

Politecnico di Torino

ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation
Doctoral Program in Chemical Engineering (36^{th}cycle)

Process Safety Data Modelling for Human-in-the-Loop Configurations in Process Control:

Experimental results of alarm response procedure optimisation and a real-time support strategy

By

Chidera Winifred Amazu

******

Supervisor(s):

Prof. Davide Fissore, Supervisor
Prof. Micaela Demichela, Co-Supervisor
Dr. Maria Chiara Leva, Co-Supervisor

Doctoral Examination Committee:

Prof. Sirin A. Okyayuz, Referee, Hacettepe University, Turkey
Prof. Alessandro Tugnoli, Referee, University of Bologna, Italy
Prof. Myrto Konstantinidou, National Centre for Scientific Research Demokritos, Greece
Prof. Emeritus Joaquim Casal, Universidad Politecnica de Catalunya, Spain
Prof. Marco Vanni, Politecnico di Torino, Italy

Politecnico di Torino
2025

Declaration

I hereby declare that, the contents and organization of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

Chidera Winifred Amazu
2025

* This dissertation is presented in partial fulfillment of the requirements for **Ph.D. degree** in the Graduate School of Politecnico di Torino (ScuDo).

I would like to dedicate this thesis to God Almighty and my loving parents.

Acknowledgements

The journey of completing this PhD has been a transformative experience, one that has shaped not only my academic and professional life but also my personal growth. Over these past three years, I have been fortunate to be surrounded by remarkable individuals whose support, encouragement, and wisdom have made this achievement possible.

First and foremost, I am deeply grateful to God for His guidance, strength, and provision throughout this journey. Every challenge, triumph, and lesson along the way has been part of His plan, and I am profoundly blessed to have experienced His grace during this phase of my life.

In no particular order to anyone mentioned or referred to:

I owe my deepest gratitude to my parents, Prof. and Dr. (Mrs.) L. U. Amazu. Their resilience, determination, and unwavering commitment to their academic pursuits laid the foundation for my belief in this path. Watching them overcome challenges to earn their PhDs inspired me and instilled in me the conviction that I, too, could persevere. Their spiritual and emotional support, as well as their continuous words of encouragement, have been my anchor throughout this journey. To my husband, Anthony, thank you for your unwavering support and ever-ready listening ears. Even though you joined me during the latter part of this journey, you caught up ever so quickly on my research. Your encouragement and belief in my abilities have been a driving force in these final years. Your presence in these times has made this process meaningful. To my siblings, Steffi, Triumphant, Sonia and Preciousgift, thank you for standing by me, for listening when I needed to vent, and for always being a source of love and understanding. Your support has been invaluable. - *I love you all!*.

I extend my heartfelt thanks to my academic and scientific supervisors. To Prof. Davide Fissore, thank you for your guidance and supervision these past years. Your expertise was a guiding light when I needed clarity. To Prof. Micaela Demichela, your thoughtful guidance and words of advice ensured that I never felt overwhelmed by the

PhD process. You helped me see the challenges as opportunities for growth, and I will always be grateful for your support and mentorship. To Dr. Maria Chiara Leva, your passion and proactive approach inspired me to push my boundaries in certain ways. Thank you for being such a supportive co-supervisor and for introducing me to Dr. John McGrory. I am immensely grateful to you, John and the other colleagues for making my research time in Dublin a memorable and enriching experience. John, your mentorship in writing my thesis transformed what initially felt daunting into an enjoyable process. You not only helped me appreciate and refine my work but also taught me to see myself someday as a mentor to others. To Dr. Gabriele Baldissone, your technical assistance and practical advice were invaluable, and your encouragement made a world of difference. To my Examination Committee, my deepest appreciation for your time and special thanks to Prof. Sirin A. Okyayuz and Prof. Alessandro Tugnoli for reviewing my work. - *It has been an honour! Thank you!*.

I also want to acknowledge the incredible network of colleagues (some turned friends) and collaborators I have had the privilege of working with. From administrative staff to fellow researchers and industrial partners, each of you has contributed to my journey in unique and meaningful ways. Thank you for making these three years not just a period of hard work but also a time of growth and camaraderie. I am also incredibly grateful to my friends and family from the International Church of Turin, who have been my confidants, cheerleaders, and sources of comfort during the more trying moments of this process. Thank you for letting me lean on you when I needed it most, for sharing in my struggles, and for celebrating my victories. You know yourselves - *You all are the best!*.

Finally, I extend my gratitude to the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skodowska-Curie Actions, for the grant awarded (Agreement No. 955901) to the Collaborative Intelligence for Safety-Critical Systems network (CISC) in which I was an early-stage researcher (ESR) and a beneficiary of the funding. This financial opportunity has allowed me to contribute to groundbreaking (with a multidisciplinary appeal) and to develop as an early-stage researcher - *Thank you!*

To everyone who has been part of this journey, including those whose research works shaped mine, I can go on and on, but I would like to say, thank you for your impact on my life. Your support, guidance, and encouragement have made this achievement possible - *God bless you!*.

Abstract

This thesis explores the shortcomings in evaluating human-in-the-loop (HiTL) performance and providing adequate situational awareness and decision support within process control room HiTL settings. It examines how decision support tools, specifically, alarm response procedures, can influence operators' cognition, behaviour, performance and consequently, safety, during human-system interactions. Secondly, it explores how operators can be supported in real-time with an understanding of their potential performance and risks. Therefore, two research phases and objectives are defined for this thesis. The first phase, which focusses on situational awareness and decision support through alarm response procedures, explores the benefits of transitioning to digital procedure options, specifically electronic procedures (EP), from current paper-based procedures (PBP). The key problem explored is how a setup with these two different procedural formats influence operators with implications for process safety in the process industry. The second phase explores a framework with a set of methods to predict human performance in real-time, understand and guide operators' behaviour based on risk-informed decisions.

The main research objectives include: (1) analysing the impact of transitioning from paper-based and digital procedure options on operator performance, behaviour, and cognition. This will involve identifying factors that affect human performance in current and near-future human-in-the-loop configurations and factors that can be used to describe the impact of alarm response procedures on operators. (2) developing a real-time, risk-based framework (R²HiTL) to predict operators' performance and make risk-informed decisions to guide operator behaviour.

The methodology was split into two phases, one focused on the experimental performance analysis to address objective one and the other on the development of a novel framework. For the first phase, the outcomes from both PBP and EP experimental study groups were compared using workload, situational awareness, behavioural, and

performance metrics. Furthermore, a hybrid method was developed for the second phase that integrates logistic regression and Bayesian Network standardised plant analysis risk-human reliability for real-time prediction of operator failure probabilities, allowing adaptive decision support tailored to individual needs.

Key findings reveal that while paper-based procedures were associated with a higher perceived workload, they lead to better situational awareness and performance in the most complex scenario. In contrast, electronic procedures, though they reduced perceived workload, resulted in lower situational awareness and more errors in high-stress situations. However, the electronic procedure from further analysis with the eye tracking provided some cognitive and behaviour benefits compared to PBP's. The research also demonstrates that the real-time, risk-based decision support framework can be applied for operator performance predictions by integrating expert knowledge with live performance data, enhancing decisions on how to guide operators' behaviour and provide recommendations.

This work contributes to the field of alarm handling and process control for process safety, decision making, human centred design, human performance and interaction modelling. Potential applications include improving alarm management systems, improving operator training programmes, and designing adaptive support tools for high-risk industries such as process, manufacturing, energy, and healthcare. The research offers valuable implications for improving safety, reducing operational risks, and optimising human-machine interactions in critical environments.

Contents

List of Figures **xii**

List of Tables **xvi**

Nomenclature **xxi**

1 Introduction **1**

1.1 Introduction 1
1.2 Background & Motivation 1
1.3 Challenges 4
1.4 Ethics 5
1.5 Objectives 6
1.6 Methodology of the research 6
1.7 Summary of Subsequent Chapters 8

2 Literature Review **10**

2.1 Introduction 10
2.2 Human Senses and Cognitive Functioning 11
2.2.1 Human sensory system 12
2.2.2 Brain Functioning 14
2.2.3 Cognitive functioning 15
2.3 Physical Process and Control Room 25
2.3.1 Physical Plant Process Control 25
2.3.2 The Process Control Room 28
2.3.3 The Human System Interfaces 29
2.4 Procedures 34

2.4.1 Procedural Guidance in Control Rooms 34
2.5 Human Reliability Analysis 41
2.5.1 Human Performance 41
2.5.2 Methods, Application and Data Sources 43
2.5.3 Challenges with HRA methods 46
2.5.4 Current Advances in Process Control Room HPA 48
2.6 Summary of Gaps 53
2.6.1 Procedures: Alarm Response Procedures 53
2.6.2 Human Reliability Assessment for PBA 54
2.6.3 Other Performance Assessment Techniques 55
2.6.4 Contribution 55

3 Design Methodology I 68
3.1 Introduction 68
3.2 Research Design 70
3.2.1 Selected Plant 73
3.2.2 Design of Tools 80
3.3 Experiment Design 87
3.3.1 Data Collection Methods/Measures and Procedures 92
3.3.2 Variables and Measures 98
3.3.3 Reliability, Validity, Sensitivity and Normality 108
3.3.4 Ethical Consideration 112
3.3.5 Data Analysis and Result Presentation Plan 113
3.3.6 Limitations of the Design 114
3.3.7 Summary 115

4 Design Methodology II 117
4.1 Introduction 117
4.2 Research Framework and Methods 119
4.2.1 Framework 119
4.2.2 Methods 124
4.3 Data Collection and Variables 138
4.3.1 HTA Task analysis and SPARH 138
4.3.2 LR and BN SPARH 139

4.4 Validity and Reliability . . . 139
4.4.1 SPAR-H . . . 139
4.4.2 Logistic Regression . . . 140
4.4.3 BN SPARH . . . 141
4.5 Data Representation plan . . . 142
4.6 Limitations of the Design . . . 142
4.7 Summary . . . 143

5 Results - Methodology 1 145
5.1 Introduction . . . 145
5.2 Descriptive Statistics and Initial Analysis . . . 146
5.2.1 Overview of Collected Data . . . 146
5.2.2 Descriptive Statistics . . . 148
5.3 Evaluation of the measures . . . 173
5.3.1 Reliability of the Subjective Measures . . . 173
5.3.2 Validity of the Measures . . . 174
5.3.3 Sensitivity Analysis . . . 202
5.4 Summary . . . 204
5.4.1 Correlations and Constructs . . . 204
5.4.2 Groupwise Comparison . . . 208

6 Results Methodology II 210
6.1 Introduction . . . 210
6.2 Phase 1: Perception (Risk Perception) . . . 211
6.3 Phase 2: Understand (Risk Understanding) . . . 213
6.3.1 Step-wise Logistic Regression . . . 213
6.3.2 Hierarchical Task Analysis . . . 215
6.3.3 SPAR-H Analysis Results . . . 219
6.4 Phase 3: Project (Risk Projection) . . . 227
6.4.1 BN SPARH . . . 227
6.4.2 Consequence and Risk Analysis . . . 243
6.5 Summary . . . 244

7 Discussion **246**
7.1 Introduction 246
7.2 Experiment: Human-in-the-loop in alarm handling and process control . 247
7.2.1 Workload, Situational Awareness, Perception of Support 247
7.2.2 Behaviour and Performance 251
7.2.3 Eye-tracking measures 253
7.2.4 Discussion on Group-wise Comparison 254
7.3 Phase 2: Real-time Risk-Based HPA and Decision Support 262
7.3.1 Discussion on Phase 1 and 2 of the R²HiTL framework 262
7.3.2 Phase 3 Discussion 264
7.3.3 Validation outcome 266

8 Conclusion **267**
8.1 Research Objective 1 267
8.2 Research Objective 2 270
8.3 Potential for generalisation and Future Work 273
References 275

Appendix A Tables **292**

Appendix B Figures **300**

List of Figures

1.1 The Process Plant and Key Components. Label 1 contains the physical plant with its systems, subsystems and perhaps a field operator or those on the field. Label 2 contains the IT infrastructure for connectedness. Label 3 is the process control room with the human system interfaces, while label 4, also part of this control room, is the human-in-the-loop. Label 5 is the environment. 9

2.1 The Human in the loop. 12

2.2 An Model of Human Information Processing Stages [33]. 16

2.3 Information Processing Model of Decision Making [33]. 24

2.4 Process Safety Pyramid and layers [75], [76]. 26

2.5 Basic Process Controls and Network. 27

2.6 Demonstrating a Control Room Setup with both paper and electronic procedures. 29

2.7 Alarm TAG Notation. 31

2.8 Hierarchy of Plant Operating Procedures [95]. 36

2.9 Procedure Format [5]. 37

3.1 Experimental Design Framework Observing the Impact of Task-Based PSFs (Alarm Response Procedures) on Internal and External Individual Factors, Including Interactions with Task Complexity, Time, Training, Experience, Alarms, and Displays. 70

3.2 The Human System Interfaces. A. Overview display providing for overall situational awareness on plant operation. B. Extra display for projection of specific plant section equipment (2), and C. the Support Display for Event Diagnosis using alarm list support, procedure support, and trends [18]. 73
3.3 The Overview display: Process Flow Diagram of the Plant. The plant sections are Tank, Methanol, Compressor, Heat Recovery, Reactor, and Absorber. 74
3.4 Tank Mimic showing a methanol tank, controllers, meters, process trends and alarms. 75
3.5 Methanol section of the plant. 77
3.6 Compressor. 77
3.7 Heat Recovery. 78
3.8 Reactor. 79
3.9 Absorber. 80
3.10 The Support Display. A. for those using paper procedures, and B. for those using electronic procedures. At the top left corner is the alarm list, while the graph trends of critical systems and production volume occupy other sections of the display [155]. 81
3.11 The Mimic Screen with an example from the Tank Plant section. 82
3.12 Electronic Procedure with the Navigation Menu. (a) The procedure for the low-pressure alarm (PAL01) details steps 1 to 6. (b) The continuation of PAL01 covering steps 8 to 14. The left sidebars illustrate the sequence of alarm arrangements to aid in navigation [156]. 86
3.13 The Paper-based procedure. (a) Cover page of the paper procedure. (b) Example of the table of contents. (c) Procedure for PAL01 (low-pressure alarm) covering steps 1 to 14. (d) The actual appearance of the procedure during testing, with a colour tag marking the beginning of each new plant section [156]. 86
3.14 Alarm-related initiating events, general operator task, and think-aloud points in scenario 1. 90
3.15 Alarm-related initiating events, general operator task, and think-aloud points in scenario 2. 91

3.16 Alarm-related initiating events, general operator task, and think-aloud points in scenario 3. 92
3.17 Order of saving the recorded simulator logs after each scenario is completed [158]. 97
3.18 Reaction time, Scenario 1. 107
3.19 Response Time, Scenario 1. 107

4.1 Real-time Risk-based Human-in-the-loop (R^2HiTL) Performance Assessment and Decision Support Framework. 120
4.2 Dependency model for subtask dependency analysis in SPAR-H. 131

5.1 Age distribution by scenario for each group. 149
5.2 Gender distribution for by scenario. 150
5.3 Scree plot to determine the optimum number of factors for the objective and subjective measures combined *(Eye-tracking excluded)*. 177

6.1 Monitoring HFE: Beta Distributions for LR, SPAR-H and LR+BN SPARH. Participant outcome = Success. 241
6.2 Planning HFE: Beta Distributions for LR, SPAR-H and LR+BN SPARH. Participants' eventual outcome = Success. 242
6.3 Intervention HFE: Beta Distributions for LR, SPAR-H and LR+BN SPAR-H. Participants' eventual outcome = Success. 243
6.4 Event tree for test operator outcome. 244
6.5 Risk Matrix. 245

7.1 Alarm list to Procedure Navigation human factor issues. 1. Incorrect layout 2. Proposed correct layout. 257
7.2 Head movement for those using the electronic procedure. 258
7.3 Alarm list showing two alarms: FAL01 and PAL01. 261

B.1 Recruitment Fyler: first part 301
B.2 Recruitment Fyler: second part 302
B.3 Participation Approval Email 303
B.4 Participation Denial Email 304

B.5 NASA-SART Questions with preliminary questions of participant ID, Group and Scenario, followed by NASA-TLX dimensions: Mental Demand, Physical Demand and Temporal Demand. 305
B.6 NASA-SART Questions showing NASA-TLX dimensions: Mental Demand, Physical Demand and Temporal Demand, and SART dimensions: 306
B.7 NASA-SART Questions showing SART dimensions: 307
B.8 Consent Form 308
B.9 PSF Evaluation Sheet for Action Task [113]. 309
B.10 PSF Evaluation Sheet for Action Task [113]. 310

List of Tables

2.1 Comparison of Warning, Alarm, and Prompt. 30
2.2 Levels of Procedural Support Systems and their Descriptions [15]. . . . 38
2.3 Comparison of Risk Assessment and Human Performance across different aspects. 58
2.4 First Generation HRA Methods. 59
2.5 Second Generation HRA Methods. 60
2.6 Third Generation HRA Methods. 61
2.7 Summary of Eye-Tracking Metrics and Definitions [18]. 62
2.8 Comparison of NASA TLX, SPAM, and SART across different aspects. 63
2.9 Summary of Statistical Methods, Purpose, and Applications part I. . . . 64
2.10 Summary of Statistical Methods, Purpose, and Applications part II. . . . 65
2.11 Summary of Statistical Methods, Purpose, and Applications part III. . . 66
2.12 Summary of Statistical Methods, Purpose, and Applications part IV. . . 67

3.1 Key performance shaping factors for comparison. These include alarm design, procedure, Display/HMI, and task complexity, which differentiate the groups and their different levels. 87
3.2 General experiment study groups [18]. 87

4.1 Inputs and Outputs for each Phase of the Risk Assessment Framework. . 121
4.2 Evaluation of Different Task Analysis Methods Based on Some Key Specified Criteria. P = Partially fulfilled, N = No, and Y = Yes. [166]. . 127
4.3 Performance Shaping Factors (PSFs) and their Descriptions. 129
4.4 Summary of Advantages and Disadvantages of the SPAR-H Method [113]. 133

5.1 Subjective Measures and Objective Measures. 148

5.2 NASA-TLX workload measures for different scenarios and the study groups: PBP vs EP. 151
5.3 SART measures for different scenarios and groups: PBP vs. EP. 153
5.4 Descriptive statistics for SPAM measures across scenarios and groups. . 155
5.5 Descriptive statistics for NASA-TLX, SPAM, SART, and Task Load measures across scenarios and groups. 157
5.6 Descriptive statistics for Alarm Support, Procedure Support, Training, and Familiarity across scenarios and groups. 160
5.7 Significant Differences for Table 5.5. T-test values are highlighted in blue. Value marked * shows significant difference $p < 0.05$. 161
5.8 Significant Differences for Table 5.6. T-test values are highlighted in blue. Value marked * shows significant difference $p < 0.05$. 162
5.9 Descriptive statistics for Recovery Time and Accuracy (Mean Square Error) across scenarios and groups. 164
5.10 Summary of Consequences, Performance, Errors, and Success/Failure Counts across Scenarios and Groups. 165
5.11 Descriptive statistics for Reaction Time, Response Time, Alarms Silenced, Acknowledgements, Mimics Opened, Number of Procedures, and Number of Alarms across scenarios and groups. 166
5.12 Descriptive statistics for eye-tracking metrics in Scenario 3 for PBP and EP groups. 168
5.13 Shapiro-Wilk and T-test/Mann-Whitney results for Recovery Time and Accuracy across scenarios and groups. 169
5.14 Shapiro-Wilk and T-test/Mann-Whitney results for Reaction Time, Response Time, Alarms Silenced, Acknowledgements, Mimics Opened, and Number of Alarms across scenarios and groups. 170
5.15 Shapiro-Wilk and T-test/Mann-Whitney results for eye-tracking metrics in Scenario 3 across PBP and EP groups. 172
5.16 Subjective Measures, their areas of focus, number of questions, and Cronbach's Alpha values. 173
5.17 Factor Loadings for Subjective and Objective Measures (without eye-tracking measures). 175
5.18 Factor Loadings for the four participants in each group under scenario 3. 179
5.19 Correlation Outcomes Between Measures and TLX dimensions. 182

5.20 Significant difference of TLX measures. p < 0.05. 184
5.21 Correlations with situational awareness measures. p < 0.05. 185
5.22 Significance of situational awareness measures. p < 0.05. 187
5.23 Correlation with other subjective measures. 190
5.24 Significance with other subjective measures. 191
5.25 Correlation with objective measures. 195
5.26 Significance of objective measures. 196
5.27 Correlation with Eye-tracking Measures. 200
5.28 Significant differences of Eye-tracking Measures. 201
5.29 Statistical Test Results for TLX and SART Measures Across Scenarios. 203
5.30 Statistical Test Results for Other Subjective Measures. 203
5.31 Statistical Test Results for Behavioural and Performance Measures Across Scenarios. 204
5.32 Constructs and Associated Variables. 204
5.33 Associating the eye-tracking measures obtained from when specifically reading through the procedure in scenario 3 with that of other variables in the same scenario. 205
5.34 Summary of Significant correlations (positive and negative), $p < 0.05$. . 206
5.35 Summary of Significant correlations between Behaviour and Performance measures (positive and negative), $p < 0.05$. 207
5.36 Summary of Significant correlations with eye-tracking (Scenario 3 only), $p < 0.05$. 207

6.1 Selected Input variables for the Framework. 212
6.2 Key process control room tasks. 213
6.3 Failure prediction for each group per scenario and task /HFE (human failure event). 214
6.4 Task Steps and Subtasks for Scenario 1. 216
6.5 Task Steps and Subtasks for Scenario 2. 217
6.6 Task Steps and Sub-tasks for Scenario 3. 218
6.7 Task Steps and Sub-tasks for Scenario 3. 220
6.8 Experts PSF assignment of multipliers for Group PBP. 221
6.9 Experts PSF assignment of multipliers for Group EP. 222

6.10 NHEP, HEP adjustment, dependency calculation and final HEP for Group PBP. . . . 224
6.11 NHEP, HEP adjustment, dependency calculation and final HEP for Group EP. . . . 225
6.12 Final HEP per Human Failure Event and the alpha and beta priors for Group PBP. . . . 226
6.13 Final HEP per Human Failure Event and the alpha and beta priors for Group EP . . . 227
6.14 Monitoring HFE: discretisation of variables. . . . 228
6.15 Planning HFE: discretisation of variables. . . . 229
6.16 Intervention HFE: discretisation of variables. . . . 230
6.17 Monitoring HFE: contexualisation of variables. . . . 232
6.18 Planning HFE: contexualisation of variables. . . . 233
6.19 Intervention HFE: contexualisation of variables. . . . 234
6.20 Monitoring HFE: count of context - error pairing. . . . 235
6.21 Planning HFE: count of context - error pairing. . . . 235
6.22 Intervention HFE: count of context - error pairing. . . . 236
6.23 Monitoring HFE: recalling priors for group PBP. . . . 237
6.24 Monitoring HFE: recalling priors for group EP. . . . 237
6.25 Planning HFE: recalling priors for group PBP. . . . 237
6.26 Planning HFE: recalling priors for group EP. . . . 237
6.27 Intervention HFE: recalling priors for group PBP. . . . 238
6.28 Intervention HFE: recalling priors for group EP. . . . 238
6.29 Monitoring HFE: Group PBP Data with Posterior-Parameters. . . . 238
6.30 Monitoring HFE: Group EP Data with Posterior-Parameters. . . . 239
6.31 Planning HFE: Group PBP Data with Posterior-Parameters. . . . 239
6.32 Planning HFE: Group EP Data with Posterior-Parameters. . . . 239
6.33 Intervention HFE: Group PBP Data with Posterior-Parameters. . . . 240
6.34 Intervention HFE: Group EP Data with Posterior-Parameters. . . . 240
6.35 New Operator data for validation. . . . 240
6.36 Monitoring Failure Probability from LR prediction. . . . 241
6.37 Planning Failure Probability. . . . 242
6.38 Intervention Failure Probability. . . . 243

A.1 HMI Design Requirements 293
A.2 APR Requirement: Access and Location 294
A.3 ARP: Content 295
A.4 Summary of Requirements for Alarm System Display [98] 296
A.5 Sample raw simulator log features by plant sections and alarms. Showing for Alarm, Tank, and Methanol [158]. 297
A.6 Sample raw simulator log features by plant sections and alarms. Showing for Compressor [158] 298
A.7 Sample raw simulator log features by plant sections and alarms. Showing for Reactor [158] 298
A.8 Sample raw simulator log features by plant sections and alarms. Showing for Heat Recovery [158] 299
A.9 Sample raw simulator log features by plant sections and alarms. Showing for Absorber [158] 299

Nomenclature

Roman Symbols

h	Planck constant
HRA	Human Reliability Assessment
AI	Artifical intelligence
AIC	Akaike information criterion
All	Alarm
ANOVA	Analysis of variance
ANS	Autonomous nervous system
ANT	Attention network task
AOI	Area of Interest
AOP	Abnormal operating procedures
ARP	Alarm response procedure
ATEX	Explosive atmosphere
ATHEANA	A technique for human event analysis
BIC	Bayesian information criterion
BN	Bayesian networks
CADET	Critical action and decision evaluation technique

CAH Concentration alarm high

CBP Computer-based procedure

CBP-PBA CBP-Procedure based automation

CFA Common factor analysis

CISC Collaborative intelligence for safety critical

CPC Common performance conditions

CREAM Cognitive reliability and error analysis method

D Demand

DAG Directed acyclic graph

DCS Distributed control systems

DoE Design of experiment

DoF Degree of freedom

DT Decision trees

EEG Electroencephalogram

EFA Exploratory factor analysis

EOP Emergency operating procedures

EP Electronic procedures

EPC Error producing conditions

EPRI Electric power research institute

ET Eye tracking

FAL01 Flow alarm low

fMRI Functional magnetic resonance imaging

FSH Flow switch at high flow

G Group

GTP Gaze transition probabilities

HEART Human error assessment and reduction technique

HERA Human event repository and analysis

HF Human Factors

HFE Human Failure Event

HiTL Human-in-the-loop

HMI Human machine interface

HoTL Human-out-of-the-loop

HPA Human Performance Assessment

HRA Human Reliability Assessment

HRV Heart rate variability

HSI Human system interface

HTA Hierarchical task analysis

ICC Intra-class correlation coefficient

IMAS Influencing modelling and assessment systems

KNN k nearest neighbours

LAH Level alarm high

LAL Level alarm low

LNG Liquified Natural Gas

LR Logistic regression

MATB-II Multi-attribute task battery

MCR Main control room

MLP Multilayer perceptron

MSE Mean Square Error

NASA-TLX NASA Task Load Index

NB Naive Bayes

NN Neural networks

NOP Normal operating procedures

OAET Operator action event tree

ONCP Off-Normal Condition Procedures

OPERA Operator performance and reliability assessment

P Proportional controllers

PAL01 Pressure alarm low

PBA Performance-based analysis

PBP Paper-based procedures

PCA Principal component analysis

PD Proportional-derivative controllers

PDF Portable document format

PFD Process flow diagram

PI Integral controllers

PIC Pressure indicator controller

PID Proportional integral derivative

PIF Performance influencing factors

PLC Programmable logic controller

PNS Parasympathetic nervous system

PRA Probabilistic risk assessment

PSA Process safety assessment

PSF Performance shaping factors

PSLL Pressure switch at low low pressure

PSV Pressure safety valve

RBPS Risk-based process safety

REC Heat recovery system

RF Random forest

RQ Research questions

S Scenario

S Supply

SA Situational awareness

SACADA Scenario authoring, characterisation, and analysis data archive

SAGAT Situation awareness global assessment technique

SART Situational Awareness Rating Technique

SCADA Supervisory control and data acquisition

SEEV Salience, effort, expectancy and value

SLIM Success likelihood index methodology

SNS Sympathetic nervous system

SPAM Situation Present Assessment Method

SPAR-H Standardised Plant Accident Risk-Human Reliability Assessment

SVM Support vector machine

TAH Temperature alarm high

TAHH Temperature alarm high high

TAHH Temperature alarm low low

TAL Temperature alarm low

THERP Technique for human error rate prediction

TIC Temperature indicator controller

TOI Time of interest

U Understanding

UAV Unmanned aerial vehicle

VAH Volume alarm high

VOC Volatile organic compounds

VR Virtual reality

WAH Power alarm high

XGB Extreme gradient boosting

Chapter 1

Introduction

1.1 Introduction

Chapter 1 discusses the motivation for this thesis, providing a broad overview of the key aspects that inform the motivation and research gaps (Section 1.2) and the objectives (Section 1.5). It also presents the challenges identified (Section 1.3) and discusses the ethical considerations relevant to this thesis (Section 1.4). Additionally, it briefly mentions the methods to be applied and the basis behind choosing these techniques (Section 1.6). The chapter concludes with a summary of the contents to be covered in subsequent chapters (Section 1.7).

1.2 Background & Motivation

Process industries, such as chemical manufacturing, petrochemicals, etc., are distinct process sectors that include high-risk, safety-critical operation facilities. One reason is the type of materials used in these facilities and the high cost of potential consequences to the environment, health, economy, and life in the event of an accident. Therefore, measures to identify hazards and risks and, if possible, predict risks and proactively manage them are paramount in these plants. The process safety management system is one such strategy for achieving this. Although this has in some cases failed to holistically fulfil its role due to inadequacy of the systems within some companies [1]. However, except for management systems, control room operators are essential to ensuring safety in such facilities.

These operators from their workstations, in at least manual and semi-automated setups, can be involved in various functions, which can be classified into three primary roles: Monitoring, Monitoring and control, and Fault Detection and Isolation [2]. They perform plant operations such as setting process parameters, monitoring plant state, troubleshooting faults, and responding to alarms. The operators can do these tasks based on their experience, training, and, in some instances, using procedures that differ depending on the process situation, such as start-up, shutdown, or alarm response. They interface with the plant through multiple monitors and displays showing not only a high-level detail of the overall plant process (level 1 displays) but also human interfaces showing more specific systems, subsystems, and other information such as alarm lists, procedures, trends, and more (level 2 - 4 displays).

The process control room serves as the hub for the operator-system interaction. The layout of this hub and the design of the systems for the operators are said to make or mar the operators' performance and their ability to ensure safety. Tools such as supervisory control and data acquisition (SCADA), distributed control systems (DCS), prioritised alarms, alarm response procedures, etc., support operators in these settings [3], [4], [5]. The combination of these tools and the resulting setup to perform this sole function is called the Human-in-the-loop (HITL). However, their ability to perform their functions individually or combined has been challenged, particularly in procedures. In a study of main accident contributors within the process industry from among 770 cases, [6] recorded that 17% of accidents were due to operating procedures just after 19% from the process failures.

Procedures are also critical for operations in nuclear and aviation sectors and have been significant causes of major accidents [7], [8], [9]. The Three Mile Island accident was a notable loss for the nuclear sector. One of the reported causes was an incorrect operating procedure followed by the operators [7]. The national transport safety port for the aviation sector [8] reported a cargo flight crash, which was due to a missing step in the checklist, leading to the death of all three on board. Despite these concerning task-related and organisational failures, attention has not been paid much to studying the impact of these tools or tests done to verify and validate their prospects.

Instead, greater blame is assigned to the operators as the primary contributors to human error rather than to a joint contribution of individual and organisational discipline towards the design and support of operators. However, discussing the error contribution of individuals such as pilots or control room operators is still essential. Stats in the

process industry have shown that the operators' errors make up about 70% of accidents [10], [11], while the aviation sector recorded that 85% of general aviation crashes were due to pilot error [12]. A counter-reaction by these industries has been to exclude humans as much as possible from the loop and introduce more process automation to plant processes, thereby complicating the existing operators' dilemma. This response is becoming the status quo in the nuclear and process industries, with very advanced automated systems, autopilots in planes, and autonomous driving systems in cars. These dynamic system advances are promising, but how do they, in turn, change the dynamics for the operators? One common issue is the human-out-of-the-loop situations.

In a human-out-of-the-loop (HoTL) scenario, the automation takes control while the frontline operator monitors. Initially, operators usually exhibit limited awareness of the ongoing process, display higher levels of complacency, and experience mental fatigue due to prolonged monitoring tasks [13]. Over time, their situational awareness can be significantly more compromised. As can be seen, the impact is less physical than cognitive for control room operators and mostly frontline operators in general. The operators receive visual or audible information from the displays as a cue to a process change or upset. Due to the heightened direct automation in the plant and the extra cognitive demands it places on operators, these operators are already juggling numerous responsibilities. As a result, they are likely to face challenges such as attention lapses, higher cognitive burdens, and diminished situational awareness. All of these challenges can impair their decision-making and response actions. They have to deal with systemic factors like the comprehensiveness of the information from the interaction interfaces or interface usability. To alleviate such cognitive efforts and support operators adequately, ideally,

- A human-centred approach to designing tools such as procedures and appropriately designed alarm systems should be the norm. Testing and Validation by the relevant users is vital for such designs. There should be a process to test several areas of impact, internally and externally, for operators.
- Automation or Artificial intelligence (AI) should likewise be directed to the design or update of these support tools as much as it is for the plant operations.
- Testing for impact can also be done beyond just the product implementation, commissioning, and training phases, as data from plants during operations can be harnessed to provide support on demand/real-time.

The nuclear and process sectors are advancing in addressing the human-centred gaps in alarm systems [14]. However, the nuclear sector has seen more advances in the design and development of procedures. Researchers in the nuclear sector are already exploring level 2, 3 and 4 type procedures for the nuclear control rooms and taking a human-centred approach to check for their impact on operators' cognitive states [15], [9]. These procedures allow for visualising the procedures in a portable document format (PDF) or embedded files on display with the same procedure lists or flowchart format as that of the paper procedures in the case of Level 2. Level 3 procedures, on the other hand, are computer-based, provide more precise procedure steps, and allow for some automation between the alarm message list and procedures. Finally, level 4 procedures have the advantages of level 3 but include embedded soft controls to issue control commands to plant equipment. [9] explored how such type two procedures in different formats or a combination of two formats impact operators' workload using subjective and objective techniques. On the other hand, [15] has compared levels 1, 2, and 3 procedures on how they impact situational awareness and workload using subjective techniques. Procedure levels, formats and types are further discussed in Section 2.4. However, studying for procedure advances and applying a more human-centred approach remains a gap in the process industry.

1.3 Challenges

In the previous section, concerns about the issues stemming from current human-in-the-loop (HITL) configurations were raised, and some recommendations were made. Implementing some of these more human-centred approaches in the process industry is rooted in the limited data available for human performance assessment, as often pointed out by human reliability assessment (HRA) experts [16]. Secondly, situational awareness and workload assessment techniques, such as NASA Task Load Index (NASA-TLX) and situation awareness rating technique (SART), and other techniques like interviews, are mainly subjective [17]. In addition, situational awareness and workload assessment techniques are often used as standalone during implementation, commission or workload reviews, limiting the assessment's robustness.

Close to the previously highlighted challenge is the unavailability or poor exploration of tools for assessing operators' internal states or behaviour during human-system interaction. This unavailability can be due to the cost of such tools, the capacity required

for data processing and poor data recording that neglects operator-system interaction records. Last but not least is the difficulty of perhaps empirically grouping multiple factors at once and the interplay between these factors, that would ideally impact the human-in-the-loop configuration. Finally, traditional reliability and risk assessment methods already being impacted by some of the earlier mentioned points relating to data have limitations for a more dynamic real-time human-in-the-loop study. This further limits the potential for having targeted real-time adaptation of support for operators.

1.4 Ethics

An attempt to address these gaps, as would be detailed in this thesis through subsequent chapters, would require human subjects for an experimental study, male and female, across chemical, process and other process safety-related domains. To ensure ethical compliance, participants in an experimental study in this work are briefed before the study begins with a high-level explanation of what the experimental research would involve. A consent form is to be signed by each participant before their participation. The consent form and other documents of importance for the study have been approved by the Collaborative Intelligence for Critical Safety Systems Ethical Committee after authorisation by the Ethics Committee of the Technological University of Dublin (TU Dublin), Ireland, with approval number REC-20-52 [18].

These documents are kept confidential, and there is no possibility to link them to the database containing their personal data from the experiment. The researcher gives each participant an anonymised hash code, which is also oblivious to them. After that, there is no way an external can match the code names to the actual names except by using the consent form, which is only available to the researcher. Their participation is voluntary, and in no situation is a participant coerced to participate. If a student attends due to extra points that can be added, it is verified that this is done voluntarily and does not impact the standard grading format to avoid coercion by style.

Also, absolute respect and integrity are maintained throughout this thesis's development, implementation, data collection, analysis, and writing. The participants are not discriminated against as the study is open to everyone of different backgrounds, colour, race, age colours, races, and ages. Contributors to this research through supervision, data collection, analysis, and co-authorship have also been acknowledged.

1.5 Objectives

Following the discussion of the motivation behind this thesis and a high-level overview of challenges presented for specific tools like the procedures and the methods for a human-in-the-loop performance assessment, this thesis addresses two (2) objectives as broadly presented below:

1. Objective 1: The first objective of this thesis is to understand the state-of-the-art alarm response procedures and to compare the effectiveness of the transition from paper versus electronic alarm response procedures for process control room operators in current HiTL configurations while handling abnormal situations.

2. Objective 2: Secondly, this thesis seeks to create a risk-based real-time framework to evaluate human performance, improve situational awareness, and support decision-making within current human-in-the-loop process control and safety configurations. This provides a basis for guiding operators' behaviour with an understanding of its contribution to risky outcomes. The framework provides a more robust process for streamlining alarm response procedures and optimising other human system interfaces for effective decision support in process control rooms.

The contribution section of Chapter 2 outlines the research questions associated with each objective.

1.6 Methodology of the research

The methodologies employed in this work are presented in phases as they address the objectives.

1. The first phase (Phase 1) involves a comprehensive literature review of the concepts discussed in the early sections of this chapter. The aim is to clearly understand the concepts and current status quo, identify gaps, and highlight the strengths of the methods implemented in phase 2.

2. Inspiration from Phase 1 motivates the methods used in Phase 2. The second phase addresses the second part of objective 1. Objective 1 covers a Top-down

analysis of the influence of the ecosystem, Physical plant operations, IT, and the human system interfaces in the control room, like the alarm response procedures, essentially the organisational and technical factors, on the human in the process control loop (See 1.1). The following methods are applied:

- A design of experiment to model the human-in-the-loop configurations of a process control room as is typical in a process plant or likely to be, where the human system interfaces are combined at different levels; different types of alarm response procedures, alarm designs and consequently slight display differences. Each configuration is uniquely assessed for its impact on operators. The following techniques, as described below on a high level, are employed:
  - Subjective methods introduced to address objective 2 involve using standard questionnaires to evaluate operators' situational awareness, mental workload, and preferences while they handle three safety-critical tasks. This work has developed a new set of questionnaires based on the situation present assessment method (SPAM) specifically designed to assess situational awareness.
  - Objective methods in this thesis involve simulators and eye-tracking to acquire quantitative data on behaviour, performance, and cognition.
- Data analysis is performed for the comparison of the four configurations against these methods.

3. In Phase 3, a bottom-up approach is applied. The results from assessing operators' behaviour, situational awareness, and other characteristics in Phase 2 are now integrated to evaluate their impact on the ecosystem and environment (see 1.1). Figure 1.1 illustrates the overall plant and some of the key components in its design. (1) Is the physical plant where the chemicals, machines, programmable logic controllers (PLCs), and plant engineers create the desired compound or widget. (2) is the middleware that communicates data from the field equipment to the interfaces in the control room and vice versa. (3) is the control room and its elements, that is, the monitors, displays, procedures, etc., (4) is the control room operator and (5) the ecosystem, including the environment, etc. This 3rd phase

combines the characteristics of both Human-in-the-Loop (HiTL) configurations and other control room components. The methods employed include step-wise logistic regression, the third-generation Human Reliability Assessment (HRA) method (specifically the SPAR-H Bayesian Network), consequence analysis, and dynamic event trees.

Chapter 3 provides a detailed explanation of the rationale behind selecting these methods, with additional context in Chapter 2.

1.7 Summary of Subsequent Chapters

The thesis is sectioned in the following way: Chapter 2 discusses in detail literature on the human in the loop and their characteristics with highlights on how this applies during human system interaction in control rooms. Then, it discusses the plant process and its key components, such as the systems and the control room. Subsequently, in Chapter 2, one of the fundamental elements of the control room and central to this thesis, the support procedures, is discussed in detail. This chapter further discusses HRA methods, challenges with these methods, and new techniques in application. It also provides a literature comparison of related experimental human performance studies. This Chapter ends with a summary of the identified gaps in the literature discussed so far and the research questions arising.

Chapters 3 and 4 discuss the methods proposed to achieve the objectives of this thesis. Chapter 3 focuses more on methods in the different phases that address Objective 1, while Chapter 4 focuses on the phases that address Objective 2. Subsequently, in Chapter 5, the application of the methods for Objective 1 is illustrated, accompanied by the results, while that of Objective 2 is discussed in Chapter 6. The discussion of the results for both objectives is done in Chapter 7, with a concluding chapter and future work presented in Chapter 8.

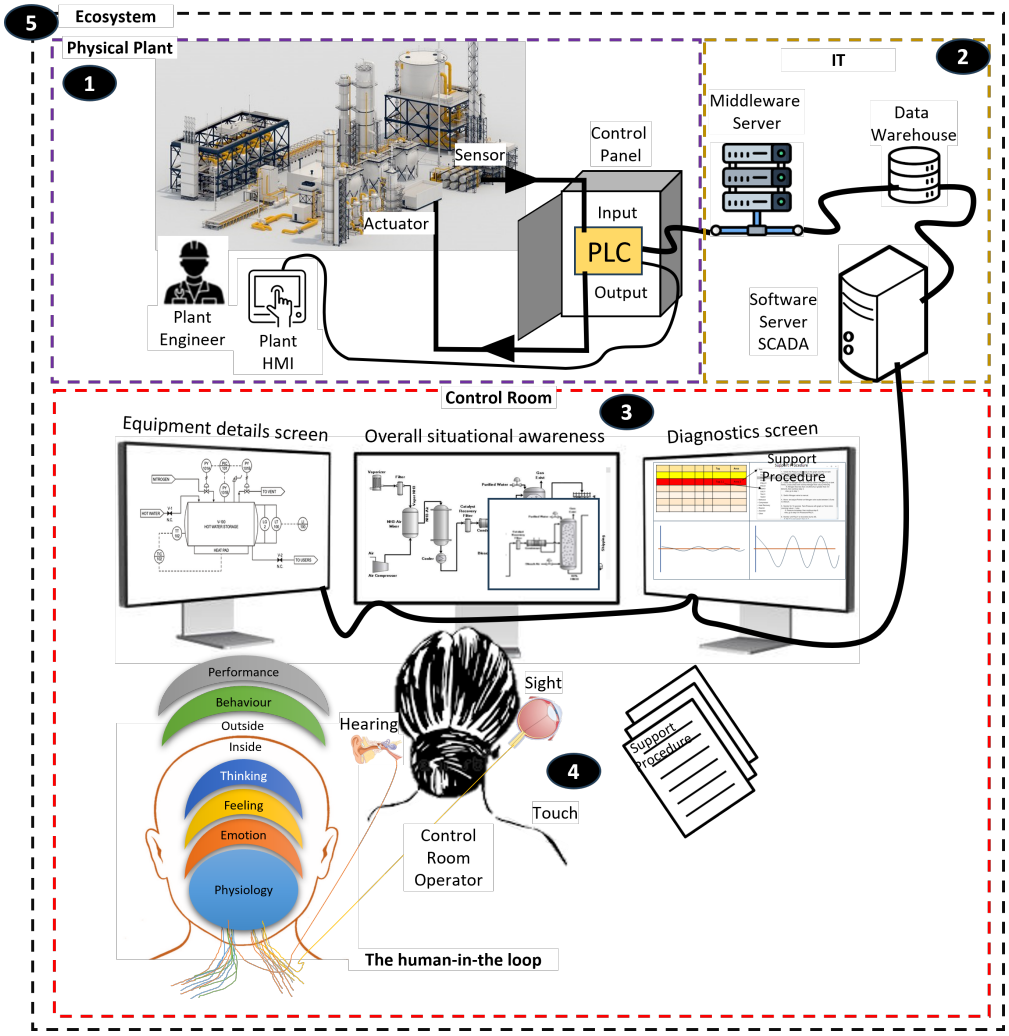


Fig. 1.1 The Process Plant and Key Components. Label 1 contains the physical plant with its systems, subsystems and perhaps a field operator or those on the field. Label 2 contains the IT infrastructure for connectedness. Label 3 is the process control room with the human system interfaces, while label 4, also part of this control room, is the human-in-the-loop. Label 5 is the environment.

Chapter 2

Literature Review

2.1 Introduction

This chapter explores the intricate interplay between human cognitive functioning and the control room environment within process industries. The chapter is structured to progressively build an understanding of the human sensory and cognitive systems, the physical process and control room setup, and the human reliability analysis and performance assessment methodologies, culminating in identifying gaps and formulating research contributions.

Section 2.2, Human Senses and Cognitive Functioning, begins with a focus on the human-in-the-loop (HiTL), providing an overview of the human sensory system and brain functioning. This section lays the groundwork by examining how sensory inputs are processed and how cognitive functions such as memory, attention, situational awareness, task switching, and decision-making are facilitated, thereby establishing the basis for understanding the operator's behaviour in control rooms. On the other hand, Section 2.3, Physical Process and Control Room, transition to the physical process and control room environment. This section outlines the essential components of physical plant process control, including *Programmable Logic Controllers* (PLC's), *Supervisory Control and Data Acquisition* (SCADA), *Distributed Control System* (DCS), and those of the process control rooms; alarms, procedures and display.

Section 2.4, Procedures, addresses the role of procedural guidance in control rooms, exploring different types, formats, and levels of procedures. Past accidents and subsequent recommendations are reviewed to highlight the critical role of procedures

and identify existing gaps. Following this, Section 2.5, Human Reliability Analysis, discusses the evolution and application of HRA methods, spanning first, second, and third-generation techniques and the challenges these methods face. Furthermore, Section 2.5 discusses current advances in the Process Control Room *Human Performance Assessment* (HPA). Objective and subjective techniques used in the process industry are compared, including the integration of AI in human performance studies. This discussion in Section 2.5 sets the stage for identifying specific gaps in current practices.

Section 2.6, Summary of Gaps, consolidates the identified shortcomings in procedures, particularly alarm response procedures, and human reliability assessments for *performance-based analysis* (PBA). This section underscores the necessity for improved performance assessment techniques. The chapter concludes with the contributions, presenting the research questions and outlining the goals of the study. This paves the way for subsequent chapters, setting a clear direction for the research undertaken to address the identified gaps.

2.2 Human Senses and Cognitive Functioning

The operator in the control room, referred to as 'the human in the loop', is central to overseeing operations in manual and semi-automatic controlled industries across multiple domains (e.g., petrochemical, nuclear, and aviation). The term human in the loop encapsulates a holistic view of the human and their dynamics during interaction (human, automation or artificial intelligence) [2]. An operator plays an important role in ensuring safety in these safety-critical environments. They interact with these plants or systems using interfaces where they monitor process trends and alarms, read procedures, plan, make decisions, and perform direct controls over continuous periods. These cognitively intense activities are possible given the dynamics between the human senses and require input from memory. The optimal interplay between these parts of the human is pronounced when dealing with alarmed situations, given that proper information filtering by the attention mechanism helps reduce cognitive load and enhance situational awareness.

A recent study on the 100 most significant losses in the hydrocarbon industry mentioned several individual (human-related) factors that led to accidents. One of these is a pipe rupture in an *Liquified Natural Gas* (LNG) plant in the United States [19]. One of the mentioned causes of this rupture was operator fatigue from working overtime.

