab. A3. 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 [35]).

Model (variability function)	Dataset (group-specific $E[p_c]$ for hierarchical model)	Mean	5th	50th	95th	EF
Continuous (lognormal PVC)	Case study (Tab. 6)	4.9e-01	1.2e-01	4.9e-01	9.1e-01	2.8
Continuous (logistic-normal PVC)	Case study (Tab. 6)	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" (Tab. A2) ($E[p_1]$: 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" (Tab. A2) ($E[p_1]$: 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" (Tab. A2) ($E[p_1]$: 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" (Tab. A2) ($E[p_1]$: 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" (Tab. A2) ($E[p_1]$: 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" (Tab. A2) ($E[p_1]$: 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" (Tab. A2) ($E[p_1]$: 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" (Tab. A2) ($E[p_1]$: 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