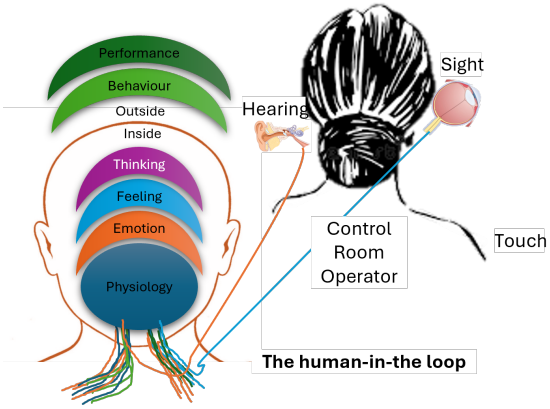


Fig. 2.1 The Human in the loop.

Mental fatigue, which is the type of fatigue experienced mainly by operators in control rooms across domains, results from prolonged cognitive effort [13]. Further on in this section, the human senses, memory, the dynamics between them and how this leads to or not effective alarmed situation response or successful human system interaction, situational awareness and cognitive load are discussed.

2.2.1 Human sensory system

The human senses, seeing, touch, hearing, taste, smell, and balance, work dynamically in some form of cross-talk [20]. According to [21], there are mentions of 12 or even 20 senses, vastly due to these cross-talks, the number of specialised cell types, the type of signals that activate them, and the types of responses they trigger. Each of these senses works by collecting information from the environment and sending signals to the brain. Such signals include motion, chemicals, temperature and light.

In most cases, the operator in the control room, like in Figure 2.1, engages or relies primarily on three (3) senses during interaction with the interfaces: touch, hearing, and seeing. Touch with the hands using input devices such as a mouse, keyboard, or direct contact with the displays, and hear and balance with the ears that listen to acoustic signals from the environments such as alarms, etc. Touch and hearing, including balance, rely on motion detection. The eyes make it possible to interpret visible light and make sense of what is displayed. The inner head layers illustrated as part of the human-in-the-loop image shown in Figure 2.1, proposed graphically by McGrory [22], build on the research

by Beck in the 1960s for Cognitive Behavioural Therapy [23]. These 3 senses and their mechanisms are further discussed below.

1. See: The eye is a vital sensory system for basically light detection, sensing light directions, and forming connections to motor structures to create movement in response to light [24]. The ability to detect visible light with the help of opsins in the photoreceptor cells enables humans to understand their environment visually, including telling shapes better, telling colours apart with the help of cones, and identifying images. Given that the cones are concentrated near the centre, close to the optic nerve, one can see objects clearly within the central field of view and blurry on the periphery. Where objects of salience are placed becomes important for optimal visual attention. The concept of attention is further discussed in later subsections. Other structures of vital importance for seeing include lenses to focus light onto photoreceptors and neurons that could process signals and coordinate behaviour.

2. Touch: Unlike the other senses, the sense of touch has receptors (mostly mechanoreceptors) distributed around the body - the skin, and not located like the others on a specific part of the body [25]. The mechanoreceptors are particularly sensitive to physical contact with the skin and deformation such as bending and stretching [26]. The skin has many sensory endings that detect things like touch, pain, itch, etc. It is possible to understand these so-called sensual feelings because the sensory endings of the skin send signals (temperature, movement, chemical) in the form of electric impulses that travel through neurons to the brain, which then talks with multiple other signals to help understand what is happening on the skin [27]. In a control room setting, this allows for proper feedback to an operator during mouse movement, touch on screens or clicking of buttons.

3. Hear and Balance (equilibrium): hearing is possible due to the sound waves picked up by the ear. The ear is also essential for balance as it contains in its inner part the vestibular system responsible for balance [28]. This system constitutes three semi-circular canals and two otolith organs, which makes head movement possible (upward, downward, right, left, sideways head movements) or the noticing of changes in turning speed and speed of linear movements (otolith organs) after nerve signals are sent to the brain by the sensory hair cells in the ear. The information

sent to the brain then cross-talks with other organs that need to respond, the eyes or muscles [28]. In a control room setting, the operator can hear an alarm sound, move the head or control motion in response to an alarm, and other audible cues as a result of these mechanisms. It is important to note that there is a limit to the sound a human can hear, which informs how loud a system presented to an operator should be. Humans can hear sounds ranging from 12 to 28,000 Hertz [29].

There is continuous communication among the organs of the human body, including the skin, ears, and eyes, with the brain serving as the central hub for this interaction. The functioning of the brain is discussed further in the following subsection.

2.2.2 Brain Functioning

The brain constitutes three systems; the Repililian Brain (responsible for instinctual behaviour), the Neocortex (perception, motor commands, spatial reasoning, language) and the Limbic System (emotional regulation, motivation, learning, and memory) made up of areas like the hippocampus, amygdala, the cingulate cortex, etc. As mentioned previously, the human senses are in a constant state of cross-talk, which happens in the brain. Each sends signals to the brain and is taken to specific areas of the cortex related to each sense. Signals from the multiple senses are then sent to an association region to be combined, creating cross-talk.

For example, when a situation of danger, say red alarms, is presented to an operator, the senses, eyes, ears, or both send information to the amygdala, which is responsible for emotion processing and regulation and interprets what this means. This distress signal is then sent to the hypothalamus, which communicates with the rest of the body through the *autonomous nervous system* (ANS) to control involuntary actions such as heartbeat, breathing, and more. The ANS comprises the *sympathetic nervous system* (SNS) and the *parasympathetic nervous system* (PNS), which trigger flight, fight or freeze responses and rest or digest, respectively [30]. Specifically, the sympathetic nervous system controls pupil dilatation, activating the “fight-or-flight-or-freeze” response, while the parasympathetic nervous system controls the constriction of the pupil and the relaxing or reducing of your body’s activities. In this case of alarm annunciation, researchers observed that a simultaneous combination of both light and sound causes more visual cortex activation than having the flash alone [27]. The combined activation increases

salience and visual attention, as discussed earlier on the eye sense. Such strong stimuli also foster faster reaction times [31]. The cingulate cortex also has a role to play with emotions as it links information that elicits emotional responses to behaviour and actions and decision making [32], while the hippocampus helps to recall and form new memories. The brain functioning and its ability to create new memories from practice or training, recall procedures, and control emotional triggers are essential for control room operators in safety-critical domains, as would be discussed later on, its role in cognitive functioning and the avoidance of accidents.

2.2.3 Cognitive functioning

Cognitive functioning refers to a person's way of acquiring knowledge, and processing information and reasoning [33]. As previously mentioned, sensory processing areas in the brain become activated when we see (eye-movement), hear (auditory), feel, smell things in our environment or remember things in our mind. This new information is perceived and selectively processed alongside old information pulled up selectively from memory when new information is given or requested respectively. An example of the human cognitive functioning stages is illustrated in Figure 2.2 [33].

To give a more practical illustration of what happens to an operator in a control room, an operator hears the sound of an alarm and likewise sees a red colour flash and retains this very briefly in his sensory memory. The accuracy of what is then perceived or understood might not be 100% because they might lose some level of details from what was initially sensed or might have been selectively discarded by the attention function. In particular, the brain finds it difficult to perform critical thinking on simultaneously relevant pieces of detailed knowledge [34]. So, it tends to simplify the level of detail into a pattern or arrangement. Therefore, critical thinking is simplified to an intuitive pattern-recognising process, not the lower-level detail. At perception, the operator can perceive that it is a red-colour alarm, which means that it is a critical alarm. The formation of such meaning is supported by past experiences, which are stored in long-term memory. In this case, the semantic and procedural memories. Perception can then be followed by an immediate silencing or muting of the alarm's audio sound, which is a decision that might be based on a flight emotion response triggered by the sound or sense of danger. This act can engage other senses and brain functioning activities in a manner that can increase workload and stress and consequently affect reaction and response times. Therefore, to

support an operator in such cases, a well-designed alarm response procedure that does not require highly focused attention can help reduce workload and stress and improve reaction time.

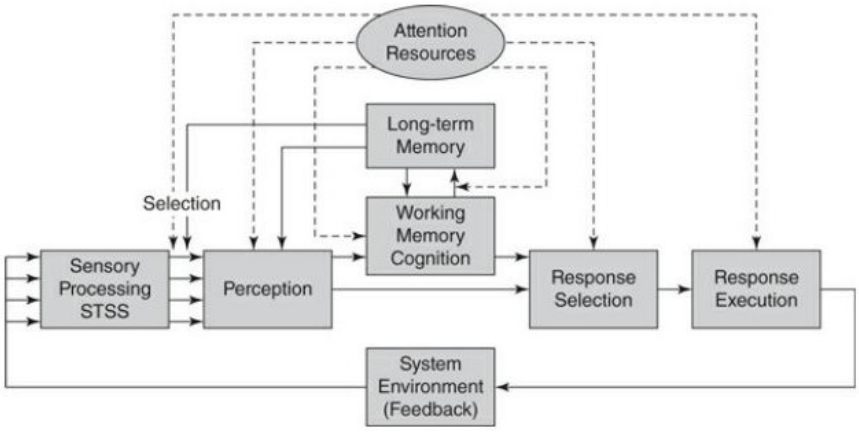


Fig. 2.2 An Model of Human Information Processing Stages [33].

Alternatively, the operator may use working memory to temporarily retain information about the alarm state. During this time, they can check the alarm list for an alarm tag and cross-reference the corresponding alarm response procedure. These tasks are managed in working memory, aided by attention, which helps filter relevant information. Attention functions to allocate cognitive resources to the processes or response stages that require them the most. However, this resource is limited, potentially hindering activities such as optimal task switching, multi-tasking, and other routine control room tasks. These cognitive functions are discussed in more detail below.

Memory

The quantity of information stored or retrievable and the ability to hold it for short or long depends on one's memory and the learning processes: attention, encoding, storage and retrieval [35]. Our sensory memory can only hold information for up to about a second [35]. The working memory, though also limited, is fed by the sensory memory. The working memory is more like the active machine that helps us do tasks with new information or information retrieved from long-term memory. It controls our attention levels, with one function being to filter the information received or recalled [33] [35]. Working memory actively works with short-term memory to temporarily store what it

works with, lasting from a few seconds to 1 minute [35]. This short-term memory holds an average of about 4 - 9 things at a time depending on the level of detail (that is, more logos, but less complex formula) [33], [35]. Subsequently, what is filtered with the right amount of attention is stored in the long-term memory and can stick for about hours to decades.

On the other hand, long-term memory can be further split into three parts: episodic (event-based) memory, semantic (fact-based) memory, and procedural (rules and skills-based) memory [35]. The earlier is one's life-based or daily memories, which are transported short to long-term using the hippocampus. Semantic memory is, however, not connected to our personal experiences and is based on facts, function, and form, such as knowing that a square has four corners or that Nigeria is in Africa. The movement of this type of information into and out of long-term memory is possible using some parts of the hippocampus and mainly by the parahippocampus in the brain. Procedural memory uses brain regions other than the others [35]. Specifically, it uses the Cerebellum, which coordinates movement. This memory includes motor-based skills and tasks, e.g. an operator who has regularly, by practice, learnt to respond to red alarms first.

In general, when recalled, these memories activate the different senses connected to the memory and the attention function, an important learning feature of the working memory, needed to filter for what is needed or not, for example, in a red mist situation. A red mist refers to a fit of extreme anger that temporarily clouds a person's judgement An operator in a control room needs to manage their attention in safety-critical complex situations. Their attention resources are needed for perception, comprehension, projection, decision-making, and even response, as already seen in Figure 2.2. Poor attention, situational awareness/out-of-the-loop situations and high mental workloads have led to accidents in safety-critical industries. Hence, these are essential concepts to discuss, as shown below.

Attention

William James stated that attention is the taking possession by the mind, in a clear and vivid form, of one of several simultaneous possible objects or trains of thought [33]. And perhaps more importantly, he also noted that it implies withdrawal from some things in order to deal effectively with others. As discussed earlier, this withdrawal can be explained by the brain's cognitive functioning. The brain finds it difficult to think

critically about relevant detailed knowledge simultaneously. Attention for a control room operator plays a vital role, given the amount of information taken in visually by the eye. Understanding the mechanism of the eye, as previously discussed, is important for understanding visual stimuli and visual attention. During interaction with the interface displays; monitoring, planning, action, etc., where the working memory is in use, the operator would need to filter across different information and would need to recall information from the long-term memory. Also, the operator sometimes needs to synthesise information from various sources. Effectively doing these would depend on his level of attention to what is relevant, which can be supported by certain tools or cues available to them. In addition, with proper attention functioning, an operator in training can store as much relevant information in long-term memory.

An operator deploys attention in 4 different aspects; selective, sustained, focused, or divided attention [33]. Each has pros and cons, depending on the situation in question. The operator can be so focused on a specific item feature of a display that they fail to notice some other important item - again, the red mist. Sustained attention can come with the cost of missing a target event due to prolonged monitoring of such rare events.

1. Selective attention: For example, selecting which alarm from an alarm list is more important. Selective Visual attention is common among control room operators and pilots and is essentially in our everyday lives. It is engaged during scene scanning, supervisory control, noticing, reading, searching, confirming, and, in many cases, in hybrids. This type of attention is further selected in two ways [33]:

- Top-down attention (filtering theory): top-down attention is a voluntary allocation of attention toward specific features or objects initiated by internal stimuli, for example, when you are able to focus on just one thing and not mindful of other things in your environment [13]. It can also be explained as inductive reasoning where a premise or, say, event is defined, and the possible action and outcomes leading to a consequence or several consequences are explored. Expectancy and Value (the importance of an area of interest to fulfilling a task), two factors that determine where the eye is looking, foster top-down attention and response times [33].

- Bottom-up attention (feature integration theory) is the opposite. Here you exhibit involuntary attention due to salient or pronounced objects in the environment, thus initiated by external stimuli. This involuntary attention situation allows for

multitasking, as an operator can be drawn to new information and tempted to do both concurrently. Salience (the extent to which an area of interest stands out from the background) and effort (the cost of moving attention from one area of interest (AOI) to another) are bottom-up type influences on selective attention.

2. Focused attention: An example is focusing on a specific alarm response procedure.

3. Divided attention: For example, simultaneously visiting the alarm list and the alarm response procedure.

4. Sustained attention / Vigilance: For example, monitoring a process state for a prolonged period.

Attention can be hampered in different ways, such as by noise, fatigue, cluttering and task duration [36], and can lead to situations such as *change blindness* [37] and *inattentional blindness* [38] [39]. According to researchers, inattention blindness (looking at something but not cognitively focusing on those things) is influenced by intoxication [39] or talking on the phone while walking, which affects attention [38]. On the other hand, change blindness — a phenomenon where individuals fail to notice significant changes in their environment is affected by display design [37]. Adverting this entails increasing the salience of critical information on such displays, for example, through effective use of warnings and perceptual cues [37]. Focused attention can be affected by noise from a second operator chattering in the background or by physical or on-screen clutter [40]. Researchers have found that inattention blindness and top-down attention (voluntary focusing on a specific thing) are negatively impacted by mental fatigue [41] [42]. Also, extensive scanning activities can impact selective attention [33].

As briefly discussed, attention is impacted by prolonged tasks and fatigue, leading to several accidents across transport, aviation and chemical industries [13]. *Attention*-based failures result in a type of human error known as slips (intending to do something but doing it incorrectly) [43]. This type of error and accidents as a result, including the cognitive functions behind is discussed further in Subsection 2.5. Attention has been measured over the years with techniques such as the *Attention Network task* (ANT) [44], the *salience, effort, expectancy and value* (SEEV) model [33], and more recently using physiological measures such as eye-tracking for visual attention [10], [13]. Attention Network task (ANT) is a technique based on a flanker task by [45], where one is presented with certain letters in a target location and must respond only to these while ignoring others. It has been used to assess bottom-up and top-down attention by interpreting the behavioural outcomes [13]. The SEEV model estimates the proportion of time the eye

will spend in each Area of Interest (AOI). For example, in a driving scenario, it might predict that the driver spends 50% of the time looking forward at the road, 30% on road signs, and 20% on the dashboard [33]. The eye-tracking technique is briefly discussed here, and its current applications within the process industry will be discussed in section 2.5.4 with detailed characteristics and tools in Chapter 3.

Using current techniques like eye-tracking, measurement of visual attention is possible with an assessment of eye movements, saccades, and the sequences in which the eyes are fixed on an object at a particular time, that is, fixations. These studies have been on since 1878 when Louis Emil Javal observed people's eyes describing the saccades and fixations [46]. The average fixation is between 150 - 600 ms but can also be as low as 50 ms when reading texts in sentences, while saccades can be better at an average of 10 - 100 ms [47]. It is said that humans spend 90% of the time actually seeing (fixating) objects, with the remaining 10% being the blind spot (saccades and blinks) because, at this point, no information runs from the optical nerves to the brain [48]. Other early eye-movement studies have involved studying people as they read (silently or aloud), doing specific search tasks and search patterns [49]. Researchers have also used *Electroencephalogram* (EEG) to monitor attention in drone operators because such signals are affected by fluctuations in attention [50]. EEGs record brain electrical activity using electrodes placed on the scalp. The EEG technique is not discussed in detail as it is not directly relevant to this thesis.

Situational Awareness

The concept of situational awareness is prevalent throughout safety-critical domains of safety. One of the famous figures in this field defines *situational awareness* (SA) as "the *perception (selective attention and noticing)* of the elements in the environment within a volume of time and space, the *comprehension (diagnoses or understanding)* of their meaning, and the *projection (anticipation or prediction)* of their status in the near future" [51]. Wickens [33] further elucidates two schools of thought on SA: a). the one-time situation assessment (a state of knowledge or product and the cognitive processes used to achieve that state) in decision-making and b). a continuous process of acquiring and maintaining awareness in a dynamic environment. SA for an operator is derived not only from the operators' cognition but also from support tools, such as the procedures. This is better understood with the concept of 'distributed cognition', where one does not

need to remember every information received and just needs to refer to the information if required [52]. An understanding of how the operators' cognition and these procedures shape situational awareness during tasks in different scenarios fosters robust design of systems that can better support operators.

As discussed earlier, cognitive functions, attention, and memory work together to make up our ability to acquire and maintain SA. Once information is perceived, with the help of the attention function, it is held in the working memory to comprehend a situation. The duration with which such data is held is significant for an operator in a control room or other front-line operators prone to active error [53]. However, this is difficult as our capacities are limited, and the level of limitation varies per individual. For such operators, optimal situational awareness levels (perception, comprehension, projection) are essential to performing effectively during normal operations and, even more importantly, given abnormal events.

In addition to our already limited capabilities, one's ability to maintain situational awareness can be impacted in several ways or factors. For example, longer time on tasks increases preservation (tasking longer to adapt to rule changes), decreases planning and evaluation capabilities [54], [55], and increases alert fatigue [13]. Situations of alert fatigue are prevalent in the control room where there is prolonged alarm response with ensuing alarm floods [13]. SA is also impacted by automation or adaptive automation [56], [57]. Automation referred to by the researchers is process-driven automation leading to system complexity or information overload and adaptive automation or interfaces. Chen & Barnes [56] in experimental studies observed that time-based and behavioural performance increased with automation, but this was paired with a decrease in situational awareness. According to Onnasch [58], this decrease in situational awareness when using automation is because automation keeps the operator out-of-the-loop when all is going well but fails to re-engage the operator adequately when it fails. The discussed SA impacts pose serious consequences leading to losses across safety-critical domains such as aviation, nuclear energy, and hydrocarbons.

A notable example in the aviation sector is the crash of Eastern Air Lines Flight 401, an L1011 Lockheed Tristar, in Florida [53]. The inattentiveness of the pilots, who were preoccupied with a malfunctioning indicator light, resulted in their failure to notice the loss of altitude, ultimately leading to the crash. Another tragic example is Air France Flight 447, which crashed into the Atlantic Ocean, resulting in the loss of all on board

[52]. A key factor identified was the crew's inability to diagnose and respond effectively to the situation.

Despite the many day-to-day factors out there that can impact SA, SA can be improved with experience or expertise where the long-term memory, specifically the procedural and semantic, becomes more active. Supporting the working memory, however, would be a better approach for novices [59]. Also, regular simulation training can be used where the operator completes tasks that are not normally occurring but are needed to be fresh in the working memory. For example, pilots who have similar regular training to recall how to handle emergencies safely. Some researchers have also proposed tools to support situational awareness of operators. These include the use of *virtual reality* (VR) [60], presentation of playbacks or timelines of events with bookmarks [61], and temporal image/object presentation and removal as effective to increasing SA [62]. However, to be able to control the success of a support mechanism it must first be accurately and consistently measured.

Various methods and metrics have been used by researchers and industry to assess situational awareness. Subjective methods, such as the *Situation Awareness Global Assessment Technique* (SAGAT) referenced as far back as the late 18th century, have been used for situational awareness assessment [63]. Given its shortcomings, an alternative method known as the *Situation Present Assessment Method* (SPAM) was introduced. Implicit performance-based measures have also been used to assess situational awareness with a focus on specific outcomes and how they classify good or bad SA. More recently, objective techniques such as eye-tracking have and are still being explored in capabilities for accessing SA [10]. However, each of these methods provides unique capabilities for assessing either subjective or objective SA. A general recommendation is to create a more robust assessment by combining both objective and subjective methods [64], [9]. The SA assessment techniques to be applied in this research are further explained in Chapter 3.

Cognitive flexibility or Task switching

The concept of cognitive flexibility often used interchangeably with task switching (the actual process) is briefly discussed here to highlight another vital concept that operators exhibit and have led to accident in the safety-critical domain. Cognitive flexibility is the ability of an individual to change rapidly from one task to another [13]. It is one of

those vital interaction concepts that operators exhibit due to the increasing complexity of processes and interfaces. However, this sometimes leads to mental fatigue, which consequently increases reaction times while decreasing accuracy, as mentioned by [65]. A similar effect on cognitive flexibility is evident during prolonged task times depending on the type of support available, like cues [66] or enough preparation time [67].

This concept of cognitive flexibility, as defined here, differs from another task engagement mechanism that operators seem to engage in often — multitasking. Multitasking entails concurrent task engagement, unlike task switching [13]. A typical example is lacing a shoe and talking. The ability to multiply our limited attentional resources for such tasks becomes inhibited when one of these tasks becomes complex, for example, a sudden need to count backwards from 99 in 7 seconds while lacing the shoe.

Decision Making

Perception, attention, diagnosis or situation assessment all lead the road to decision-making. Decisions are essentially the choices produced by mapping a few or many pieces of information perceived to responses [33]. Figure 2.3 illustrates the key cognitive functioning areas leading to this making of 'choice'. Using the process control room operator as an example, the operator picks a cue, perhaps an alarm, perceiving it as a critical alarm and then diagnoses the fault with the help of selective attention (bottom-up) and long-term memory. The latter provides the possibility to hypothesise, evaluate causes, make probabilistic estimations, etc., leading to her 'choice'. However, this choice is not guaranteed optimal because of our limited cognitive resources and perhaps the cues given.

To expand on the above decision-making model, it is worth noting that decisions can be made based on heuristics/biases (which can save time but lead to errors). Secondly, decisions are made intuitively (naturalistic decision-making). This intuition is based on skill and expertise and can be recognition-primed or based on pattern matching as earlier discussed [33] [34]. These decision-making patterns are common amongst experts, including expert front-line operators and is evident that the former comes with risks. Nevertheless, for experts and novices, in addition to familiarity and expertise, decision-making is also shaped by the degree of uncertainty of the consequence and the time involved [33].

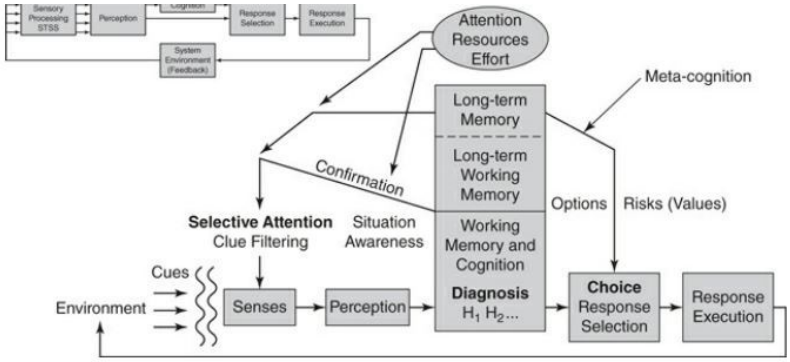


Fig. 2.3 Information Processing Model of Decision Making [33].

As already mentioned, our decision can result in positive and negative outcomes. When the latter is the case, it can come with little or severe consequences, depending on what is at stake. In safety-critical domains, front-line operators make decisions that can lead to or not to severe consequences. For example, in the aerospace sector, the Challenger Space Shuttle exploded during launch due to a misjudged decision to launch at cold temperatures, which eventually destroyed the seals. [33].

The good news is that decision-making capacity can be improved. Some researchers have used and proposed methods such as training [68] [69], procedurisation [70], and displays [71]. The first involves systematic approaches aimed at reducing biases by promoting critical thinking, providing comprehensive feedback, presenting data effectively, and emphasising probabilistic reasoning. Procedurisation enhances decision-making by offering clear guidelines and structured methodologies [70]. It involves breaking down decisions into manageable components, applying formal techniques, and promoting systematic analysis to improve objectivity and effectiveness. Unlike debiasing, which directly addresses and mitigates cognitive biases [69] and procedurisation provide a framework for more structured decision-making.

Additionally, the effective display of information is crucial in supporting the early decision-making stages. Well-designed displays optimise the processing of information and enhance decision-making effectiveness by facilitating the integration of signals, minimising cognitive effort, and reducing biases. Whether using pictorial or graphical representations, as opposed to text, numbers, or verbal formats, these displays are structured to reduce attentional effort [72] [71]. In doing so, they improve the way information is processed and integrated, leading to more effective decision-making [73].

Facilitating Cognitive Functioning

As discussed earlier, understanding how humans process information is essential for improving the design of systems that support users across various domains. The format in which a message is presented significantly influences attention and comprehension [72]. Stone et al., [72] found that combining icons with text enhances readability and recognition compared to using icons or text alone. Additionally, auditory cues, such as flashes and sounds, have effectively captured attention and improved response times in critical situations [71]. Moreover, shared cues between the primary task and unexpected stimuli can support situational awareness and task performance [74].

This section has thoroughly addressed the complexity of human nature and the consequences of overlooking this crucial aspect. The goal is to emphasise the intricate interactions between humans, the systems they engage with, and the surrounding processes—whether these interactions are direct or indirect. This discussion underscores the impact of individual factors on safety. The next section will explore these systems of interaction, with a particular focus on those relevant to the process industries

2.3 Physical Process and Control Room

2.3.1 Physical Plant Process Control

Process plants, including those in chemical manufacturing, oil and gas, pharmaceuticals, and refineries, are safety-critical facilities that handle highly regulated materials. Consequently, the operating metrics of these plants are meticulously monitored and controlled for aspects such as: within specific limits for pressure, temperature, flow, and level. The primary objectives are to optimise production, ensure quality, reduce costs, and, most importantly, prevent risks that could lead to catastrophic events like explosions, fires, loss of containment, or loss of life. The materials are stored in systems designed to suit their properties and control their volumes. These materials pass through various systems throughout the production process, such as valves, reactors, and separators, to achieve production goals.

The central system involved in the transportation, storage separation, etc., of the materials function with the help of specific equipment that records (recording pressure gauge, etc.) and controls their functioning from single loops to larger control loops. This

equipment also serves as a safety barrier as it can perform various functions, from cutting off flow to shutting down the plant in high to very high abnormal situations. Figure 2.4 shows at what stage these basic process controls are on the process safety pyramid, and one can see their foundational importance in preventing tier 1 and 2 events. There is other equipment, such as the measuring indication instrument (gauges, etc.), which, when faulty, can lead to poor situational awareness and inadequate response by operators, inadvertently leading to an accident. These instruments requiring direct interfacing will be discussed later in this section.

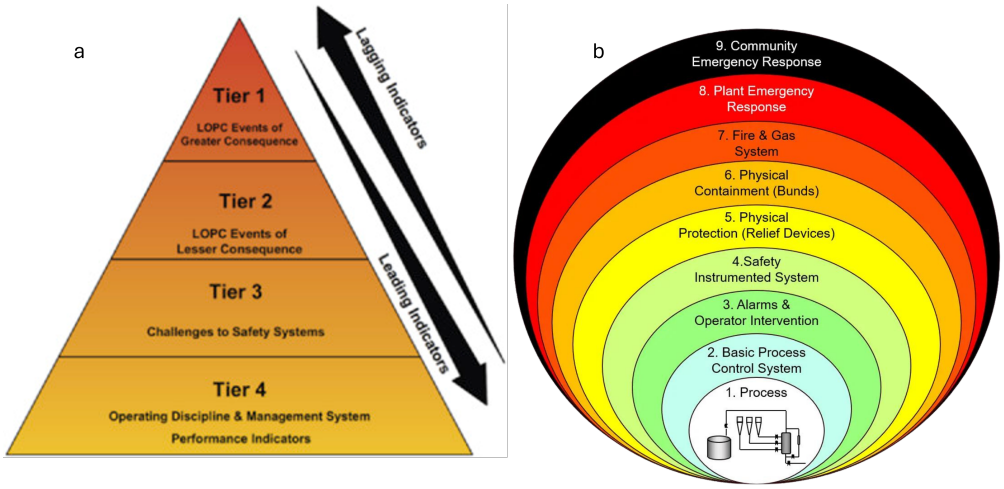


Fig. 2.4 Process Safety Pyramid and layers [75], [76].

Process plants, along with the control systems that underpin their operations, have become increasingly sophisticated over the years. This evolution reflects the growing complexity of process architecture and the rising demands for advanced control systems [77]. Early use of feedback controls in manufacturing dates back to the 18th century, with examples such as the infamous centrifugal governor invented by James Watt to regulate steam turbines [78]. This developed further to on/off controls in the 1920's [79], [80]. Such on/off controls, mainly temperature and pressure controls were primarily built and operated as air-actuated or self-contained. The earlier versions used air to snap open and close valves or switches, allowing control within a narrow temperature range. The later system, actuated by the medium it controlled, gradually opened and closed to maintain the desired set-point, typically between 370° and 390°F [80]. Throughout the 20th century, these systems evolved into modern automatic controllers, beginning with *proportional*

controllers (P) , followed by the addition of *Integral controllers* (PI) , *proportional-derivative controllers* (PD) , and ultimately, *Proportional-Integral-Derivative* (PID) controllers. Today, a more advanced controller in use is the *Programmable Logic Controller* (PLC), which will be discussed in the following subsection along with other higher-level systems as illustrated in Figure 2.5.

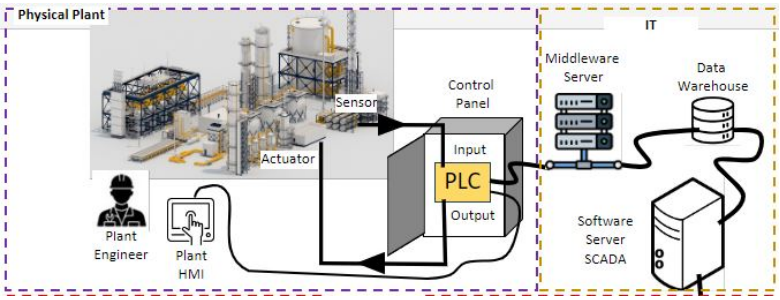


Fig. 2.5 Basic Process Controls and Network.

PLC's, SCADA, DCS's

Programmable Logic Controllers (PLCs), developed in the 20th century, can perform complex control functions. They can be programmed using various languages, such as Ladder Diagrams, which employ logical rule-based clauses in the form "If <Condition> then <Action>" [81]. PLCs are directly connected to machinery to read variable statuses, compute control outputs, and adjust actuator settings. Compared to Proportional-Integral-Derivative controllers (PIDs), PLCs offer greater robustness in handling multiple simultaneous control tasks and operations. They transmit information from machines to Supervisory Control and Data Acquisition (SCADA) systems or Distributed Control Systems (DCS). SCADA systems consolidate data from multiple PLCs for higher-level control, data acquisition, and supervision, often from a centralised location. In contrast, DCS structures are more decentralised, with control functions distributed near the processes they manage. Both SCADA and DCS provide interfaces between machines and humans, allowing operators to monitor processes and alarms. Alarms notify operators of malfunctions requiring action and are configured to trigger when sensor readings exceed allowable process parameter limits, such as high or low-pressure levels. Although the human-machine interface, including alarms, is discussed in the next subsection, it is crucial to highlight accident scenarios arising from these technical systems or controllers, also known as technical factors.

Mechanical failures and other technical issues from the plant have led to accidents resulting in explosions, fires, injuries, and fatalities, as documented in [19]. Notable examples in hydrocarbon plants include autoignition in filters at Hammerfest, Norway; seal leaks at Cactus Reforma, Mexico; pipe ruptures in Port Neches, Texas, and Flixborough, UK; and structural failures in Deer Park, Texas (1997), Pajaritos, Mexico (1991), and Vohburg, Germany. Additionally, blowouts and mechanical failures have occurred in Enchova, Brazil (1988), and Jubilee Field, Ghana, with corrosion cited as a contributing factor. Natural disasters have also impacted systems, causing accidents, failures, shutdowns, and significant financial losses. Examples include lightning strikes in Komo, Papua New Guinea (2018); flooding in Cedar Bayou, Texas, US (1994); storms in Gryphon, North Sea, UK (2011) and near Newfoundland, Canada; and the earthquake in Sendai, Japan [19].

To prevent such accidents, plant processes and systems are monitored from process control rooms, where control room operators and their supervisors can detect failures or issues through alarms that alert them to take action. The process control room and its components are discussed below.

2.3.2 The Process Control Room

A process control room is a dedicated area within an industrial plant where operators monitor and control the various processes that maintain the effective running of the facility. Control room operators monitor process parameters in the plant, such as pressure, concentration, temperature and flow rates. They can also control these parameters from the control room and change their set points when necessary. These rooms are equipped with advanced technology, which can include control panels, computers, networks and monitoring systems to oversee and manage the complex operations of the plant, as shown in Figures 2.5 and 2.6. These systems facilitate communication between the operator, other personnel or different parts of the plant. Data loggers are also in place to store the data for future retrieval, analysis, and process optimisation.

The process control rooms are central pivotal locations in a plant and are highly visual safety-critical environments, as Figure 2.6 demonstrates. They date back to the early 20th century, during the second industrial revolution [82]. The room started as a central location for overseeing plant operations, performing manual operations that affect the plant, resetting values as required, etc. The layout and the design of the human-machine

interfaces or elements, graphic displays, keyboards, screens, etc, necessary for these types of tasks have evolved over the years. This evolution has not only denoted a change to the systems of interaction but also to how the operators now have to interface with the plants, their cognitive processing and how the systems are designed to adapt to operators. The critical human system interface elements of relevance in this thesis are discussed in the following subsection. They have been identified by researchers and included in standards as key for managing process operations, especially considering their impact on ensuring safety [5], [83], [84], [85].

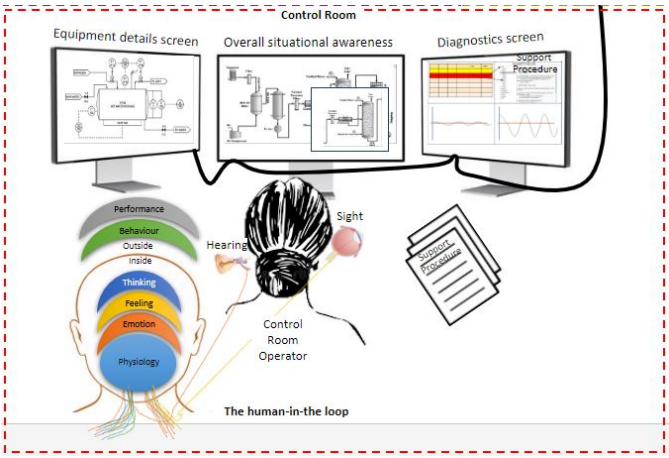


Fig. 2.6 Demonstrating a Control Room Setup with both paper and electronic procedures.

2.3.3 The Human System Interfaces

The human-system interfaces described here, including alarms, displays, and procedures (whether display embedded or printed), are essential elements in control rooms and key factors influencing human-system interaction. Control room operators use these interfaces to monitor plant conditions, respond to alarms, and perform control actions based on the feedback they receive. This subsection provides a brief overview of alarms, with displays and procedures discussed in the following sections.

Alarm System

As already mentioned, operators are alerted by alarms that are visually seen or audible. Alarms direct operators' attention towards a plant state that has deviated from the set

Table 2.1 Comparison of Warning, Alarm, and Prompt.

Description	Warning	Alarm	Prompt
Purpose	A warning is intended to draw the operator's attention to a condition/parameter approaching a problematic threshold. It serves as an early notice that attention may be needed soon.	An alarm is designed to stimulate the operator's attention to a condition that has already reached a critical level and requires immediate attention. Action is needed to prevent safety hazards, equipment damage, or significant operational disruptions.	A prompt provides guidance, reminders, or instructions to operators to assist them in performing tasks correctly and efficiently. Prompts help ensure that procedures are followed and tasks are completed as needed.
Urgency	Moderate urgency. Indicates that there could be an issue if the current trend continues, but immediate action may not be required.	High urgency. Indicates that a parameter is outside safe operating limits and that immediate intervention is necessary.	Low to moderate urgency. Prompts are generally about enhancing efficiency and accuracy rather than responding to imminent danger or critical conditions.
Response Required	Operators should monitor the situation closely and prepare for potential action. The goal is to prevent the situation from escalating into a more critical state.	Immediate action is required to address the issue and bring the condition back within safe limits. This might involve emergency shutdowns, manual overrides, or other urgent corrective measures.	Operators are expected to follow the guidance or instructions provided by the prompt. This could involve entering data, performing a routine check, or following a procedural step.
Examples	Temperature nearing the upper threshold in a process. Pressure getting close to maximum safe levels. Low fuel levels approaching the minimum operational limit.	Pressure exceeding safe maximum limits. A pump or motor failure. Detection of a gas leak or fire.	A reminder to perform a scheduled maintenance check. Instructions to enter data into a log. Guidance on the next step in a complex procedure.
Nature	Indicates a potential future problem. It's a cautionary notice.	Indicates the actual detection of a current problem. It's a cautionary notice.	Provides guidance or reminders. It's informational and supportive.

limit, has reached a critical level, and requires the operators' action [86]. Based on this definition and before further discussion on alarms, it is important to differentiate between some common words used in association with alarms (see Table 2.1). One of these concepts is the alert. The alerts replaced alarms from previous alarms that did not require operators' action; high/low notification of deviations. Alerts/warnings differ from alarms because they do not require a response. Likewise, messages do not require operators to take action and provide the status of a normal operation. Prompts, on the other hand, are similar to alarms and require operator action and serve as reminders or notifications of instructions for doing a task [86].

Alarms are set to different levels (high, high-high, low, low-low) for the various process parameters in each system where they are present. A breakdown of the alarm tag notation is shown in Figure 2.7. This taxonomy is standard in the industry. These alarms are subsequently prioritised for the sake of operator response. The priority on which to be resolved first includes the consideration of possible resulting consequences and the time available. Alarms that have higher consequences and require a quick response are assigned higher priorities than the others. Those with the least priorities are set as warning alerts, prompts or messages depending on the goal.

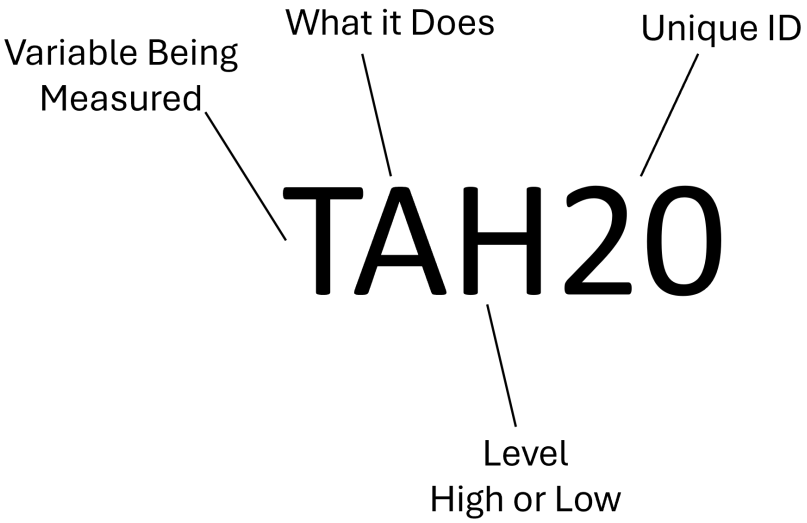


Fig. 2.7 Alarm TAG Notation.

Accidents in process industries due to poor alarm systems have led to significant losses. Notable incidents include the BP Texas Refinery US and the Buncefield Storage

Depot fire in Hertfordshire in 2005, as well as the three-mile island incident. Some of the precipitators of these accidents were false alarms, lack of an effective alerting system, and alarm floods, respectively [87], [1]. In addition to nuisance alarms or their absence and alarm floods [18], several other alarm management problems have been identified as contributing to accidents. These include missing or neglecting significant alarms, improper alarm shelving, inadequate alarm installation, poor design [86], [18]. In an experimental study by Amazu et al. (2024), the researchers used black text for the alarm tag (an abbreviation given to an alarm for easy recognition of the process parameter and condition being alarmed) name on a red background [18]. This action reduced text salience and led to two operators responding to a wrong alarm tag. A recent incident in 2022 entailed the rupture of an *Liquified Natural Gas* (LNG) pipeline, which remained obstructed for an extended period due to ineffective alarm management [19]. Efforts to address alarm-related issues have involved alarm rationalisation and redesign of the alarm philosophy [86].

Screen Displays

The displays on the screens of monitors, SCADA or DCS systems allow cross-talk/interfacing between the operator and plant. They provide control room operators with information on trends of process variables and real-time plant conditions with the possibility of direct control in manual or semi-automated plants. Display screens in control rooms were in the 20th-century analogues displays which were arranged on so-called 'Control walls' [88]. Driven by the need for less difficulty in interaction, error reduction, and the general push from the industrial revolution, the interfaces have evolved over the years. From the late 20th century to digital technologies for SCADA and DCS, with the introduction of cathode ray tubes, and finally to digital graphical displays. These digital graphical displays have themselves become more sophisticated over the years [88].

The design of these displays has evolved over the years, initially focusing on functionality. However, accidents resulting from poor design and ergonomics highlighted the need to integrate human factors into their development. Today, human-centred design is essential, taking into account both the physical and cognitive limitations of users. Concepts such as memory, attention, situational awareness, and

decision-making are now integral to the entire lifecycle of interface displays, from specification and design to implementation and operation [89].

Displays are typically categorised into four levels [89]. Level 1 provides general information, such as a process flow diagram of the plant, while Level 4 offers more detailed and specific information, such as displays focused on particular systems, procedures, and other critical details. Higher-level displays (Levels 2–4) allow operators to access finer details and include alarm lists. In more advanced control rooms, alarm response procedures are integrated into these displays in computerised formats. The next section will further discuss the various types of procedures. However, it is important to note that display issues have contributed to several accidents.

A famous one is the Three Mile Island incident. Although alarm floods have already been mentioned as one of the causes, the accident was also precipitated by a faulty interface design, which made room operators unable to correctly control the issue [90]. In the railway sector, a deadly train accident occurred in England (the Ladbroke Grove rail incident) because it took the supervisors considerable time to search through a traffic display and identify which train was causing a collision alarm to sound [4]. The need to support operators by designing systems that are better tailored to the users and that better support situational awareness, attention, and reduced workload cannot be overemphasised. Standards such as IEC 63303 (Human-Machine Interfaces for Process Automation Systems) exist to guide those involved in the *Human Machine Interface* (HMI) role on how to design such systems. IEC 63303 focuses on providing a standard for the design, implementation, operation and continuous improvement of an HMI. It has as its basis an HMI philosophy to guide the design and use of the HMI and a style guide and toolkit for the philosophy implementation.

Despite such available standards for HMI development and improvement, a recent industrial study showed that some of the discussed issues and concerns related to HMI still persist. From an industrial survey by Amazu et al. (2023) on process industries, a couple of recommendations were given by Distributed Control System (DCS) experts and operators on what must, should, and could be incorporated for a human-centred design of the displays [91]. These can be summarised briefly: must be simple to operate, provide suggestions of decisions, and ease navigation. They should reduce fatigue, be clear, consider the *Human Factors* (HF) good practices in standards; could be more interactive and have simple colours. Here, one not only sees the priority of some of these

concerns but also sees that some of the issues prevalent some years ago are still concerns today.

To enhance operators' cognitive functioning and performance, an experimental study on *Unmanned Aerial Vehicle* (UAV) operators provided them with notifications on the display when their attention levels decreased. This approach successfully helped operators regain focus during tasks [50]. Similarly, in the air traffic control sector, a study found that covering the display in a red hue when there was a risk of mid-air collision significantly improved detection rates, understanding, and overall performance [92].

The next section will discuss another critical human-machine interface: procedures, which, in advanced *Human-in-the-loop* (HITL) configurations, are often integrated with displays.

2.4 Procedures

2.4.1 Procedural Guidance in Control Rooms

Procedures, whether operating procedures or alarm response procedures, are crucial performance-shaping factors and fundamental to safety management. The health and safety Executives define the procedure as "the written description of steps you need to follow to perform a task" [87]. In human reliability studies, they are classified as technological, task-related, or organisational factors, all of which are critical to human-machine interaction, particularly in safety-critical situations [93], [94]. These procedures support control operators in managing process plant operations and safety, including startup and shutdown of a plant, normal operations, and emergency operations (see Figure 2.8). Operators rely on these procedures for guidance and information to ensure safe operations, which can be computerised or paper-based. Ideally, these procedures should assist operators within the available time provided by early warnings. They must be easily accessible and user-friendly to avoid wasting valuable time, especially in modern process control configurations where operators are often out of the loop.

The design of procedures, levels, format, types, etc. of procedures have not come without challenges. Such challenges identified around this 20th-century period included procedure interfacing issues where operators had to use multiple procedures simultaneously, inconsistency in the level of details written, incomplete procedures, poor navigation support, poor procedure adaptation for diverse experience levels and task

complexity. These issues have resulted in operators following wrong procedures, feeling overwhelmed, missing relevant information, and poor task performance for inexperienced operators and non-compliant behaviours [95], [96]. Other researchers have explored design constraints concerning procedural formats (medium of presentation and style) and their implications on operator workload, particularly during emergencies. As discussed on procedure levels, Industries have transitioned from paper-based to screen-based and computerised procedures over the years [9], [97]. However, the impact of this transition, the pros and cons from both sides, have been scarcely investigated in the nuclear sector with no investigation to the best of the author's knowledge within the process industry. Accidents resulting from using these procedures over the years are discussed later in this section.

The procedures discussed in this section are those used by the front-line operator for active operations and not, for example, management procedures. It is essential to remember that this thesis focuses on alarm response procedures. Notwithstanding this, discussing the different types, categories of procedures and alternatives is necessary. The procedure hierarchical scheme is shown in Figure 2.8[95].

Type of Procedures

There are different types of procedures depending on the expected function, as shown in Figure 2.8. These include *abnormal operating procedures* (AOP), *normal operating procedures* (NOP), and *emergency operating procedures* (EOP) [95]. The abnormal operating procedures comprise the *alarm response procedure* (ARP) and *Off-Normal Condition Procedures* (ONCP). Alarm response procedures (ARPs) are abnormal operating procedures. They are used in response to alarms notifying operators of a plant process upset or equipment failure requiring their response. They do not include alarms that are not part of the primary process alarm systems and will need simple responses [98]. While ARPs focus on immediate responses to alarms, off-normal condition procedures handle a broader range of scenarios where conditions deviate from normal but are not yet emergencies. They include comprehensive instructions to prevent these conditions from escalating. Normal operating procedures are likewise broken down into start-up/shutdown procedures and equipment operating procedures/maintenance. NOPs provide instructions for the plant's general operations. Startup and shutdown procedures provide instructions for start-up and shutdown procedures, respectively, as well as

equipment maintenance procedures. Emergency operating procedures are those used by operators to reduce the consequences of a broad range of accidents or equipment failure [99].

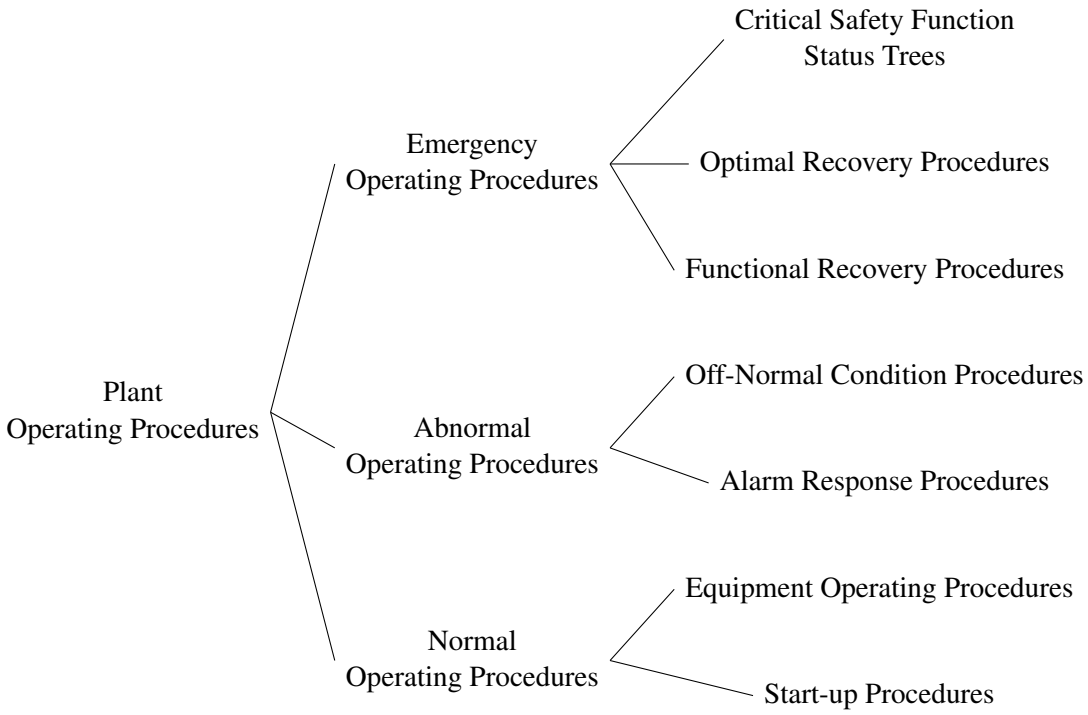


Fig. 2.8 Hierarchy of Plant Operating Procedures [95].

The procedures discussed differ from policies, standards, and guidelines, and it is recommended that these documents not be combined. Policies outline strategies for achieving business objectives, while standards set mandatory requirements to support these policies; guidelines offer detailed references and support for procedures and standards [5].

Formats

Depending on the task complexity, task familiarity and frequency of use, these procedures can come in the form of job aids, step-by-step guides, diagrams or narrative procedures [4]. Step-by-step guides are primarily used for medium/high complexity tasks with medium/high familiarity and in infrequent/rare situations. Job aids can be Checklists (low task complexity, high familiarity), which are used to check all important elements

systematically have been considered, or checklists plus visual aids or decision trees (high task complexity, high familiarity) or diagrams used to illustrate the layout of a unit or flow charts which are used to show how control of work system operates or as narrative/no written procedures. Given the mentioned criteria, a recommendation of which format to use is better explained in Figure 2.9.

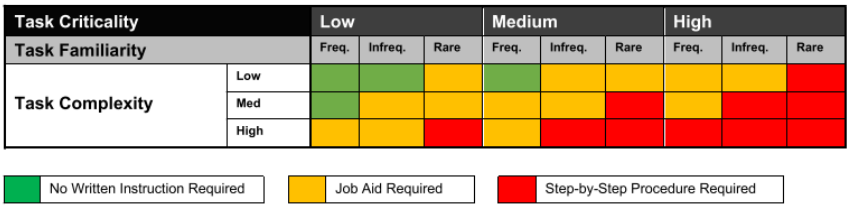


Fig. 2.9 Procedure Format [5].

Procedure Levels

In the nuclear sector, there is an ongoing transition from *paper-based procedures* (PBP) (level 0) to computerised support procedures across various functional levels: Level 1 involves *Electronic Procedures* (EP), Level 2 encompasses *Computer-Based Procedure* (CBP) Systems, and Level 3 integrates CBPs with *Procedure-Based Automation* (PBA) as detailed in Table 2.2 [15]. Depending on the level, operator support ranges from basic functions, such as selecting and displaying procedures on a computer screen, to advanced integrations, allowing a single operator to execute multiple coordinated steps through soft controls [100]. This transition has necessitated accidents over the years spurred by poor procedures.

The advantages and disadvantages of the different procedure levels on situational awareness and mental workload have been assessed by researchers within the nuclear sector. Computerised emergency operating procedures, for example, have been identified to significantly impact operators in digital *main control rooms* (MCR) [101]. Yang et al. (2012) in an experimental research observed that workload (assessed using *NASA Task Load Index* (NASA-TLX)) was higher when using PBPs, followed by EPs and low mental workload with CBP [15]. This is also evident in the situational awareness (assessed using SART) outcome, in that those using the paper-based procedure had lower situational awareness than others. Some other researchers in the nuclear domain have mostly focused on assessing different electronic-based formats [9], [97]. Gao et

Table 2.2 Levels of Procedural Support Systems and their Descriptions [15].

Level	**Description**	**Purpose**
Level 0: Paper-based Procedures (PBP)	Procedures provided in hard copy/print formats are still used in some industries like chemical manufacturing and aviation.	Used as checklists or manuals to guide operators or pilots during routine operations, such as take-off preparation or standard operating procedures.
Level 1: Electronic Procedures (EP)	Text procedures are presented on displays in the form of lists or flowcharts, closely resembling paper-based procedures. They include navigation links to other procedures or displays.	• Provides an electronic format of procedures. • Offers ease of navigation and access to related information or alarms, enhancing operational awareness and efficiency.
Level 2: Computer-based Procedures (CBP)	Computer-based procedures (CBPs) offer significant advantages over Paper-Based Procedures (PBPs) and Electronic Procedures (EPs) in that they automatically retrieve and display necessary information, integrate relevant indications and process step logic with result displays.	• CBPs check prerequisites across multiple steps, provide soft controls for procedure actions, and offer context-sensitive aids for decision-making. • Additionally, they issue cautions or warnings based on current plant conditions, enhancing operational efficiency and safety.
Level 3: CBP with Procedure-Based Automation (C-PBA)	include embedded soft controls to issue control commands to plant equipment and an automatic sequence of steps along with other sophisticated functions. Procedure-based automation is a form of automation applied to the procedures by grouping multiple deterministic steps into predefined sequences.	• Provides similar benefits as CBP. • Additionally, automates control actions that directly impact the plant and makes decisions regarding the execution of steps, in contrast to CBP. Control take-over can be initiated by the operator

al. (2013) investigated the impact of two computerised emergency operating procedures with different complexity levels on mental workload.

On the other hand, Xu et al. (2008) performed an ergonomic study of two emergency operating procedure flow chart styles. Though the results in these studies are clear on the impact of these types of procedures focusing on workload, it has not been considered what other factors might have constituted these outcomes. It is essential to consider the influence of other potential *performance shaping factors* (PSFs) during such assessments [102]. The NASA-TLX and SART questionnaires introduce subjectivity in assessment and are known to highly correlate with each other [17]. Therefore, a combination of other methods with an overview of behaviour and performance outcomes, in comparison with the subjective assessment, would provide a more robust insight. In the process industry and to the best of the author's knowledge, there have not been investigations on these types of procedures to compare their benefits and impact in a case of transition. Experimental studies in the industry have mostly presented the procedures as pop-ups to the operators or test participants [103].

Resulting Accidents and Recommendations

Investigations into major incidents have shown that procedures are a key contributing factor [95], [96]. Reasons range from the irrelevance of the procedures to addressing actual situations, procedure overload, poor presentation, and contradictory procedures. Such inadequacy in procedural support can impact operators' performance and, consequently, process safety. Other reasons include frequent or no procedure references and unclear exit and entry conditions [95]. In a study done by Baker et al. (2017) to identify process safety management elements that are main accident contributors (from amongst 770 major accident cases) in chemical process plants, operating procedures were found to be the second most contributors to accidents in these plants at 17% just after process failures at 19% [6].

Notable examples of major incidents influenced by procedural issues include the Piper Alpha Platform explosion and the Longford Gas Plant 1 explosion and fire. Additionally, outdated and incomplete procedures contributed to the BP Texas and Buncefield Depot fires and explosions. Inadequate procedures were also a factor in an LNG line rupture in the US in 2022, leading to non-compliance issues [19]. Similarly, a refinery fire in Europe in 2022 was caused by procedural non-compliance. Even after the recent

accidents and despite some of the standards and guidelines available, it is evident that more work has to be done regarding the design and writing of these procedures in use by front-line operators.

The different formats and levels have to be investigated in terms of specific and overall impacts on attention, situational awareness, and even the behaviour of the operators. New proposals can then be given with such insight. In addition to experimental investigations, Decisions for optimisation can be made using a real-time understanding of operator-system interaction or work-as-done in process operations, as also recommended in [5]. This is to understand, in reality, the gaps in the usage of procedures and the impact of the procedural design alongside other performance-shaping factors on operators. Also, in an industrial survey conducted by Amazu et al. (2023a), some HMI experts and process control room operators provided recommendations for effective procedures, emphasising that procedures must ensure clarity and auto-detection upon deviation [91]. Auto-detection here refers to the automatic detection and update of changes to process parameters originally specified. They should undergo quality testing and evaluation, be accessible and interactive for all users, and incorporate a reminder system to ensure adherence. In addition, procedures could benefit from flexibility and adaptability to operations, improved design, and could provide next-step guidance displays.

Gaps

The pros and cons and another potential impact of a transition from paper to digital or computer-based procedures, have been scarcely investigated in the nuclear sector with no investigation to the best of the author's knowledge within the process industry.

Though researchers and data from accident reports have summarised the role procedures have to play on situational awareness, mental workload and error, other factors that might have contributed to these outcomes have not been considered. According to Zarei [102], it is essential to consider the interactions among variables and potential relationships during human performance assessments. Experimental studies of process control room setups that have used or projected procedures in their studies have presented them as pop-ups to the operators or test participants [103].

2.5 Human Reliability Analysis

In the last section, procedures were extensively discussed as they play a fundamental role in human reliability and risk analysis. They serve as input for task analysis during human reliability or critical safety assessments. Subsequently, the content of job aids, including instructions, warnings, and cautions, are also updated from these task analyses, walkthroughs, and results from hazard identification activity and risk assessments. Human reliability assessment generally involves three steps: task analysis, identifying factors contributing to error in each task, and modelling and quantifying these errors. By quantifying the likelihood of error, the potential risk posed by the human element is captured and included in the case of human participation in process operation during the probabilistic *process safety assessment* (PSA) [104]. This section discusses human performance, old and new HRA methods to assess performance, and the gaps in using these methods.

2.5.1 Human Performance

Human errors, including slips, lapses, mistakes, and violations, are cited as the cause of approximately 75-80% of accidents in process plants (Ref.). This is because each layer of protection against potential hazards often involves a human element, who may inevitably make mistakes. Be it during the alarm design process, the procedure design and writing, or the safety instrumented system design. In the case of front-line operators like process control room operators, each error is influenced by one of the types of human performance: skill-based, knowledge-based or rule-based performance of operators [105]. Likewise, these types of human performance are not influenced by personal factors but have certain task-related and organisational factors that influence human performance. This is evident in the accident scenarios discussed earlier, including the alarms, procedures, and individual capabilities. Table 2.3 provides a comparative summary between risk assessment and human performance in order to understand their unique contributions despite the impact of the human aspects on risk.

A summary of the type of human performance, the human error it leads to, and the factors or support that impact it is presented below.

1. Skill-based task performance refers to completing a task at a high level of accuracy and reliability with a low level of attention. A basic example would be pedalling for a

cyclist. This type of task-based behaviour can lead to execution errors. One is where the human forgets to do something even if intentions were right (lapses), and the other is where the task is carried out incorrectly. The former is linked to the individual's memory, for example, forgetting a step in a procedure. In contrast, the latter is related to attention, for example, misreading an alarm TAG in an alarm list. Slips and lapses can be precipitated by interruptions, fatigue, task overload, stress, noise, etc., [105].

2. Rule-based task performance involves carrying out unfamiliar tasks by memorising and adhering to set rules or written instructions, such as following procedural steps. This type of behaviour is where a person does what they intended to do but does not work, in that the goal or planning was wrong (mistakes). For example, when an operator recognises and applies a bad procedure to resolve a situation. This type of error is precipitated by missing, confusing and incomplete procedures or information and divided attention [105].

3. Knowledge-based performance refers to completing tasks that require information gathering on a situation, analysing and making decisions based on artefacts. Therefore, errors can occur when knowledge is lacking in accuracy, precision, presence or timeliness. Like rules-based task performance, this category leads to planning failures and, thus, mistakes. This type of error is precipitated by a lack of knowledge, information, time, fatigue and stress [105].

Accidents resulting from human error are common and can be attributed to a range of error-contributing factors, including personal, task-related, organisational, and technical issues, some of which have not been previously discussed. While incidents related to procedures, alarms, displays, and personal factors have been examined, numerous older and more recent accidents stem from issues such as poor management practices, staff shortages, communication breakdowns, and inadequate training. Notable examples include the incidents at Longford (Victoria, Australia) and Port Neches (Texas, US) [19]. Other reported accidents involve improper chemical storage, as seen in China (2019) and Toulouse, France, and maintenance errors, such as those at Piper Alpha (UK, 1988), Baker (US, 1989), and Mina Al-Ahmadi (Kuwait, 2000). Additionally, specific procedural failures, such as inadequate reactor purging during shutdowns, have led to accidents, as exemplified by the Pampa, Texas incident in 1987.

From discussions and analysis of accident scenarios across various safety-critical industries in this thesis, it is evident that humans are not the sole actors responsible for accidents. These incidents result from human, technical, task-related, and organisational

factors. Therefore, measuring each factor and its combined effects to understand potential errors is a recommended approach to human performance assessment. Over the years, several methods have been implemented, commonly referred to as human reliability methods, to achieve this, with new tools currently being used alongside these techniques. The next subsection discusses these methods, their successes, and gaps.

2.5.2 Methods, Application and Data Sources

Human Reliability Analysis (HRA) methods include first-generation methods such as the *Success Likelihood Index Methodology* (SLIM), the *Human Error Assessment and Reduction Technique* (HEART), and the *Technique for Human Error Rate Prediction* (THERP). Second-generation methods include the *Standardised Plant Analysis Risk-Human Reliability Analysis* (SPAR-H), the *Cognitive Reliability and Error Analysis Method* (CREAM), and *A Technique for Human Event Analysis* (ATHEANA). Third-generation methods incorporate *Bayesian Networks* (BN), such as BN-CREAM, BN-SLIM, and BN-SPAR-H, to estimate the error probability of an action or diagnosis task [102] [106]. These methods have evolved, with each generation addressing the gaps identified in its predecessors.

Generally, the foundation for using these methods remains consistent. They begin retrospectively by understanding human failure events and the related tasks and sub-tasks, which can be derived from procedures. These tasks provide the context for analysis. However, the specific analysis process varies across different methods. This process often involves selecting a list of specified *Performance Shaping Factors* (PSFs), or what is referred to as *Common Performance Conditions* (CPCs) or *Error Producing Conditions* (EPCs), for each task step. These factors may change depending on the task context, as seen in first-generation methods. The PSFs, CPCs, EPCs, and similar factors are then weighted based on their likely contribution to the task in question.

It is expected to see a combination of different HRA methods or quantification approaches like in the THERP used for the subtasks of single *Human Failure Event* (HFE) or different HFEs [16]. However, some methods have been proposed for use by specific industries or have adapted the methods they apply from other industries. For example, ATHEANA, THERP, HEART, SPAR-H, BN-SPARH in the case of the nuclear sector; SLIM, SPAR-H, Petro-HRA, Phoenix-PRO, CREAM in the case of the process industry

(with many of its early methods adapted from the nuclear industry) [107], [108]; HEART, CREAM in the manufacturing sector [109]; and aviation [110].

There is also a preference for the choice of method depending on the type of operation, such as emergency, maintenance, or control room operations. Tables 2.4, 2.5, and 2.6 show the HRAs and the type of operation they are used for. It also shows the distinguishing features between the methods. For example, HEART, SLIM, SOHRA (a specialised version of HEART for the maritime field), and THERP are mostly used for maintenance operations, especially in the marine sector, SPAR-H, BN-CREAM for control room operations; in more recent years, BN-CREAM, BN-SLIM, BPL, BN-SPARH, HFACS, Fuzzy- TOPSIS, BN for emergency operations, and Bayes-SLIM for lifting operations, etc., [102] [111] [112].

This thesis focuses on methods to assess safety-critical control room operations and process industries. In this respect, some of the most important and commonly used methods are discussed and compared below.

First Generation Methods

The first of these methods, THERP, was introduced for control room applications as early as 1983 [113]. This was followed by the SLIM (USNRC, 1984), INTENT [114], and HEART [115] methods. These methods, though so far observably applied in the nuclear control rooms, allow for a breakdown of tasks and the potential impact of the *Performance Influencing Factors* (PIFs) on each task. They are known to focus on the skill and rule base level of human actions [116], and the PSFs are mainly derived by concentrating on environmental impacts on operators [117]. There is also a general gap in consideration of context, cognition, organisational factors, and commission errors when using these methods [104] [116].

Though no studies show the application in process control rooms, these methods have been used in the process industry to quantify the performance of maintenance operators [118]. The lack of generalisation of these methods for the practical settings was pointed out in a recent plant maintenance study by [119]. The authors applied the THERP method to study the influence of human and organisational factors on maintenance operators in an explosive atmosphere (ATEX) human-machine interaction (HMI) setting. They opted to apply an alternate second-generation method. This was due to the limited number of PSFs in the THERP method, thus lacking consideration of context and organisational

factors, the complexity of assessment requiring more expertise and the time-consuming nature of evaluation. Some of these limitations are not peculiar to the THERP method but to other first-generation methods and have driven the push for innovations in methods to address these challenges, so-called second-generation methods. The second-generation methods applied in control rooms or process plant operations are discussed below.

Second generation methods

Second-generation methods such as ATHEANA and CREAM were introduced to include cognitive and contextual aspects missing in the first-generation techniques, as seen from the PSFs. However, CREAM and ATHEANA have been criticised for being complex and time-consuming and have had limited application in the nuclear sector as mentioned [104] with a few applications in the process industry like in [119]. The SPAR-H method compensates for the complex nature of the methods above [104]. It also provides a detailed description of steps for quantifying the error component, which some first-generation methods lack. This method is now widely used in nuclear, aviation, healthcare, and chemical plants[120]. It is a helpful method to address both pre- and post-initiator HFEs, unlike Petro-HRA used for post-initiator errors [108]. SPAR-H has been criticised for having no explicit guidance for using a more comprehensive range of PSFs in practical settings [112]; this issue is not peculiar to the SPAR-H method. Researchers have sought to address the rigidity of the precious methods by incorporating AI techniques such as the Bayesian Network (a graphical model that represents a set of variables and their probabilistic dependencies using a *directed acyclic graph* (DAG)) into some of the methods [94] [112], [121].

Third generation HRA methods

As already discussed, despite the improvements brought about by the second-generation methods, there is a gap concerning accommodating new factors or so-called PSFs during analysis. To address this gap, methods such as phoenix HRA [121], Bayes SLIM [112], BN-CREAM [122], BN-SPAR-H [106] have been applied in process and nuclear industry applications. These methods are marked by their use of the Bayesian Network machine learning algorithm. Also, Mosleh [121] developed a new HRA framework for inclusion in *Probabilistic Risk Assessment* (PRA) and used BN as a method for the PIF quantification stage as initially developed by [94]. The PIFs were derived from new observations in

the literature, meaning the BN allowed for the integration of new factors. Another good potential of the BN is that it enables the building of dependencies between the PIFs. However, in this study, data was still based on expert judgment in their study.

In a similar industry, Groth [106] was able to utilise expert judgement for the scoring of the eight (8) traditional SPAR-H PSFs, which was subsequently used to populate the conditional probability table of each node together with the SPAR-H equations. This BN-SPARH, unlike the SPAR-H method, made it possible to create conditional independence relationships between the factors. This is a generally cited benefit of the BN as it addresses an important HRA issue of lack of consideration of the interaction between factors. It made it possible to add extra nodes/factors not previously considered in the SPAR-H method or that better capture data on an already defined factor. This was further extended and tested for cases where simulator data is available, although applying a different approach for estimating the priors [123]. BN has also been used with the CREAM method in a petrochemical setup [122]. This method made it possible to combine different levels of the so-called common performance conditions, which is another way of calling PSFs. It made changing prior conditional probabilities possible from deterministic to probabilistic ones.

The first-generation method SLIM has also been used jointly with the BN method by [112] in the manufacturing sector. The SLIM method does not utilise much expert judgement and is usually not done formally. Using the BN-SLIM approach, it was possible to incorporate the expert information coherently. Just like the method already discussed, it allowed for the incorporation of context-specific and relevant factors.

Though these so-called third-generation methods have so far shown prospects in addressing the gaps in HRA, they are mostly still new, thus, they also lack the breadth of real-life implementations that can identify weaknesses and strengths too. Also, there is an observable gap in advances within the process industry regarding HRA in comparison with the nuclear sector. The next subsection presents a summary of current HRA challenges still prevalent in process industries.

2.5.3 Challenges with HRA methods

Data Limitation

There is limited data for a quality human reliability analysis, particularly to make predictions [104]. Attempts to address this challenge or limitation have involved the

building and population of databases from simulator data within the nuclear domain. Such databases include the *Operator Performance and Reliability Assessment* (OPERA), *Human Event Repository and Analysis* (HERA), *Scenario Authoring, Characterisation, and Analysis Data Archive* (SACADA), and CORE-DATA [102]. However, this gap remains significant in the process industry, mainly due to the challenges posed by the absence of, or limited access to, data for Human Reliability Analysis (HRA) and behaviour or performance prediction [104].

Modelling Limitation

For the Nuclear sector, one of the HRA modelling concerns with the first and second-generation methods has been a lack of integration of information (qualitative and quantitative) from other domains. Also, the use of very traditional techniques limits the modelling of the complex nature of human performance, the human system interaction and factor dependencies [106]. Modelling is also done retrospectively (post-initiation) using past accident scenarios. This shows that a lot of work is still to be done with regard to advances in this sector.

No Real-time updating capability

In addition to the challenges discussed above, the use of these very specific PSFs and their respective multipliers, especially for those adapted from the nuclear industry, assume that these types of factors are also prevalent for process industry operators and are always prevalent for every control room operations at such range of weighted scores, which of course will vary depending on the assessor. They also are not updated or reviewed following work-as-done; thus, initial calculations pre or post-commission are still used over time. Therefore, the role of the dynamic nature of the operator and their environment is neglected in real-time error probabilities.

Missing Cognitive or behavioural elements

Another critical gap with the discussed HRA methods is the lack of insights into the cognition and behaviour of the operator and how these aspects of the 'human', as extensively discussed in an early part of this chapter, can increase or decrease error probabilities. It has been evident from the discussions that the cognitive states of the

operators and the other factors, task-related and organisational, are not mutually exclusive. Hence, understanding this dynamic aspect of humans and integrating them in some way into the analysis, not with some static values that hinder a robust analysis, would make for a robust HRA. Progress has been made in recent years by researchers through empirical studies to understand these critical aspects of the human and the result on performance and error by using tools like eye trackers, electroencephalogram, operator-system data, etc. [103], [124]. These human performance study approaches applied for the process control room are discussed in the following subsection.

2.5.4 Current Advances in Process Control Room HPA

The human performance techniques used in more recent times aim to mitigate the data availability gap and provide in-depth insight into operators' cognition. Thus, compared to the more probabilistic outcomes of the HRA using some performance shaping factors, these techniques investigate operators' reliability and performance using cognitive and behavioural factors [103], [124]. Such methods have now been experimentally applied in several safety-critical domains such as aviation [62]; nuclear [124], and process plants [103], [11], [125]. These techniques include *eye-tracking* (ET) and electroencephalogram (EEG), health monitoring using *heart rate variability* (HRV), *Functional magnetic resonance imaging* (fMRI), etc. Though they do not realise outcomes such as error probability scored for direct inclusion in HRA, they can be used to validate HRA methods, give a good grounding for training proposals or development of other support tools, provide less subjective insight on variables such as situational awareness, workload and stress etc. [102]. They can also be observed for correlations with state-of-the-art PSFs to identify their potential as objective substitutes for traditional PSFs.

Objective techniques in Process Industry

Amazu et al., [2] in a recent systematic review of the process industry and current advances in HPA within the process industry, grouped the application of these techniques based on the role of operators(monitoring and control, monitoring, failure detection and isolation) as detailed in the reviewed experimental studies. They identified how performance is being measured during this study, including the tools and metrics used and the factors assessed using these tools. To shed more light, Das et al., [126], during a control room monitoring and control experiment, evaluated the impact of HMI

information and required action on the mental state of control room operators using gaze transition probabilities (GTP), a known eye-tracking metric. The GTP allows quantifying the extent to which the operator examines the relationship between two variables.

In a similar monitoring and control task [10] and [127] studied the impact of multiple component use and task complexity, respectively, on cognitive workload using eye-tracking metrics like mean pupil values, Pupil dilation [10] and EEG metrics like Power spectral density of theta waves [127]. The pupil is reported to dilate significantly in a case of high workload and before taking a decision [10], while higher power spectral density of theta waves is associated with higher mental activity, specifically attention and mental processing. On the other hand, [128] used eye-tracking metrics like Gaze transition entropy (the extent of disorder in the operators' gaze and transitions) and dwell time entropy (assessing the distributed dwell duration on areas of interest within a time frame) to observe operators' situational awareness. A summary of the common eye-tracking metrics is provided in Table 2.7.

In addition, a control room monitoring task set-up [11] used eye-tracking metrics, fixation duration means, and fixation/sprint ratio to assess the impact of display types on cognitive processing and attention. Longer fixation durations are known to indicate substantial information and deliberate processing [11] [129], while higher fixation to saccade ratios indicate more processing than searching for targets [129]. Jung [130], while studying the impact of multi-task complexity on cognition, used the ratio of pupil dilation, which is closely related to the metric used by Bhavsar et al., [10] during a monitoring and control task.

Other objective metrics exist for human performance assessment. These include time-based and behavioural metrics. Behavioural metrics like response time [11] [131] [132], and time-based metrics like time to complete the task, and time to detect anomaly [14], the ratio of time spent on the task (primary and secondary) [130] have been used by researchers in the process industry domain. Simonson [14] studied the impact of different alarm designs and levels of automation on time to complete tasks and time to detect anomalies during a control room monitoring and control task. On the other hand, during a control room monitoring task, Shi et al., [11] studied the impact of different display types on response time while [131] instead studied the impact of display element shapes and densities.

Behaviour metrics for human performance assessment are an area of growing interest, given the shortcomings of using eye-tracking or EEG measures in real-time practical control room settings. Braarud and Chen emphasised the advantages of using objective data from real-time interactions between operators and systems to classify operator workloads in their research [124], [56]. Their studies identified several vital features, including the number of alarms per station and the interactions with the *human-system interface* (HSI), such as navigation activities to process images, selection of objects, input of values, and various speech characteristics.

Integrating multiple factors through Bayesian Networks (BN) in traditional Human Reliability Analysis (HRA) methods illustrates how metrics from both objective and subjective techniques can be combined. Similarly, as discussed in the following subsection, these metrics—often referred to as variables—can be integrated with other Performance Shaping Factors (PSFs) using artificial intelligence (AI) methods.

Subjective techniques in Process Industry

There are obvious benefits from using these objective techniques to assess human performance on a cognitive and behavioural scale, specifically for person-related factors like workload, situational awareness, stress, etc. This is done in a few cases alongside subjective techniques like NASA-TLX for workload and SPAM, SART, SAGAT, etc, for situational awareness. For example, Shi & Rothrock, [125], evaluated the effectiveness of the SART questionnaire using eye-tracking metrics. The study by Yang et al., [15] from the nuclear domain, comparing the different levels of procedures based on their impact on workload and situational awareness, used only questionnaires. One can observe from the previous paragraphs that many more studies have been conducted on the objective assessment of workload. This is understandable due to the type of metrics obtainable from the eye-tracking and EEG techniques. Hence, for situational awareness assessment, subjective questionnaires are still very much in use in most safety-critical domains.

The NASA TLX is a widespread technique used to assess workload. The NASA-TLX technique has also found application in other sectors like the process industry [133], [14]; nuclear [17], [15] and health [134]. In a recent process control room study, Simonson et al., [14], assessed the impact of alarm design and level of automation on workload using the NASA Task Load Index (NASA-TLX). Likewise, in their experiential studies,

[131] and [135]. A higher NASA-TLX index means a high workload. However, the scores can be split into low, medium, and high levels or, in some cases, low and high. The *Multi-Attribute Task Battery* (MATB-II) is another workload assessment technique commonly in use in safety-critical domains [130], [11].

Several techniques are used for situational awareness (perception, understanding and projection). These include the SART, SPAM, and SAGAT [64]. Higher scores, in contrast to the NASA-TLX, signify better situational awareness. The SART technique has gained popularity for situational awareness assessment compared to the other methods. SART allows participants to rate their SA on different dimensions, such as perception, comprehension, and projection. However, the suitability of such a technique for SA evaluation has been questioned [17]. One of the main challenges has been that it is usually completed post-scenario, prompting the reliance on memory recall [64]. It has also been criticised as highly correlated with the NASA-TLX on its demand and dimensions, therefore likely assessing workload more than situational awareness [17]. SAGAT, a direct objective situational awareness measure in contrast to SART, has proven to be an effective measure of situational awareness [63]. However, it has been criticised for its intrusiveness due to periodic freezes during the simulated tasks [63]. Hence, SPAM, which objectively scores participants without freeze probes, has been favoured over SAGAT. SPAM is a direct objective situational awareness technique that allows objective scoring while the test continues. In this case, there is no need to stop the test to evaluate the participants. Therefore, the participants are assessed using a concurrent think-aloud approach. A comparative summary of these techniques is shown in Table 2.8.

Statistical methods in the process industry

Common statistical methods used in the process domain, their purpose and when used are summarised in Tables 2.9, 2.10, 2.11 and 2.12. Highlighting this aspect is crucial, especially in human performance studies given their use for analysing subjective and objective data. For example, for checking significant differences between study groups, data description, reliability analysis, and so on.

AI in Human Performance Studies

While machine or deep learning techniques have not been used to quantify or predict human performance, they have been applied to complement the HRA assessment process using traditional or new performance assessment techniques. As already discussed, one ubiquitous machine learning algorithm in such applications is the Bayesian network (BN). According to Groth, the BN technique is compatible with PRA methods, which is required for new HRA methods [106]. It can quantify probabilities, interface with PRA models, and handle uncertainties. In comparison, other AL algorithms used in human performance-related studies across industries have primarily been applied to predict workload. For example, Gao et al., [9] predicted workload in an emergency operating procedure situation, [151] used the fusion of eye-tracking and EEG data, [124] used a combination of speech data and operator system logs, and [56] predicted workload in a human-agent teaming scenario. In a few cases, situational awareness has also been predicted using such AI algorithms [125], [152].

Braaud [124], in a study on a Halden facility which is close to process control room operations, classified cognitive workload using speech, system data (operator & crew alarms, acknowledged operator & crew alarms, human system interface activity and commands) with *Random Forest* (RF), *Neural Networks* (NN), *Support Vector Machine* (SVM), *Extreme Gradient Boosting* (XGB), *k-Nearest Neighbors* (KNN), *Naive Bayes* (NB) algorithms. For the data used in the crew system, RF XGB group performed better than the others at an accuracy of 61% and an average of 72% when trained for a single operator. The important features were the number of alarms and the amount of operator-system interaction. In a different domain, Rahman [140], while classifying the cognitive load of non-contact-based drivers, used Linear Discriminant Analysis (LDA), SVM, *Logistic Regression* (LR), and NN, with LR outperforming the other methods. Similarly, while classifying workload using driver performance and psycho-physiological data, Solovey [150] reported that logistic regression and naive Bayes classifiers outperformed the others (*Decision Trees* (DT), *Multilayer Perceptron* (MLP), NN).

Some of the algorithms mentioned so far were likewise used for workload or situational awareness classifications by researchers whose inputs were solely from eye-trackers and EEG [152], [151]. Though the application of AI for human performance assessment has been on workload and situational awareness predictions and classification, similar algorithms could be explored for error or behaviour predictions.

Chapter 3 builds on the techniques and process introduced in this chapter and further applies an HRA method, some new and old cognitive and behaviour performance techniques, and LR and BN machine learning algorithms to address the research objects and some of the gaps identified in the sections of this chapter.

Gaps

Though the discussed new HPA techniques have made it possible to investigate cognition and behaviour, researchers within the process industry have focused on observing single factor impact on cognition and behaviour, such as different types of displays or interface layouts on operators, at a time. Also, operators' behaviour has been chiefly discussed without an interpretation of how this translates to error.

The NASA-TLX and SART questionnaires introduce subjectivity in assessment and are known to correlate highly with each other [17], so a combination of other methods and performance and an overview of behaviour and performance in comparison with the subjective assessment would provide a more robust insight. However, the referenced experimental studies have, in most cases, except Ikuma et al., [135] and Shi & Rothrock [125], used these objective or subjective techniques as standalone without providing comparable insights, combinations or validations of one with the other [64], [9]. The need for multi-modal measures has been emphasised through findings by researchers in different domains. According to Hinss et al., [13], while studying the impact of prolonged monitoring on mental state, some researchers found that a combination of measures, with the addition of a reliable estimation method, outperforms single measures for their assessment.

2.6 Summary of Gaps

2.6.1 Procedures: Alarm Response Procedures

1. The pros, cons, and potential impacts of transitioning from paper to digital or computer-based alarm response procedures on cognition, behaviour, performance and safety have been scarcely investigated in the nuclear sector and not in the process industry.

2. While procedures' roles in situational awareness, mental workload, and error have been noted in the literature, other contributing factors to these outcomes have not been thoroughly explored.

3. Experimental studies on process control rooms have mainly used procedures as pop-ups for operators or test participants, lacking comprehensive analysis of their integration and impact [103].

4. It is crucial to consider the interactions among variables like the procedures, alarm, systems and potential relationships during human performance assessments, as emphasised by Zarei [102]. Still, this aspect has been scarcely investigated in process industries.

2.6.2 Human Reliability Assessment for PBA

1. Limited Data for Quality HRA:

    - Insufficient data for reliable human reliability analysis (HRA) and prediction in the process industry.

2. Modelling Limitations:

    - Lack of integration of qualitative and quantitative information from other domains in the nuclear HRA models currently used by process industries.
    - Use of traditional techniques that do not capture the complex nature of human performance and system interactions.
    - Retrospective modelling based on past accident scenarios.

3. No Real-time Updating Capability:

    - PSFs and multipliers adapted from the nuclear industry assume static applicability across process industry operations.
    - Scores are not updated post-commissioning, ignoring dynamic changes in operators and environments.

4. Missing Cognitive or Behavioural Elements:

- Current HRA methods lack insights into operators' cognitive and behavioural aspects influencing error probabilities.
- Cognitive states and other factors are not mutually exclusive and should be dynamically integrated into HRA.

2.6.3 Other Performance Assessment Techniques

1. Researchers in the process industry have focused on the impact of single factors (e.g., display types, interface layouts) on cognition and behaviour without considering how these behaviours translate to errors.
2. NASA-TLX and SART questionnaires introduce subjectivity and are highly correlated.
3. Behavioural metrics are yet to be explored in process industries on how they translate to error.
4. AI methods have not been extensively explored for error prediction in human performance assessment studies.
5. Most experimental studies use objective or subjective techniques as standalone measures, lacking comparable insights, integrations, or validations. There is a need for multi-modal measures to provide a more comprehensive assessment.

2.6.4 Contribution

This thesis contributes to human factors and process safety knowledge in process industries by addressing identified procedural gaps, human reliability and performance assessment and AI application.

Research Questions: Objective 1

A fundamental aspect of human reliability assessment for probabilistic safety analysis or safety-critical task analysis is identifying performance-shaping factors that impact the likelihood of successful performance and the interactions or relations between these factors. Therefore, the first research question in this thesis, which has been somewhat addressed through the reviewed literature, is:

- RQ1: What are the most critical factors and variables — both human, organisational, and technical — for decision and situational awareness support and human performance assessment of current dynamic human-in-the-loop process control configurations? Why are they important?

This question is further validated with Objective 2.

In addition, the pros and cons and other potential impacts of a transition from paper-to-digital- or computer-based procedures have scarcely been investigated in the nuclear sector, with no investigation conducted to the best of the author's knowledge within the process industry. Hence, the research question:

- RQ2: How does the transition from paper to digital or computerised procedure options impact operators' general behaviour, cognition, and safety? Any difference? If any, how significant are these differences?

Research Questions: Objective 2

This thesis also aims to develop a dynamic, risk-based framework for human-in-the-loop assessment, addressing the following research questions,

- RQ3: What operator behavioural metrics are the best predictor of a successful recovery of a plant alarmed state?
- RQ4: How can live and legacy data be integrated for a real-time human-in-the-loop performance assessment framework, decision support, and procedure optimisation?
- RQ5: What are the benefits of integrating operator-system data and expert knowledge for risk-based real-time decision and situation awareness support?

These questions are expanded upon in Chapter 3.1 and 4.1, with specific techniques applied to address each one. The outcomes of the methods discussed and applied in Chapters 3 and 4 contribute to advancements in *Risk-Based Process Safety* (RBPS) within the 'Understand Hazards and Risk' block and the 'Hazard Identification and Risk Analysis' pillar [1]. Also, by applying such frameworks, support is not only provided for operators by the identification of key factors that impact likely performance outcomes (for example, does the procedure type or support or the frequency with which they are

opened affect operators), but this framework further support supervisors and management to adequately optimise tools or better adapt automation - in and post real-time.

An essential contribution of this thesis to the body of knowledge is the experimental study developed to address the above research questions and validate the reviews. This enriches the database for control room applications and makes data available for other researchers within and across domains. It further provides a platform for validating PSFs and other variables not studied so far in literature for their impact and importance in process control rooms.

Table 2.3 Comparison of Risk Assessment and Human Performance across different aspects.

Aspect	**Risk Assessment**	**Human Performance**
Focus	Identification and evaluation of hazards, risks, and mitigation strategies	Evaluation of human capabilities, limitations, and performance under various conditions
Objective	To estimate potential risks and the likelihood of negative outcomes in a system or process	To improve and optimise human performance, identify factors affecting efficiency and safety
Method	Qualitative or quantitative analysis, risk matrices, fault tree analysis, FMEA (Failure Mode Effects Analysis), etc.	Psychometric tests, observational techniques, task analysis, performance simulations, physiological measurements
Application	Safety-critical industries (e.g., aviation, healthcare, manufacturing, nuclear) to reduce accidents, injuries, and losses	Design of work environments, training programs, and safety protocols in industries like aviation, military, healthcare, and manufacturing
Key Measures	Likelihood of occurrence, severity of consequences, risk exposure, hazard identification, risk mitigation strategies	Task performance metrics (e.g., speed, accuracy, efficiency), cognitive workload, fatigue, stress, error rates, physiological responses (e.g., heart rate, eye movement)
Outcome	Risk reduction, decision-making about safety measures, prioritisation of resources for mitigation	Optimised human performance, reduced errors, increased efficiency, better task management under stress or fatigue

Table 2.4 First Generation HRA Methods.

HRA Method	**Distinguishing Criteria**	**Application and Data Sources**	**Type of Operation**
SLIM	Focuses on skill and rule-based levels of human actions, with limited consideration of context, cognition, and organisational factors.	Nuclear sector, maintenance operations; data often based on environmental impacts on operators.	Maintenance operations, general operational tasks.
HEART	Uses Error Producing Conditions (EPCs) to quantify human error probabilities; simple and quick application.	Manufacturing sector, maintenance operations; data derived from operator actions and environmental conditions.	Maintenance operations, emergency procedures, general operations.
THERP	Breaks down tasks into sub-tasks with associated Performance Influencing Factors (PIFs); focuses on procedural adherence.	Nuclear sector, control room applications; process industry for maintenance operations; data involves procedural and environmental factors.	Control room operations, maintenance, and emergency scenarios.

Table 2.5 Second Generation HRA Methods.

HRA Method	**Distinguishing Criteria**	**Application and Data Sources**	**Type of Operation**
SPAR-H	Includes cognitive and contextual aspects missing in first-generation methods and detailed quantification steps.	Widely used in nuclear, aviation, healthcare, and chemical plants; data involves both pre- and post-initiator Human Failure Events (HFEs).	Control room operations, emergency procedures.
CREAM	Integrates cognitive modelling with error analysis; considers cognitive factors and their influence on error rates.	Limited application in the nuclear sector, some use in the process industry; data involves cognitive task analysis and environmental conditions.	Process control operations, emergency operations, routine checks.
ATHEANA	Focuses on complex and time-consuming analyses; includes cognitive and organisational factors.	Mainly used in nuclear sector; limited use in other industries; data involves complex scenarios and cognitive influences.	Complex operational scenarios, emergency situations.

Table 2.6 Third Generation HRA Methods.

HRA Method	**Distinguishing Criteria**	**Application and Data Sources**	**Type of Operation**
BN-CREAM	Utilises Bayesian Networks to estimate error probabilities and integrate cognitive reliability models.	Used in petrochemical setups; data derived from simulator data and expert judgment; integrates common performance conditions.	Control room operations, emergency procedures.
BN-SLIM	Incorporates expert judgment into Bayesian Networks; allows for context-specific factors integration.	Manufacturing sector; data involves expert elicitation and probabilistic modelling.	Emergency operations, maintenance tasks, lifting operations.
BN-SPAR-H	Uses Bayesian Networks to create conditional dependencies between Performance Shaping Factors (PSFs); allows for new factor integration.	Process and nuclear industries; data based on SPAR-H PSFs, expert judgment, and simulator data.	Emergency scenarios, control room operations, complex tasks.

Table 2.7 Summary of Eye-Tracking Metrics and Definitions [18].

Metric	Definition
Heat Map	A visual representation showing areas of high visual attention. Brighter spots indicate regions with more frequent or prolonged focus. When combined with Areas of Interest (AOIs), it highlights specific regions of engagement across different phases of the experiment.
Number of Visits	The total count of times participants revisit specific AOIs, revealing patterns of attention and whether certain areas consistently draw interest or shift over time.
Visit Duration	The time spent by participants during each visit to an AOI. It complements "Number of Visits" by showing how long participants engage with specific areas.
Number of Fixations	Tracks how often participants focus on specific points of interest. This provides insight into visual exploration and attention consistency within AOIs and across experimental phases.
Fixation Duration	The time spent focusing on a particular point. It identifies elements that capture sustained attention, offering insights into cognitive engagement during various scenarios, such as alarms or critical events.
Saccade Duration	Measures the time taken for rapid eye movements between fixations. It reflects the efficiency and fluidity of visual scanning, providing insights into cognitive processing speed under different conditions.
Pupil Diameter	Indicates changes in cognitive load and emotional arousal. Variations in pupil size reveal mental effort and emotional responses linked to visual stimuli and specific AOIs over time.

Table 2.8 Comparison of NASA TLX, SPAM, and SART across different aspects.

Aspect	**NASA TLX**	**SPAM**	**SART**
Focus	Workload	Situational Awareness	Situational Awareness
Type of Measurement	Subjective self-report measure	Subjective assessment, scored by assessor	Subjective self-report measure
Real-time/Post-task	Post-task	Real-time, during task	Post-task
Dimensions Assessed	Mental demand, physical demand, temporal demand, performance, effort, frustration	Situation perception, understanding and projection	Situational awareness supply, demand and situation understanding
Invasiveness	Non-invasive	Invasive	Non-invasive
Advantages	Well-established, reliable, comprehensive, easy to use	Well established, Provides a real-time measure and based on a joint assessment	Well-established, reliable, comprehensive, easy to use
Disadvantages	Introduces subjective judgment, can be time-consuming, done post-task	May be difficult for complex tasks because it can be invasive, may lead to cognitive load	Introduces subjective judgment, can be time-consuming, done post-task

Table 2.9 Summary of Statistical Methods, Purpose, and Applications part I.

Method	**Purpose**	**Reference**
Shapiro-Wilk Test	Test for normality of data distribution.	[136], [137]
Mann-Whitney U Test	Compare differences between two independent groups (non-parametric).	[133]
Cronbach's Alpha	Assess internal consistency of a questionnaire or survey.	[138], [139], [17]
T-Test (Independent)	Compare means of two independent groups.	[133]
Paired T-Test	Compare means of paired data. Used for within-subject comparisons	[37], [9]
Analysis of Variance (ANOVA) (One-Way)	Compare means across multiple groups.	[17], [131]
Mean Squared Error (MSE)	Measure accuracy of predictions. Commonly applied in predictive modeling or when comparing model performance	[139], [140], [141]
Correlation (Pearson, Spearman)	Measure the strength and direction of relationships between variables. Used to identify associations between variables. Spearman is used for ordinal or non-linear data	[142], [133]

Table 2.10 Summary of Statistical Methods, Purpose, and Applications part II.

Method	Purpose	Reference
Factor Analysis	Identify underlying structures in data. Used to explore relationships or to reduce dimensionality while retaining important information	[139], [143]
Common Factor Analysis (CFA) is	a technique to identify latent variables that explain the correlations between observed variables.	[144]
Principal Component Analysis (PCA)	A dimensionality reduction technique that transforms correlated variables into a smaller set of uncorrelated components.	[144], [139], [145]
Sensitivity Analysis	Assess the effect of input variations on outputs.	[17], [146]
Sphericity Test (Mauchly's)	Test whether variances of differences are equal (for repeated measures). Required before running repeated-measures ANOVA to ensure valid results	[97]
Chi-Square Test	Test relationships between categorical variables. Used to analyze associations.	[144], [147]

Table 2.11 Summary of Statistical Methods, Purpose, and Applications part III.

Method	Purpose	Reference
Kruskal-Wallis Test	Compare multiple groups for non-parametric data. Alternative to ANOVA when normality assumptions are not met.	[133], [61]
Regression Analysis	Predict dependent variable based on independent variables.	[140], [133], [146]
Wilcoxon Signed-Rank Test	Compare two related samples (non-parametric).	[148]
Baseline Correction	Normalis data against initial measurements. Used to adjust for shifts in the baseline of a signal, especially in time-series data.	[149]
Tukey's Test	Post-hoc test used after ANOVA to identify specific group differences while controlling for Type I error. Assumes equal variances and is robust for multiple comparisons.	[131], [17]

Table 2.12 Summary of Statistical Methods, Purpose, and Applications part IV.

Method	Purpose	Reference
Factorial Design	Explore multiple independent variables and their interactions in a single experiment. Allows testing main effects and interaction effects simultaneously. Commonly used in experimental psychology and behavioural sciences.	[138], [58]
Within-Subject Design	Participants are exposed to all experimental conditions. Reduces variability due to individual differences, increasing statistical power. Used when comparisons across conditions are the focus.	[131]
Between-Subject Design	Participants are assigned to separate groups, each experiencing one condition. Eliminates carry-over effects but requires more participants. Useful for studies with independent group comparisons.	[150], [18]
G*Power	A statistical software tool to calculate sample size and power for various study designs. Ensures sufficient sample size to detect significant effects while minimising over-recruitment.	
Degree of Freedom (DoF)	Represents the number of independent values in a calculation. Used in hypothesis tests like t-tests, ANOVA, and chi-square tests to determine critical values.	[142], [129]
Effect Size	Measures the magnitude of an effect or relationship, regardless of sample size. Examples include Cohen's d, partial η^2, and r^2. Essential for interpreting practical significance.	[149]

Chapter 3

Design Methodology I

3.1 Introduction

Chapter 3 of this dissertation outlines the design methodology used to achieve the second part of the first research objective (Section 3.1). The research focuses on key sections of a formaldehyde production chemical plant: tank, methanol, compressor, heat recovery, reactor, and absorber sections (Section 3.2.1). These areas of the plant are crucial to understanding the operational dynamics and were carefully selected for detailed analysis.

To effectively assess the performance and interaction within these plant sections, a suite of Human-System Interface (HSI) tools was designed. These tools include screen displays that present crucial information to the participants. Furthermore, there is an alarm list display for monitoring and managing alarms. In addition, structured alarm response procedures guide participants through appropriate actions in the event of an alarm. These tools were developed to simulate a realistic environment and observe how participants interact with the system under different scenarios. Importantly, observing the impact of alarm response procedures and other organisational and technical elements on operators' cognition, behaviour, and performance. (Section 3.2).

The experiment is meticulously structured, starting with selecting participants with varying levels of experience and training (Section 3.3). The experimental scenarios are carefully crafted to reflect realistic operational challenges within the plant, ensuring that the data collected is both relevant and robust. The backgrounds of the participants in terms of training and experience are documented, as these factors are expected to influence their performance and interaction with the system.

Data collection in this study employs a variety of methods to capture both subjective and objective measures. Situational awareness is assessed using the Situational Awareness Rating Technique (SART) and the Situation Present Assessment Method (SPAM). Workload is measured using the NASA Task Load Index (NASA-TLX), offering a quantitative assessment of the mental and physical demands placed on participants. Additionally, eye tracking technology is used to monitor where participants focus their attention, while simulator recordings capture their interactions with the system in real time. The study also considers key variables that could affect the outcomes, ensuring that these are carefully measured and analysed (Section 3.3.1). Reliability, validity, sensitivity, and normality are evaluated to ensure the integrity and relevance of the data (Section 3.3.3).

Ethical considerations are crucial in this research to protect participants' rights and ensure integrity in data collection processes (Section 3.3.4. The chapter outlines the data analysis plan, detailing how both objective and subjective data will be processed (Section 3.3.5). The chapter ends by acknowledging the study's design limitations, followed by a chapter summary.

Research Objective 1

As previously stated, this chapter investigates the first objective of the thesis and answers the central research questions. The first objective is to understand the state-of-the-art alarm response procedures and to compare the effectiveness of the transition from paper versus electronic alarm response procedures for process control room operators handling abnormal situations.

The literature review shows that alarm response procedures are critical for the safety of the process control room. However, gaps are associated with their design, application, and advances in process industries (see Section 2.6.4). Therefore, the following *research questions* (RQ) are to be addressed in this research:

1. RQ 1: What are the most critical factors and variables — both human, organisational, and technical — for decision and situational awareness support and human performance assessment of current dynamic human-in-the-loop process control configurations? Why are they important?

2. RQ 2: How does the transition from paper to digital or computerised procedure options impact operators' general behaviour, cognition, and safety? Any performance differences, and how significant are these differences?

3.2 Research Design

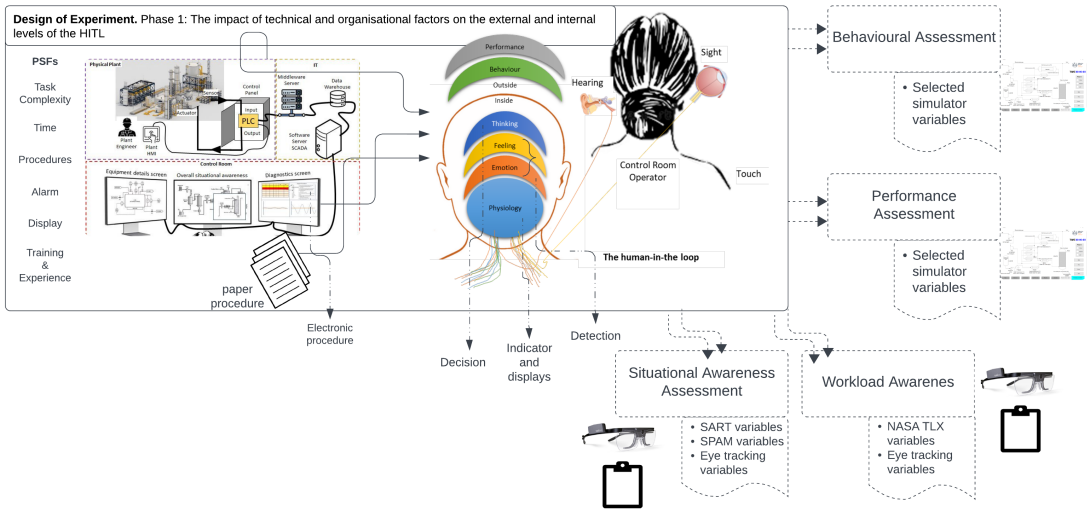


Fig. 3.1 Experimental Design Framework Observing the Impact of Task-Based PSFs (Alarm Response Procedures) on Internal and External Individual Factors, Including Interactions with Task Complexity, Time, Training, Experience, Alarms, and Displays.

Objective 1 involves the use of a mixed method research design that combines qualitative and quantitative techniques for data collection, insights, and observation validation. A mixed method approach is used because of the advantages of using multiple data inputs to derive a more holistic view, as emphasised by Gao et al., [9], Endsley [64], given that one approach may compensate for the lack of the other, as demonstrated in studies like Yang et al., [152]. This approach helps address a gap in the literature review specifically focusing on the issue of using subjective or objective measures as stand-alone in control room experimental studies.

Experiment Justification and Contribution

As shown in Figure 3.1, the central method chosen is the design of the experiment (DoE). The design of an experiment is a common approach when investigating control room-related issues. As the goal of this first phase involves control room alarm response

procedures, this approach is therefore applicable for the study of the impact of alarm response procedures on operators, including identifying possible design issues. However, the study design in this thesis provides three beneficial distinctions from other designs in the following ways;

- Multiple PSFs: This study introduces classical performance variables, including task complexity, procedures, alarm design, time availability, training, experience, and displays. These variables closely resemble the performance shaping factors (PSFs) used in the SPAR-H method, as will be discussed later in this thesis. In contrast, many studies reviewed typically focus on specific PSFs or a limited combination of two in their study design. Das et al. compare the impact of interface displays on workload [126], while Simonson et al. and Shi et al. examined how automation affects workload [14] or performance [11], respectively.

- Procedures: Secondly, this study examines the alarm response procedure (ARP) as applied in a process control room during abnormal situations. The ARP is typically detailed on paper and sometimes provided as a portable document format (PDF) on the displays. This study addresses a gap in research on procedure transitions within the process industry as discussed in section 2.4 of Chapter 2, an area that has been largely overlooked despite its critical role in preventing accidents. Previous experimental studies in this sector have not focused on this aspect. When mentioned, procedures often just provided operators with simple pop-up tasks, as seen in studies like Kim & Yang [130], Shi et al., [11], and Bhavsar et al., [10].

- No of Screens: This study describes the use of three screens (see Figure 3.2). There is a central/overview display, commonly found in control rooms and positioned between two other displays at a slightly higher level (level 1). To the right is a support display for alarm lists, procedures, and trend presentations (level 2). However, to the left is a blank display for opening additional displays, such as system mimics. Most experimental studies in this domain use two (2) or single displays for their experimental study [10], [130].

Justification of Techniques

The qualitative methods used in the experimental study include NASA-TLX for workload assessment. This tool has been commonly employed in assessing the workload of control room operators [17], [14], [15]. SART and SPAM are also used to assess

situational awareness [17], [15], [125]. According to Chen et al., subjective measures like NASA-TLX have provided reliable results over the years [145]. By using both SPAM and SART, we can validate the claim that SART correlates highly with NASA-TLX due to the strong similarity and correlation of SART's demand and supply dimensions with the dimensions of NASA TLX.

Furthermore, employing a second situational awareness technique, SPAM, offers the benefit of acquiring responses as tasks progress, just like the Situation Awareness Global Assessment Technique (SAGAT). This allows experts to observe responses and assess situational awareness based on feedback, response time, and other factors, potentially more accurately than the participants themselves.

The questions in the study were designed to capture operators' perceptions of support when comparing the two types of procedures. A human factors expert reviewed these questions to ensure their accuracy and usefulness in collecting the necessary data. Although subjective measures like these are often criticised for being post hoc and introducing subjectivity, the inclusion of objective measures helps to counterbalance this issue.

In this study, three objective measures were introduced: eye tracking, performance, and behaviour measures. Eye tracking is commonly used to evaluate human cognition, particularly in relation to attention and situational awareness [145]. This technique has also been employed in studies of workload and stress using pupillometry-based metrics [10], [9]. Eye-tracking is currently being used in control room studies in the process domain for various research questions [103], [10]. However, it is important to note that eye-tracking can be intrusive, particularly in long-term, real-world applications. Considering this limitation and its emphasis on cognitive assessment, additional measures—namely, performance and behavioural measures—were incorporated into the study.

Performance and behavioural measures in this study were derived from simulator logs. These measures provide valuable information for comparing outcomes and making improvement decisions. Behavioural metrics are collected in real-time and offer insights into cognitive states, particularly mental workload [56], [17], and fatigue [153]. On the other hand, the performance measures are retrospectively evaluated.

Section 3.3, Experiment design, discusses the variables of this study and subjective and objective measures. However, it is important to examine the selected plant, system input and outputs and control room elements (key design inputs into the DoE) before

diving into the experiment properly. The control room elements are mainly human system interfaces (HSI) developed for the experiment setup. They include the procedures on paper and screen, the screen displays and the alarm design. A quick overview of the experiment setup is shown in Figure 3.2.

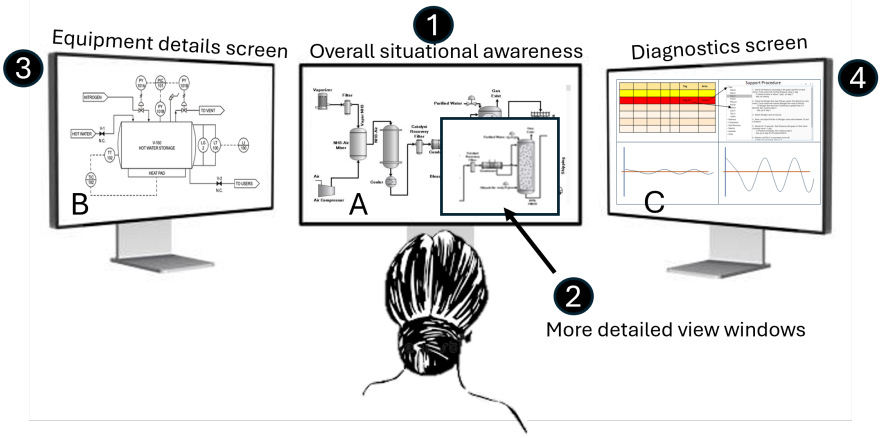


Fig. 3.2 The Human System Interfaces. A. Overview display providing for overall situational awareness on plant operation. B. Extra display for projection of specific plant section equipment (2), and C. the Support Display for Event Diagnosis using alarm list support, procedure support, and trends [18].

3.2.1 Selected Plant

The present study uses a simulated interface residing on the MATLAB platform of a formaldehyde production plant, with modifications based on the design originally presented by [154]. The plant produces roughly 10,000 kg/hr of a 30% formaldehyde solution by partially oxidising methanol with air (see 3.1).

$$CH_3OH + \frac{1}{2}O_2 \rightarrow_{r1} CH_2O + H_2O \tag{3.1}$$

Furthermore, a secondary reaction 3.2 occurs that completes the oxidation of methanol to carbon monoxide, thus reducing the formaldehyde yield:

$$CH_3OH + \frac{1}{2}O_2 \rightarrow_{r2} CO + H_2O \tag{3.2}$$

In a refinement to the design presented by Demichela et al. [154], which was based originally on three sections, this study divided the instead into six sections (see Figure 3.3). These sections represent the critical components of the plant for process operation, which operates under the partial oxidation process to produce formaldehyde. Figure 3.3 includes a button for Emergency Shutdown and an alarm box to indicate critical, medium, or low-level alarms as presented on the alarm list. Extra buttons, time, start, stop, pause, and save, were added for the experiment only.

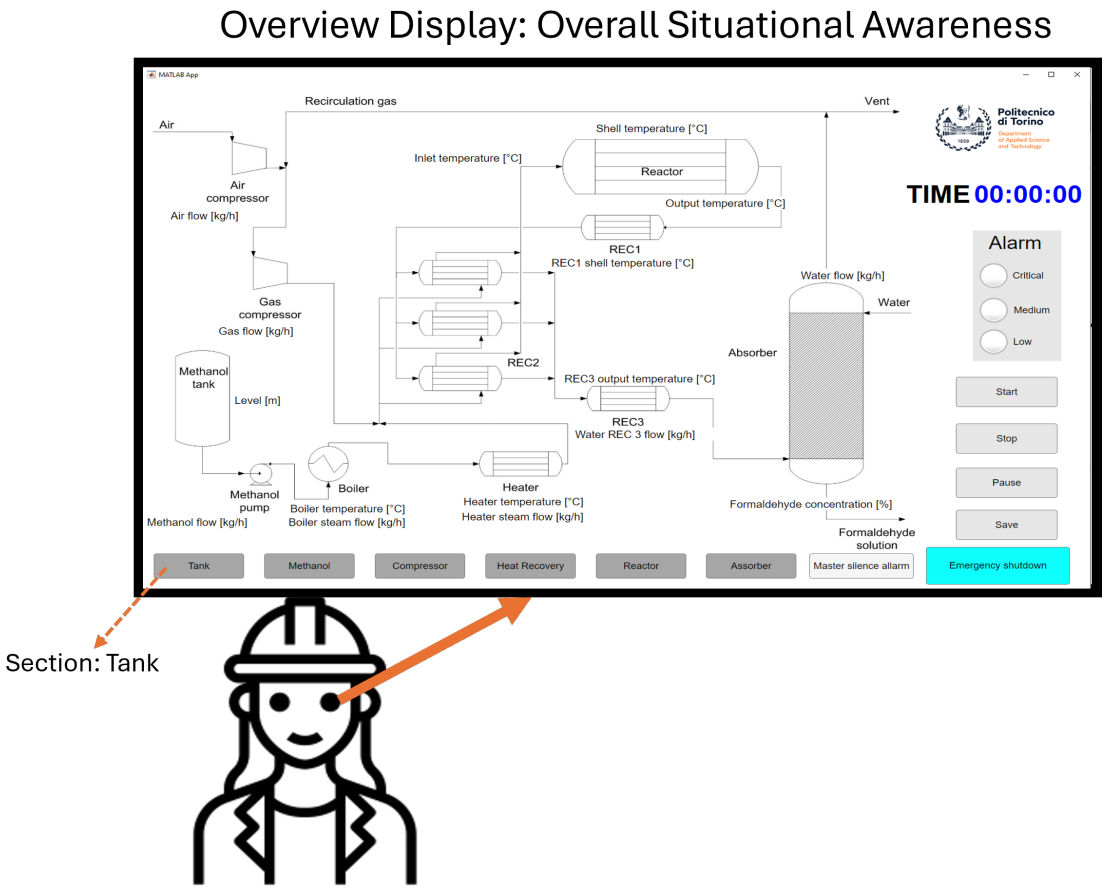


Fig. 3.3 The Overview display: Process Flow Diagram of the Plant. The plant sections are Tank, Methanol, Compressor, Heat Recovery, Reactor, and Absorber.

Tank

The plant layout starts with the Tank Section shown in Figure 3.4. In this section, a methanol tank pressurised with nitrogen is equipped with alarms to alert operators of any deviations. The alarms are grouped into red alarms, which are critical; orange alarms, which are medium; and yellow alarms, which are low. This colour scheme was verified by the control room and distributed control system (DCS) experts in EPRI Ireland and Yokogawa UK. The alarm design format for this study is discussed in detail in section 3.2.2. The *low-pressure alarm* (PAL01) alerts the operator of a critical conditions 'Red alarm' when the pressure reaches 0.99$[ata]$. The 'Yellow alarm' *low flow alarm* (FAL01) is triggered by low Nitrogen flow, possibly caused by a primary nitrogen valve or pressure indicator controller failure. The valve opens automatically when the *Pressure Indicator Controller* (PIC) communicates with it on a need to open or close.

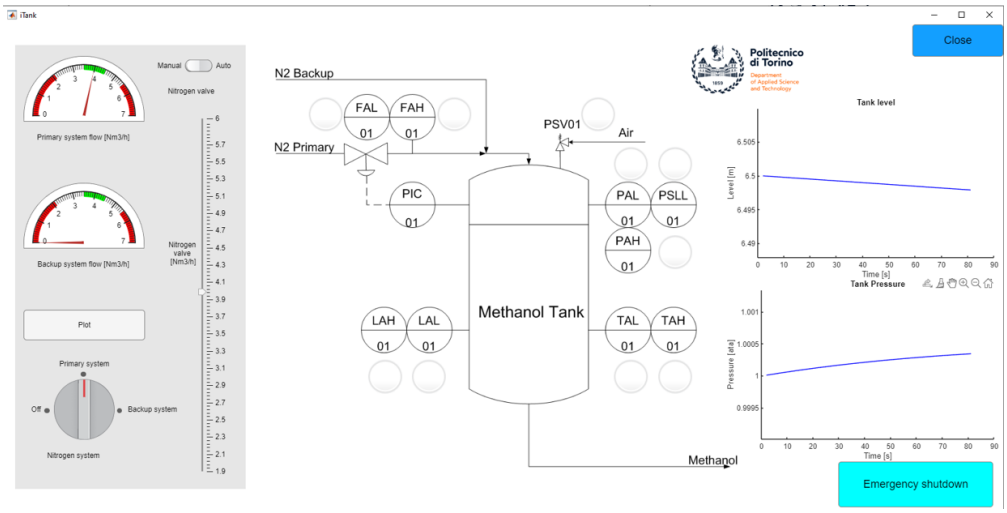


Fig. 3.4 Tank Mimic showing a methanol tank, controllers, meters, process trends and alarms.

When FAL01 or PAL01 alarms are triggered, the operator can manually control the plant by switching to the backup system in case of a faulty valve. In the case of a faulty PIC, the operator can manually control the valve with the manual primary valve scale. If the operator cannot intervene to control the plant manually at 0.98 [data], the *pressure valve* (PSV01) is activated. PSV01 allows the recovery of tank pressure by allowing the entrance of air from the atmosphere. If the pressure exceeds 0.97 [ata], the sensor triggers the *pressure switch* (PSLL01), leading to an emergency shutdown within a minute. It is

also at the operators' discretion to shut the plant down when necessary. The PSV and PSLL alarms are also critical in this section.

The consequence of low nitrogen flow, and thus low pressure, can result in the tank imploding. This can lead to spillage, likely creating a vapour cloud and increasing the likelihood of an explosion. Leaving the PSV open for an extended period can lead to the formation of impurities and oxygen in the tank, creating a flammable atmosphere.

To reflect a real control room environment, some additional nuisance alarms and stimuli were introduced during the exercise to increase the complexity of the system, with the intent to distract the operators' attention and disrupt their thought process flow. These include the low and high-level alarms, LAL01 and LAH01, respectively. The high-pressure and flow alarms are opposite their low alarms. The temperature alarms, TAL01 and TAH01, are annunciated when the tank temperature goes low or high, respectively, beyond a stipulated threshold.

Methanol

The Methanol Section follows in Figure 3.5, where liquid methanol is extracted using a pump and taken to a boiler at a controlled rate. The methanol in the boiler is vapourised at 70°C and controlled by the steam flow rate. This vapourised methanol is further heated to around 190°C in the heater. The critical alarm here is the high-flow rate switch (FSH01) that puts the plant in an emergency state during a high flow rate. The heated vapour further mixes with compressed air from the Compressor Section. The alarm nomenclature is similar to those mentioned for the Tank and as would be later explained in Section 3.2.2.

If there is no flow of methanol, possibly due to pump failure, the boiler can overheat, damaging the system and stopping production. Extra alarms, such as WAH02 for high power and VAH02 for high volume, indicate high power and volume, respectively.

Compressor

Fresh air is required for the chemical reaction in the reactor. This is obtained from the atmosphere and compressed by the air compressor in the compressor section in Figure 3.6. This air is compressed simultaneously as the recirculated gas from previous production, if any. The compressed air goes through a second compressor to attain the

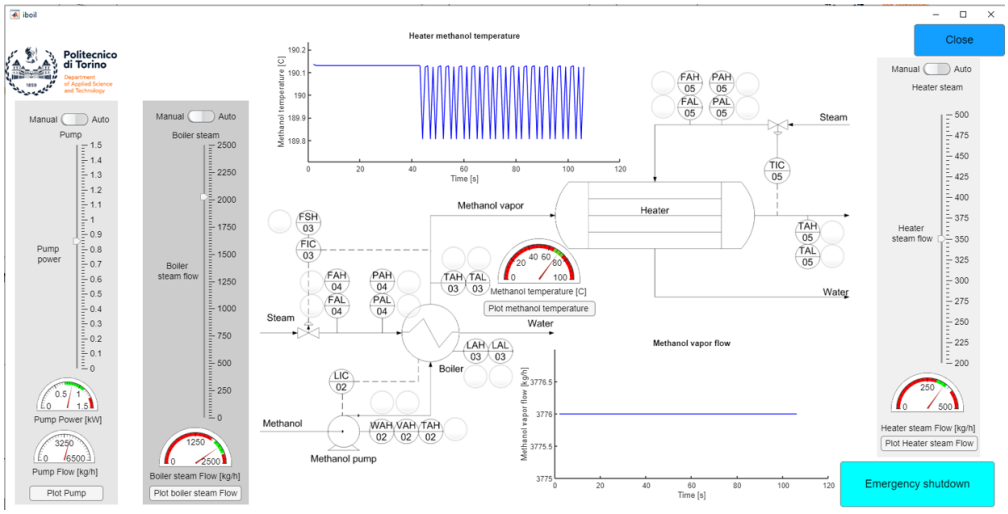


Fig. 3.5 Methanol section of the plant.

required pressure before it goes to the heat exchangers. There are high power and volume alarms for each compressor.

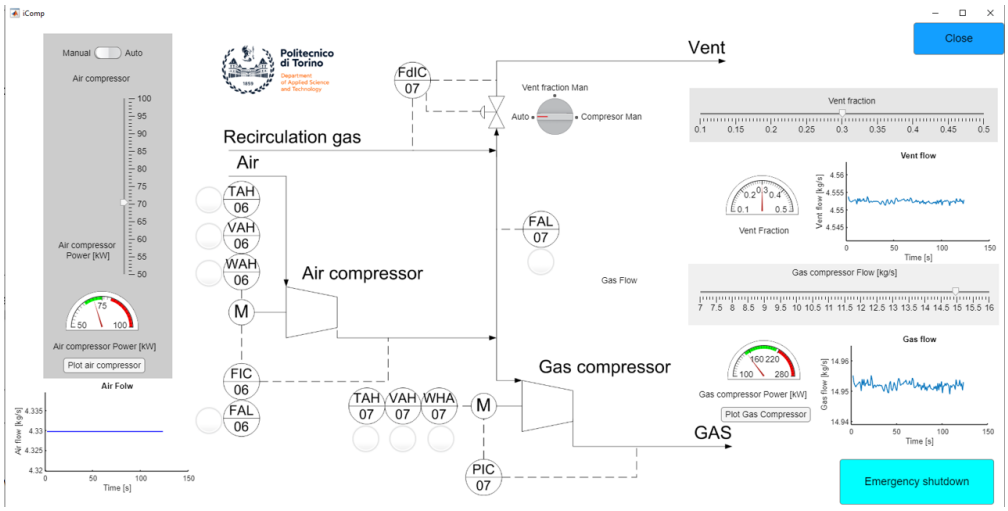


Fig. 3.6 Compressor.

Heat Recovery

The gas mixture passes through a heat exchanger (REC2) in Figure 3.7. Here its temperature is raised to around 200°C before entering the reactor in the Reactor Section.

In the Heat Recovery Section, heat exchangers REC1, REC2, and REC3 are employed, with REC1 and REC3 tasked explicitly with cooling the reactor output. The heat recovered from these exchangers is utilised for processes such as boiling and cooling water within the plant. REC3 cools the product to an absorption temperature of around 67°C before it enters the Absorber Section. Then it is absorbed by water flowing in the opposite direction.

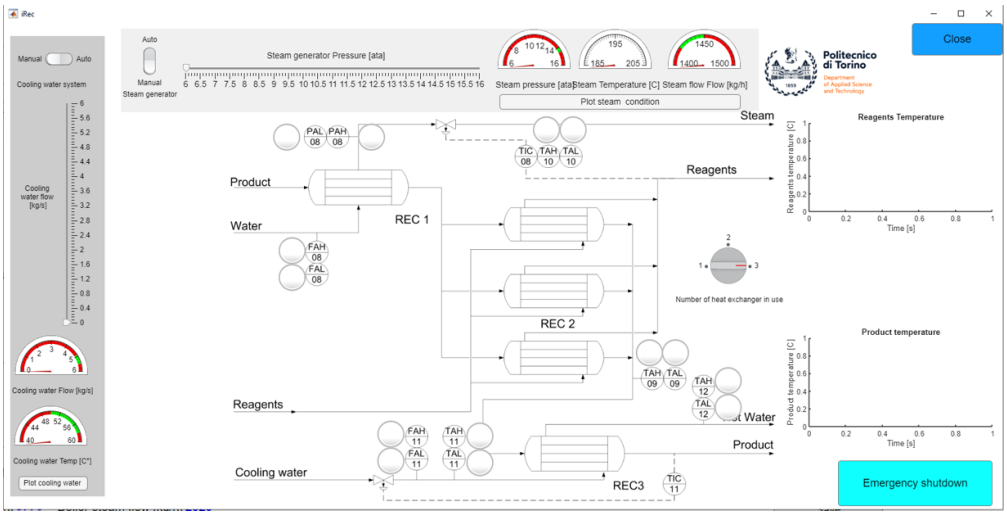


Fig. 3.7 Heat Recovery.

Reactor Section

The chemical reactions 3.1 and 3.2 take place in the reactor. The reactor in Figure 3.8 is a shell and tube heat exchanger where the chemical reaction occurs within the tubes using an iron and molybdenum oxide catalyst. The gases enter the reactor at approximately 220°C and exit at around 290°C . The exothermic oxidation reaction generates heat, partially removed by producing steam. A pressure controller manages the reactor's temperature by adjusting the steam conditions.

On the reactor's input line, CAH13 and CAH14 detect high concentrations of oxygen and methanol, respectively. Temperature is monitored using the following thresholds: TAH13 at temperatures above 335°C, TAHH14 at 345°C, TAL13 below 315°C, and TALL14 below 305°C. These alarms help prevent catalyst damage and avoid disruptions in downstream absorption processes. Additionally, if the steam pressure in the reactor's shell exceeds safe limits, a rupture disk (D5) activates, putting the plant in an emergency

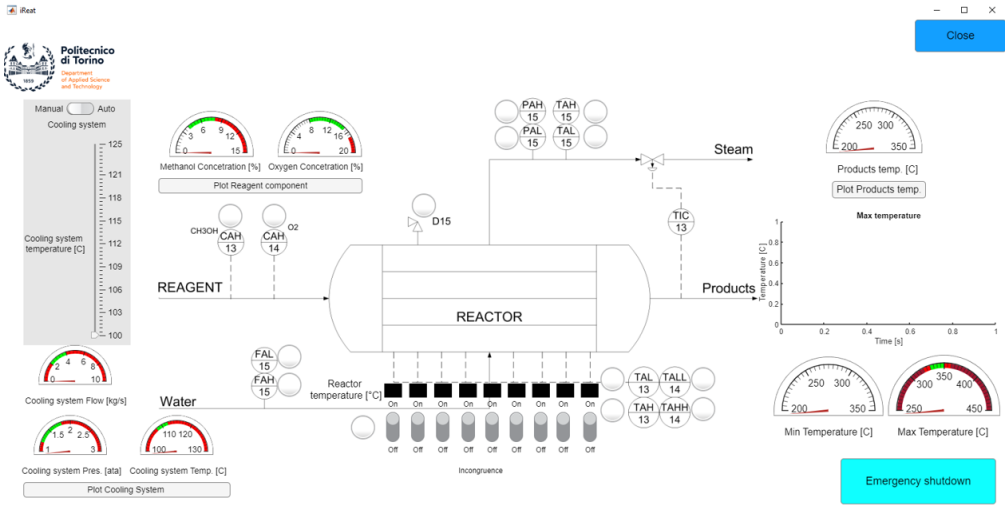


Fig. 3.8 Reactor.

state. After the reaction, the product is directed to the heat recovery section, where it is cooled before moving on to the absorber.

Absorber Section

In the absorber (See Figure 3.9, the cooled reaction product is processed to extract formaldehyde by absorbing it in water. The absorber is a packed tower where water flows from the top while gases move in the opposite direction (counter-current flow). Approximately 70% of the exhaust gas from the absorber is recirculated. The remaining 30% is directed to a combustor for the removal of volatile organic compounds (VOC).

The concentration of formaldehyde in the output liquid phase is primarily controlled by the input water flow. The absorber can be overheated at high temperatures. Temperature control is managed by alarms TAH16 and TAL16. Flow rates are also monitored, with FAH16 and FAL16 alarms responding to 120% and 80% of the nominal water flow rate, respectively. Temperature variations within the absorber column trigger the TAH18 or TAL18 alarms and other alarms like the CAL20, FAL16 and TAH18. If the operator takes no action, alarms from various sections of the plant (TAH20, TAH12, TAL15, TAH09, WAH07, PAL15, PAL08, FAH08, and TAH13) will be triggered. Followed shortly by an emergency shutdown. At the column's base, liquid levels are tracked by LAH19 and LAL19 alarms. The developed simulator includes over 80 alarms

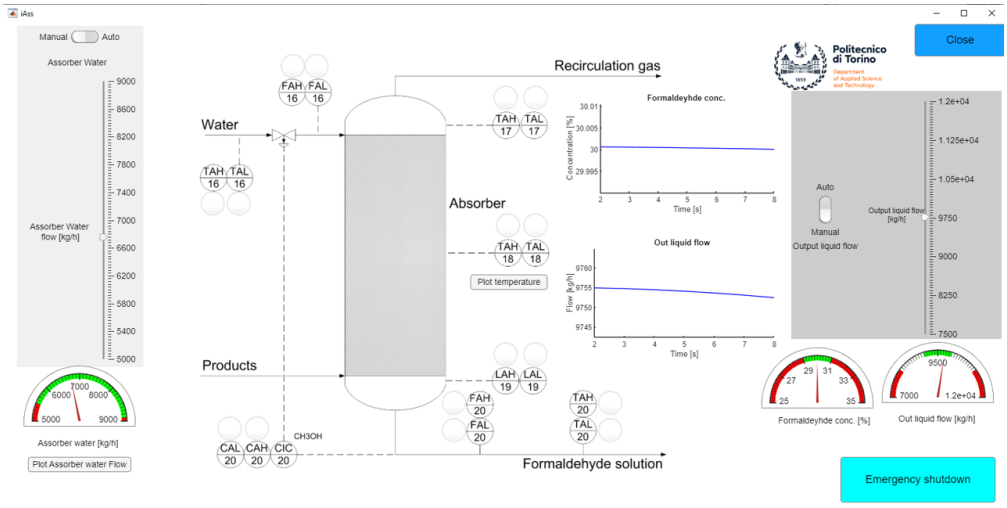


Fig. 3.9 Absorber.

in total - 20 temperature, concentration and flow alarms (high and low) each, 19 level alarms, 15 pressure alarms and some controllers.

3.2.2 Design of Tools

HSI: The Screen Displays

Requirements for the displays, pop-ups, alarm list, and trends were collected based on IEC standards 63303 [89] and [84]. This was done in addition to input from human factors and DCS experts. In addition to the general human factors evaluation (HFE) concepts for the HMI design in IEC 63303 (Human-Machine Interfaces for Process Automation Systems), Appendix A show some of the additional requirements gathered, those implemented or not implemented in this study with the appropriate references.

According to the IEC 63303 standard, the display hierarchy for process automation systems should consist of a maximum of four levels. That is, from broad coverage at level 1 to more detailed coverage at level 4. The displays developed in this work are grouped into levels 1 and 2. In this study, we have designed three types of displays. The first type is the support displays (see Figure 3.2), which contain the alarm list, procedures, and critical system graphs (see Figure 3.10). The second type is the mimic displays, which show specific plant systems and subsystems and allow for operator manual control (see Figure 3.11). The third type is the overview/central display, which includes the

plant's process flow diagram (PFD). Some buttons for navigating plant sections and for experimental purposes only were added.

The Support display 3.10 can be described as a combination of level 1 and 2 displays. It is split into four sections. The alarm message lists are in the top left corner, and the electronic procedure list is in the top right corner. The bottom left corner houses the reactor temperature graph and, in some cases, the absorber input water flow graph and formaldehyde production trend. The bottom right corner houses the absorber input water flow graph and formaldehyde production.

The mimics (see Figures 3.4 to 3.9) are a level 3 display, and the overview display is level 1 (See Figure 3.3).

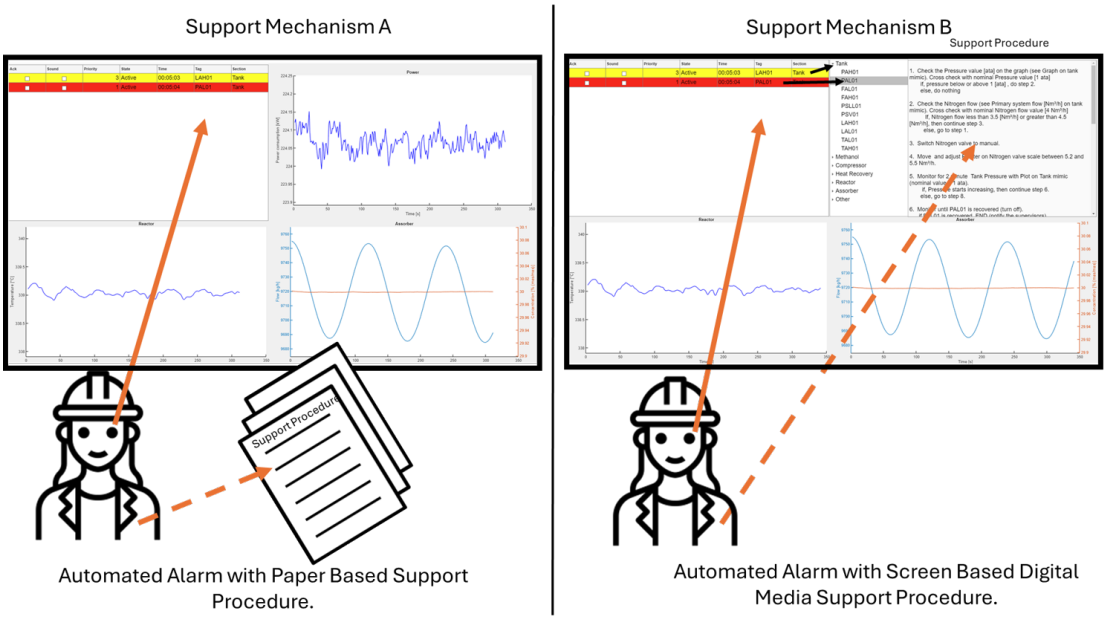


Fig. 3.10 The Support Display. A. for those using paper procedures, and B. for those using electronic procedures. At the top left corner is the alarm list, while the graph trends of critical systems and production volume occupy other sections of the display [155].

HSI: Alarm List Display

An alarm through the alarm HSI essentially serves three purposes. The system alerts the operator to deviations or issues, communicates their priority and nature, and can notify the operator of successful resolutions [98]. Requirements for the alarm system display and presentation were extracted from [98] as summarised briefly in Table A.4.

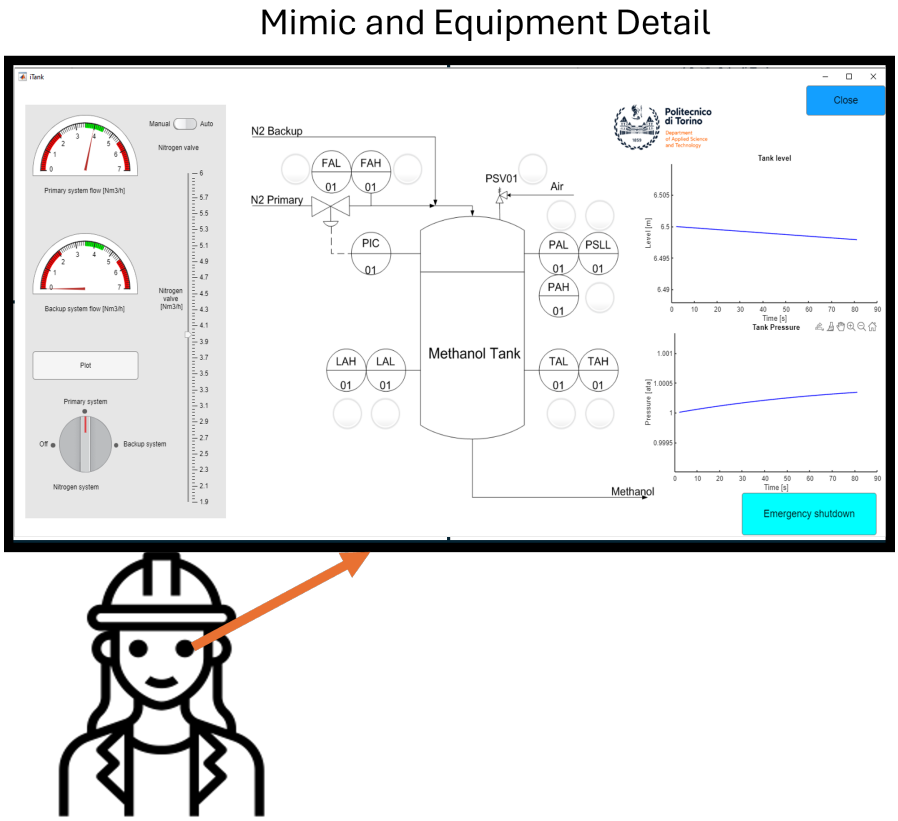


Fig. 3.11 The Mimic Screen with an example from the Tank Plant section.

The following details the alarm display as designed for the study:

1. Naming Convention: The naming convention for the alarm TAGs is SASXX. The first S is the process parameter: pressure, level, temperature, concentration, and flow. A refers to alarm, and the last S is low, low-low, high, or high-high. The XX refers to the number of the tag counting from 01 - 18 in this plant (see Figure 2.7). Such visual aids (a combination of colour and text) guide operators' attention to the relevant alarm.

2. Alarm priority: Alarm prioritisation refers to assigning importance to an alarm based on specific criteria. These criteria include the need for operator action, challenges to the safety system, threat to critical safety function, and post-risk analysis [98]. Two alarm types were developed for this study: alarms without prioritisation and alarms with prioritisation.

   The alarm prioritisation has been statically determined for this study and assigned before the system is implemented. Each alarm was assigned a risk-based priority of 1, 2, or 3, where 1 was a more critical alarm and 3 was less critical. Also, colour codes indicate priority levels: red (priority 1) for 'Critical', orange (priority 2) for 'Medium', and yellow (priority 3) for 'Low'.

3. Alarm sound: Alarm sound and visuals provide operators with attention support. They serve as a cue to direct operators' attention to a change in the plant. In the formaldehyde plant simulated design plant, three different sound levels were encoded. The alarm is silenced upon annunciation by operators, signifying a recognition of the alarm and removing the associated noise.

4. Alarm Acknowledgement Acknowledging an alarm is an essential step in alarm handling. By acknowledging an alarm, the operator communicates their intention to resolve it. In this study, a column on the alarm message box allows the operator to check any alarm. Most importantly, it is important to attend to the critical alarms first.

5. Alarm Message The study presents the alarm message as a list of active alarms for both Mechanism A (Digital Procedure) and Mechanism B (Electronic Procedure) (see Figure 3.10). Message lists display only active alarms and provide more information than specially dedicated, continuously visible (SDCV) displays [98].

In this study, the message lists are integrated into the support display. It is situated just before the electronic procedure display or, in the case of Mechanism A, before the graph of the plant power system. The message box first displays a tab to acknowledge the alarm, the next to silence the alarm. It also shows the alarm priority levels (1, 2, and 3) and the respective colour codes (yellow, orange, and red). The alarm state, whether active or inactive, is followed by the alarm TAG and Section (alarm location in the plant).

HSI: Alarm Response Procedures

Procedures are crucial in safety-critical industries. They are vital for safety management [6], improving situational awareness [15], impacting workload [9], and aiding decision-making [15]. As mentioned earlier in Chapter 2, different types of procedures, formats, and levels exist. For the study, alarm response procedures are of specific interest in a detailed list format (see Figure 3.10, Mechanism B). This study compares two levels of procedures: paper-based and electronic. Also, one ARP was written for each of the 80 alarms in the plant.

The electronic/on-screen procedure's position was considered in the recommendations by [98]. According to [98], the operator should not be required to leave the location at which the alarm message is displayed to access the ARP information. Hence, the ARP for those having electronic support (Mechanism B) has been developed beside the alarm list, as shown in Figure 3.10. However, this changes for those using paper-based procedures, as is the status quo in most process industry control rooms. Operators who use paper will have to shift their gaze farther from the alarm list for interfacing, as shown in Figure 3.10. A couple of key features were added to the procedure in compliance with standards and guidelines, as shown in Appendix A.2 - A.3. However, the detailed list format was similar for both the paper and electronic procedures. Some features are described below:

1. ARP Access: The TAGs and section names for the alarm response procedure indexing correspond to that of the alarm list (see Appendix A.4). This access rule applies to levels 1 (paper-based procedure) and 2 (electronic procedure).

2. ARP Content: Requirements for the procedure content include grouping format, alarm information recording, follow-up tasks, and references/abbreviations. Also,

the procedure core process content should be available, such as setpoints and limits of alarms or process parameters.

3. Consistency with the screen displays: According to [98], the information on the procedure, terminologies, conventions, colour codes, standards, etc., should be consistent with what is used on the other displays, overview, mimics, alarm support.

4. Format: It is recommended that procedures be presented in a format that facilitates quick and effective searches [5]. The procedure format designed for this study is a list type format as shown in Figures 3.12 and 3.13. A hierarchical, rule-based format was used to structure the procedures, outlining each step needed to address plant alarms. One version of this representation is shown on a display interface, while the other is presented on-screen [18]. The procedures are further described below:

- Digitised Procedure:

  - Each alarm procedure is arranged for more straightforward navigation in the digitised format displayed on screens, as illustrated in Figure 3.12a. These procedures are grouped by plant section and sorted by alarm number from 01 to n. Figures 3.12a and 3.12b show procedures related to the tank section of PAL01 (low-pressure alarm).

- Paper Procedures:

  - The paper-based procedure presented in Figure 3.13 aims to improve organisation and simplify searching. It includes a table of contents at the beginning (Figure 3.13b) and colour-coded tabs at the start of each plant section (Figure 3.13d). The format and writing style are similar to the digitised version, as depicted in Figure 3.13c. The final steps for the PAL01 procedure in Figure 2c continue from the previous ones and are shown in a red colour compared to in Figure 3.12.

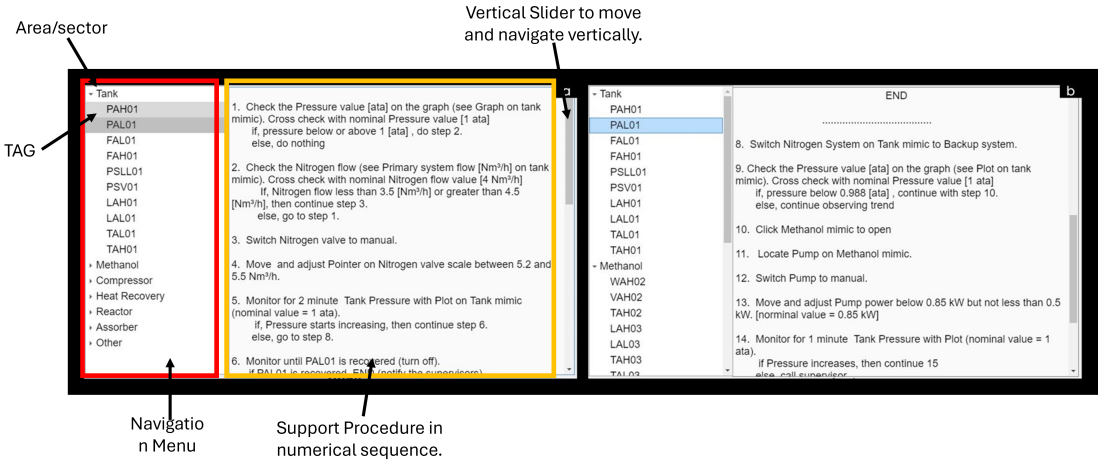


Fig. 3.12 Electronic Procedure with the Navigation Menu. (a) The procedure for the low-pressure alarm (PAL01) details steps 1 to 6. (b) The continuation of PAL01 covering steps 8 to 14. The left sidebars illustrate the sequence of alarm arrangements to aid in navigation [156].

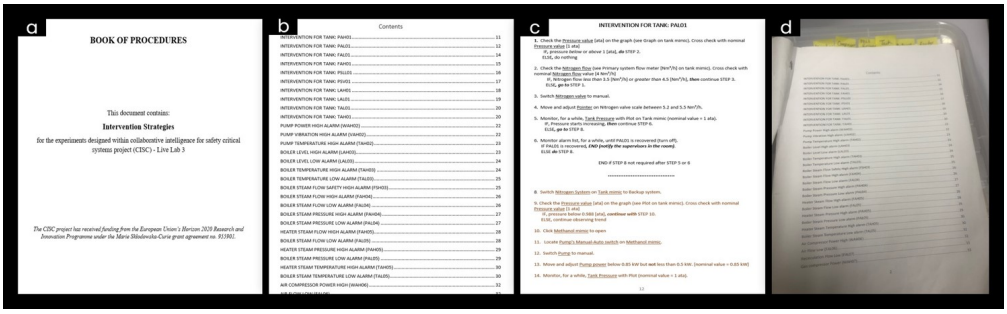


Fig. 3.13 The Paper-based procedure. (a) Cover page of the paper procedure. (b) Example of the table of contents. (c) Procedure for PAL01 (low-pressure alarm) covering steps 1 to 14. (d) The actual appearance of the procedure during testing, with a colour tag marking the beginning of each new plant section [156].

3.3 Experiment Design

The experimental study in this thesis compares two HiTL configurations: 1. using paper procedures during alarm handling and process control and 2. electronic procedures as shown in Table 3.1. However, this study is part of a more extensive experimental study investigating 4 HiTL configurations regarding alarm rationalisation, procedures, and displays as shown in Table 3.2.

Table 3.1 Key performance shaping factors for comparison. These include alarm design, procedure, Display/HMI, and task complexity, which differentiate the groups and their different levels.

Group	Procedure Type	Display Mechanism
PBP	Paper	Mechanism A (Figure 3.10)
EP	Electronic	Mechanism B (Figure 3.10)

Table 3.2 General experiment study groups [18].

Group (G)	Group Initial	Description
Group 1	G1	Lacks alarm rationalisation, uses paper procedures, and lacks AI support.
Group 2	G2	Implements alarm rationalisation but does not utilise on-screen procedures or AI support.
Group 3	G3	Utilises alarm rationalisation and on-screen procedures but does not use AI support.
Group 4	G4	Incorporates alarm rationalisation, on-screen procedures, and AI support.

Design

Given the focus of this thesis, the experiment focuses on a 2×2 factorial design considering the following factors: Procedure Type (Paper vs. Electronic) and Display

Mechanism (Mechanism A vs. Mechanism B). Both are between-subject variables, given that each group uniquely tests them. However, the task and alarm design are classed as within-subject variables since both groups progressively test all three levels, including the alarm design. The alarm design stays the same as it is not the focus for comparison in this study.

Participants

Participants were selected using a purpose-sampling approach. The criteria are as follows: must be 18+, be proficient in English, and have a normal or corrected-to-normal vision. The study includes 92 participants (23 in each HiTL configuration), 36 females and 56 males, from the Politecnico di Torino and a few external individuals. Participants range in age from 21 to 61, with a mean age of 25 and a standard deviation of 5.4. Participants in each group were randomly assigned by time of day to be distributed equally between morning, afternoon, and evening. The group exhibits diverse experience levels. It mainly comprises junior process engineers who volunteer among the master's students in chemical engineering at Politecnico di Torino. The group includes experienced professionals such as Ph.D. candidates and faculty members.

However, for this thesis, where only two groups out of the original four are of importance, the total number of participants is thus 46. This number is lower than the estimated sample size of 277 using the G-Power Tool. Using the g power tool, the effect size, f, is set to 0.2. Effect size (ES) represents the smallest difference a researcher aims to identify between groups in a study [157]. A small effect size of 0.2 is chosen based on the slight difference in means reported by [15] in their study. The further use of 46 was approved because related control room studies mostly involved 3 to 4.. participants in their studies [11], [103], [10], [130], [15], [17].

α error probability was set to 0.05. Probability at 0.05 is commonly used in control room studies in industries [15], [17]. Power (1 - β) err probability is set to 0.8 using equation 3.4, A β of 0.2 is set as the probability of making a Type II error (failing to detect a difference between the groups), and is the value commonly used in clinical studies [157]. Also, high power reduces the chances of such errors. There are no reported values for the control room studies.

The numerator df to 3 is calculated using equation 3.3, and the number of groups to 4 is calculated using equation 3.5.

The degree of freedom of the numerator for 2×2 ANOVA is calculated by:

$$\text{df_numerator} = (a-1) + (b-1) + (a-1)(b-1), \tag{3.3}$$

$$\text{df_numerator} = (2-1) + (2-1) + (2-1)(2-1) = 1 + 1 + 1 = 3,$$

$$\text{Power} = 1 - \beta \tag{3.4}$$

$$\text{Power} = 0.8 \implies \beta = 0.2,$$

$$\text{Number of Groups} = a \times b, \tag{3.5}$$

$$\text{Number of Groups} = 2 \times 2 = 4,$$

where a represents the number of levels in the first factor of the factorial design and b represents the number of levels in the second factor of the factorial design.

Scenarios (S)

To compare the factors to be studied practically, three critical safety scenarios were identified from the case study of the formaldehyde process plant. All participants completed three scenarios, each lasting between 15 and 17 minutes: 15 minutes for scenarios 1 and 2 and 18 for scenario 3. Scenario 3 had a longer duration because of the task complexity. The scenarios are detailed below:

1. Malfunction of Pressure Indicator Control (PIC01) in the Tank Section:

    - Scenario: The automated tank pressure management system fails, requiring manual adjustment of the nitrogen inflow by the operator to maintain pressure. If nitrogen supply halts, the pressure will drop as methanol continues to be transferred to the plant.
    - Hazardous Event: Potential for the methanol storage tank to implode.

- Objective: Prevent the activation of the pressure switch PSLL01 within 7 minutes of the initial alarm.
- Complexity criteria: No alarm flood scenario and interaction with 1 ARP with less than 7 task steps.

This scenario assumes that the PIC01, FAL01, and PSV01 are faulty. The alarms annunciated, task flow and the time they occur are shown in Figure 3.14.

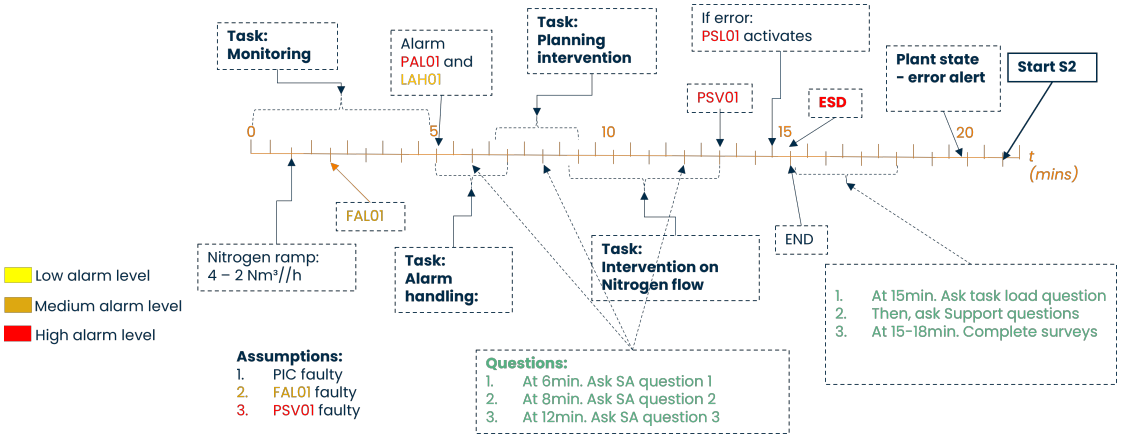


Fig. 3.14 Alarm-related initiating events, general operator task, and think-aloud points in scenario 1.

2. Failure of the Primary Nitrogen Supply Valve in the Tank Section:

- Scenario: The main nitrogen source fails, which requires the operator to activate a backup system. The backup system activates slowly, so the operator must adjust the pump to stabilise the tank pressure.
- Hazardous Event: Risk of the methanol storage tank imploding.
- Objective: Prevent the tank from imploding.
- Complexity criteria: No alarm flood scenario and interaction with 1 ARP of more than 7 task steps.

This scenario assumes that the FAL01, PSV01 and the Nitrogen flow manual control scale are. The alarms annunciated, task flow and the time they occur are shown in Figure 3.15.

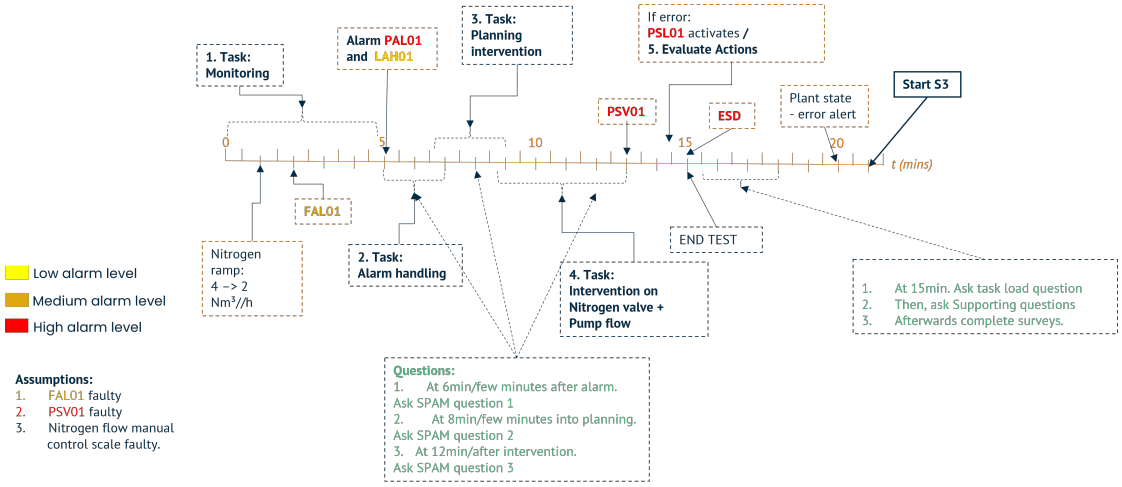


Fig. 3.15 Alarm-related initiating events, general operator task, and think-aloud points in scenario 2.

3. Temperature Control (TIC11) Malfunction in the Heat Recovery Section (REC3 system):

   - Scenario: A malfunction in the temperature control could lead to the reactor overheating, which may result in increased pressure. The goal is to prevent the pressure switch (PSL01) from being triggered within 18 minutes after the alarm is raised.
   - Hazardous Event: Reactor overheating.
   - Objective: Avoid overheating in the reactor.
   - Complexity criteria: Alarm flood scenario and possible interaction with two or more APRs.

This scenario assumes the TIC11 and cooling system control (heat recovery) is faulty. The inlet water flow to the absorber should not be touched. Lastly, the operator can only maintain and manage reactor temperature until the 18th minute as the cooling is simulated to recover only at that time. The alarms annunciated, task flow and the time they occur are shown in Figure 3.16

Participants in each group completed scenarios 1 to 3 in order of increasing complexity. The 'Scenario' or 'Task Complexity' factor was not randomised to avoid introducing additional variables. This approach allowed us to observe the effects of

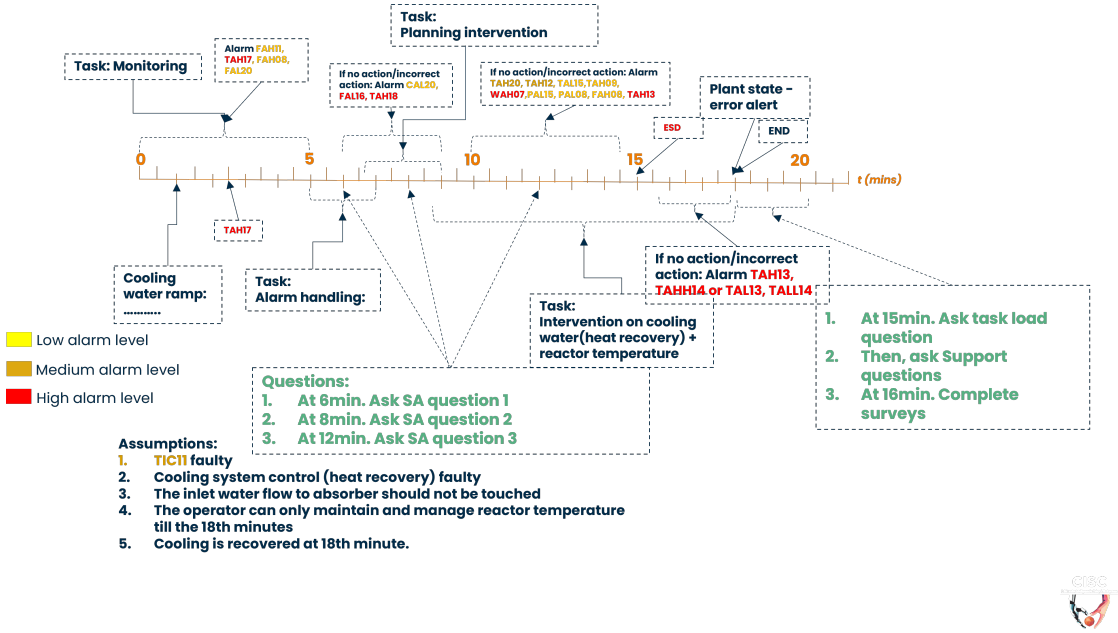


Fig. 3.16 Alarm-related initiating events, general operator task, and think-aloud points in scenario 3.

increasing complexity, similar to what might occur in a typical control room where situations escalate. Randomisation was also impractical due to the small pool of participants.

3.3.1 Data Collection Methods/Measures and Procedures

In this study, four subjective techniques/methods were used: a questionnaire, SART, SPAM, and NASA-TLX tests. In addition, two objective methods— a simulator and eye tracking— were employed. These methods have been briefly reviewed for their application in control room-related studies across industries. Their importance for use in this study has also been discussed in Section 3.2 of this chapter.

A carefully crafted questionnaire was developed to gather information on the operators' training, experience, age, gender, and other preliminary details before the study began. The supervisors assigned scores for training and experience. After each scenario, a second questionnaire was used to assess operators' perceptions of procedural support, alarm priority, and task load. Additionally, NASA-TLX, SPAM, and SART were administered post-task. Both SPAM and SART aimed to complement each other,

ensuring a thorough evaluation of situational awareness. In that, SPAM focuses on decision-making from a managerial context, which includes social, personal and analytical sides, whereas SART focuses on situational awareness from human factors, ergonomics interaction and cognitive load to maintain recognition in a rapidly changing situation. Operators and system interactions, such as procedure openings and alarm acknowledgements, were recorded on the objective side. Technical data on alarms, such as the alarm TAGS and the parameters like pressure, temperature, and so on, were also collected. The introduction of eye-tracking allowed for the observation of cognitive processes while using either paper or electronic procedures.

Training and Familiarity

After completing consent forms, participants watched a brief 5-minute video summarising the plant process and study expectations. This was followed by a longer training between 35 - 45 minutes. Training is a key performance-shaping factor in control rooms and was essential to factor in for the study. However, two questions on familiarity were asked before the study: their familiarity with control rooms and their familiarity with chemical industry process operations. The questions asked are shown in Subsection 3.3.2.

Perceived support, alarm design and Task Load/Complexity

Participants were asked questions about their perception of support or how well the procedures and alarm design supported them in each of their groups. They were also asked to rate the task complexity of each scenario upon completion. The questions asked are shown in Subsection 3.3.2.

Situational Awareness: SART

The Situation Awareness Rating Technique (SART) is a self-assessment tool for evaluating situational awareness across demand, understanding, and supply after each scenario. Situation Understanding (U) comprises Information Quantity, Information Quality, and Familiarity. Situation Demand (D) encompasses the situation's Instability, Complexity, and Variability. Meanwhile, the 'Supply' of attentional resources (S) includes Arousal, Arousal, Concentration, Division of Attention, and Spare Capacity. The variables for each dimension are added to achieve the overall score for each

dimension. For example, scores on Information Quantity, Information Quality, and Familiarity are added to achieve the Situation Understanding score. An earlier part of this chapter, Section 3.2, discusses the justification for choosing this method. The SART Index is calculated using the equation:

$$U - (D - S). \tag{3.6}$$

A sample of the SART SA questionnaire employed in this study is available in the Appendix B.6, B.7. The participants rated their perceived demand, understanding, and supply for each scenario on a scale of 1-7. "1 represents 'very low' situational awareness, while 7 represents 'high' situational awareness." The variables from this method and the questions used to assess them are shown in subsection 3.3.2.

Situational Awareness: SPAM

This study also employed a SPAM-based method to evaluate situational awareness. SPAM is considered a direct subjective technique that enables scoring without interrupting the ongoing test [64]. However, some researchers also refer to it as a direct objective method [64]. As mentioned earlier, a key advantage of this approach is that it doesn't require pausing the test for participant evaluation. Instead, participants are questioned using a concurrent think-aloud method. In this case, the supervisors rated the participants, unlike in the SART method, on a scale of 1 – 5. 1 means 'very low' and 5 means 'high' situational awareness.

The chosen levels for assessing situational awareness (SA) align with Endsley's SA framework [45], Endsley SA was based on a hierarchical three-level model of situational awareness. Level 1 is the lowest level, and the perception of the environment is level 2, the intermediate level, and the comprehension of the significance of the current situation. Finally, Level 3, the highest level, Projection of future status, in particular forecasting outcome, with or without intervention to aid in the formation of a decision. These levels are designed to capture broader aspects of SA beyond what is typically covered by SART. Questions were crafted and posed to participants at three key points during each scenario: during alarm handling, while planning the intervention or initially reading procedures, and finally, after the intervention. These questions were generally asked around the 6th, 8th, and 12th minutes of each scenario (see Figures 3.14, 3.15, and 3.16).

A SPAM index is calculated to evaluate SA using the equation:

$$f_{(a_1,\ldots,a_3)} = \sqrt{\frac{1}{3}\sum_{i=1}^{3} a_i}, \tag{3.7}$$

a_i denote the SPAM dimensions of Perception, Understanding, and Projection, which are elaborated through the questions posed to the participants. The variables from this method and the questions used to assess them are shown under a subsequent subsection - variables and measures.

Workload: NASA-TLX

The raw-NASA-TLX is a widely recognised tool for evaluating subjective workload. Participants rated their mental, physical, and temporal demands, as well as their performance, effort, and frustration, on a scale from 1 to 7. Higher scores reflect more significant workload levels. These ratings offer valuable insights into participants' personal workload experiences and help us gauge the perceived demands of the task. Like in the SART, NASA TLX was scored by the participants. Also, it was scored post-task together with the SART, as both were designed under the same questionnaire.

The NASA TLX index is computed using the following formula:

$$f_{(a_1,\ldots,a_6)} = \sqrt{\frac{1}{6}\sum_{i=1}^{6} a_i}, \tag{3.8}$$

a_i represents the dimensions of the NASA TLX. These include: a) perceived mental, physical, and temporal demands of the tasks; b) effort and perceived performance on the task; and c) perceived effort expended and frustration levels. The variables from this method and the questions used to assess them are shown under a subsequent subsection - variables and measures.

Eye-tracking

Eye-tracking technology was employed for data collection to improve our understanding of the operator's cognitive processes using physical measurements. That is, information processing, attention allocation, mental workload, situational awareness and, in summary, visual behaviour and cognitive performance [151]. It allows for capturing various eye-related pupillometry and gaze-based metrics. Pupillometry involves studying changes

in pupil diameter related to cognitive processing. At the same time, gaze tracking is performed using lens cameras that monitor eye movements and record fixations, saccades, and periods of inactivity [151]. The measures that comprise the eye-based variables in this study are discussed in the latter subsection, variables and measures. The goal is to examine the cognitive disparities among the groups during the intervention, particularly in relation to the most challenging tasks, whether using paper or electronic procedures.

Tobii Glasses 3, operating at 50Hz, was used in the experimental study. A frequency of 50 Hz is commonly selected for control room-related studies across industry [9]. However, higher frequencies provide detailed and accurate measurements of eye movements. Tobii glasses were chosen for their flexibility with eye and body movement. The experiment in this thesis is designed with three screens, meaning that the participants would need to move their heads often, and Tobii glasses provide such affordance. In terms of analysis, Tobii provides a platform, Tobii Pro Lab software, for raw data analyses.

Simulator recording

Simulator data were recorded and saved for each experimental scenario, capturing information from plant systems, alarms, ARP, and display elements. This data is categorised based on different plant sections as shown in Appendix A.5 to A.9. The goal is to analyse the recorded features to derive performance and behavioural variables and insights. The derived insights will aid in comparing various Human-in-the-Loop (HiTL) configurations.

The benefits and justification for such behavioural and performance measures were discussed in Section 3.2. The variables and how they were derived are discussed in the subsequent subsection titled "Variables and Measures.

Before delving into the specifics of the variables and measures, a crucial discussion is the application of these methods during the test. Also, how the scenarios were conducted and the preparatory steps involved.

Procedure

Procedure Format:

This study involved 46 participants who engaged in three scenarios, each lasting 15 to 18 minutes (depending on the scenario, as already stated). The scenarios simulated

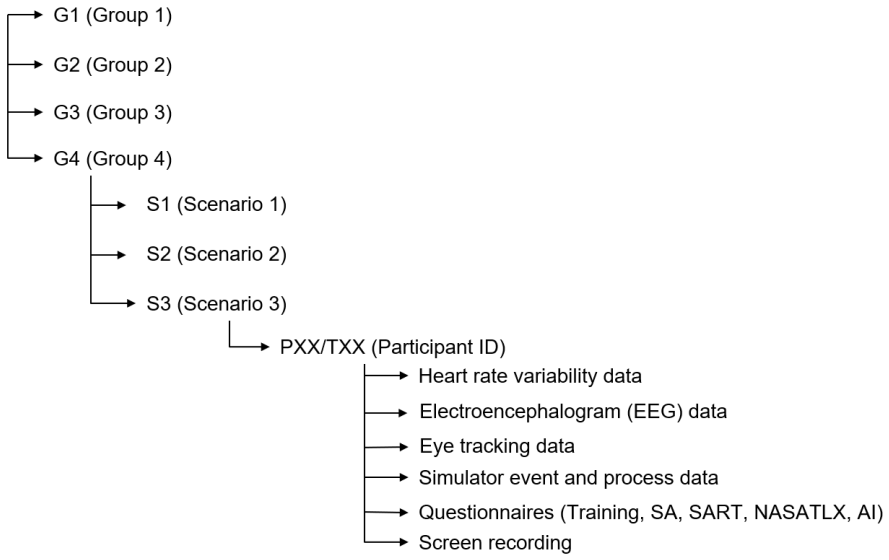


Fig. 3.17 Order of saving the recorded simulator logs after each scenario is completed [158].

typical control room tasks: monitoring, alarm handling, planning, and intervention. Scenario 3 involved the supervisor communicating extra instructions to the operator. The running order of scenarios for all participants is shown in Figure 3.17.

The data for each participant and scenario is recorded and logged. The experiment participants underwent the following steps:

- Step 1: Prior to the day of the experiment, hopeful participants had received emails with a two-page information sheet about the exercise before registering. On the day of the test, registered participants were given a short introduction to the experiment. They read a two-page information sheet that outlined the test procedures, signed the consent form upon arrival, and completed a demographic questionnaire covering age, gender, and academic background. Participation in the experiment was voluntary. The participants underwent similar scripted training conducted by the researchers to ensure no participant had any specific advantage or bias as a result of the training element. To this end, participants watched a brief 5-minute video introducing the test protocol, plant process, and interface displays before the training. The training session included 30-45 minutes of instruction on plant processes and dynamics, as well as guidance on using the operating procedures. The operating procedures consisted of a paper document outlining process parameter limits and operator expectations. Additionally, the session

covered the ARP related to their study group, PBP, or EP. A 5-10 minute break was provided after the training before the setup of monitoring equipment.

- Step 2: During the test, participants wore the eye tracker. Each scenario included collecting subjective, behavioural, cognitive, and performance data. The eye-tracking device was set up after the 5-minute break. Situational awareness was assessed using the SPAM methodology while the test was ongoing. The questions were asked at specific times-typically the 6th, 8th, and 12th minutes (see Figures 3.14 3.15, 3.16. These questions aimed to evaluate the three levels of situational awareness: perception, comprehension, and projection. The eye trackers were removed from the description of the participants after each scenario and after filing the questionnaires in case they needed a break. The eye-tracking record is saved after each scenario. Also, each simulator record is saved after each scenario in a designated folder in the format shown in Figure 3.17.

- Step 3: After each scenario, participants completed the NASA-TLX, SPAM, and support rating questionnaires (procedure, alarm design, and task load). In scenario 3, an extra question on 'Communication' after 'task load' was asked to ascertain how helpful the supervisor's communication was. Finally, the eye tracker was removed at the end of the three scenarios. Then records of the last scenario from the eye-tracker and simulator were saved.

3.3.2 Variables and Measures

Variables

Several variables have been considered in this study. Some of these variables are clustered under the different methods they represent, such as the SART, SPAM, and NASA-TLX. The variables and the way they are defined or collected are presented below:

Training and familiarity: the supervisors asked the following questions to the participants to assess familiarity:

- What is your Level of familiarity with the chemical industry processes (1: Low, 5: High)

- Level of familiarity with the process control rooms (1: Low, 5: High)

The supervisors asked the participants questions to assess their training. The questions were designed to assess their past and current knowledge post-training on basic control tasks, alarm handling and intervention tasks, as follows;

- Can you show us how you change a value on a scale? Below is the instructor's assessment of how well the participant is capable of demonstrating. (1: Low, 5: High)
- Can you show us how you identify what section of a plant to open? Below is the instructor's assessment of how well the participant is capable of demonstrating. (1: Low, 5: High)
- Can you show us how you acknowledge an alarm? Below is the instructor's assessment of how well the participant is capable of demonstrating. (1: Low, 5: High)
- Can you show us where you can find a procedure? Below is the instructor's assessment of how well the participant is capable of demonstrating. (1: Low, 5: High)
- Can you show us how you acknowledge a suggestion by the "AI enhanced decision support system"? Below is the instructor's assessment of how well the participant is capable of demonstrating. (1: Low, 5: High)
- How would you rate your current knowledge of the simulator? (1: Low, 5: High)
- How would you rate your current knowledge of the operating procedure? (1: Low, 5: High)
- How would you rate the quality of information provided by the video during your training? (1: Low, 5: High)
- Any comment/ suggestion on how to improve the training?

Perceived Support and Task Load:
Task load:

- How complex did you perceive the task to be? (1: Low, 5: High)

Alarm design:

- How clear was the information received on the interface for the list of alarms, and how to acknowledge, silence or follow up on them? (1: Low, 5: High)
- How helpful was the support received in differentiating the alarm priorities? (1: Low, 5: High)

Procedure:

- How helpful was the support of consulting procedures during the scenario? (1: Low, 5: High)

SART: the standard SART variables are instability, variability, complexity, arousal, spare capacity, concentration, attention division, quantity, quality and familiarity. The first three are added up to form the SART demand dimension. The following four form the SART supply dimension, and the last three are for the SART understanding dimensions. The following questions were adapted from the traditional SART questionnaire to assess each variable, and some wording was adjusted for clarity:

- Instability: How unstable/ changeable was the task? (1: Low, 7: High)
- Variability: How many variables were changing with the task? (1: Low, 7: High)
- Complexity: How complicated was the task? (1: Low, 7: High)
- Arousal: How aroused/thrilled were you in the task? (1: Low, 7: High)
- Spare Capacity: How much mental capacity did you have to spare in the task? (1: Low, 7: High)
- Concentration: How much were you concentrating on the task? (1: Low, 7: High)
- Attention Division: How much was your attention divided in the task? (1: Low, 7: High)
- Quantity: How much information have you gained about the task? (1: Low, 7: High)

- Quality: How good is the information you have gained about the task? (1: Low, 7: High)
- Familiarity: How familiar are you with the task? (1: Low, 7: High)

These questions also help with understanding what the variables mean. The same questions were used in each scenario.

NASA TLX: the NASA-TLX variables include: Mental demand, Physical demand, Temporal demand, Performance, Effort and Frustration [17]. The following questions were asked to assess each dimension, and some words were adjusted for better understanding:

- Mental Demand: How mentally demanding was the task? (1: Low, 7: High)
- Physical Demand How physically demanding was the task? (1: Low, 7: High)
- Temporal Demand: How hurried or rushed was the pace/speed of the task? (1: Low, 7: High)
- Performance: How successful were you in accomplishing what you were asked to do? (1: Low, 7: High)
- Effort: How hard did you have to work to accomplish your level of performance? (1: Low, 7: High)
- Frustration: How insecure, discouraged, irritated, stressed, and annoyed were you? (1: Low, 7: High)

The same questions were used in each scenario.

SPAM: the SPAM dimensions or variables are the classical Perception, Understanding and Projection dimensions [51] [17]. This technique allows for writing the questions that best suit the experimental study. The following questions were asked to assess each dimension in scenarios 1 and 2:

- Perception: Which of these alarms, in your opinion, requires to be verified first $[FAL01, PAL01]$? and why?
- Understanding: Why do you think the PAL01 alarm is activated? And what do you intend to do?

- Projection: Now that you have done this, what do you think will change in the system? Why?

Scenario 3 involved a different section of the plant compared to scenarios 1 and 2. Hence, the questions on perception and understanding were different from previous scenarios, but that of projection stayed the same, as shown below:

- Perception: Which of these alarms, in your opinion, must be verified first [FAL11, TAH17]? and why?
- Understanding: Why do you think the TAH17 alarm is activated? And what do you intend to do?
- Projection: Now that you have done this, what do you think will change in the system? Why?

Eye-tracking: Eye tracking measures are some of the more commonly used in control room-related studies. They include gaze-based metrics for AOI analysis: fixation duration, visits, saccade amplitude and duration, and pupillometry-based metrics like pupil diameter. These variables are described below:

- Fixation duration (ms): This metric represents when the eye remains focused on a single point [159]. Typically ranging from 50 to 600 ms, fixation duration indicates the cognitive effort involved. Longer fixations suggest more intense cognitive processing [160].
- Number of visits or revisits: This metric counts how often the area of interest (AOI) is revisited. It reflects the level of attention or attraction an area receives, whether beneficial or detrimental.
- Saccade Amplitude (°): Saccades are rapid eye movements between fixations. The amplitude of a saccade, which measures the distance travelled by the eye, varies based on the task. For instance, during reading, it typically spans about 2° for 7-8 characters in standard font size [160]. Saccade amplitudes generally decrease as task complexity and cognitive load increase [159], [125], meaning that larger amplitudes often suggest lower difficulty in tasks like searching or reading and lower cognitive load.

- Saccade Duration (ms): The duration of a saccade ranges from 20 to 40 ms [160]. It reflects the time taken for the eye to move between fixations.

- Pupillometry: This technique measures changes in pupil size, which are influenced by the autonomic nervous system. Pupil dilation typically results from sympathetic activation, while constriction is associated with parasympathetic activation. Various factors can affect pupil size, including arousal levels, light conditions (low or high luminance), exposure to bright or dark stimuli, workload intensity, and the presence of expected or unexpected stimuli [161]. Table 1 provides insights from studies on how this metric correlates with these factors.

Simulator-based variables: The raw simulator data features are as shown in Appendix A.5 to A.9. However, some variables will be extracted from these raw features to observe the behavioural states and performance. These variables are defined below:

- Recovery Time: This metric measures the time taken by operators to resolve the pressure alarm PAL01, which was triggered 60 seconds into the experiment. In scenarios 1 and 2, this time is measured from the fault's onset to the clearing of alarm annunciation PAL01. However, in scenario 3, recovery time is irrelevant because the task involves maintaining a stable reactor temperature until the scenario concludes.

- Reaction Time: Reaction time is the period between the fault's initiation and when the operator manually overrides a control. Scenarios 1 and 2 involve switching the Nitrogen valve from automatic to manual mode. Scenario 3 pertains to the time taken to manually switch the cooling water system (FAL11) from automatic to manual.

- Response Time: This metric assesses the time operators need to adjust system settings correctly after a fault occurs. Scenario 1 measures the time taken to adjust the nitrogen valve. Scenario 2 captures the duration necessary to switch from primary to backup systems, with a failure to switch labelled as an error. In scenario 3, response time is represented by the time of the last action before a spike in the cooling system's temperature trend. Also, in this scenario, it is referred to as 'survival time' and is expected to last 1080 seconds.

- Accuracy: Accuracy is calculated by comparing the operator's adjustments with the target range specified in the procedure. This difference is measured using the Mean Squared Error (MSE) in all scenarios. Mean squared error (MSE) is commonly used to measure the average of the squares of errors of a variable or model prediction. That is, the difference between the actual (observed) values and the predicted/new values generated by a model or in this case, the operators' control action. In scenario 1, accuracy pertains to the nitrogen valve adjustments. In scenarios 2 and 3, it relates to the pump power valve (MWpopOld_l) and the temperature cooling system (Toutrec2c_l), respectively.

- Overall Performance: Overall performance is categorised based on recovery time. In scenarios 1 and 2, performance is categorised as optimal, good, or poor based on recovery time percentiles. In scenario 3, optimal performance means a recovery time above 900 seconds, good performance is between 850 and 900 seconds, and others 'poor'.

- Consequence: This metric evaluates the outcome of different system states. In scenarios 1 and 2, the consequence depends on whether specific alarms (e.g., PSV01 or PSLL) are triggered, with possible outcomes ranging from 'Safe' to 'Implosion'. In scenario 3, if the reactor temperature exceeds 400°C, it's labelled as 'Reactor Overheat' with a score of 5.

- Alarms Silenced: This metric counts the number of alarms silenced by the operator during the scenario, consistent across all three scenarios.

- Acknowledgement: It tracks the number of alarms acknowledged by the operator throughout scenarios 1, 2, and 3.

- Mimics Opened: This counts the number of mimic diagrams opened by the operator during the scenario, again consistent across all scenarios.

- Procedures Opened: This tracks the number of procedures the operator opened during the scenario, with no variation across the scenarios.

- Alarms Triggered: This metric counts the total number of alarms activated during the scenario, consistent across all scenarios.

- Error: The error metric indicates whether the scenario ended with unresolved issues. Scenarios 1 and 2 refer to situations where pressure rises or stabilises normally without initiating a shutdown. Scenario 3 relates to instances where the reactor's high-temperature alarm was triggered.

Performance and System Variables Extraction: Performance-based metrics evaluated in this study encompass recovery time, accuracy, overall performance, and consequences. While the number of alarms activated is a system feature, it also reflects the outcomes of the operators' actions, indicating whether those actions were correct or incorrect. It was clear that extracting certain metrics in Scenario 2 was similar to Scenario 1 but differed significantly in Scenario 3 due to different objectives. Scenario 3 focused on the reactor and absorber sections, while scenarios 1 and 2 were related to the tank section.

In scenarios 1 and 2, recovery time was measured based on the 'All21' feature (each alarm in the plant was recorded with a feature labelled 'All' followed by a number, such as 21, corresponding to the Pressure Alarm Low (PAL01)). However, in scenario 3, the recovery time did not apply to the control room operator. This was because the faulty cooling water system issue in the heat exchanger REC3 unit was intentionally introduced and could only be resolved in the field (see Figure 3.7). For this reason, scenario 3 did not have a recovery time per se. In this scenario, survival time was considered in place of response time. Survival time indicated the operator's ability to prevent a shutdown or reactor overheating until the end of the test, which lasted 18 minutes. The study assumed that by the 18th minute, the field operator would have resolved the cooling system issue with REC3.

Accuracy was assessed based on different features in each scenario. That is, 'FN2serb1O1' (Primary Nitrogen Flow) for scenario 1, 'MWpopOldl' (Manual Pump Power Value) for scenario 2. Finally, 'Toutrec2cl' (Temperature Out RECl) for scenario 3. Accuracy was calculated as the mean squared error (MSE) between the operator's adjusted value and the mean of the specified range outlined in the procedure.

The distribution of recovery times among participants was analysed to determine overall performance. Operators whose recovery times fell within the 25th percentile were classified as 'optimal,'. Those below the 50th percentile as 'good,' and the rest as 'poor'. The consequence of each scenario was determined based on raw features such as 'PSERB1,' 'TMAXREATORE1,' and 'emmerO1,'. These conditions led to one of several outcomes, which are ranked based on the magnitude of impact to process safety:

- Safe (rank 1),
- Air Impurity in the Tank Atmosphere (rank 2),
- Air in the Tank Atmosphere (rank 3), Shutdown (rank 4),
- Implosion (rank 5, applicable only to scenarios 1 and 2), and Reactor Overheated (rank 5, applicable only to scenario 3).

Finally, all alarms were data logged during each experiment and from that, various analyses could be completed, including the frequency of alarms, the number of different alarms, and so on. Each alarm in the plant was recorded with a feature labelled 'All' followed by a number, such as All21 for the Pressure Alarm Low (PAL01). When an alarm was triggered, its value changed from 0 to 1. The feature extraction process counted every instance where the value changed to 1 during the scenario. Also, it included any subsequent activations after previous deactivations and summed these occurrences.

Behavioural Metrics Extraction: The behavioural metrics include reaction and response time, number of procedures and mimics opened, silenced and acknowledged alarms. These metrics reflect the operators' behaviour during the task, which various underlying factors can influence. For instance, these behaviours can indicate situational awareness; an operator may open more mimics than necessary, suggesting a lack of clarity regarding the task. Similarly, acknowledging numerous alarms may imply a failure to prioritise critical alarms, while opening multiple procedures can signal poor situational awareness. Metrics similar to these, capturing operator-system interaction, have been used in previous studies, such as [124], to classify workload.

Reaction time was calculated based on the raw feature 'MNitsel1' (Manual Nitrogen Valve Switch) for scenarios 1 and 2. 'REC3WMO1' (Heat Recovery 3 Cooling Water System Switch) was used for scenario 3.

Response time measurements are based on the features 'MmanNit1', 'Nitsel1', and 'Toutrec2c1' for scenarios 1, 2, and 3, respectively. 'MmanNit1' means Manual Nitrogen Valve. 'Nitsel1' means Primary to Backup System Control. 'Toutrec2c1' means Reactor Cooling System. Figures 3.18 and 3.19 present example graphs showing the extraction of reaction time and n response time for scenario 1.

The feature 'proclis1' (Procedures) tracks the number of procedures opened. This metric was primarily recorded for those using the Electronic procedure due to the ease of recording operator-system interactions on the display for this group. Additionally, the

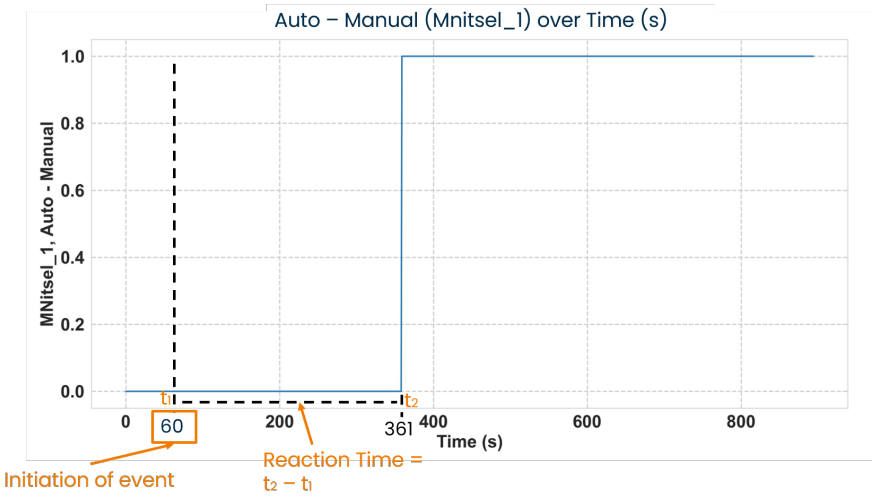


Fig. 3.18 Reaction time, Scenario 1.

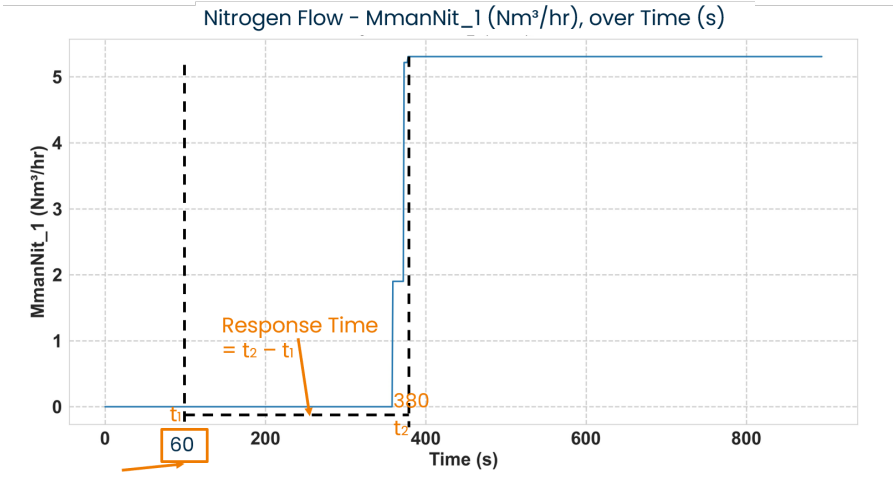


Fig. 3.19 Response Time, Scenario 1.

number of mimics or displays opened, the number of alarms silenced, and the number of alarms acknowledged were recorded for all groups. They are based on the features 'intop' (Mimics Opened), 'sAll' (Silenced Alarms), and 'AckAll' (Alarm Acknowledgments), respectively.

3.3.3 Reliability, Validity, Sensitivity and Normality

Reliability (Consistency)

The previous subsection discussed eye-tracking, questionnaire-based methods, and simulator variables. However, the reliability of these methods for this study has to be determined. Cronbach's alpha, introduced in 1951, is a reliability metric that assesses the internal consistency of items within a test [139]. It is estimated to be 1 to 10, with 0.7 or 0.75 often used as the cut-off point [139]. Cronbach's alpha is calculated using the following equation [139], [138]:

$$\begin{aligned} \alpha^C &= \frac{p}{p-1}\frac{\operatorname{Var}\left(\sum_{j=1}^{p} Y_j\right) - \sum_{j=1}^{p}\operatorname{Var}\left(Y_j\right)}{\operatorname{Var}\left(\sum_{j=1}^{p} Y_j\right)}, \\ &= \frac{p}{p-1}\frac{\sum\sum_{j\neq k}\sigma_{jk}}{\sum\sum_{j,k}\sigma_{jk}}. \end{aligned} \tag{3.9}$$

σ_{jk} is the covariance of $\left(Y_j, Y_k\right)$. Let Y_j represent a series of items, expressed as $Y_j = T_j + \varepsilon_j$ for $j = 1, \ldots, p$. Here, T_j denotes the true, unobservable item scores, and ε_j represents the associated errors. These errors are assumed to be independent of the true item scores and have a distribution with a mean of zero.

Some researchers have used Cronbach's alpha to verify consistency, like [17] in a nuclear control room study and [162] in a driving study. [17] checked for the reliability of the SART, NASA TLX and another SA questionnaire in two test situations (post-session and post-video). The reliability scores for SART ranged between 0.71 and 0.85. That of NASA TLX was (0.83 and 0.86) and SA3 (their SA method) between 0.88 and 0.95. Likewise, [162] in their research used Cronbach's alpha to determine the internal consistency of psychophysiological measures and driving performance for their scenarios studied.

Another reliability step in this thesis would be to check the data for inter and intra-participant consistency from the tools. However, the effect of inter-test repeatability for

the same subject cannot be analysed for this thesis. This is because the experiment was a between-subject study, and fewer subjects were used.

Validity

Ensuring both content and construct validity is crucial for confirming that the methods used in this thesis effectively measure what they are intended to measure. A valid test should accurately reflect the concept it is designed to assess. For this study, the relevant variables for the NASA-TLX, SART, and SPAM techniques have been fully integrated, ensuring compliance validity. However, context validity, which verifies whether the questionnaire truly measures the intended theoretical construct, has yet to be confirmed.

Factor Analysis: Factor analysis is a useful method for assessing construct validity [144]. It is a statistical technique that identifies the underlying theoretical constructs (latent factors) explaining the patterns in a data set. It evaluates how well these latent constructs represent the original observed variables [144], [143]. It can be calculated with the following equations:

According to Ford [143], factor analysis can be approached in two ways. That is, *common factor analysis* (CFA) and components analysis (CA), both of which are valid factor models. CFA is more appropriate when the observed variables are believed to be influenced by underlying latent factors. This approach models these unmeasured factors, examining how they relate to the observed variables and how they load onto the latent factors. Researchers such as [17] have utilised this approach to assess the NASA-TLX and SART loadings and their SPAM variables.

On the other hand, the component model (*Principal Component Analysis*, PCA) is more suitable for maximising the explained variance of observed variables. It summarises the data by creating components that capture the most variance without assuming underlying latent constructs, thereby reducing the dimensionality of the data. Since this thesis stage aims not to reduce data dimensionality but to assess validity, the CFA approach will be applied.

Correlations: Correlation is a measure of the strength of a linear association, but not of its nature. It can also be used to assess the reliability of measurement tools [163]. Pearson correlation is one of the most commonly used correlation methods in the safety-critical domain because of its simple and intuitive interpretation and its broad applicability across fields [162], [163]. However, other methods include Spearman's rank

correlation, Kendall's Tau, and Intra-class correlation coefficient (ICC) [164], [142]. The Pearson correlation coefficient is symbolised by r in a sample. It measures the strength of linear association between two variables in a range from -1 to +1 [164].

The Pearson correlation coefficient is symbolised by r in a sample and measures the strength of linear association between two variables [164] in a range from -1 to +1. In the entire population from which the sample is derived, it is denoted by ρ. r is calculated based on the equation 3.10 [164]:

$$r = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{\sqrt{\left[n\sum x^2 - \left(\sum x\right)^2\right]\left[n\sum y^2 - \left(\sum y\right)^2\right]}}, \tag{3.10}$$

where n is the total number of paired data points or observations in the sample., $\sum x$ is the sum of all values of the first variable x across all observations. $\sum y$ the sum of all values of the second variable y across all observations. The numerator $n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)$ calculates the covariance between the two variables, adjusted for the size of the sample. The denominator $\sqrt{\left[n\sum x^2 - \left(\sum x\right)^2\right]\left[n\sum y^2 - \left(\sum y\right)^2\right]}$ normalises this covariance by the product of the standard deviations of x and y. This ensures that the correlation coefficient r is dimensionless and falls within the range of -1 to +1.

When the data is determined to have a non-parametric distribution, or it is unclear what the distribution is, Spearman's rank correlation will be used. Other researchers have used Spearman's on such occasions [142], [163], [133]. Spearman correlation is a non-parametric test similar to Pearson correlation in its assumptions. However, it does not require the data to follow a normal distribution [141].

Combining factor analysis with correlation analysis provides a more robust validation of both the methods and the variables under consideration. This approach helps to clarify which variables should be retained for further analysis, ensuring that the final results are reliable and relevant.

Sensitivity

This study will ascertain the sensitivity of the tools/methods and their variables to the scenarios/task complexity for each study group. This further ensures their accuracy and effectiveness under different conditions. ANOVA with a Turkey's Test is used for this analysis. This paired test method has been used in related control room studies [17]

and [9]. Other methods like the already described correlation analysis[162] and linear regression have been used for such sensitivity analysis [133], [125]. [9] has used it for both continuous and categorical data, while [17] used it for categorical data.

Normality

A Shapiro-Wilk L1 Test is conducted to assess the normality of the data within each group [18]. In the case of non-normal data distribution, a Mann-Whitney U Test is used to further evaluate the statistical significance of the mean data in group scenario comparisons [18]. On the other hand, a t-test is used in the case of normal distribution [165].

Shapiro-Wilk Test Statistic The test statistic W is calculated using the formula:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{3.11}$$

where:

- $x_{(i)}$ is the i-th order statistic (sorted data),
- $\bar{x}$ is the sample mean,
- a_i are the coefficients derived from the expected values of the ordered normal quantiles.

The Mann-Whitney U statistic is calculated as follows:

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2}, \tag{3.12}$$

$$U_2 = R_2 - \frac{n_2(n_2+1)}{2}, \tag{3.13}$$

$$U = \min(U_1, U_2), \tag{3.14}$$

where:

- R_1 and R_2 are the sums of ranks for the first and second samples, respectively,
- n_1 and n_2 are the sample sizes for the first and second samples.

The t-statistic for independent samples is given by:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \tag{3.15}$$

where s_p is the pooled standard deviation, calculated as:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}, \tag{3.16}$$

where:

- $\bar{x}_1$ and $\bar{x}_2$ are the sample means,
- s_1 and s_2 are the sample standard deviations,
- n_1 and n_2 are the sample sizes.

3.3.4 Ethical Consideration

This research received approval from the internal ethics committee of the Collaborative Intelligence for Safety Critical (CISC) project. The approval was then supported by that of the Technological University of Dublin. Participants were selected based on specific criteria, including 18 years or older, fluency in English, and normal or corrected to normal vision. Upon completing data collection, participants were awarded certificates of participation. Appendix B includes a copy of the approved flyer, information sheet, consent form, training questionnaire, and NASA TLX, SART, and SPAM questionnaires. The registration form can be accessed in Jotform [1].

The consent form addressed: the study goals, right to withdraw, data type and use, confidentiality, anonymisation, and data security (see Appendix B.8). Participants were required to answer Yes or No to these questions. If participants said no to any questions, they would not have participated in the study. In cases where there were concerns about the clarity of research goals or other doubts, participants were encouraged to ask questions for better understanding.

[1] https://form.jotform.com/231382873722358

3.3.5 Data Analysis and Result Presentation Plan

Data Analysis Plan: Eye Tracker

Data analysis for eye tracking data is performed using Tobii Pro Lab software. For each participant, the analysis begins with a first look at the intervention procedure after initiating the first critical alarm within the 18-minute session. The analysis window, or time of interest (TOI), spans from 6 minutes to approximately 27 minutes and 28 seconds. This period is chosen because it is deemed crucial for intervention. Also, this specific time frame is considered sufficient due to the extensive time required to analyse the entire recording.

Area of interest (AOI) analysis is performed to extract relevant eye-tracking metrics. For this study, the paper and screen procedures are designated as essential AOIs, as illustrated in Figure 3. These AOIs are specified before or during assisted mapping for each participant. Unlike the mapping process, completed once and applied to all participants, the heatmap aggregates participants' eye movements over the TOI.

Further Tobii Pro Lab data analysis is performed using Python (version 3.9.18) to obtain descriptive statistics (mean +/- standard deviation). Group comparisons are performed using ANOVA in RapidMiner (free version 10.1.003), with a p-value of <0.05 for significance. The mean and standard deviation for pupil diameter is calculated based on the difference between each participant's eye movement data during the selected TOI and their collective median baseline values (i.e., the median eye movement value within the first 1 minute and 30 seconds of scenario 3). This subtractive baseline correction approach is applied as [149]) recommended. Other metrics are analyzed without baseline correction.

Data analysis plan Survey

The SART, NASA TLX, SPAM, demographic, training, experience, personal data, etc., are collected using questionnaires developed in Microsoft Forms. The forms allow for the questions and associated responses to be downloaded in Excel format, facilitating further preprocessing and analysis. The survey data are merged using Python and all subsequent analyses are also performed using Python in the Jupyter Notebook.

Data analysis plan Simulator

The extraction of performance and behavioural variables using the specified features is discussed in Subsection 3.3.3. As illustrated in Figure 3.17, the simulator files are initially saved to the desktop used for simulation in .mat (MATLAB) format. These files are converted to .xlsx (Excel) format before being stored on the CISC project's SharePoint cloud repository. The script for this conversion is written in MATLAB. The plant sections (e.g., alarm, tank) and their respective features are saved anonymously and separately for each participant per scenario. These files are then concatenated into a single Excel file for each participant per scenario using Python in Jupyter Notebook. The variables are extracted from the concatenated file for further analysis.

Further analysis and Presentation plan

Further data analysis will involve pre-processing the data which includes cleaning to address missing data and other abnormalities. It also involves integrating the different measures, and transforming the data through scaling and normalisation. Following this, exploratory data analysis will be conducted, where the descriptive statistics of the measures will be analysed and presented, along with an assessment of normality. The validity, reliability, and sensitivity of the variables will also be evaluated, as already described in Sections 3.3.3, 3.3.3 and 3.3.3.

A more comprehensive analysis of the results will be conducted to address the research questions related to Objective 1. This analysis will involve examining the key measures in this study. This analysis will facilitate group-wise comparisons between paper-based and electronic procedures for each scenario, addressing Research Questions (RQ) 1 and 2. The final step will be to assess the significance of the variables for overall performance, which will address Research Question (RQ) 2.

3.3.6 Limitations of the Design

It is important to acknowledge several limitations of this study. Firstly, the reliance on self-reported measures and the lack of detailed questions to assess the procedure formats are notable constraints. In this study, just one query was posed about the perceived support for the procedure.

The study involved various chemical engineering master's students, some processors, and Ph.D. students. An Actual control room operator would have been ideal. However, to compensate for this, the selected scenarios were sufficient for the participants' level of expertise.

Due to its controlled environment, the study may not fully reflect real-world safety-critical situations. The setting could affect the operator's self-reported ratings, possibly due to pressure to perform accurately within the timeframe. Suggestions for improvement include incorporating more complex scenarios and alarm conditions and conducting the study in a properly equipped facility.

The study's reliance on a single simulated formaldehyde production facility limits the generalisability of the findings to other industrial processes and control room environments. In addition, the immediate effects of interfaces and tools were evaluated without considering long-term impacts such as fatigue, burnout, or lasting effects. Consequently, the longer duration of the test in the final scenario may affect the participants' perceptions of system support, workload, and situational awareness.

3.3.7 Summary

Chapter 3 of the dissertation outlined the design methodology employed to achieve the second research objective, focusing on a simulated formaldehyde production plant. The study meticulously examined key sections of the plant, such as the tank, methanol, compressor, heat recovery, reactor, and absorber sections. These sections, which included process-focused, technological and performance components, were crucial for understanding the plant's operational dynamics and were chosen for detailed analysis to assess performance and interaction within the plant. Human-System Interface (HSI) tools were designed, including screen displays, alarm lists, and structured alarm response procedures, to simulate a realistic environment and observe participant interactions under various scenarios. The experiment was carefully structured, with participants selected based on predefined criteria.

The study used qualitative and quantitative approaches for data collection, such as NASA-TLX, SART, and SPAM, in conjunction with eye-tracking technology and simulator recordings. Key variables such as ARP level, alarms, task complexity, experience, etc., were factored in for the study and for comparison between a group using paper-based procedures (PBP) versus electronic (EP). The collected data were

analysed for reliability, validity, sensitivity, and normality. Ethical considerations were also taken into account to maintain the research's integrity. The chapter concluded by outlining the data analysis strategy and acknowledging the study's constraints.

The next chapter, Chapter 4, discussed the thesis's second methodology, essential to address objective three (3) and its research questions.

Chapter 4

Design Methodology II

4.1 Introduction

This chapter focusses on the research framework and the methods employed to achieve the second objective of this dissertation. The second objective is to develop a real-time risk-based human performance assessment and decision support framework. This is crucial in deepening the understanding of the top-down effect of the interactions between the human-in-the-loop (label 4) and the systems (label 3) on the plant (label 1) and environment (label 5) (see Figure 1.1). The first methodology in Chapter 3 focused on a bottom-up effect, while the framework in this Chapter took the opposite approach. To address the research objective, a carefully structured research framework was developed, introduced, and detailed in the initial section of this chapter (Section 4.2). The framework is then broken down into its core components, where the specific methods chosen and how they will be applied are explained (Section 4.2.2). These methods were selected for their prospects for achieving the research objective, as discussed in Chapter 2. The reason for selecting them is further justified in this chapter.

Chapter 4 then delves into the data collection techniques used for each method, including an overview of the variables required for the analysis (Section 4.3). These variables are categorised according to their use in task analysis, *Standardised Plant Analysis Risk* (SPAR-H) Human Reliability Analysis, and more advanced models such as *logistic regression n* (LR) and *SPAR-H Bayesian Networks* (BN SPARH). Next, the validity of these methods is critically assessed (Section 4.4). This includes a detailed examination of the approaches taken to ensure the robustness of the SPAR-H method,

the Logistic Regression model, and the BN SPARH. The methods used to validate these approaches and the criteria for assessing their precision and reliability are explained in detail. Chapter 4 then delves into the data collection techniques used for each method, including an overview of the variables required for the analysis (Section 4.3). These variables are categorised according to their use in task analysis, *Standardised Plant Analysis Risk* (SPAR-H) Human Reliability Analysis, and more advanced models such as *logistic regression* (LR) and *SPAR-H Bayesian Networks* (BN SPARH). Next, the validity of these methods is critically assessed (Section 4.4). This includes a detailed examination of the approaches taken to ensure the robustness of the SPAR-H method, the Logistic Regression model, and the BN SPARH. The methods used to validate these approaches and the criteria for assessing their precision and reliability are explained in detail.

Also, the chapter outlines the result representation plan, starting with task analysis and SPAR-H, followed by the application of Logistic Regression and SPARH BN (see Section 4.5. This plan ensures that the analysis is systematic and coherent, aligned with the research objectives. This lays the groundwork for presenting the results in the next chapter.

Finally, the chapter addresses the limitations of the research design in Section 4.6, considering how these may have influenced the findings. It concludes with a summary in Section 4.7, recapping the main points discussed and how they address the research objective and questions.

Research Objective

Alongside research objectives 1, as discussed in Chapters 2 and 3, this thesis seeks to create a risk-based real-time framework to assess human performance (R^2 HiTL), improving situational awareness and decision-making support within current human-in-the-loop process control and safety configurations. This provides a more robust process for streamlining alarm response procedures and optimising other human system interfaces for effective decision support in process control rooms. The following questions are asked to achieve this:

RQ3: What operator behavioural metrics are the best predictor of a successful recovery of a plant alarmed state?

RQ4: How can live and legacy data be integrated for a real-time human-in-the-loop performance assessment framework, decision support, and procedure optimisation?

The tasks below are applied to answer this research question:

- Identify and apply suitable HRA and AI techniques that align with the goals of a dynamic risk-based framework.
- Investigate the application of machine learning and data analytics to enhance the accuracy of performance assessments.

RQ5: What are the benefits of integrating operator-system data and expert knowledge for risk-based real-time decision and situation awareness support?

The methods and steps, as detailed in the developed framework (see Figure 4.1, are discussed next.

4.2 Research Framework and Methods

4.2.1 Framework

The framework presented in Figure 4.1 outlines a comprehensive methodology for a real-time risk-based human-in-the-loop performance assessment in process control room operations. It consists of three main stages: Risk Perception, Risk Understanding, and Risk Projection. The inputs and outputs of each phase are summarised in Table 4.1. This framework integrates multiple analytical techniques using live and legacy data to provide a robust approach to understanding operators' behaviour, guiding it, and mitigating human error in high-risk environments. In addition, it applies to various fields, including the nuclear, automotive, and aviation sectors.

The three key phases of the framework in Figure 4.1 as summarised below:

1. Phase 1: Perception (risk perception) In this phase, the system and human-in-the-loop behaviour are observed during real-time interaction. From a back-end perspective, the variables derived from Methodology 1 serve as critical inputs. These variables are also inputs for the subsequent phase 2 - Understand. The focus is on the operator's perception of key alarms and the relevant procedure (tasks). This phase primarily involves observation.

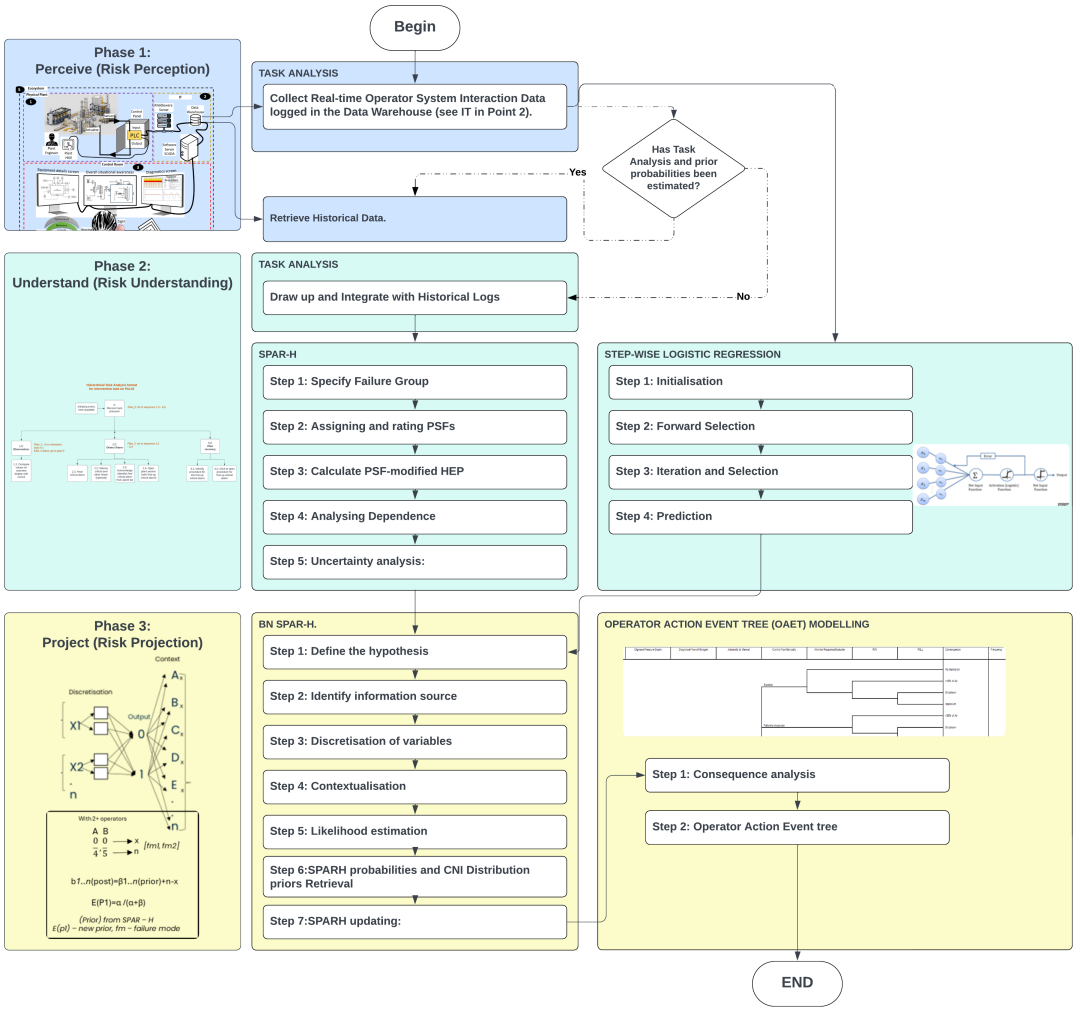


Fig. 4.1 Real-time Risk-based Human-in-the-loop (R^2HiTL) Performance Assessment and Decision Support Framework.

Table 4.1 Inputs and Outputs for each Phase of the Risk Assessment Framework.

Phase	Input	Output
Phase 1: Perceive (Risk Perception)	Real-time operator system interaction data, Historical data logs.	Perceived risk factors, recalled expected task and task analysis.
Phase 2: Understand (Risk Understanding)	Recalled Task analysis data, recalled error probability from logs.	Task analysis and Modified Human Error Probabilities (HEPs) (if not recalled in phase 1), dependence analysis, Uncertainty analysis, predicted outcome and predictors.
Phase 3: Understand (Advanced Risk Understanding)	Output from phase 2	Discretised variables, Operator context, Likelihood estimates, Posterior probabilities, Refined risk predictions, Operator Action Event Tree (OAET) outcomes.

2. Phase 2: Understand (Risk Understanding) The Understand phase involves identifying the key variables that predict the operator's likelihood of success or failure in the observed task. Factors such as the type of procedure employed, the number of alarms, and other relevant elements contributing to the operator's specific setup are considered. A stepwise logistic regression method is applied, with justification provided in Section 4.2.2. This method allows for the selection of output variables based on their significance, optimising accuracy, precision, and recall. The variables are categorised under the types of tasks/Human Failure Events (HFE) they represent: monitoring, planning, or intervention. In some instances, a variable may pertain to more than one task category. The predicted outcome, whether 'Failure' or 'Success', guides the analyst in determining the likely result, thereby facilitating the effective application of the method in Phase 3.

   The second part of Phase 2 involves understanding the pre-assessed error probabilities for the current critical task. If such probabilities already exist for that task, it is simply recalled in phase 1. The SPAR-H method is employed here for a first-time prior estimation, with justification provided in Section 4.2.2. This

second-generation method is typically utilised at the commissioning stage of configurations or during HRA reviews and updates, if applicable. Experts evaluate the tasks for a specific safety-critical scenario as derived form the procedure or job aid, incorporating the eight SPAR-H variables and their multipliers (values for scoring). The output consists of estimated error probabilities for each task, which serve as the prior input for Phase 3.

3. Phase 3: Project (Risk Projection) In this phase, outputs from the Stepwise Logistic Regression and SPAR-H methods are utilised to project error probabilities and associated risks through *Bayesian Network SPAR-H* (BN SPAR-H) and the operator action event tree (OAET) analysis. Unlike the pre-estimated probabilities with SPAR-H's fixed eight (8) performance shaping factors, the selected variables are those collected from real-time interaction and identified by Logistic Regression as predictive of error. Traditional SPAR-H methods are limited in their application for such dynamic assessments. The BN in the SPAR-H compensates for this limitation, comprising seven (7) steps, as similarly applied by [123]. The first step involves discretising variables using entropy, information gain, and optimal split points. The discretised variables are then contextualised for each operator. The realised contexts are then paired with analysts' assumed error (backed up by the Logistic Regression prediction). This pairing/likelihood estimation identifies the number of previous failure cases within the same context (x) and the new total number of opportunities for this failure in this context (n). Step 6 entails obtaining SPAR-H prior values along with the initial CNI alpha and beta parameters (utilised in a beta distribution for assessing uncertainty). The concluding step comprises calculating the posterior alpha and beta parameters as well as determining the posterior error probability.

   The posterior probabilities are then projected within the category of the task they represent. The event tree displays paths for failure and success, cumulative error probabilities, and the resultant consequences for each path.

The primary goal is to embed this real-time, risk-based human-in-the-loop (R^2-HiTL) performance assessment phase into an interface or software application for enhancing situational awareness and decision-making. Ultimately, the goal is for an integrated software at the IT level (see image in Phase 1 of Figure 4.1) to gather data on human-system interaction characteristics at time, t, during the monitoring or managing of a

process after a *critical* initiating event, such as PAL01 mentioned in this thesis within scenario 1 of the experiment (refer to Chapter 3). It will automatically update predefined variables, model them using stepwise logistic regression, and apply them to the BN SPAR-H technique with the recalled SPAR-H expert-estimated priors. In this study, data is gathered from participants divided into two groups: paper-based procedures and electronic procedures. The logistic regression (LR) model is developed using the combined datasets of both groups. However, before the modelling stage, an operator is randomly excluded from either the paper-based or electronic group. This excluded dataset will be subsequently utilised to test the real-time LR and BN SPARH methodologies in Phases 2 and 3 of Figure 4.1.

The framework provides the following benefits for the operator, management, HRA analysts and procedure experts:

- Supports the understanding of the factors or operators' behaviour leading to a likelihood of success or failure. For example, type of procedure, response time, frequency of opening same procedure, number of procedures opened within a time frame, number of mimics open, etc.

- With insights on these procedures, event or individual-driven behaviours and the weights they carry in predicting error, real-time recommendations can be provided to the operators on what to do. For example, reduce the number of mimics by X number.

- To tie back to the alarm response procedures, the framework benefits not only operators but also those who write up alarm response procedures and for management of change post events. This is so that, post events, the procedures can be reviewed to determine how they contribute to some of the factors impacting operators or necessitating a certain behaviour from the operators. For example, if following a specific procedure at a time, t, (identified from the records), leads to opening many other procedures or many mimics within a short time frame, this can be understood to perhaps contribute to poor situational awareness. Also, depending on the procedure-based variables (e.g. type of procedure, frequency of opening a specific procedure) derived and their weight in predicting error, insights can be obtained on other challenges posed by the procedure in use by an operator. Essentially, it can be checked for in what time window this behaviour was

observed, what actions were performed at that time which led to this behaviour, and the specific task step in the procedure to enable targeted re-writing of this procedure step.

4.2.2 Methods

The methods introduced in the framework are further detailed below.

Stepwise Logistic Regression (LR)

Stepwise logistic regression is a technique for developing predictive models by progressively selecting the most significant variables until certain criteria, such as model fit or statistical relevance, are satisfied. The logistic regression (model calculates the probability of a binary outcome based on one or more predictor variables as shown in Equation 4.1. It offers straightforward and interpretable outputs that highlight the effect of each predictor on the predicted outcome. Due to its simplicity and computational efficiency, Logistic Regression is often the preferred initial approach when analysing large datasets or potential predictors. By applying a stepwise logistic regression, the researcher progressively selects the most impactful variables, ensuring that the final model remains efficient and reliable. Thus, this research utilises a stepwise Logistic Regression method to incrementally choose significant variables from those gathered in Chapter 3 (Experiment Design). This method is used to forecast errors [Failure, Success], thereby validating the analyst assumptions in Phase 3 BN SPARH of the Research Framework (refer to Figure 4.1).

The logistic function used in this model is represented by:

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n)}}, \tag{4.1}$$

where:

- Y is the binary response variable (e.g., occurrence of an error),
- $X_1, X_2, \ldots, X_n$ are the predictor variables,
- $\beta_0, \beta_1, \ldots, \beta_n$ are the coefficients in the model.

This stepwise procedure includes the following steps:

- *Step 1 (Initialisation)*: Begin with either an empty model or a model containing a single predictor variable.
- *Step 2 (Forward Selection)*: Incrementally add predictors to the model. At each step, introduce the predictor that provides the greatest improvement to the model, typically assessed using criteria such as the *Akaike Information Criterion* (AIC) or *Bayesian Information Criterion* (BIC).
- *Step 3 (Iteration and Selection)*: Continue adding and removing predictors until no further significant improvements can be made according to the chosen criteria. Select the significant variables and the suggested variables from Step 2.
- *Step 4 (Prediction)*: Predict the likelihood of success or failure using the variables selected from Step 3.

Step-wise logistic regression aims to balance model simplicity with predictive performance, thereby minimising the risk of over-fitting by including only those predictors that make a meaningful contribution to the model, as determined by statistical significance (e.g., p-values) and overall model accuracy. In our analysis, this method was employed to pinpoint key variables influencing the probability of human error. By systematically adjusting the variables, we developed a streamlined model that accurately predicts the occurrence of errors based on various behavioural metrics.

The findings of the step-wise logistic regression provide valuable information on the significance of different predictors, enhancing our understanding of the factors that contribute to human errors in control room settings.

Hierarchical Task Analysis (HTA):

Task analysis is a fundamental method applied for human error assessment when using any HRA method. Defining task analysis will depend on the approach taken for the analysis. Two methods are available: 1. An action-oriented method emphasises documenting observable elements of operators' actions at different task stages (actual task execution), including the task structure. They include: *Hierarchical Task Analysis* (HTA), *Operator Action Event trees* (OAET), the *Decision/action flow diagrams* (D/A) etc. These methods are the most frequently used. 2. A second set of methods, known as the cognitive-based methods, concentrates on the mental processes that lead to errors.

Methods here include *Critical Action and Decision Evaluation Technique* (CADET), the *Influencing Modelling and Assessment Systems* (IMAS), etc. Information from the procedure or checklists and from observing work as done are used as input for the task analysis.

The HTA method differentiates itself from action-oriented and cognitive-based methods by emphasising observable behaviours while detailing the interactions between humans and control systems. Table 4.2 succinctly highlights the strengths and shortcomings of the HTA method compared to the others. The performance model incorporating human interaction, illustrated in Figure 4.1, is designed around operators interfacing with control systems or interfaces (as relevant) for process management or various tasks, necessitating an approach such as HTA. It is important to highlight that although this thesis aims to address the frequently overlooked cognitive aspects in performance evaluation, it also examines behaviour and performance. This examination is conducted by considering the interaction between the operator and the system as a whole rather than focusing solely on the operator. Also, HTA is used to allocate functions. For example, when an operator and another agent (as an AI assistant system) have to share responsibilities during a task.

HTA starts with identifying a higher-level goal and then the tasks and subtasks needed to achieve that goal. The level of task detailing depends on the analysis. For each task or subtasks are conditions for fulfilling them. Such conditions need not be defined if not needed. The HTA for scenarios 1, 2 and 3 are shown under Results in Chapter 5. This detailed HTA structure would be later on reduced to the OAET structure in the last phase of the framework to capture the success and failure probabilities of the key tasks or so-called human failure events: Monitoring, Planning and Intervention.

Standardised Plant Analysis Risk Human Reliability(SPAR-H) Method:

SPAR-H is an HRA method used to evaluate the error probabilities of specific tasks or sub-tasks generally categorised as diagnoses or action tasks. The US developed it initially in the nuclear sector to support factoring in human errors during probabilistic risk assessment and as a simplified method to achieve this. Its refinement through tests was mentioned in other domains such as aerospace [167]. Furthermore, compared to some HRA methods, it was developed based on a precise human performance information processing model rooted in nuclear sector activities [113]. Moreover, it provides an

Table 4.2 Evaluation of Different Task Analysis Methods Based on Some Key Specified Criteria. P = Partially fulfilled, N = No, and Y = Yes. [166].

Method Evaluation	**HTA**	**OAET**	**DA Charts**	**CADET**	**IMAS**
1. Focus on observable aspects of operator behaviour?	Y	Y	Y	N	N
2. Focus on the mental processes underlying behaviour?	N	N	N	Y	Y
3. Identify points where critical decisions have to be made?	Y	Y	Y	Y	Y
4. Identify important information on the control panel?	Y	N	Y	Y	Y
5. Describe the temporal characteristics of the task?	P	N	P	N	N
6. Identify interactions between task-steps and possible side effects?	Y	N	Y	Y	Y
7. Describe interactions between people and control systems?	Y	P	N	P	P
8. Describe communication requirements between teams?	Y	N	N	N	N
9. Classify tasks into different categories?	N	N	N	Y	N
10. Provide a qualitative description of the technical system?	N	N	Y	N	Y

improved foundation from the technique for the human error rate prediction (THERP) method to establish task dependencies. One good thing about this method is that it can be used for different purposes (build models, analyse tasks, and so on) and provides detailed guidelines for applying every step. The key steps in the SPAR-H method include:

- *Step 1: Failure groups*

  Following the task analysis, the task or so-called human failure events (HFE) and their sub-tasks are then assigned a specific failure type: diagnosis or action based on the task - diagnosis task or action tasks. Diagnosis tasks have a base/nominal probability value of 1.0 E-2, while the action task has a nominal probability of 1.0 E-3. The SPARH assigns a higher nominal probability to the diagnoses tasks. These nominal values were calibrated against other HRA methods before its adoption. The HFEs are the monitoring (m), planning (p) and intervention (i) tasks.

- *Step 2: Assigning and rating PSFs*

  Before discussing how scores are assigned to *Performance Shaping Factors* (PSFs), it is important to highlight these PSFs and what they are. The performance shaping factors are factors said to influence performance in a given context. The SPAR-H developers have adopted eight PSFs that are said to impact human performance and are included in the HRA analysis. These PSFs are: Stress and stressors, available time, experience and training, complexity, ergonomics /HMI, work processes, support procedures and fitness for duty [113]. Table 4.3 provides a brief definition of each PSF. The SPAR-H method provides unique worksheets for both diagnosis and action tasks (see Appendix B.10 and B.9) where information is found on each PSF, PSF levels, and the multipliers that can be used to score them by the expert.

  A PSF can have between 4 - 6 PSF levels, and these levels or their multipliers can vary depending on if it is for analysing a diagnosis or action task. For Example, For the action task, Experience/training has 4 PSF levels: "Low" with a multiplier of 3, "Nominal" - 1, "High" - 0.5 and "Insufficient Information" - 1. However, stress/stressors have the same 4 PSF levels for the diagnosis task but a different multiplier for the low level. The "Low" level has a multiplier of 10, "Nominal" - 1, "High" - 0.5 and "Insufficient Information" - 1. The analyst has to choose one of

Table 4.3 Performance Shaping Factors (PSFs) and their Descriptions.

PSF	**Description**
Stress	Level of undesirable conditions impeding task completion, including mental, physical, or environmental factors.
Available Time	The amount of time available to diagnose and act on an abnormal event.
Training/Experience	Refers to the experience and training of operators in the specific task and scenario.
Complexity	Difficulty of performing the task considering the task and environment context.
Procedures	Existence and adequacy of formal operating procedures for tasks.
Ergonomics/HMI	Quality of equipment design, layout, and human-machine interaction for task execution.
Fitness for Duty	Physical and mental readiness of an individual to perform a task.
Work Processes	Aspects related to planning, communication, and execution of work affecting performance.

the levels for the PSF. For this step of analysis, it is recommended that a minimum of 5 experts or HRA analysts participate [121].

- *Step 3: Calculate PSF-modified HEP*

  At this stage, the final HEP can be calculated for the diagnosis and action tasks using Equation 4.2 and 4.3, respectively. This equation is simply the multiplication of the nominal human error probability (NHEP) with the PSF multiplier [113]. For *diagnosis* HEP, the equation is:

$$(\text{HEP}) = (\text{NHEP})1.0 \times 10^{-2} \times _ \times _ \times _ \times _ \times _ \times _ \times _ \times _ \times _. \quad (4.2)$$

  The gaps are to be filled by the chosen multipliers for each PSF. That of *action* task is given as:

$$(\text{HEP}) = (\text{NHEP})1.0 \times 10^{-3} \times _ \times _ \times _ \times _ \times _ \times _ \times _ \times _ \times _. \quad (4.3)$$

When the negative PSF influence is present, that is, when 3 or more PSFs for a single task or sub-task have values more than 1, then a correction equation is applied for the HEP estimation as shown in Equation 4.4 below:

$$\text{HEP} = \frac{\text{NHEP} \cdot \text{PSF}_{\text{composite}}}{\text{NHEP} \cdot \left(\text{PSF}_{\text{composite}} - 1\right) + 1}. \tag{4.4}$$

Equation 4.4 applies for both diagnosis and action tasks.

- *Step 4: Analysing Dependence*

  Dependence refers to a situation where one event or sub-tasks, success or failure, affect the probability of a second event. The *Technique for Human Error Rate Prediction* (THERP) method provides the basis adopted by SPAR-H for dependence analysis. THERP evaluates dependence on a subtask level while SPAR-H on an HFE level [168].

  To calculate a task/HFE's HEP without formal dependence $P_{\text{w/od}}$, that is, in situations where both action and diagnosis exist for one task, the HEPs calculated for each are added (see Equation 4.5.

$$P_{\text{w/od}} = \text{Diagnosis HEP} + \text{Action HEP}. \tag{4.5}$$

  However, Figure 4.2 dependency model is used for formal dependencies (explicit dependency) calculation as given by the SPAR-H method and as similarly applied by [169]. $P_{\text{w/od}}$ becomes the estimated or adjusted HEP from Step 3. For this research, formal dependencies are calculated amongst the subtasks as it is assumed that failure or success in the former has an effect on the latter. Dependence is verified in that the operators' diagnosis and action tasks follow each other as per training given to the participants.

$$\text{HEP}_{m_0} = 1 - \prod_{i=1}^{n} \left(1 - \text{HEP}_{\text{Subtask}_i}\right). \tag{4.6}$$

- *Step 5: Uncertainty analysis:* Unlike earlier versions of the SPAR-H method and other HRA methods, error factors or the log-normal distribution assumption do not apply to the current SPAR-H method [113]. SPAR-H uses a beta distribution

Dependency Condition Table

Condition Number	Crew (same or different)	Time (close in time or not close in time)	Location (same or different)	Cues (additional or no additional)	Dependency	Number of Human Action Failures Rule ☐ - Not Applicable. Why?_____
1	s	c	s	na	complete	When considering recovery in a series e.g., 2nd, 3rd, or 4th checker If this error is the **3rd error in the sequence**, then the dependency is at least **moderate**. If this error is the **4th error in the sequence**, then the dependency is at least **high.**
2				a	complete	
3			d	na	high	
4				a	high	
5		nc	s	na	high	
6				a	moderate	
7			d	na	moderate	
8				a	low	
9	d	c	s	na	moderate	
10				a	moderate	
11			d	na	moderate	
12				a	moderate	
13		nc	s	na	low	
14				a	low	
15			d	na	low	
16				a	low	
17					zero ✔	

Using $P_{w/od}$ = Probability of Task Failure Without Formal Dependence (calculated in Part III):

For Complete Dependence the probability of failure is 1.
For High Dependence the probability of failure is $(1+ P_{w/od})/2$
For Moderate Dependence the probability of failure is $(1+6 \times P_{w/od})/7$
For Low Dependence the probability of failure is $(1+19 \times P_{w/od})/20$
For Zero Dependence the probability of failure is $P_{w/od}$ ✔

Calculate $P_{w/d}$ using the appropriate values:

$P_{w/d}$ = (1 + (_____ * _____))/ _____ = **N/A**

Reviewer: aw

Fig. 4.2 Dependency model for subtask dependency analysis in SPAR-H.

capable of approximating normal and lognormal distributions, with probabilities extending from 0 to 1. The Prior distribution used is based on Atwood [170]. Atwood's constrained non-informative prior provides a minimum or no specific information about the parameter being estimated while also satisfying the boundary constraint of 0 to 1. This type of prior is useful when the goal is to ensure that the prior distribution is as uninformative as possible to avoid biasing the results as would be needed later when including priors from this stage to the BN SPARH. It is also helpful because it reflects the constraints of the boundary from 0 to 1.

After successfully calculating the dependencies, the new subtask HEPs are multiplied using equation 4.6 to get the failure probabilities of each HFE. Here, $HEP_{Subtask_i}$ represents the probability of failure of a subtask in an HFE / Key task. The key tasks in this study are Monitoring, Planning, and Intervention. Their prior distribution on m is represented as m_0, p as p_0 and i as i_0. Using m as an example, the formula for estimating m_0, after estimating the dependency of the subtasks, is Equation 4.7 and is interpreted as the mean value of m_0, represented as $E(m_0)$ for a context given new data. In this case, a context represents the unique combination of PSFs or discretised PSFs. The beta distribution is then applied using the equation 4.7:

$$E(\text{Beta}(\alpha,\beta)) = \frac{\alpha}{\alpha+\beta} = E(m_0). \tag{4.7}$$

where α is the number of failures in the prior distribution and $(\alpha+\beta)$ the number of demands over which these failures occurred.

$E(m_0)$ is already estimated using SPAR-H. However, the values of α and β are unknown. That of α can be found in the table by Atwood (1886) [170] and using the estimated $E(m_0)$. β is then estimated using the Equation 4.7.

Over the years, the SPAR-H method has been used by the US Nuclear Regulatory Commission and researchers in this sector to support specific plant PRA model development, event and task analysis [113], [106]. Its benefits have had more far-reaching prospects than the downsides as it has over the years been adopted by other sectors like the process industry [169], construction [147], Maritime [171], health [172], aviation [136] etc. The merits and demerits of the SPAR-H method are summarised in Table 4.4.

In the process industry, [169] adapted the method for a pigging maintenance operation in a gaps facility and has likewise calculated for dependencies using Figure 4.2. In their work, they further compared the traditional SPAR-H method with one of the BN SPAR-H approaches, which is different from that to be applied later on in this research. The Bayesian Network (BN) is one major technique now commonly used at the PSF analysis stage in SPAR-H as seen in works by [106], [136]. On the other hand, [171], while assessing error probability during emergency response drills on ships, implemented fuzzy logic for the PSF rating stage due to the constraint of using multipliers derived from domain-specific data. Likewise [147] adopted the DEMATEL-ANP multi-criteria decision method for the PSF stage to introduce new PSFs derived from text mining accident reports.

Table 4.4 Summary of Advantages and Disadvantages of the SPAR-H Method [113].

SPAR-H Method: Advantages and Disadvantages	
Advantages	**Disadvantages**
• Suitable for pre-initiator events. • Based on a clear human performance model. • Considers positive PSF effects, lowering error rates. • Accounts for dependencies in HEP calculation. • Provides flexibility in application.	• Lacks clarity on positive dependencies in HEP. • Risk of double counting due to PSF overlap. • PSF evaluations can exceed a value of 1 in extreme conditions. • Does not distinguish between active and latent failures. • Assumes same PSFs and base rates for errors of omission and commission.

This SPARH method has thus been selected for its versatile application in the safety-critical domain and its merits, especially its flexibility in integrating other approaches. The method defines the fundamental stage for the estimation of priors and distribution for use in the next stage of the research framework.

The BN SPARH

SPAR-H experts have assumed that SPAR-H PSFs account for most of the performance influences observed in events and are the PSFs most applicable to support model development [113]. However, researchers' current integration of new techniques for PSF assessment highlights that the dynamic nature of human-machine interactions in certain applications necessitates the precise use of variables beyond those contained in SPAR-H. In process control rooms, for example, the literature review in Section 2.5.4 has shown all indications that behavioural variables during *Human-System Interaction* (HSI), workload, situational awareness, etc, are critical factors in current human-in-the-loop configurations. However, this is not explicitly captured in the SPAR-H PSFs. In addition, the available predefined multipliers in SPAR-H, for which researchers have found alternatives, pose challenges when developing a model for a dynamic control room setting.

As mentioned, researchers have integrated other techniques with SPAR-H either to estimate PSFs and mitigate the limitations of using the SPAR multipliers [169], [136], or they have applied it to refine priors coming from the SPAR using new sparse data from field [123]. The most widely used technique, alongside SPAR-H, among others, has been the Bayesian Network (BN) [123] [106], [136]. BN SPARH addresses the gap in using multipliers defined in a particular industry and was assumed to apply/generalise in all cases. Bayesian Inference is used within a Bayesian Network for such tasks. Realistically, a technique suitable for current human-in-the-loop configurations in process control rooms should:

- Allow continuous data integration from real-time HSI.
- It should allow for integrating new PSFs, which is unfortunately still missing in most HRA techniques.

Groth et al. (2014) [123] has applied BN SPARH to refine error probabilities using performance data on the 8 SPAR-H PSFs. Their approach holds potential for adaptation in this research in that:

- It facilitates the use of limited data from experimental studies to enhance error probability estimates derived from experts. This approach provides a comprehensive and reliable estimation of both sets of probabilities.

- This method offers a way to incorporate new PSFs or various combinations of HRA data and methods combinations.

- Like many HRA methods, this method allows for assigning HEPs based on the context of PSF (performance, cognitive, etc.) outcomes.

However, the simulator data used in their application were classified based on operators' ratings on only the 8 SPARH PSFs. Though their implementation of the method does not use other possible PSFs, it forms the foundation for its adaptation in this thesis. The steps applied for the posterior probability estimation with the BN SPARH method by [123] and adapted for this thesis is described below:

1. Bayesian inference (BI)

   Bayesian inference (BI) is a method used to assess and rank hypotheses by taking into account both prior knowledge and the evidence at hand [123]. This technique relies on Bayes' Theorem, which combines existing beliefs about a hypothesis (D) with new data (A), allowing for an updated understanding of the hypothesis. Through Bayes' theorem, the updated probability $P(D \mid A)$ is calculated, which shows how probable the hypothesis is, given the new evidence. As more data becomes available, this probability can be revised, with the updated probability from one step serving as the starting point for the next. This process enables a structured way to integrate various pieces of information and arrive at informed conclusions. The initial belief about the hypothesis is expressed as the prior, which is then refined in light of new evidence.

2. BN SPARH Steps: The BN SPARH steps described below can be applied to a single operator or a crew. The goal is that at a time, t when an operator or a crew is monitoring or controlling a process following an initiating event (predicted or not), software at the IT level (see Figure 4.1) can collect data on the human system interaction features, generate the variables automatically, model them using the Logistic Regression and apply the variables. The steps are described below:

   - *Step 1: Define the hypothesis:* the aim of the BI is to convey a belief regarding the hypothesis that 'the HEP for the HFEs (Monitoring, Planning and Intervention), equals m, p and i, respectively.

- *Step 2: Identify information source:* For this thesis, data from the experiment in Chapter 3 is used together with expert ratings (estimated priors, alpha and beta) from the SPAR-H method.

- *Step 3: Discretisation of variables:* The variables are discretised following steps a to d below.

  a. Calculate Entropy: Entropy is a measure of the uncertainty or impurity in a dataset. The formula for entropy $H(X)$ is:

  $$H(X) = -\sum p(x)\log(p(x)), \tag{4.8}$$

  where $p(x)$ is the probability of each class or outcome within the dataset. The summation is over all possible classes.

  b. Determine Information Gain Information Gain (IG) measures the reduction in entropy achieved by splitting the dataset on an attribute. The formula for Information Gain $IG(S,A)$ when splitting on attribute A is:

  $$IG(S,A) = H(S) - \sum \frac{|S_v|}{|S|} H(S_v), \tag{4.9}$$

  where:

  - $H(S)$ is the entropy of the original dataset S,
  - S_v represents the subsets of S after splitting on attribute A,
  - $\frac{|S_v|}{|S|}$ is the proportion of the subset relative to the entire dataset.

  c. Select Optimal Split Point: The optimal split point is the value that maximises the Information Gain. There isn't a single equation for this step; instead, you iterate over all possible split points, calculate the Information Gain for each, and select the split point that gives the highest Information Gain.

  The optimal split point is chosen where $IG(S,A)$ is maximised.

  These equations guide the process of discretising a continuous variable into two classes by evaluating possible split points and selecting the one that provides the most informative division of the data.

d. Discretise Continuous Variables: Assume you have two continuous variables A and B, which are discretised into two groups ("Low" and "High") based on the entropy-based binning algorithm.

Let A' and B' be the discretise versions of A and B respectively, where:

$$A' = \begin{cases} \text{"Low"} & \text{if } A \leq \text{Split}_A \\ \text{"High"} & \text{if } A > \text{Split}_A \end{cases} \tag{4.10}$$

$$B' = \begin{cases} \text{"Low"} & \text{if } B \leq \text{Split}_B \\ \text{"High"} & \text{if } B > \text{Split}_B \end{cases} \tag{4.11}$$

- *Step 4: Contextualisation*: The SPAR-H technique essentially has 19,440 contexts [123]. That is, 19,440 unique combinations of PSF as in:

$$(P(\text{Error} \mid \text{PSF}_1 \cap \text{PSF} \cap \cdots \cap \text{PSF}_8)).$$

In this thesis, the same contextualisation approach is employed on the selected variables by LR under each HFE/Task using $P(\text{Error}_{\text{Cm0}} \mid \text{A'} \cap \text{B}' \cdots \cap \text{n})$. Where Cm0 represent a context on the monitoring HFE, and A' and B' discretised from A and B variables.

Given that each data point has a context label C and an associated error, e, as assumed from the LR prediction, the next step is to group the data by these two variables. The goal is to count how many times each combination of C and e occurs. Note that a context$C$might appear twice because of the error [0 or 1]. Hence, a group of operators might fall under context C, but some fail [1] and others succeed [0].

- *Step 5: Likelihood estimation:* This refers to the likelihood of observing failure on demand. Two parameters are of the essence here: the number of human failures, x and the number of independent opportunities for human failure, n. For instance, if a new operator in the monitoring task estimation is categorised under context C with e = 1, this context information contributes to the existing data in the database from other operators in STEP 3. Assuming e was 1 from the operators' predicted likelihood of failure and under context

C only 3 out of 10 operators had previously failed, the updated *x* becomes 4, while *n*, initially counted as 10 in context *C*, becomes 11 with the addition of the new operator.

- *Step 6: SPAR-H probabilities and CNI Distribution priors Retrieval:* The prior probabilities from the SPARH encodes details on the hypothesis earlier defined. Also, the calculation of the CNI alpha and beta priors needed in this method have been earlier on discussed.
- Step 7: SPARH updating: The Bayesian updating is done using the new data: x, n as in Step 5, and the CNI priors: α_prior and β_prior for each HFE. Alpha_posterior and beta_posterior are then calculated using Equations 4.12 and 4.13, respectively. Finally, the posterior probability E for *m*, *p* and *i*, that is E_m_1, E_p_1, and E_i_1 are calculated using Equation 4.14.

$$\alpha_{\text{post}} = \alpha_{\text{prior}} + x_{\text{posterior}}, \tag{4.12}$$

$$\beta_{\text{post}} = \beta_{\text{prior}} + n_{\text{posterior}} - x_{\text{posterior}}, \tag{4.13}$$

$$E(m_1) = \frac{\alpha_{\text{posterior}}}{\alpha_{\text{posterior}} + \beta_{\text{posterior}}}. \tag{4.14}$$

4.3 Data Collection and Variables

4.3.1 HTA Task analysis and SPARH

Information on the tasks and sub-tasks for the hierarchical task analysis (HTA) is obtained from the alarm response procedures. The HTA chart is developed by the researcher using Lucid Chart software (footnote). For the SPAR-H, the PSFs and associated multipliers were obtained from the SPAR-H guideline in Appendix B.10 and B.9. One human factors and automation expert performed a blind rating of the PSFs/variables using an Excel template developed by the researcher. According to [167], in HRA, analysts can be HRA specialists with many years of experience or HRA analysts with one year of experience, and finally, they can be analysts of probabilistic safety assessment (PRA) competent in engineering analysis following the stipulated instructions.

The PSF/variables as defined in Section 4.2.2 are: Available Time, training/experience, HMI(display/alarm system), Work processes, Stress/stressors, procedures, Task Complexity, and Fitness for duty, and are briefly explained below.

4.3.2 LR and BN SPARH

Data for the stepwise logistic regression are obtained from the collected data in the experiment in Chapter 3. However, since the aim of the framework in this Chapter is to build a real-time model, the variables to be considered from the collected data are those that are readily available and would not likely change in the short to medium term. It also includes those that can be collected or assessed during actual human-system interaction (HSI). These include behavioural variables derived from simulator logs as described in Section 3.3.2 of Chapter 3 and readily available data on training, experience, subjective situational awareness, and operator workload.

BN SPARH takes as input priors calculated using SPARH, the predictive variables after LR implementation and their respective data, and the predicted outcome, y. These are subsequently used to calculate the likelihoods x and n and the posterior probabilities for the monitoring, planning and intervention tasks.

4.4 Validity and Reliability

4.4.1 SPAR-H

Face validity is the norm in HRA methods [167]. Validity can be checked for both error identification and HEP estimation stages. However, similar to THERP, CREAM, and some other methods, SPARH does not include specific error identification guidelines. However, its HEP estimation has been independently validated by the method developers by comparing HEP results to base rates [167]. Reproducibility, on the other hand, refers to the extent to which different individuals achieve consistent results when performing the same task [167]. According to [167], SPAR-H is known to have very high reproducibility and low sensitivity (variations in numerical results due to alterations in input variables). This aspect of low sensitivity is important considering the variable updates to be implemented using BN SPARH.

4.4.2 Logistic Regression

Validation for machine learning algorithms, such as the stepwise logistic regression used in this framework, can be performed in various ways. One approach involves splitting the data into training (usually 70 - 80%) and test sets (typically 20 - 30%) [124]. This method enables the calculation of the Brier score to assess the model's accuracy by comparing the predicted outcomes (E_p) with the actual results (E) using Equation 4.15 [141]. This is a common model validation approach in safety-critical domains [124], [151], [141].

Additionally, metrics such as precision, recall and the f1 score have been employed for evaluating model performance in the health sector [141]. The study by Rahmatinejad et al. (2024) implemented cross-validation by partitioning the dataset into k folds (in this study, 5 folds). The model was trained on K - 1 of these folds, with the remaining fold used for validation. Subsequently, the precision, recall, and F1 score of the model were calculated [141]. In this thesis, the calculation of Brier Score/accuracy is employed for validating logistic regression. Furthermore, precision, recall, and F1 scores are additionally utilised and computed using equations 4.16, 4.17, and 4.18, respectively, serving as supplementary validation methods.

$$\text{Brier Score} = \frac{1}{n}\sum_{i=1}^{n}(E_i - E_p)^2, \tag{4.15}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}, \tag{4.16}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}}, \tag{4.17}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{4.18}$$

where

- TP: True Positives is the number of cases where the model correctly predicts the positive class.

- FP: False Positives is the number of cases where the model incorrectly predicts the positive class.

- TN: True Negative is the number of cases where the model correctly predicts the negative class.
- FN: False Negative is the number of cases where the model incorrectly predicts the negative class.

4.4.3 BN SPARH

In some sense, the logistic regression-based BN SPARH (with its objective HEPs), as applied in this research, provides internal validation for the SPAR-H method. It provides a technical basis for how well it measures all the factors it asserts to measure. However, evaluating such newly developed methods, as in the case of this thesis, would require comparison with the original expert evaluation for validation. This type of comparison has been done by other researchers to validate their evaluations, paying attention to the closeness of posteriors to priors evaluated by the experts [123] [169]. Furthermore, for reproducibility, two (2) approaches are possible:

1. A paired t-test using Equation 4.20 is applied to compare the HEP of the BN SPARH and SPARH if the HEPs of the tasks have a normal distribution while in a case of non-parametric distribution, a Wilcoxon Signed-Rank Test is applied using the following steps:
    - For each pair of observations (X_i, Y_i), compute the difference using Equation 4.19.

$$d_i = X_i, Y_i, \tag{4.19}$$

    - Calculate the absolute differences $|d_i|$ for each pair.
    - Rank the absolute differences $|d_i|$ from smallest to largest. Assign average ranks in case of ties.
    - Assign the original sign of the difference $|d_i|$ (positive or negative) to each rank
    - Sum the ranks corresponding to the positive differences W^+ and the ranks corresponding to the negative differences W^-.
    - The test statistic W is the smaller of W^+ or W^- in Eqaution 4.21.

Later, the p_values from the tests are used to check for significant differences. Ideally, there should be no significant difference between the two.

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}, \tag{4.20}$$

where: $\bar{d}$ is the mean of the differences between paired observations. s_d is the standard deviation of the differences and n, the number of pairs.

$$W = \min(W^+, W^-), \tag{4.21}$$

where: W^+ is the sum of ranks for positive differences and W^- is the sum of ranks for negative differences.

2. An alternative approach could involve excluding the data of a single participant before LR prediction to compare the BN SPARH result when this new participant data is fed into the framework against the actual outcome for that participant. This would be a form of gold-level validation step. An ideal scenario is testing the validity of the entire framework via a new set of studies.

4.5 Data Representation plan

The first step would be the SPAR-H analysis which starts with task analysis of the most complex scenario from the experiment. This analysis is followed by the presentation of the PSF assessment by the expert and all subsequent steps of the SPARH analysis process as discussed in Section 4.2.2 of this Chapter. Then, the LR results are presented, showing the selected variables and predicted outcomes. The model validation result is also shown. The selected variables to be used subsequently in BN SPARH are also specified, along with a recap of the reason for this choice. Finally, results on each step of the BN SPARH framework and the validation and reproducibility outcomes are presented.

4.6 Limitations of the Design

The constraint described in Chapter 3 regarding the experimental design, specifically the limited number of participants involved in this study, reduces the model's robustness

within this framework. Similarly, the experimental setup, though not an authentic control room, mimics one to a certain extent. This choice was made to align the complexity of scenarios with the participants' training and experience levels. It is a future plan to set up a similar study in an industrial setting and tailor the scenarios to the skills of actual operators while recruiting as many as possible. This number will, regardless, add to the data provided by this thesis and for use in the HRA and process safety community.

It should be highlighted that the framework has not yet been deployed as a support tool or platform and, therefore, has not been tested for how it works in real-time safety-critical human-in-the-loop settings. This would be the next step for future research. Also, another limitation is that the framework involves a lot of steps. In addition, some of these steps would require the recall of historical data that have been collected from experts, such as the prior probabilities etc. It would also require individuals with expertise at the beginning, but in the long run, it would not be necessary when the steps have been automated and the data or outcomes communicated in a user-friendly way on not just the error probabilities but the variables that account for this and recommendations to be followed by the operator.

Furthermore, an assumption in the framework is that the scenarios can be used as the PSF task complexity and Group as the procedure level. That is, there is a predefined label for the complexity level of the critical alarm procedure being followed. Also, there might be different procedure levels in use at the facility. The model is not limited by the simultaneous use of two procedure levels, as this can also be given a certain label. Another assumption is that the operator responds to the following procedure expected with the first critical alarm attended to first since the priors estimated in SPARH are procedure-based.

Finally, the variables factored in at this stage of the framework development, particularly the behavioural variables, are not exhaustive as they do not account for all the data variability. Several other factors in operator-system behaviour might potentially explain these gaps. Exploring other potential behavioural variables that can be obtained in real-time and how they correlate with other variables or describe operators' behaviour is a plan for future work.

4.7 Summary

In summary,

The stepwise LR also provides the backbone to analyse how the integration of variables from multiple data sources improves prediction accuracy. Also, it helps to determine the most influential data sources and their contributions to performance predictions. Therefore, it answers the research question:

- RQ3: What operator behavioural metrics are the best predictor of a successful recovery of a plant alarmed state?

Applying both SPARH and stepwise logistic regression (LR) in the framework helps identify and prioritise key variables for assessing human performance. Also, integrating the LR with the BN SPARH technique and steps, as described in this thesis, provides the platform to explore the potential of combining machine learning and HRA techniques for a real-time risk-based performance assessment. The statistic tests and data mining techniques applied provide for good integration of data science and human factors engineering. Therefore, the last research question in this thesis, as shown below, can be answered.

- RQ4: How can live and legacy data be integrated for a real-time human-in-the-loop performance assessment framework, decision support, and procedure optimisation?
- RQ5: What are the benefits of integrating operator-system data and expert knowledge for risk-based real-time decision and situation awareness support?

Chapter 5

Results - Methodology 1

5.1 Introduction

In Chapter 5, the results from Methodology 1 are presented, with a particular focus on the measurement and classification of both subjective and objective aspects of the research. Subjective relates to operators' behaviour influenced and triggered by personal feelings, opinions, interpretations, training or biases. On the other hand, an objective is a measurement of behaviour that is a measurable fact or an unbiased observation. When a property is measured directly, based on a fundamental physical principle (e.g. eye movement). These are both important characteristics to measure an operator's behaviour and performance. The chapter begins by introducing the data collected in Section 5.2, providing an overview of the dataset to set the context for the ensuing statistical evaluations. Descriptive statistics are then discussed in the same Section, covering subjective measures (such as self-reported responses) and objective measures (such as behaviour, performance and eye-tracking outcomes). Key differences in performance, behavioural, and eye-tracking measures between the study groups are identified and presented.

The chapter then moves to an evaluation of the measures in Section 5.3, starting with presenting the results from the assessment of the reliability of subjective measures to ensure the consistency of the data. This is followed by the results of the validity testing, where factor analysis is presented in two stages: first, without eye-tracking data, and then with the inclusion of these eye-tracking measures, using scenario 3 only, for a more comprehensive understanding. Scenario 3 was chosen due to its complexity and the

cumulative impact garnered from the initial two scenarios. This level of effect is likely to yield more notable differences between the two study groups. The correlation analysis outcomes are shown as well to explore the relationships between various measures, offering insights into how subjective and objective data align and interact.

The outcome of the sensitivity analysis is presented in Section 5.3, showing how responsive the different measures are to variations in task complexity. The aim is to identify which measures are most sensitive to changes in the experimental scenarios. The chapter concludes with a summary of the key findings in Section 5.4, focusing on the observed correlations, factor loadings and underlying constructs. A summary of the group-wise comparison is presented, exploring differences between participant groups. This chapter establishes a solid foundation for the in-depth discussion presented later in Chapter 7, concentrating on how it achieves objective 2 and addresses its research questions.

5.2 Descriptive Statistics and Initial Analysis

5.2.1 Overview of Collected Data

The results presented in this chapter are based on subjective and objective data collected during the human-in-the-loop experimental study on alarm handling and process control as discussed in chapter 3.3.1, Figure 3.1. The data collected were from two study groups. Study Group PBP managed regime handling using the paper-based format of alarm support procedures, whilst Study Group EP managed an identical alarm handling regime using an electronic on-screen-based format. To recap, both paper and on-screen support procedures were presented in a similar font, format and order. The paper-based support consisted of a stack of sequentially organised single-sided printed pages featuring a table of contents at the start and coloured tabs to demarcate different sections. On the other hand, on-screen support allowed users to access information through a layered navigation menu on the left pane, facilitating easy movement between sections and utilising specific tags to locate the relevant procedures. The procedures for both groups were organised as a numbered list.

Regardless of group, each participant completed three scenarios with increasing levels of complexity, from scenario 1 to 3, without randomisation. The participants were randomly selected by time of day: morning, afternoon, and evening. Table 5.1

shows a recap of the gathered data collection type. Subjective data was collected after the scenario, with the exception of training, experience, and demographics, which were obtained prior to the test. They include:

- Workload was subjectively assessed using the NASA Task Load Index (NASA TLX) questionnaire. These questions were based on the six NASA TLX dimensions: Mental demand, temporal demand, physical demand, performance, effort, and frustration. These dimensions were averaged to derive the NASA-TLX index. The performance variable here can be compared later with final performance objective measures.

- Situational awareness (SA) assessment was subjectively measured using SPAM and SART questionnaires. The SPAM questions cover situational awareness in three dimensions: perception, understanding, and projection. These dimensions are averaged to derive the SPAM index. The SART measure likewise had three dimensions, like in SPAM, but these dimensions are calculated based on SART's sub-dimensions. The SART dimensions are SA Supply, SA Demand and SA Understanding. SA supply is calculated by adding the subdimensions Arousal, concentration, division of attention, and reduced capacity. SA Demand comprises the subdimensions Instability, Complexity, and Variability. Finally, SA Understanding comprises information quantity, information quality, and familiarity. The SART and SPAM techniques were discussed in subsection 3.3.1.

- Questions were asked about the operators' familiarity with the control room and chemical plant operations. After an orientating trial test, candidates were asked questions to evaluate their training. Their familiarity and training scores were averaged to derive the familiarity and training variables, respectively. This familiarity variable can be compared with the familiarity variable of the SART Understanding dimension.

- Data was collected to evaluate the participants' perceived performance on the support of the alarm system (alarm list, etc.) and the alarm response procedure. In addition, they were asked to rate the complexity of the task after each scenario.

The objective data were collected during each scenario, in contrast to the subjective data. They include:

Table 5.1 Subjective Measures and Objective Measures.

Subjective Measures	**Objective Measures**
NASA TLX	Eye-tracking
SART	Simulator log: Performance Measures
SPAM	Simulator log: Behavioural Measures
Demographics, Training and Experience	
HMI assessment and task load	

- Eye-tracking data provides an objective measure of attention as a base for situational awareness and pupillometry measures, that is, the pupil diameter, for objective workload assessment. The attention-based measures were fixation duration (ms), number of visits or revisits, saccade amplitude (°), and saccade duration (ms).

- Performance measures collected include overall performance classified as optimal, good or poor. In addition, data on the resulting outcome better called Error, was recorded as either "failed" or "success" depending on the goal of each scenario. This classification is common in the domain of process safety and for reliability assessment. Finally, the resulting consequence of the operator's task outcome, such as an overheating of the reactor or a plant shutdown, and so on, was analysed. Other performance measures include the accuracy of the critical control action (calculated as the mean square error of expected vs actual value) and the recovery time. Other performance measures include the accuracy of the critical control action (calculated as the mean square error of expected vs actual value) and the recovery time.

- The behavioural metrics collected consist of: reaction time (in seconds), response time (in seconds), count of alarms triggered, count of alarms acknowledged, and count of alarms silenced. Additionally, the number of process-related alarms was incorporated.

5.2.2 Descriptive Statistics

Below are the descriptive statistics for the subjective and objective measures applicable to both study groups for each experiment scenario (S):

- (Group with Paper-based procedures): This group has an average participant age of around 23 years, comprising 10 females and 13 males.

- (Group with Electronic Procedures): The participants in this group have an average age of about 25 years, with 10 females and 13 males.

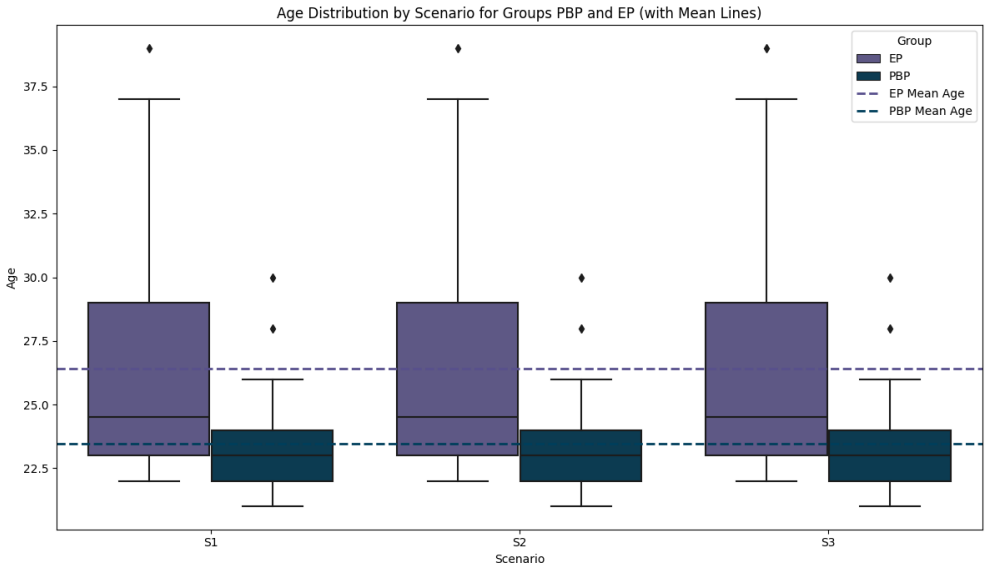


Fig. 5.1 Age distribution by scenario for each group.

Subjective Measure Descriptive Data

NASA TLX Variables (Subjective Measures):

Table 5.2, 5.3, 5.4, 5.5, and 5.6 display the standard deviation, mean, and 95th percentile values for the subjective measures NASA-TLX, SART, SPAM, perceived support, and training. In the analysis, 3 participants were excluded from each group due to the absence of some records and, moreover, to ensure balance between the groups. Consequently, the final selection for group PBP included 9 females and 11 males, the same as for group EP, as illustrated in Figure 5.2. The mean ages of these participants were 23 for PBP and 26 for EP, depicted in Figure 5.1.

The NASA-TLX measures in Table 5.2 across both groups (PBP and EP) highlight several patterns as scenarios increase from S1 to S3.

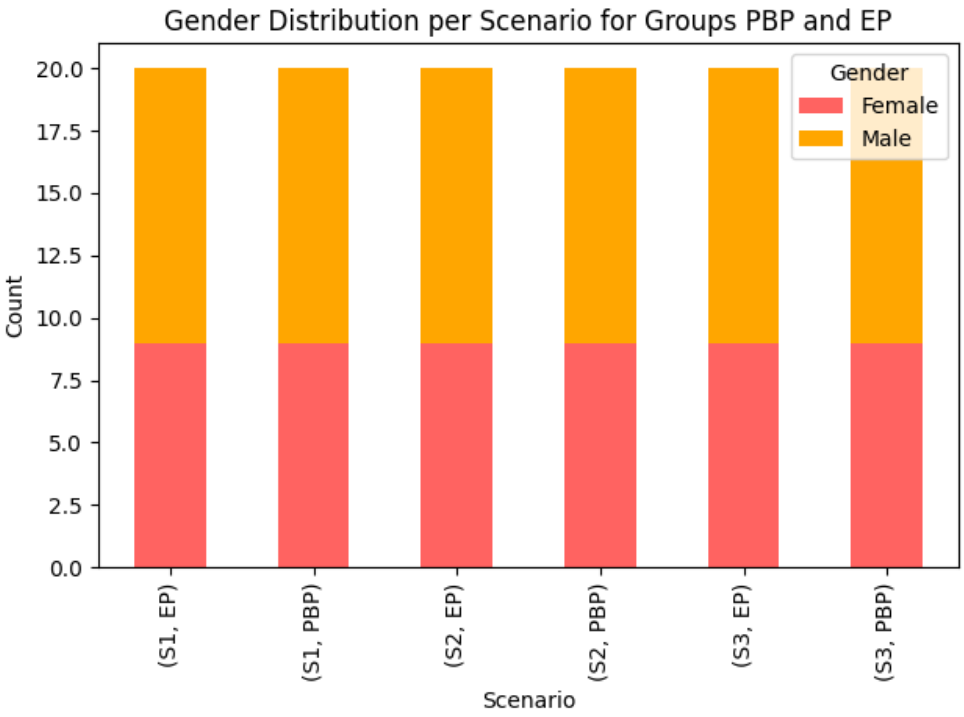


Fig. 5.2 Gender distribution for by scenario.

Table 5.2 NASA-TLX workload measures for different scenarios and the study groups: PBP vs EP.

Scenario	**S1**		**S2**		**S3**	
Group	**PBP**	**EP**	**PBP**	**EP**	**PBP**	**EP**
Mental_demand						
mean	3.400	3.150	3.650	3.600	5.700	5.300
std	1.046	1.461	1.226	1.465	1.261	1.455
95%	5.000	6.000	5.050	6.050	7.000	7.000
Physical_demand						
mean	1.850	1.500	1.900	1.850	2.450	2.750
std	1.348	0.607	1.252	0.875	1.504	1.410
95%	3.200	2.050	3.150	3.000	4.150	5.050
Temporal_demand						
mean	3.600	3.000	3.650	2.600	5.150	4.600
std	1.536	1.717	1.226	1.667	1.461	1.501
95%	6.000	6.000	5.050	5.100	7.000	7.000
Performance						
mean	5.300	5.450	5.400	4.900	3.550	3.150
std	1.490	1.572	1.314	1.744	2.212	1.954
95%	7.000	7.000	7.000	7.000	7.000	7.000
Effort						
mean	3.600	2.950	3.850	3.450	5.550	4.750
std	1.635	1.234	1.424	1.234	1.317	1.482
95%	6.000	5.000	6.000	5.000	7.000	7.000
Frustration						
mean	2.350	2.850	2.700	2.700	4.250	4.050
std	1.309	1.843	1.525	1.302	1.585	1.504
95%	4.050	5.100	5.050	5.050	6.050	6.050

- Mental Demand: Both groups (PBP and EP) show a notable increase in mental demand from S1 to S3. For the PBP group, the mental demand increases from 3.400 (S1) to 5.700 (S3), and for the EP group, from 3.150 to 5.300. This increase suggests that as scenarios become more complex, participants perceive a greater cognitive load, particularly in the PBP group.

- Physical Demand: Physical demand remains relatively low in S1 and S2, but both groups experience an increase in S3. In PBP, it increases from 1.850 (S1) to 2.450 (S3), while in EP, the demand increases from 1.500 to 2.750. This indicates that later scenarios are perceived as more physically demanding compared to earlier ones.

- Temporal Demand: There is a slight increase in temporal demand, with PBP increasing from 3.600 to 5.150 and EP from 3.000 to 4.600 across scenarios. The increase suggests that time pressure becomes more prominent in the latter scenarios, especially for PBP participants.

- Performance: Performance ratings generally decline in the scenarios, particularly for the PBP group, which drops from 5.300 to 3.550, and the EP group, from 5.450 to 3.150. This decrease implies that participants perceived that they performed worse in the more complex later scenarios.

- Effort: Both groups show an increase in effort required as the scenarios progress. In PBP, the effort increases from 3.600 (S1) to 5.550 (S3), and in EP from 2.950 to 4.750. This indicates that the perceived effort to complete tasks becomes more significant as the complexity of the scenarios increases.

- Frustration: Frustration levels generally increase over time, with the PBP group experiencing frustration increasing from 2.350 to 4.250 and the EP group from 2.850 to 4.050. The higher frustration in S3 suggests that tasks in this scenario are perceived as more frustrating or challenging.

SART Variables (Subjective Measures):

The SART (Situational Awareness Rating Technique) measures in Table 5.3 provide insights into situational awareness across both groups (PBP and EP) and scenarios (S1 to S3).

Table 5.3 SART measures for different scenarios and groups: PBP vs. EP.

Scenario	**S1**		**S2**		**S3**	
Group	**PBP**	**EP**	**PBP**	**EP**	**PBP**	**EP**
Instability						
mean	2.550	2.400	2.850	2.950	5.500	5.550
std	1.146	1.536	0.933	1.356	1.395	1.234
95%	4.050	5.050	4.000	5.050	7.000	7.000
Variability						
mean	2.400	2.250	2.700	2.750	5.700	5.700
std	0.883	0.910	0.801	1.251	1.342	1.455
95%	4.050	3.100	4.000	5.050	7.000	7.000
Complexity						
mean	2.450	2.250	3.050	2.950	5.850	5.350
std	1.050	1.251	1.050	1.356	1.496	1.348
95%	5.000	5.000	4.100	6.000	7.000	7.000
Arousal						
mean	3.450	2.900	3.550	3.200	5.350	4.350
std	1.468	1.210	1.395	1.240	1.424	1.531
95%	5.100	5.000	5.100	4.100	7.000	6.050
Spare_capacity						
mean	3.850	3.450	3.650	3.500	5.200	4.500
std	1.565	1.669	1.040	1.433	1.399	1.732
95%	7.000	6.050	5.000	6.000	7.000	7.000
Concentration						
mean	5.450	5.400	5.300	5.150	6.100	6.000
std	1.234	1.231	1.081	1.496	1.518	1.170
95%	7.000	7.000	7.000	7.000	7.000	7.000
Attention_division						
mean	4.000	4.350	3.900	4.300	4.900	4.900
std	1.654	1.755	1.483	1.780	1.619	1.861
95%	6.050	6.050	6.000	6.050	7.000	7.000
Quantity						
mean	5.000	5.100	4.900	5.200	5.000	5.100
std	1.257	1.553	1.410	1.240	1.414	1.619
95%	7.000	7.000	7.000	7.000	7.000	7.000
Quality						
mean	5.700	5.400	5.300	5.350	5.100	5.050
std	1.081	1.353	1.081	1.309	1.334	1.538
95%	7.000	7.000	7.000	7.000	7.000	7.000
Familiarity						
mean	4.050	3.850	4.550	4.300	4.250	3.800
std	1.986	2.007	1.669	1.922	1.682	1.852
95%	7.000	7.000	7.000	7.000	6.050	7.000

- Instability: Instability shows a gradual increase across scenarios. In PBP, it rises from 2.550 (S1) to 5.500 (S3), while in EP, it rises from 2.400 to 5.550. This indicates that participants perceive more instability or unpredictability in later scenarios, particularly in S3.
- Variability: A similar result pattern to instability is observed where variability increases across scenarios. PBP group report an increase from 2.400 to 5.700, and EP from 2.250 to 5.700. This suggests a growing sense of variability or fluctuations in situational conditions in the later scenarios.
- Complexity: Both groups show a significant increase in perceived complexity over time. In PBP, complexity increases from 2.450 to 5.850, while in EP, it rises from 2.250 to 5.350. This trend highlights that as scenarios become more advanced, the perceived complexity of the tasks increases as well, particularly in the PBP group.
- Arousal: Arousal levels remain relatively stable, although they peak slightly in S3 for both groups. PBP group shows an increase from 3.450 to 5.350, and EP from 2.900 to 4.350. This reflects heightened activation or alertness in the later stages but less drastically compared to other measures.
- Spare Capacity: Spare capacity shows decreases over time, suggesting participants feel they have fewer resources available as scenarios progress. In PBP, it decreases from 3.850 to 5.200, and in EP from 3.450 to 4.500. The reduction in spare capacity implies that more demanding scenarios require greater mental resource allocation.
- Concentration: Both groups report high levels of concentration, with slight variation across scenarios. PBP starts at 5.450 and ends at 6.100, while EP starts at 5.400 and ends at 6.000. This suggests participants are able to maintain focus despite increasing scenario complexity.
- Attention Division: Attention division remains relatively stable, with a slight increase in S2 and S3. For PBP, it starts at 4.000 and rises to 4.900 in S3, while in EP, it starts at 4.350 and stays consistent at 4.900. This reflects that participants need to divide their attention more as tasks grow more complex.
- Quantity and Quality: Both groups report consistently high levels of perceived quantity and quality of performance across all scenarios. PBP's quantity starts

at 5.000 and remains stable, while EP shows similar stability at 5.100. Likewise, Quality remains high, with PBP starting at 5.700 and ending at 5.100 and EP ranging between 5.400 and 5.050. This suggests that despite increasing task complexity, participants perceive their quantity and quality of performance as relatively stable.

- Familiarity: Familiarity shows moderate decreases over time, suggesting participants feel less familiar with tasks as scenarios increase. PBP group starts at 4.050 and drops to 4.250, while EP drops from 3.850 to 3.800. This may indicate that the more complex scenarios (e.g., S3) introduce less familiar challenges.

SPAM Variables:

Table 5.4 Descriptive statistics for SPAM measures across scenarios and groups.

Scenario	**S1**		**S2**		**S3**	
Group	**PBP**	**EP**	**PBP**	**EP**	**PBP**	**EP**
SA1						
mean	4.900	4.700	4.800	4.600	3.850	4.450
std	0.308	0.923	0.523	0.754	1.424	0.999
95%	5.000	5.000	5.000	5.000	5.000	5.000
SA2						
mean	4.250	3.700	4.050	4.000	3.200	2.950
std	0.851	1.380	0.945	1.124	1.735	1.638
95%	5.000	5.000	5.000	5.000	5.000	5.000
SA3						
mean	4.500	4.050	4.450	4.200	2.900	3.450
std	0.889	1.504	1.146	1.399	1.683	1.572
95%	5.000	5.000	5.000	5.000	5.000	5.000

Table 5.4 presents three key dimensions of situational awareness (SA1: Perception of critical alarm, SA2: Comprehension of the situation, and SA3: Projection of future states) across three scenarios (S1, S2, S3) for two groups: PBP and EP.

- SA1 (Perception of the critical alarm and Why): SA1 remains relatively stable across S1 and S2, indicating that participants' ability to perceive the critical alarm is not significantly impacted by scenario progression in the earlier stages. However, by S3, the PBP group shows a significant drop in their perception of the alarm

(from 4.9 in S1 to 3.85 in S3), while the EP group experiences a smaller but noticeable decline (from 4.7 to 4.45). Those in group PBP had smaller standard deviations from the mean except for in scenario 3. This means that situation perception varied a lot more in scenario 3 for those in this group, which was also the case for group EP in scenarios 1 and 3. As scenarios grow more complex, the ability to quickly and accurately detect the critical alarm weakens, particularly in the PBP group, suggesting that cognitive overload or task complexity impacts perception in later stages.

- SA2 (Comprehension/Understanding of what to do and Why): SA2 shows a steady decline across all scenarios for both groups, with a sharper drop for the PBP group. The PBP group starts at 4.25 in S1 but falls to 3.2 by S3, while EP drops from 3.7 to 2.95 over the same period. The standard deviation in the situation understanding case increased compared to in scenario 1, showing obviously more variations in situation understanding among participants in the same group, especially in the most complex scenario. As the task complexity increases, participants' ability to comprehend the situation and know what actions to take declines. This suggests that the increasing complexity or ambiguity in the scenario makes it harder for participants to rely on their understanding or reference procedures, particularly in PBP participants.

- SA3 (Projection/predicting the consequences of their actions): In like manner, SA3 declines across scenarios, especially in S3. The PBP group drops from 4.5 in S1 to 2.9 in S3, while EP decreases from 4.05 to 3.45. This sharp decline in projection suggests that participants find it increasingly difficult to predict the consequences of their actions as the scenarios become more complex.

Group EP showed a higher standard deviation than Group PBP in the first two scenarios, and the reverse was the case in scenario 3. This indicates that those in Group EP had more varying perceptions and understanding of the situation within their study group. In contrast, those in Group PBP had more varying projections of the situation. Generally, this variation increased as task complexity increased. However, those in group EP had low variability in scenario 2 compared to scenario 1.

NASA-TLX Index, SART dimensions, SART and SPAM Index and Task load (Subjective Measures):

Table 5.5 Descriptive statistics for NASA-TLX, SPAM, SART, and Task Load measures across scenarios and groups.

Scenario	**S1**		**S2**		**S3**	
Group	**PBP**	**EP**	**PBP**	**EP**	**PBP**	**EP**
NASA-TLX index						
mean	3.350	3.150	3.525	3.183	4.442	4.100
std	0.880	0.745	0.867	0.709	0.881	0.718
95%	4.567	4.358	4.692	4.342	5.408	5.175
SPAM Index						
mean	4.550	4.150	4.433	4.267	3.317	3.617
std	0.423	0.834	0.593	0.792	1.111	1.039
95%	5.000	5.000	5.000	5.000	4.367	5.000
SART Demand						
mean	7.400	6.900	8.600	8.650	17.050	16.600
std	2.393	3.093	2.062	3.216	3.873	3.068
95%	11.150	11.200	11.050	14.100	21.000	20.050
SART Supply						
mean	16.750	16.100	16.400	16.150	21.550	19.750
std	4.459	4.217	3.676	4.392	4.186	4.191
95%	24.000	21.000	22.000	22.050	26.000	24.100
SART Understanding						
mean	14.750	14.350	14.750	14.850	14.350	13.950
std	3.582	3.588	3.143	3.329	3.631	3.471
95%	20.050	19.100	19.100	21.000	20.050	19.100
SART Index						
mean	24.100	23.550	22.550	22.350	18.850	17.100
std	5.553	6.724	5.296	4.771	5.324	5.505
95%	29.450	31.250	31.000	28.250	26.100	24.200
Task Load						
mean	2.250	2.400	2.800	2.650	4.600	4.150
std	1.020	1.095	0.834	0.875	0.503	1.089
95%	4.050	4.050	4.000	4.000	5.000	5.000

Table 5.5 shows the NASA Index calculated based on the NASA-TLX measures using Equation 3.8 and the SPAM Index calculated based on the SPAM measures using Equation 3.7. The SART Supply, Demand and Understanding were calculated using the measures that describe them in Chapter 3 and briefly introduced in the early part of this Section. SART Index is then calculated based on these three SART dimensions using Equation 3.6. An important part of the results at this stage is the significant difference between the groups on each measure and for each scenario. To achieve this, a Shapiro test was run to check for the normality of the measures and to determine the type of statistical test to be used to calculate significant differences. The result of the Shapiro-Wilk test is shown in Tables 5.5 and 5.8.

If the value for a group is $> p$, that is 0.05, then the data for that group, under a specific measure being tested, is normally distributed. Else the data is said to have a non-parametric distribution. Those with parametric data were tested using a student t-test (see Section 3.1) and coloured in blue in Tables 5.5 and 5.8. A t-test is used only when both groups have values $> p$. If this is the case for only 1 group, then a Mann-Whitney test is used. A Mann-Whitney test was used for cases of Non-parametric distribution as shown in the results in Table 5.7. The results of Table 5.7 show that there is no significant difference between the two groups on any of their perceived workload based on NASA-TLX, situational awareness based on SPAM and SART Index and Task Load.

Support Mechanism, Training and Familiarity: Table 5.6 presents mean scores, standard deviations, and 95% confidence intervals for various support and familiarity measures (Alarm List Support, Alarm Priority Support, Procedure Support, Training, and Familiarity) across three scenarios (S1 to S3) and two groups (PBP and EP).

- Alarm List Support: The mean values remain relatively stable from S1 to S2 for both groups (PBP: 4.300 to 4.200, EP: 4.200 to 4.300). However, there is a noticeable decline in S3 (PBP: 3.850, EP: 3.700), indicating a decrease in perceived support related to alarm lists as the task complexity increases in the latter scenario. The standard deviation increases in S3, particularly for the EP group, indicating more variation in participants' perceptions of alarm list support in more complex tasks.

- Alarm Priority Support: Mean scores are consistently high across the scenarios, but both groups show a decline in S3 (PBP: 4.450 to 4.100, EP: 4.400 to 3.750), reflecting lower perceived support for alarm priority in the final scenario. The

standard deviation likewise increases in S3, indicating a more diverse experience of alarm priority support across participants as the scenario becomes more complex.

- Procedure Support: The PBP group shows a decline in perceived procedure support from S2 (4.600) to S3 (3.450), while the EP group shows a similar drop from 4.100 (S2) to 3.250 (S3). This suggests that both groups feel less supported by the procedure as tasks become more demanding. The increasing standard deviation in S3, especially for the EP group, indicates that participants had varying experiences of procedure support under high-pressure conditions.

- Training Average: The training average remains consistent across all scenarios for both groups, with the PBP group maintaining a mean of 3.917 and the EP group at approximately 4.1. This stability suggests that perceived training performance does not vary significantly as task difficulty increases. The low standard deviation across scenarios highlights a relatively consistent perception of training effectiveness across participants.

- Familiarity Average: Familiarity with the process and control room remains stable across all scenarios (PBP mean = 2.650, EP mean = 2.425 to 2.375), indicating that participants' familiarity with their environment does not change significantly as tasks become more complex. The standard deviation remains relatively low, suggesting minimal variation in how familiar participants feel across scenarios.

Calculating Significant Differences for Tables 5.5 and 5.6:

The results of the Shapiro-Wilk tests in Table 5.8 indicated that most distributions were non-normal for the majority of the measures across all scenarios, leading to the application of Mann-Whitney U tests to compare the PBP and EP groups. Specifically, Alarm List Support and Alarm Priority Support exhibited non-normal distributions across all three scenarios. The Mann-Whitney U tests revealed no significant differences between the two groups in any of the scenarios for these measures (for example, S2: $p = 0.976$ for Alarm List Support and $p = 0.447$ for Alarm Priority Support). This suggests that both groups had similar perceptions of alarm support regardless of the increasing task complexity.

However, for Procedure Support, while the Shapiro test indicated non-normality, a significant difference between the PBP and EP groups emerged in Scenario 2. The Mann-Whitney U test revealed a significant result ($p = 0.021$), indicating that the two

Table 5.6 Descriptive statistics for Alarm Support, Procedure Support, Training, and Familiarity across scenarios and groups.

Scenario	**S1**		**S2**		**S3**	
Group	**PBP**	**EP**	**PBP**	**EP**	**PBP**	**EP**
Alarm_list_support						
mean	4.300	4.200	4.200	4.300	3.850	3.700
std	0.979	0.696	1.056	0.801	1.089	1.302
95%	5.000	5.000	5.000	5.000	5.000	5.000
Alarm_priority_support						
mean	4.450	4.400	4.350	4.500	4.100	3.750
std	0.686	0.821	0.745	0.761	1.165	1.372
95%	5.000	5.000	5.000	5.000	5.000	5.000
Procedure_support						
mean	4.450	4.250	4.600	4.100	3.450	3.250
std	0.999	0.967	0.995	0.968	1.146	1.552
95%	5.000	5.000	5.000	5.000	5.000	5.000
Training_average						
mean	3.917	4.113	3.917	4.091	3.917	4.091
std	0.581	0.451	0.581	0.446	0.581	0.446
95%	4.449	4.714	4.449	4.714	4.449	4.714
Familiarity_average						
mean	2.650	2.425	2.650	2.375	2.650	2.375
std	0.690	0.847	0.690	0.841	0.690	0.841
95%	3.525	3.525	3.525	3.525	3.525	3.525

Table 5.7 Significant Differences for Table 5.5. T-test values are highlighted in blue. Value marked * shows significant difference $p < 0.05$.

Scenario	**S1**		**S2**		**S3**	
Group	**PBP**	**EP**	**PBP**	**EP**	**PBP**	**EP**
NASA-TLX index						
Shapiro	0.025	0.433	0.981	0.622	0.151	0.046
T-test/Mann W.	0.578		0.181		0.135	
SPAM Index						
Shapiro	0.009	0.011	0.008	0.004	0.018	0.018
Mann W.	0.184		0.646		0.404	
SART Demand						
Shapiro	0.035	0.086	0.235	0.265	0.000	0.031
T-test/Mann W.	0.379		0.954		0.360	
SART Supply						
Shapiro	0.191	0.059	0.254	0.354	0.001	0.111
T-test/Mann W.	0.638		0.846		0.100	
SART Understanding						
Shapiro	0.403	0.779	0.676	0.256	0.557	0.966
T-test	0.726		0.923		0.724	
SART Index						
Shapiro	0.381	0.681	0.510	0.266	0.242	0.077
T-test	0.779		0.901		0.313	
Task Load						
Shapiro	0.005	0.033	0.013	0.018	0.000	0.000
Mann W.	0.646		0.623		0.280	

Table 5.8 Significant Differences for Table 5.6. T-test values are highlighted in blue. Value marked * shows significant difference $p < 0.05$.

Scenario	**S1**		**S2**		**S3**	
Group	**PBP**	**EP**	**PBP**	**EP**	**PBP**	**EP**
Alarm_list_support						
Shapiro	0.000	0.001	0.000	0.000	0.005	0.003
Mann W.	0.374		0.976		0.832	
Alarm_priority_support						
Shapiro	0.000	0.000	0.000	0.000	0.000	0.003
Mann W.	1.000		0.447		0.430	
Procedure_support						
Shapiro	0.000	0.000	0.000	0.001	0.085	0.008
Mann W.	0.367		0.021*		0.803	
Training_average						
Shapiro	0.163	0.062	0.163	0.087	0.163	0.087
T-test	0.241		0.293		0.293	
Familiarity_average						
Shapiro	0.223	0.236	0.223	0.306	0.223	0.306
T-test	0.363		0.265		0.265	

groups had different perceptions of the support provided by procedures during the second scenario. This finding suggests that the participants in the PBP and EP may have varied in how effectively they felt supported by operational procedures at this task stage.

Regarding the training average, the results of the Shapiro test indicated normality for most scenarios, leading to the application of T-tests. However, no significant differences were found between the groups in any of the scenarios (e.g., S2: $p = 0.293$), indicating that both groups perceived their training performance similarly across the scenarios.

Similarly, for the average of familiarity, T-tests were performed as the Shapiro results indicated normality. Again, no significant differences were observed between the groups (for example, S2: $p = 0.265$), suggesting that both groups felt equally familiar with the process and the environment of the control room as the scenarios progressed.

Objective Measure Descriptive Data

To this point in this chapter, the focus was on subjective measurements, that is, on the emotional self-reporting of the participant, which gives us a window into the mental state of the person as perceived by the participants themselves. But by coupling that with objective measurements, which are physical outer body observations of behaviour and performance, we can not only gain insights into behaviour and performance but also validate the outcomes of the subjective measures.

To match the data from the collected simulator log with that of the objective metrics, some participants were excluded due to incomplete simulator data from some of the participants. Seventeen (17) participants were left in each group after this cleaning process. Group PBP consists of 8 women and 9 men with an average age of around 25 years, whereas EP includes 6 women and 11 men with an average age of about 23 years.

Performance Measures (Objective Measures):

Performance measures include: Accuracy, Recovery Time, Overall Performance, and Error. The descriptive statistics for the performance data are shown in Table 5.9.

As shown in Table 5.9 for Recovery Time/Survival (s), in scenario 1, the PBP group had a mean recovery time of 337.647 seconds, while the EP group was faster, averaging 260.824 seconds. This indicates that the EP group recovered faster in the first scenario. In scenario 2, Recovery times increased for both groups, with PBP at 458 seconds and EP slightly higher at 488.235 seconds. This suggests that as the scenario increased in complexity, both groups took longer to recover. Recovery times drastically increased in

Table 5.9 Descriptive statistics for Recovery Time and Accuracy (Mean Square Error) across scenarios and groups.

Scenario	**S1**		**S2**		**S3**	
Group	**PBP**	**EP**	**PBP**	**EP**	**PBP**	**EP**
Recovery_time (s)						
mean	337.647	260.824	458.000	488.235		
std	186.875	112.108	217.011	149.412		
95%	658.600	497.600	833.000	777.600		
Accuracy (mse)						
mean	0.263	0.294	0.003	0.002	78.924	115.614
std	0.573	0.568	0.006	0.007	201.702	318.391
95%	1.492	1.412	0.019	0.013	268.635	457.485

S3 for both groups. There is no recovery time in scenario 3 because the plant recovers right at the end of the test and the participants are only expected to maintain the plant from overheating until that time.

Accuracy, measured by mean squared error (MSE) , was relatively low in S1 for both groups (PBP = 0.263, EP = 0.294). These values indicate a low level of error, and both groups performed similarly. Mean squared error (MSE) is commonly used to measure the average of the squares of errors of a variable or model prediction. That is, the difference between the actual (observed) values and the predicted/new values generated by a model or in this case, the operators' control action. When the operator moves a manual valve to the exact expected value then there is no error, the MSE equals zero. As error increases, the MSE value increases. In S2, MSE dropped significantly to very low values for both groups (PBP = 0.003, EP = 0.002), indicating a high level of accuracy in this scenario which could be due to transferred learning from scenario 1. However, MSE increased drastically in S3, with PBP rising to 78.924 and EP to 115.614. This indicates that both groups made significantly more errors in the final scenario as task complexity increased, with the EP group experiencing a slightly higher error rate compared to the PBP group.

Following this, Table 5.10 presents a condensed descriptive summary of the results pertaining to overall performance, errors, and consequences (which are all categorical variables), shedding light on the actual performance of participants in both groups.

Behavioural Measures (Objective Measures): Table 5.11 presents the mean, standard deviation, and 95% confidence interval for several behavioural measures across

Table 5.10 Summary of Consequences, Performance, Errors, and Success/Failure Counts across Scenarios and Groups.

Scenario	Group	Participant	Top Consequence	Top Overall Perf.	Success/Failure Count
S1	G2	17	Safe (17)	Good (8)	Success (17)
S1	G3	17	Safe (17)	Optimal (11)	Success (17)
S2	G2	17	Impurity of air in tank atm.(8)	Good (8)	Success (14), Failure (3)
S2	G3	17	Impurity of air in tank atm.(8)	Good (12)	Success (15), Failure (2)
S3	G2	17	Safe (11)		Success (11), Failure (6)
S3	G3	17	Reactor Overheated (9)		Success (8), Failure (9)

three scenarios (S1, S2, and S3) and two groups (PBP and EP). Behavioural measures include reaction time, response time, silenced alarms, acknowledgements, open copies, number of procedures, and number of alarms.

In the first scenario, the reaction times of the two groups were quite close, with PBP taking slightly longer at 283.588 seconds compared to EP's 270.647 seconds. In the second scenario, both groups' reaction times stayed consistent. However, in the third scenario, a difference was observed, as PBP averaged 289.765 seconds, while EP's reaction time dropped considerably to 197.118 seconds, indicating that EP participants responded more quickly as the complexity increased.

The response times were shorter in the early scenarios but drastically increased in Scenario 3, especially for PBP, which increased from 345.235s in S2 to 897.000s in S3. Group EP too, experienced an increase in response time, though not as severe as PBP. A longer response time is a good result because the participants are expected to stay as much as possible on the process until the 18th minute. For the measure Alarms Silenced, in Scenario 1, both groups silenced an average of approximately 2 alarms. However, by Scenario 3, the number of alarms silenced increased drastically, especially for PBP (12.059) compared to EP (9.941). This high number of alarms in scenario 3 is expected, given the complexity of the scenario.

Acknowledgements remained low in the earlier scenarios, averaging around 2 per group. However, in Scenario 3, PBP acknowledged significantly more alarms (6.765) compared to EP (4.588), showing that PBP participants were interacting with more alarms as the complexity increased and perhaps had a poor 'Perception.' On the other hand, for Number of Mimics Opened, both groups showed stable mimic opening behaviour across scenarios, averaging between approximately 3 and 4, indicating little change in this metric regardless of task complexity.

Table 5.11 Descriptive statistics for Reaction Time, Response Time, Alarms Silenced, Acknowledgements, Mimics Opened, Number of Procedures, and Number of Alarms across scenarios and groups.

Scenario	**S1**		**S2**		**S3**	
Group	**PBP**	**EP**	**PBP**	**EP**	**PBP**	**EP**
Reaction_time (s)						
mean	283.588	270.647	282.353	280.647	289.765	197.118
std	74.306	52.414	94.633	61.872	150.602	126.136
95%	350.200	313.600	421.200	364.800	514.600	338.400
Response_time (s)						
mean	365.000	307.294	345.235	358.529	897.000	763.588
std	145.137	120.621	144.414	104.997	191.042	207.354
95%	666.000	418.800	520.200	487.000	1049.200	1026.000
Alarms_silenced (no.)						
mean	1.588	1.471	1.882	1.529	12.059	9.941
std	0.712	0.874	0.857	0.800	5.804	7.267
95%	2.000	2.200	3.000	2.000	19.400	19.000
Acknowledgements (no.)						
mean	1.529	1.529	1.706	1.412	6.765	4.588
std	0.624	0.800	0.772	0.795	4.085	4.583
95%	2.000	2.200	3.000	2.200	13.400	11.600
Mimics_opened (no.)						
mean	3.118	2.765	2.882	2.882	3.647	3.765
std	1.691	1.954	1.576	1.833	1.367	1.251
95%	6.000	6.000	6.000	6.000	6.000	6.000
std	0.000	0.612	0.000	0.795	0.000	2.404
95%	0.000	4.000	0.000	4.000	0.000	12.000
Number_of_alarms (no.)						
mean	1.941	2.235	2.235	2.059	17.118	17.882
std	0.429	0.562	0.562	0.243	4.357	3.887
95%	2.200	3.000	3.000	2.200	23.400	24.200

The EP group opened an average of approximately eight procedures, while the record for PBP could not be ascertained because there was no recording of this information.

The number of alarms increased sharply in Scenario 3 for both groups, with PBP encountering 17.118 alarms and EP facing 17.882 alarms. This suggests a much more alarm-intensive environment in the final scenario.

Eye-Tracking (Objective Measures):

The eye-tracking analysis was examined for Scenario 3 as this scenario is believed to highlight significant distinctions between the two groups, largely due to the cognitive effort required. Since Scenarios 1 and 2 are presumed to require less cognitive effort for an in-depth analysis of eye movements compared to Scenario 3. Two female and two male participants were selected at random from each group for this analysis. The emphasis was on examining the participants' eye movements within the areas of interest pertinent to each group, specifically the procedure (Paper procedure or Electronic Procedure). The randomly selected participants had an average age of approximately 23 in each group. Table 5.12 shows the descriptive statistics of their outcome in scenario 3 based on the eye-tracking measures selected for the comparison of both groups: average fixation duration, number of visits, average pupil diameter, saccade amplitude, saccade duration, and number of fixations.

The outcomes of the eye-tracking for scenario 3 as shown in table 5.12 are briefly discussed below. For the measure, Average Fixation Duration, the EP group showed a significantly longer mean fixation duration (292.625 ms) compared to the PBP group (197.250 ms). The large standard deviation for the EP group (157.080 ms) indicates high variability in fixation durations compared to the PBP group (68.674 ms).

For the Number of Visits, both groups showed similar mean numbers of visits, with PBP at 5.125 and EP at 5.000. However, the variability in visits was slightly higher for the EP group (std = 3.697) compared to the PBP group (std = 3.198), showing a more consistent behaviour in the PBP group.

Interestingly, the EP group showed a negative corrected mean pupil diameter (-0.515 mm), while the PBP group had a positive corrected mean of 0.458 mm. This disparity may reflect different physiological responses between the groups during the task, with the negative value indicating pupil constriction in the EP group. A constriction generally shows a state of relaxation or rest as discussed in Section 2.1, meaning that this was the case of those in the EP group while group PBP was more alert. Average Saccade Amplitude: The PBP group exhibited larger saccade amplitudes, with a mean of 5.848°,

Table 5.12 Descriptive statistics for eye-tracking metrics in Scenario 3 for PBP and EP groups.

Scenario 3	**Group**	
Measure	**PBP**	**EP**
Average fixation duration (ms)		
mean	197.250	292.625
std	68.674	157.080
95%	261.200	471.875
No of visits (count)		
mean	5.125	5.000
std	3.198	3.697
95%	8.625	9.450
Average pupil diameter (mm)		
mean	0.458	-0.515
std	0.109	0.292
95%	0.589	-0.193
Average saccade amplitude (°)		
mean	5.848	3.948
std	0.891	0.804
95%	6.886	4.487
Saccade duration (ms)		
mean	37.068	32.393
std	3.770	3.286
95%	40.472	34.790
No of fixations		
mean	6.433	27.593
std	3.806	29.283
95%	10.357	60.038

compared to 3.948° in the EP group. This suggests that the PBP group tended to make larger eye movements with more cognitive efforts to shift attention or focus during the task. On the other hand, both groups had similar saccade durations, although the PBP group (mean = 37.068 ms) exhibited slightly longer durations than the EP group (mean = 32.393 ms). The standard deviations were comparable for both groups, indicating consistent saccade behaviour across participants.

Finally, there was a substantial difference in the number of fixations between the two groups. The EP group recorded a much higher number of fixations (mean = 27.593) compared to the PBP group (mean = 6.433). The high variability in the EP group's fixations (std = 29.283) suggests that some participants in EP fixated much more frequently than others. When associated with the saccade amplitude, this shows that group EP had a slight lessening of attention shifting compared to group PBP, therefore enabling more brain processing activity.

The next step would be to find out if there were significant differences between both groups on the measures, as shown in the next subsections, titled Significant Differences.

Significant Differences for Performance Measures in Table 5.9

Table 5.13 Shapiro-Wilk and T-test/Mann-Whitney results for Recovery Time and Accuracy across scenarios and groups.

Scenario	**S1**		**S2**		**S3**	
Group	**PBP**	**EP**	**PBP**	**EP**	**PBP**	**EP**
Recovery Time						
Shapiro	0.413	0.102	0.882	0.180	-	-
T-test	0.156		0.639			
Accuracy						
Shapiro	0.000	0.000	0.000	0.000	0.000	0.000
Mann W.	0.179		0.986		0.918	

Table 5.13 shows that the Shapiro-Wilk normality test reveals mixed results for recovery time and accuracy in all scenarios. In Scenario 1 (S1), the recovery time data for the PBP group follow a normal distribution (p = 0.413), while the EP group shows non-normality (p = 0.102). Despite this, the T-test did not find a statistically significant difference in recovery time between the two groups (p = 0.156). In Scenario 2 (S2), the

Shapiro-Wilk test confirmed normality for the PBP group (p = 0.882), but not for the EP group (p = 0.180). A T-test again indicated that there was no significant difference in recovery time between the groups (p = 0.639). Recovery time data for Scenario 3 (S3) were not available for analysis.

Regarding accuracy, the Shapiro-Wilk test consistently indicated nonnormality (p = 0.000) for both groups in all scenarios, requiring the use of the Mann-Whitney U test. In scenario 1, the Mann-Whitney test did not show significant differences in precision between the PBP and EP groups (p = 0.179). Similarly, in Scenario 2 (p = 0.986) and Scenario 3 (p = 0.918), no significant differences in accuracy were observed between the groups.

Significant Differences for Behavioural Measures in Table 5.11

Table 5.14 Shapiro-Wilk and T-test/Mann-Whitney results for Reaction Time, Response Time, Alarms Silenced, Acknowledgements, Mimics Opened, and Number of Alarms across scenarios and groups.

Scenario	**S1**		**S2**		**S3**	
Group	**PBP**	**EP**	**PBP**	**EP**	**PBP**	**EP**
Reaction Time (s)						
Shapiro	0.002	0.000	0.051	0.001	0.420	0.103
T-test/Mann W.	0.148		0.630		0.061	
Response Time (s)						
Shapiro	0.096	0.000	0.290	0.084	0.001	0.046
T-test/Mann W.	**0.022***		0.761		0.064	
Alarms Silenced (no.)						
Shapiro	0.000	0.005	0.001	0.000	0.152	0.025
T-test/Mann W.	0.662		0.158		0.568	
Acknowledgements (no.)						
Shapiro	0.000	0.008	0.015	0.024	0.425	0.007
T-test/Mann W.	0.969		0.279		0.069	
Mimics Opened (no.)						
Shapiro	0.013	0.029	0.007	0.025	0.001	0.002
T-test/Mann W.	0.427		1.000		0.581	
Number of Alarms (no.)						
Shapiro	0.000	0.000	0.000	0.000	0.595	0.468
T-test/Mann W.	0.095		0.211		0.593	

Table 5.14 shows that the Shapiro-Wilk normality test revealed mixed results in all scenarios for the variables reaction time, response time, silenced alarms, acknowledgements, mimics opened and number of alarms. In scenario 1 (S1), the reaction time was not normal for both groups, and the T-test did not indicate significant differences between PBP and EP ($p = 0.148$). Similarly, no significant differences were observed in scenario 2 (S2) ($p = 0.630$), while scenario 3 (S3) showed a trend toward significance ($p = 0.061$), suggesting that the reaction time may have been shorter for the EP group under more complex conditions.

For Response Time, in S1, the Mann-Whitney U test revealed a significant difference between the groups ($p = 0.022$), with EP responding faster. However, in S2 and S3, the response time did not differ significantly between the groups ($p = 0.761$ and $p = 0.064$, respectively).

Regarding the number of alarms silenced, no significant differences were observed across all scenarios, as indicated by the Mann-Whitney U test (S1: $p = 0.662$, S2: $p = 0.158$, S3: $p = 0.568$). Similarly, for acknowledgements, the Mann-Whitney U test revealed no significant differences in S1 ($p = 0.969$) or S2 ($p = 0.279$), although the result for S3 approached significance ($p = 0.069$), suggesting a possible difference in how the groups acknowledged alarms in the final scenario.

In terms of Mimics Opened, no significant differences were detected between the groups in any of the scenarios. Lastly, the number of alarms did not show significant differences between PBP and EP in S1, S2 or S3.

In summary, the results suggest that while both groups performed similarly for most metrics, the EP group showed a significantly faster response time in S1, and there were trends toward differences in Acknowledgements and Reaction Time in S3. These findings indicate that group differences may become more apparent under more complex task conditions.

Significant Differences for Eye-tracking Measures in Table 5.12

Table 5.15 shows that the Shapiro-Wilk test for normality and subsequent T-test and Mann-Whitney U tests reveal both significant and non-significant differences in eye-tracking metrics between the PBP and EP groups in Scenario 3. For Average Fixation Duration, both groups exhibited normal distributions (PBP: $p = 0.602$, EP: $p = 0.859$), with the T-test showing no significant difference between the groups ($p = 0.308$).

Table 5.15 Shapiro-Wilk and T-test/Mann-Whitney results for eye-tracking metrics in Scenario 3 across PBP and EP groups.

Scenario	**S3**	
Group	**PBP**	**EP**
Average fixation duration (ms)		
Shapiro	0.602	0.859
T-test	0.308	
No of visits (count)		
Shapiro	0.255	0.029
Mann W.	0.882	
Average pupil diameter (mm)		
Shapiro	0.013	0.968
Mann W.	**0.029***	
Average saccade amplitude (°)		
Shapiro	0.451	0.100
T-test	**0.019***	
Saccade duration (ms)		
Shapiro	0.391	0.163
T-test	0.111	
No of fixations (count)		
Shapiro	0.974	0.368
T-test	0.202	

In terms of the Number of Visits, the EP group exhibited non-normality (p = 0.029), but the Mann-Whitney U test indicated no significant difference between the groups (p = 0.882). Interestingly, the Average Pupil Diameter showed non-normality for the PBP group (p = 0.013), and the Mann-Whitney U test indicated a significant difference between the groups (p = 0.029), suggesting that pupil dilation differed significantly between PBP and EP participants.

For Average Saccade Amplitude, both groups had normal distributions, but the T-test revealed a significant difference between the groups (p = 0.019), indicating that the PBP and EP groups exhibited different saccade patterns. However, for Saccade Duration, the T-test found no significant difference between the groups (p = 0.111), and both distributions were normal.

Lastly, for the Number of Fixations, no significant differences were found between the groups (p = 0.202), with both groups exhibiting normal distributions.

5.3 Evaluation of the measures

5.3.1 Reliability of the Subjective Measures

To verify the reliability of the questionnaires, the internal consistency of the measures (that is, whether they evaluate the intended constructs) is assessed using Cronbach's alpha, as discussed in subsection 3.3.3. Alpha values less than 0.6 indicate poor consistency [139].

Table 5.16 Subjective Measures, their areas of focus, number of questions, and Cronbach's Alpha values.

Name of Measure	Measuring	No of Questions	Alpha
NASA TLX	Workload	6	0.6
SART	Situational awareness	10	0.8
SPAM	Situational awareness	3	0.5
Task load	Task load	1	-
Alarm design Support	Layout & alarm priority	2	0.8
Procedure Support	Paper or electronic support	1	-
Familiarity	Plant process & control room	2	0.5
Training	Post-training performance	8	0.4

Table 5.16 outlines the key subjective measures used in the study, each assessing different aspects of workload, situational awareness, task load, support mechanisms, training and familiarity. The NASA TLX measure assessed workload across six items and showed moderate reliability with a Cronbach's Alpha of 0.6. The SART (Situational Awareness Rating Technique) measured situational awareness across ten items and demonstrated high internal consistency, achieving a Cronbach's Alpha of 0.8, indicating robust reliability. The SPAM (Situation Present Assessment Method), designed to measure situational awareness using three items, yielded a lower Cronbach's Alpha of 0.5, suggesting weaker reliability. Similarly, the Familiarity measure, assessing knowledge of the process and control room across two items, showed low reliability with an Alpha of 0.5.

The Alarm Design Support measure evaluated the layout and prioritisation of alarms using two items and demonstrated high reliability with a Cronbach's Alpha of 0.8. Procedure Support and Task load were assessed through items specific to their respective measures, meaning their internal consistency cannot be evaluated. Lastly, the Training measure, which assessed post-training performance through eight items, yielded the lowest Cronbach's Alpha of 0.4, suggesting that the scale or content requires further refinement.

5.3.2 Validity of the Measures

Here we present a factor analysis of both subjective and objective measures. As the volume of experiment data in the captured dataset is limited, it is prudent to maximise available data to improve accuracy by undertaking factor analysis across the wider spectrum of scenarios and including all groups.

In the case of more available data, the factor analysis should be done per scenario with all groups combined regardless. The outcome of the factor analysis is later compared with the correlation analysis of the measures from all scenarios combined and later with scenarios separated. The scenario-based correlations versus the combined scenario correlations are compared with each other.

Factor analysis without eye-tracking measures

The first step was to ascertain the suitability of data for factor analysis. The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy in factor analysis showed a score of 0.775.

Table 5.17 Factor Loadings for Subjective and Objective Measures (without eye-tracking measures).

Measure	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11
SUBJECTIVE											
TLX_index	**0.86***										
Mental_demand	**0.80***										
Physical_demand						0.31				0.55	
Temporal_demand	**0.76***										
Performance					-0.43						
Effort	**0.80***										
Frustration							-0.31				
SPAM_Index				**0.97***							
SA1				**0.60***							
SA2				0.52							
SA3				0.58							
SART_Index		**0.82***						0.37			
SART_Demand	**0.82***										
SART_Supply	**0.82***							0.41			
SART_Understanding		**0.95***									
Task_load	**0.63***		0.35								
Alarm_list_support							0.38				
Alarm_priority_support					-0.39		0.41				
Procedure_support					-0.33	0.54		0.35			
Training_average		0.38									
Familiarity_average		**0.97***									
OBJECTIVE											
Reaction_time								**0.72***			
Response_time			**0.80***								
Alarms_silenced			**0.68***								
Acknowledgements			0.53								
No_of_alarms			**0.72***		0.31						
Recovery_time			**0.73***								
Overall Performance					0.40			0.48			
Consequence					**0.86***						
Error					**0.80***						

This KMO score strengthens the validity of the factor analysis results, as it confirms that the correlations among variables are strong enough to justify the extraction of factors. Had the KMO been significantly lower (e.g., below 0.5), it would have indicated that the data might not be suitable for factor analysis, as was the case when trying to separate by scenario.

The factor loadings of Table 5.17 present the results of the exploratory factor analysis (EFA) conducted on the subjective and objective measures. Values below 0.3 have been suppressed for readability, and values greater than or equal to 0.7 have been marked with an asterisk (*) to indicate strong loadings. This approach helps to focus on the most significant relationships between the variables and the factors, revealing underlying patterns in the data. That of the eye-tracking data was not included because of the number of participants analysed and because it was just for scenario 3. Therefore a separate factor analysis was done with this limited eye-tracking data and the other variables for the selected participants as shown under Table 5.18.

The scree plot, Figure 5.3, further illustrates the factor structure by displaying the eigenvalues for each factor in descending order. A steep drop is observed after the first few factors, with eigenvalues tapering off significantly after Factor 11. This aligns with the decision to retain 11 factors, as the first 11 factors explain a substantial proportion of the total variance in the data. The subsequent factors contribute little additional variance, indicating diminishing returns in terms of explanatory power. Overall, the factor analysis reveals a clear structure in both subjective and objective measures, with key factors corresponding to known constructs such as workload, situational awareness, and performance. The combination of high factor loadings in the first few factors and the sharp drop in eigenvalues in the scree plot supports the retention of 11 factors as a robust solution for this dataset.

The Subjective Measures exhibit high factor loadings under factors 1, 2 and 4. These factor loadings show that they share a common source of variation or are influenced by the same latent (unobserved) construct. is because they have strong relations with each other and describe or impact a common factor. For instance, the TLX Index shows strong correlations with Factor 1 (loading = 0.86), indicating that this factor captures a significant portion of workload-related variance, as would mental demand, temporal demand, effort, SART demand and supply and task load. The SART supply and demand loading on Factor 1 alongside the NASA TLX variables was similarly observed in a study

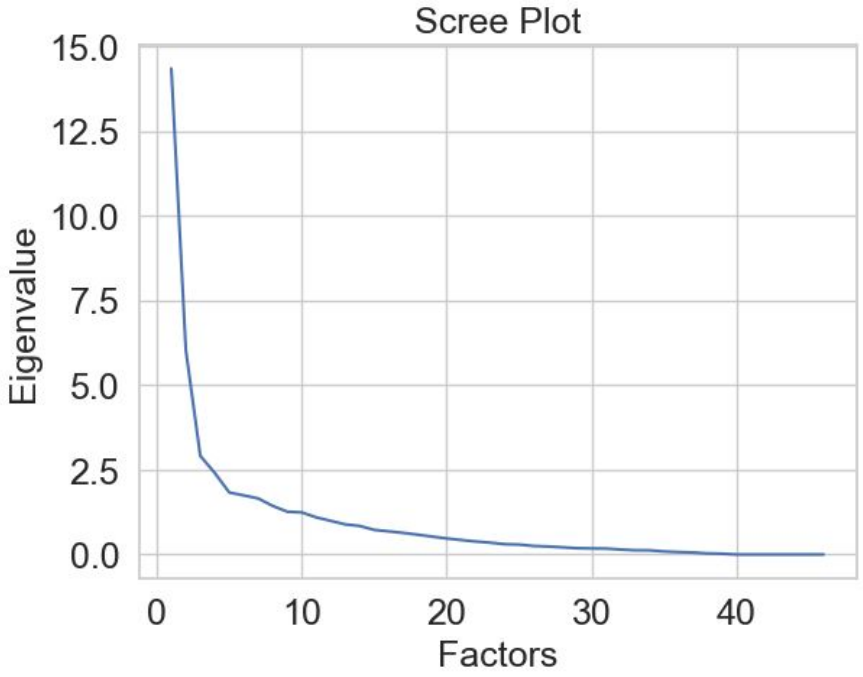


Fig. 5.3 Scree plot to determine the optimum number of factors for the objective and subjective measures combined *(Eye-tracking excluded*).

by Braarud [17]. Factor 1 and its related variables can thus be considered a *Workload construct*, independent of the scenario type and group.

The SART Index, its subcomponent SART-Understanding and familiarity average primarily loaded on Factor 2, with loadings as high as 0.95 for SART Understanding and 0.97 for familiarity. This implies that Factor 2 represents an aspect associated with situational awareness, emphasising the substantial role of process and control room familiarity on SART understanding in accurately assessing situational awareness, in contrast to SART demand and supply. Interestingly SPAM Index and its subcomponent SA1 - Perception, which measures situational awareness, loads strongly on a different Factor 4. The other two levels of situational awareness in SPAM (SA2 and SA3) closely load high on this factor, reinforcing that Factor 4 represents elements of situational awareness. It is important to note that the SPAM questions were asked during the exam, while the SART questions were given afterwards. Braaud [17] refers to the post-study evaluation as 'Non-Informed Situational Awareness', while the act of reviewing, or potentially experiencing and subsequently rating, is termed 'Informed Situational Awareness'. Thus, Factor 2 and its associated variables can be labelled *Non-Informed*

Situational Awareness, while Factor 4 can be referred to as *Informed Situational Awareness*.

Objective measures (excluding eye-tracking data) exhibited significant loadings on factors 3, 5, and 8. Reaction time had a positive loading of 0.72 on Factor 8, independent of the other factors. Response time, the number of alarms silenced, the number of alarms annunciated, and recovery time loaded on Factor 3. The number of acknowledgements likewise loaded positively on Factor 3 but did not exceed 0.6. Consequence, overall performance, and error were loaded onto factor 5, though overall performance showed only a low loading and did not appear significantly on any other factors. Factor 3, together with its associated variables, may be identified as a *behaviour construct*. Factor 5 is related to *performance*, while it is unclear what construct Factor 8 describes. Nevertheless, this classification requires further verification through correlation analysis. Factors 6 - 7 and 9 - 11 had no high loading and can be ignored.

These clusters offer an overview of the general nature of the variables and will be explored in more detail in relation to the results from each group and scenario. Additionally, comparing these findings with correlation analysis can highlight their interconnections and validate the concepts they represent or assess.

Factor analysis with eye-tracking measures

The factor loadings in Table 5.18 present the results of the exploratory factor analysis (EFA) conducted on subjective and objective measures, including the eye-tracking measures. Values below 0.3 have been suppressed for clarity, and values greater than or equal to +/- 0.7 have been marked with an asterisk (*) and in bold to indicate strong loadings. Those below +/- 0.7 were marked in asterisks. A steep drop in the scree plot in Figure 5.3 is observed after the first few factors, and the eigenvalues decrease significantly after factor 7.

The findings presented in Table 5.18 reveal that eye-tracking metrics primarily load on Factors 2, 3, and 6. Close examination of the loadings, derived from both subjective and objective measures from the simulator, highlights changes when selecting specific samples. The main objective is to verify the eye-tracking metrics in order to understand the concepts they describe while using the procedure, as well as their connection with other factors, to draw more accurate conclusions within each study group. Supplementary correlation analysis could substantiate the observed factor loadings. Average pupil

Table 5.18 Factor Loadings for the four participants in each group under scenario 3.

Measure	F1	F2	F3	F4	F5	F6	F7
SUBJECTIVE							
TLX Index	**0.92***						
Mental Demand	0.47	**-0.81***					
Physical Demand			0.42				**0.86***
Temporal Demand	0.63*				-0.64*		
Performance		0.58			0.57		
Effort	**0.93***						
Frustration	0.35		-0.64*	-0.56			
SPAM Index			**0.85***				
SA1			**0.70***				
SA2						0.61*	
SA3			**0.86***				
SART Index	0.62*	0.67*					
SART Demand			**0.75***			0.52	
SART Supply	**0.82***						
SART Understanding		**0.84***					
Task Load			**-0.88***				
Alarm List Support		0.68*					
Alarm Priority Support			**0.90***				
Procedure Support			**0.91***				
Training Average						**-0.90***	
Familiarity Average	**0.70***				0.55		
OBJECTIVE							
Reaction Time					**0.87***		
Response Time					**0.82***		
Alarms Silenced			-0.59*				
Acknowledgements		**-0.94***					
Mimics Opened				**0.81***			
No of Alarms	0.61*						
Recovery Time				**0.82***			
Accuracy							**0.71***
Consequence					**-0.85***		
Overall Performance					**-0.68***		
Error					**-0.86***		
OBJECTIVE *(eye-tracking)*							
Average Fixation Duration						**0.81***	
No of Visits		-0.65*					
Average Pupil Diameter				**0.83***			
Average Saccade Amplitude				**0.98***			
Saccade Duration				**0.73***			

diameter, saccade amplitude, and saccade duration show high positive loadings on Factor 4. Typically, pupil diameter variables indicate workload, while saccades can reflect workload, attention levels, engagement, search difficulty, or reduced processing capabilities. These variables load on the same factor as the number of opened mimics and recovery time but not with any situational awareness or subjective workload measures, except for an inverse loading of the NASA TLX measure - frustration. The combined loadings suggest relationships between these factors, showing direct associations for the primary variables and an inverse one for frustration. Recovery time correlates as a behaviour construct in Table 5.17, with the number of mimics opened not showing earlier factor loading. This may mean that for scenario 3, the number of mimics opened could describe a workload or search difficulty situation, as the three eye-tracking measures: average pupil diameter, saccade amplitude, and saccade duration. A correlation analysis will be conducted to validate these relationships.

Conversely, *number of visits* loads negatively on Factor 1 (loading = -0.65), showing a parallel relationship with mental demand (loading = -0.81) and number of acknowledgements (loading = -0.94) and an opposite relationship with SART Index (specifically from SART Understanding) (loading = 0.84) and support of the alarm list (loading = 0.68). The similarity between the number of visits and mental demand, a workload construct, indicates that this factor could describe workload but could also pertain to objective situational awareness due to the SART understanding of loading or support given the alarm list support loading. Hence, this factor might be termed a *an Objective SA construct*, offering broader conclusions on situational awareness and even workload, depending on the negative or positive loadings of the variables. The relationships among variables suggest that fewer alarm acknowledgements or visits could indicate lower mental workload and higher objective situational awareness (based on SART understanding), better alarm list support, and nearly improved perceived performance.

Furthermore, *average duration of fixation*, often used to describe attention, perceptual, and cognitive load, loads high on Factor 6 (loading = 0.81), alongside average training which loads negatively (loading = -0.90) and SPAM Question 2 on comprehension/understanding loading slightly above 0.6 (loading = 0.61). The positive correlation between the SPAM understanding variable (SA2) under the informed situational awareness construct and average fixation duration indicates that fixation duration may reflect informed situational awareness. The inverse relationship with

training suggests that poorer training performance could lead to longer fixations and better situational awareness. Typically, experts and novices who have recently undergone training exhibit longer average fixation durations [173]. Nonetheless, longer fixations are observed among less proficient readers [174]. In scenario 3, the factor loadings reveal that a poor training average leads to longer fixations, which is consistent with existing literature. However, these longer fixations suggest improved situation understanding, contrasting with the relationship seen with the number of fixations. The longer fixation duration and its positive correlations with SA2 indicate a *learning* process while using the procedure, which consequently facilitates situation understanding.

Gap: The factor analysis was conducted with all scenarios combined due to insufficient validity when evaluating factor loadings individually by scenario. As illustrated in Table 5.18, the results for the factor loadings of participants in scenario 3 indicate that the analysis is influenced by task complexity. Hence, with more data, ideally, the scenarios (task complexity) should be considered separately for factor loading.

Despite the small sample size, the substantial variations in eye-tracking metrics, including pupil diameter and saccades shown in Table 5.15, highlight the validity of these metrics for group comparisons, assuming there is a clear comprehension of the constructs they assess.

Correlation analysis

Another test for validity is performed through a correlation analysis. First, through a correlation analysis of the measures with groups and scenarios combined and later with the scenarios separated. By performing the latter, it can be confirmed if the measures are sensitive to increasing task complexity.

Considering all scenarios, the results of the Pearson correlation used in a case of normal distribution and that of Spearman used in the cases of non-parametric distribution are shown in Tables 5.19, 5.21, 5.23, 5.25, and 5.27.

TLX Dimensions:

Table 5.19 Correlation Outcomes Between Measures and TLX dimensions.

	TLX Index	Ment. Dem	Phys Dem	Temp. Dem	Perf	Effort	Frustration
TLX index	1.000	0.832	0.541	0.808	-0.132	0.851	0.586
Mental demand	0.832	1.000	0.364	0.661	-0.364	0.737	0.482
Physical demand	0.541	0.364	1.000	0.310	-0.137	0.384	0.407
Temporal demand	0.808	0.661	0.310	1.000	-0.248	0.679	0.359
Performance	-0.132	-0.364	-0.137	-0.248	1.000	-0.246	-0.464
Effort	0.851	0.737	0.384	0.679	-0.246	1.000	0.452
Frustration	0.586	0.482	0.407	0.359	-0.464	0.452	1.000
SPAM Index	-0.301	-0.362	-0.136	-0.277	0.412	-0.301	-0.403
SA1	-0.121	-0.266	-0.073	-0.128	0.249	-0.170	-0.180
SA2	-0.194	-0.241	-0.069	-0.176	0.279	-0.196	-0.263
SA3	-0.293	-0.277	-0.134	-0.308	0.313	-0.214	-0.393
SART Index	-0.046	-0.158	-0.153	-0.136	0.621	-0.048	-0.396
SART Demand	0.781	0.802	0.436	0.673	-0.468	0.725	0.561
SART Supply	0.662	0.686	0.189	0.610	-0.131	0.651	0.275
SART Understanding	0.099	-0.031	-0.003	-0.056	0.564	0.099	-0.233
Task load	0.680	0.699	0.384	0.581	-0.432	0.634	0.527
Alarm_list_support	-0.203	-0.235	-0.031	-0.316	0.332	-0.176	-0.257
Alarm_priority_support	-0.298	-0.274	-0.155	-0.329	0.390	-0.293	-0.351
Procedure support	-0.165	-0.208	-0.102	-0.118	0.508	-0.167	-0.545
Training average	0.061	0.009	0.030	-0.092	0.241	0.082	-0.071
Familiarity average	0.299	0.146	0.377	0.176	0.318	0.203	-0.008
Reaction time	-0.170	-0.116	-0.014	-0.168	-0.125	-0.066	-0.021
Response time	0.429	0.502	0.302	0.286	-0.296	0.431	0.291
Alarms silenced	0.363	0.447	0.268	0.274	-0.223	0.270	0.288
Acknowledgements	0.400	0.494	0.172	0.408	-0.277	0.373	0.298
mimics opened	0.095	0.231	-0.227	0.146	-0.128	0.042	0.041
No_of_alarms	0.527	0.580	0.379	0.424	-0.484	0.479	0.474
Accuracy	0.425	0.337	0.194	0.448	-0.224	0.375	0.325
Recovery time	0.550	0.625	0.365	0.406	-0.442	0.573	0.444
Consequence	0.220	0.305	0.149	0.187	-0.378	0.213	0.245
Overall Performance	0.255	0.312	0.213	0.166	-0.299	0.313	0.238
Error	0.323	0.403	0.202	0.340	-0.515	0.273	0.360

Table 5.19 and 5.21 present the correlation outcomes among the subjective and objective measures, revealing intricate relationships between these variables, while Table 5.20 and 5.22 check for significant correlations. There were significant correlations between NASA TLX and its dimensions, except for Performance. NASA TLX also significantly correlated with SPAM Index (r = , $p < 0.01$), SA3 - Understanding (r =, $p < 0.01$), SART demand ($p < 0.001$), SART supply ($p < 0.001$), Task load ($p < 0.001$), Alarm_list_support ($p < 0.05$), Alarm_priority_support ($p < 0.01$), Familiarity_average ($p < 0.01$), Response_time ($p < 0.001$), Alarms_silenced ($p < 0.001$), Acknowledgements ($p < 0.001$), No_of_alarms, Accuracy, Recovery_time, Consequence ($p < 0.05$), Overall Performance ($p < 0.01$), Error ($p < 0.001$).

Table 5.20 Significant difference of TLX measures. $p < 0.05$.

	TLX Index	Ment. Dem	Phys Dem	Temp. Dem	Perf	Effort	Frustration
TLX_index	1.00	**2.22E-27**	**4.34E-09**	**1.14E-24**	0.185	**8.80E-30**	**9.54E-11**
Mental_demand	**2.22E-27**	1.00	0.000	**3.92E-14**	**1.66E-04**	**9.83E-19**	**2.94E-07**
Physical_demand	**4.34E-09**	**0.000**	1.00	**0.00151**	0.170	**6.85E-05**	2.22E-05
Temporal_demand	**1.14E-24**	**3.92E-14**	**0.002**	1.00	0.0119	**4.59E-15**	**2.12E-04**
Performance	0.185	0.000	0.170	0.0119	1.00	0.0128	8.97E-07
Effort	8.80E-30	9.83E-19	6.85E-05	4.59E-15	0.0128	1.00	1.83E-06
Frustration	9.54E-11	2.94E-07	2.22E-05	2.12E-04	8.97E-07	1.83E-06	1.00
SPAM-SAGAT	0.00210	0.000183	0.174	0.00481	1.72E-05	0.00210	2.62E-05
SA1	0.227	0.00687	0.463	0.200	0.0117	0.0874	0.0696
SA2	0.0513	0.0145	0.493	0.0761	0.00448	0.0481	0.00748
SA3	0.00285	0.00487	0.180	0.00166	0.00138	0.0307	4.44E-05
SART_Index	0.644	0.113	0.124	0.171	3.43E-12	0.634	3.71E-05
SART_Demand	3.93E-22	4.38E-24	4.52E-06	8.92E-15	7.12E-07	6.83E-18	8.40E-10
SART_Supply	3.43E-14	1.75E-15	0.0577	9.81E-12	0.189	1.22E-13	0.00523
SART_Understanding	0.322	0.755	0.977	0.577	6.53E-10	0.322	0.0184
Task_load	3.93E-15	2.92E-16	6.71E-05	1.59E-10	5.85E-06	8.33E-13	1.26E-08
Alarm_list_support	0.0404	0.0174	0.761	0.00122	0.000665	0.0776	0.00917
Alarm_priority_support	0.00237	0.00537	0.120	0.000741	5.00E-05	0.00282	0.000298
Procedure_support	0.0966	0.0362	0.307	0.238	4.94E-08	0.0927	3.07E-09
Training_average	0.544	0.932	0.766	0.357	0.0149	0.412	0.476
Familiarity_average	0.00230	0.144	9.20E-05	0.0774	0.00114	0.0406	0.936
Reaction_time	0.0870	0.244	0.891	0.0922	0.210	0.509	0.836
Response_time	6.93E-06	7.67E-08	0.00202	0.00357	0.00257	6.14E-06	0.00298
Alarms_silenced	0.000173	2.42E-06	0.00652	0.00525	0.0246	0.00604	0.00335
Acknowledgements	3.17E-05	1.29E-07	0.0831	2.07E-05	0.00478	0.000111	0.00232
mimics_opened	0.340	0.0193	0.0216	0.144	0.199	0.678	0.684
No_of_alarms	1.29E-08	1.63E-10	8.73E-05	9.19E-06	2.50E-07	3.63E-07	4.94E-07
Accuracy	8.41E-06	0.000536	0.0503	2.31E-06	0.0239	0.000102	0.000852
Recovery_time	2.15E-09	2.22E-12	0.000163	2.28E-05	3.23E-06	3.12E-10	2.93E-06
Consequence	0.0265	0.00182	0.136	0.0596	8.78E-05	0.0319	0.0131
Overall Performance	0.00975	0.00143	0.0319	0.0961	0.00231	0.00133	0.0162
Error	0.000945	2.74E-05	0.0415	0.000479	3.01E-08	0.00554	0.000199

Table 5.21 Correlations with situational awareness measures. $p < 0.05$.

	SPAM Index	SA1	SA2	SA3	SART_Index	SART_Demand	SART_Supply	SART_Understanding
TLX_index	-0.301	-0.121	-0.194	-0.293	-0.046	0.781	0.662	0.099
Mental_demand	-0.362	-0.266	-0.241	-0.277	-0.158	0.802	0.686	-0.031
Physical_demand	-0.136	-0.073	-0.069	-0.134	-0.153	0.436	0.189	-0.003
Temporal_demand	-0.277	-0.128	-0.176	-0.308	-0.136	0.673	0.610	-0.056
Performance	0.412	0.249	0.279	0.313	0.621	-0.468	-0.131	0.564
Effort	-0.301	-0.170	-0.196	-0.214	-0.048	0.725	0.651	0.099
Frustration	-0.403	-0.180	-0.263	-0.393	-0.396	0.561	0.275	-0.233
SPAM Index	1.000	0.522	0.716	0.755	0.390	-0.412	-0.266	0.283
SA1	0.522	1.000	0.201	0.335	0.220	-0.294	-0.183	0.162
SA2	0.716	0.201	1.000	0.220	0.146	-0.267	-0.244	0.095
SA3	0.755	0.335	0.220	1.000	0.436	-0.328	-0.174	0.344
SART_Index	0.390	0.220	0.146	0.436	1.000	-0.352	0.312	0.803
SART_Demand	-0.412	-0.294	-0.267	-0.328	-0.352	1.000	0.652	-0.101
SART_Supply	-0.266	-0.183	-0.244	-0.174	0.312	0.652	1.000	0.246
SART_Understanding	0.283	0.162	0.095	0.344	0.803	-0.101	0.246	1.000
Task_load	-0.447	-0.279	-0.257	-0.392	-0.246	0.760	0.581	-0.058
Alarm_list_support	0.161	0.227	0.025	0.220	0.230	-0.305	-0.288	0.303
Alarm_priority_support	0.180	0.231	0.063	0.231	0.298	-0.379	-0.210	0.244
Procedure_support	0.344	0.194	0.184	0.333	0.486	-0.398	-0.079	0.315
Training_average	0.115	-0.013	0.080	0.183	0.407	-0.062	0.108	0.431
Familiarity_average	0.031	-0.120	-0.017	0.127	0.243	0.059	0.142	0.315
Reaction_time	-0.092	0.019	-0.113	-0.081	-0.036	-0.136	-0.142	-0.105
Response_time	-0.366	-0.244	-0.252	-0.234	-0.238	0.564	0.368	-0.015
Alarms_silenced	-0.344	-0.209	-0.155	-0.287	-0.247	0.380	0.254	-0.105
Acknowledgements	-0.248	-0.188	-0.068	-0.270	-0.377	0.545	0.284	-0.230
mimics_opened	-0.068	0.029	-0.026	-0.088	-0.075	0.186	0.153	0.010
No_of_alarms	-0.449	-0.224	-0.272	-0.365	-0.470	0.724	0.387	-0.190
Accuracy	-0.290	-0.103	-0.094	-0.296	-0.229	0.463	0.364	-0.052
Recovery_time	-0.401	-0.279	-0.228	-0.306	-0.265	0.705	0.480	-0.003
Consequence	-0.146	-0.082	-0.091	-0.175	-0.191	0.368	0.210	-0.094
Overall Performance	-0.053	-0.064	-0.033	-0.049	-0.034	0.311	0.175	0.024
Error	-0.263	-0.109	-0.166	-0.274	-0.353	0.516	0.292	-0.235

Situational Awareness Measures (SART and SPAM):

Table 5.23 presents the correlations between other variables and the situational awareness measures, while Table 5.24 shows if there are significant correlations. The SPAM index significantly correlates positively with its dimensions SA1 - 3 and negatively with the SART measures and procedure support. In addition, it significantly correlates negatively with the NASA TLX Index and its dimensions, except for Physical demand. It likewise negatively correlates significantly with task load, Response time, Alarms silenced, Acknowledgements, No of alarms, Accuracy, Recovery time and Error.

SA1 exhibits a p-value of 0.227 with the TLX Index, indicating no significant correlation. However, it shows a significant relationship with Mental Demand (p = 0.0069). Similarly, SA2 does not correlate significantly with the TLX Index (p = 0.051) but has a significant correlation with Mental Demand (p = 0.0145), performance, effort and frustration. SA2 significantly correlates with SART demand and supply, task load, response time no of alarms and recovery time. In contrast to SA 1 and 2, SA3 correlates significantly with the TLX Index (p = 0.0029) and Mental Demand (p = 0.0049), including other TLX dimensions except Physical demand, like with the SPAM Index. SA3 also correlates with SART Index, SART demand and Understanding, task load, alarm list support, Alarm_priority_support, Procedure_support, Response_time, Alarms_silenced, Acknowledgements, No_of_alarms, Accuracy, Recovery_time and error.

The analysis of significant correlations for the SART Index reveals several important relationships. A strong correlation is noted with Task Load (p = 0.0125) and Procedure Support (p = 2.21E-07). Additionally, Alarm Priority Support (p = 0.0023) and Training Average (p = 2.19E-05) show notable correlations. Other significant correlations include Alarm List Support (p = 0.0201), Response Time (p = 0.0161), Alarms Silenced (p = 0.0123), Acknowledgements (p = 9.45E-05), and Accuracy (p = 0.0205). Furthermore, a very strong correlation is observed between the Number of Alarms (p = 6.37E-07) and Error (p = 0.00027).

Table 5.22 Significance of situational awareness measures. $p < 0.05$.

	SPAM Index	SA1	SA2	SA3	SART_Index	SART_Demand	SART_Supply	SART_Underst.
TLX_index	0.00210	0.227	0.0513	0.00285	0.644	3.93E-22	3.43E-14	0.322
Mental_demand	0.000183	0.00687	0.0145	0.00487	0.113	4.38E-24	1.75E-15	0.755
Physical_demand	0.174	0.463	0.493	0.180	0.124	4.52E-06	0.0577	0.977
Temporal_demand	0.00481	0.200	0.0761	0.00166	0.171	8.92E-15	9.81E-12	0.577
Performance	1.72E-05	0.0117	0.00448	0.00138	3.43E-12	7.12E-07	0.189	6.53E-10
Effort	0.00210	0.0874	0.0481	0.0307	0.634	6.83E-18	1.22E-13	0.322
Frustration	2.62E-05	0.0696	0.00748	4.44E-05	3.71E-05	8.40E-10	0.00523	0.0184
SPAM Index	1.00	1.87E-08	2.85E-17	5.02E-20	5.01E-05	1.68E-05	0.00687	0.00391
SA1	1.87E-08	1.00	0.0428	0.000571	0.0263	0.00275	0.0652	0.103
SA2	2.85E-17	0.0428	1.00	0.0262	0.142	0.00670	0.0134	0.344
SA3	5.02E-20	0.000571	0.0262	1.00	4.51E-06	0.000757	0.0796	0.000393
SART_Index	5.01E-05	0.0263	0.142	4.51E-06		0.000286	0.00140	3.61E-24
SART_Demand	1.68E-05	0.00275	0.00670	0.000757	0.000286		1.21E-13	0.314
SART_Supply	0.00687	0.0652	0.0134	0.0796	0.00140	1.21E-13		0.0126
SART_Understanding	0.00391	0.103	0.344	0.000393	3.61E-24	0.314	0.0126	
Task_load	2.50E-06	0.00445	0.00908	4.52E-05	0.0125	2.19E-20	1.49E-10	0.564
Alarm_list_support	0.106	0.0218	0.806	0.0263	0.0201	0.00183	0.00336	0.00195
Alarm_priority_support	0.0701	0.0196	0.526	0.0197	0.00231	8.68E-05	0.0339	0.0136
Procedure_support	0.000395	0.0508	0.0637	0.000617	2.21E-07	3.45E-05	0.430	0.00124
Training_average	0.250	0.894	0.424	0.0661	2.19E-05	0.535	0.280	5.98E-06
Familiarity_average	0.754	0.231	0.866	0.205	0.0138	0.555	0.156	0.00124
Reaction_time	0.360	0.849	0.259	0.418	0.716	0.174	0.154	0.293
Response_time	0.000155	0.0136	0.0106	0.0181	0.0161	6.89E-10	0.000144	0.883
Alarms_silenced	0.000400	0.0349	0.119	0.00345	0.0123	8.25E-05	0.00994	0.293
Acknowledgements	0.0120	0.0590	0.498	0.00604	9.45E-05	3.28E-09	0.00382	0.0198
mimics_opened	0.495	0.771	0.797	0.381	0.456	0.0617	0.123	0.921
No_of_alarms	2.19E-06	0.0234	0.00559	0.000161	6.37E-07	7.83E-18	5.88E-05	0.0560
Accuracy	0.00312	0.304	0.346	0.00250	0.0205	9.87E-07	0.000168	0.602
Recovery_time	2.98E-05	0.00455	0.0213	0.00177	0.00710	1.28E-16	3.26E-07	0.976
Consequence	0.143	0.413	0.361	0.0790	0.0544	0.000141	0.0338	0.347
Overall Performance	0.600	0.525	0.743	0.628	0.736	0.00145	0.0793	0.811
Error	0.00756	0.274	0.0956	0.00533	0.000268	2.91E-08	0.00288	0.0173

The significant correlations for SART_Demand illustrate critical relationships affecting situational awareness demand. An extremely strong positive correlation with Task Load (p = 2.19E-20) highlights that perceived demands on situational awareness increase sharply with workload. Its strong negative correlation with Alarm List Support (p = 0.0018) and Alarm Priority Support (p = 8.68E-05) suggest that better prioritisation and support can reduce SART demands. Notably, Procedure Support shows a strong negative correlation (p = 3.45E-05), indicating that structured procedures are essential for managing demands. Conversely, both Training Average (p = 0.5349) and Familiarity Average (p = 0.5546) exhibit weak correlations, suggesting a minimal impact on perceived demands. A very strong positive correlation with Response Time (p = 6.89E-10) indicates that increased SART demands significantly increase response times. Additional strong correlations include Alarms Silenced (p = 8.25E-05), Accuracy (p = 9.87E-07), and Consequence (p = 0.00014). The positive correlation with Recovery Time (p = 1.28E-16) further emphasises that increased demands prolong recovery. Lastly, a very strong positive correlation with Error (p = 2.90E-08) suggests that higher demands are closely linked to an increased likelihood of errors. Overall, these findings highlight the significant impact of situational awareness demands on task performance and error rates.

The significant correlations for SART_Supply reveal important interactions affecting situational awareness. A strong positive correlation with Task Load (p = 1.49E-10) indicates that increased supply is associated with higher task load. Effective Alarm List Support (p = 0.0034) and Alarm Priority Support (p = 0.0339) show a negative significant correlation. However, correlations with Procedure Support (p = 0.4296), Training Average (p = 0.2808), and Familiarity Average (p = 0.1559) are weak, indicating the limited impact of these factors on situational supply. In terms of the objective metrics, a significant positive correlation with Response Time (p = 0.00014) implies that situational supply can positively influence response times. Additionally, Alarms Silenced (p = 0.0099) and Acknowledgements (p = 0.0038) demonstrate that better supply aids in effective alarm handling. A very strong correlation with Accuracy (p = 0.000168) indicates that adequate supply positively affects task accuracy. The correlation with Recovery Time (p = 3.26E-07) suggests that supply significantly influences recovery durations, while a moderate positive correlation with Error (p = 0.0029) indicates that better SART supply is associated with better outcomes.

For SART_Understanding, a weak correlation with Task Load (p = 0.5635) indicates that understanding does not significantly affect task load. However, a significant positive correlation with Alarm List Support (p = 0.0019) highlights that effective alarm support improves SART understanding. The moderate correlation with Alarm Priority Support (p = 0.0136) suggests that prioritisation aids in enhancing understanding, while a strong positive correlation with Procedure Support (p = 0.0012) indicates that clear procedures are vital for improving understanding. Additionally, a strong positive correlation with the Training Average (p = 5.98E-06) implies that training significantly enhances understanding, and familiarity shows a similar strong correlation (p = 0.0012), indicating it substantially contributes to understanding. In the objective metrics, weak correlations with Reaction Time (p = 0.2933), Response Time (p = 0.8831), and Alarms Silenced (p = 0.2935) suggest a limited impact of understanding on these areas. However, a moderate negative correlation with Error (p = 0.0173) indicates that understanding somewhat influences error, while other metrics, including Overall Performance (p = 0.8105) and Accuracy (p = 0.6016), show negligible influence.

Other Subjective Measures: Table 5.23 shows the correlations between the other subjective variables with the others while Table 5.24 shows if there are significant differences.

Table 5.23 Correlation with other subjective measures.

	Task load	Alarm list Supp.	Alarm Priority Supp	Procedure Supp	Training_average	Familiarity_average
TLX_index	0.680	-0.203	-0.298	-0.165	0.061	0.299
Mental_demand	0.699	-0.235	-0.274	-0.208	0.009	0.146
Physical_demand	0.384	-0.031	-0.155	-0.102	0.030	0.377
Temporal_demand	0.581	-0.316	-0.329	-0.118	-0.092	0.176
Performance	-0.432	0.332	0.390	0.508	0.241	0.318
Effort	0.634	-0.176	-0.293	-0.167	0.082	0.203
Frustration	0.527	-0.257	-0.351	-0.545	-0.071	-0.008
SPAM Index	-0.447	0.161	0.180	0.344	0.115	0.031
SA1	-0.279	0.227	0.231	0.194	-0.013	-0.120
SA2	-0.257	0.025	0.063	0.184	0.080	-0.017
SA3	-0.392	0.220	0.231	0.333	0.183	0.127
SART_Index	-0.246	0.230	0.298	0.486	0.407	0.243
SART_Demand	0.760	-0.305	-0.379	-0.398	-0.062	0.059
SART_Supply	0.581	-0.288	-0.210	-0.079	0.108	0.142
SART_Understanding	-0.058	0.303	0.244	0.315	0.431	0.315
Task_load	1.000	-0.185	-0.353	-0.395	-0.001	0.043
Alarm_list_support	-0.185	1.000	0.594	0.358	0.225	0.016
Alarm_priority_support	-0.353	0.594	1.000	0.555	-0.009	-0.016
Procedure_support	-0.395	0.358	0.555	1.000	0.100	0.287
Training_average	-0.001	0.225	-0.009	0.100	1.000	0.102
Familiarity_average	0.043	0.016	-0.016	0.287	0.102	1.000
Reaction_time	-0.152	0.193	0.171	-0.003	-0.078	0.103
Response_time	0.518	-0.010	0.043	-0.106	-0.087	0.045
Alarms_silenced	0.448	-0.006	-0.032	-0.115	-0.063	0.037
Acknowledgements	0.463	-0.188	-0.222	-0.226	-0.204	-0.104
mimics_opened	0.166	-0.044	0.033	-0.044	-0.159	-0.191
No_of_alarms	0.648	-0.143	-0.231	-0.398	-0.066	-0.038
Accuracy	0.420	-0.090	-0.055	-0.249	0.002	-0.014
Recovery_time	0.673	-0.113	-0.115	-0.280	-0.071	0.002
Consequence	0.370	-0.218	-0.269	-0.309	0.016	-0.127
Overall Performance	0.323	0.020	-0.059	-0.026	-0.105	-0.071
Error	0.490	-0.314	-0.378	-0.457	-0.083	-0.150

Table 5.24 Significance with other subjective measures.

	Task load	Alarm list Supp.	Alarm Priority Supp	Procedure Supp	Training_average	Familiarity_average
TLX_index	3.93E-15	0.0404	0.00237	0.0966	0.544	0.00230
Mental_demand	2.92E-16	0.0174	0.00537	0.0362	0.932	0.144
Physical_demand	6.71E-05	0.761	0.120	0.307	0.766	9.20E-05
Temporal_demand	1.59E-10	0.00122	0.000741	0.238	0.357	0.0774
Performance	5.85E-06	0.000665	5.00E-05	4.94E-08	0.0149	0.00114
Effort	8.33E-13	0.0776	0.00282	0.0927	0.412	0.0406
Frustration	1.26E-08	0.00917	0.000298	3.07E-09	0.476	0.936
SPAM-SAGAT	2.50E-06	0.106	0.0701	0.000395	0.250	0.754
SA1	0.00445	0.0218	0.0196	0.0508	0.894	0.231
SA2	0.00908	0.806	0.526	0.0637	0.424	0.866
SA3	4.52E-05	0.0263	0.0197	0.000617	0.0661	0.205
SART_Index	0.0125	0.0201	0.00231	2.21E-07	2.19E-05	0.0138
SART_Demand	2.19E-20	0.00183	8.68E-05	3.45E-05	0.535	0.555
SART_Supply	1.49E-10	0.0034	0.034	0.430	0.281	0.156
SART_Understanding	0.564	0.00195	0.0136	0.00124	5.98E-06	0.00124
Task_load		0.0629	0.000279	3.93E-05	0.989	0.664
Alarm_list_support	0.063		4.476E-11	0.0002	0.023	0.877
Alarm_priority_support	0.000279	4.48E-11		1.39E-09	0.932	0.871
Procedure_support	3.93E-05	0.000216	1.39E-09		0.318	0.00351
Training_average	0.989	0.0231	0.932	0.318		0.306
Familiarity_average	0.664	0.8767	0.871	0.004	0.306	
Reaction_time	0.128	0.0522	0.0850	0.980	0.439	0.303
Response_time	2.49E-08	0.922	0.667	0.289	0.387	0.655
Alarms_silenced	2.35E-06	0.954	0.749	0.251	0.529	0.716
Acknowledgements	9.84E-07	0.0588	0.0250	0.0222	0.0400	0.299
mimics_opened	0.0950	0.658	0.743	0.658	0.110	0.0541
No_of_alarms	1.84E-13	0.152	0.0194	3.36E-05	0.508	0.702
Accuracy	1.10E-05	0.369	0.582	0.0116	0.981	0.890
Recovery_time	8.87E-15	0.259	0.251	0.00434	0.481	0.987
Consequence	0.000128	0.0280	0.00632	0.00156	0.872	0.204
Overall Performance	0.000943	0.839	0.555	0.795	0.294	0.480
Error	1.67E-07	0.00129	8.77E-05	1.37E-06	0.404	0.132

The negative correlation between Task Load and Alarm List Support is significant (p = 0.0629). Similarly, a strong negative correlation with Alarm Priority Support (p = 0.0003) suggests prioritisation decreases task loads. The significant negative correlation with Procedure Support (p = 0.00004) underscores the need for structured procedures to navigate and reduce task loads. However, the weak correlations with the Training Average (p = 0.9894) and Familiarity Average (p = 0.6643) suggest these factors have a limited impact on managing task demands. In terms of objective measures, task load non-significant correlation with Reaction Time (p = 0.1276) and Response Time (p = 0.1276) indicates minimal effects. Conversely, it has a very significant positive correlation with the Number of Alarms (p = 1.8437E-13) and Accuracy (p = 0.00001), showing that higher task loads trigger more alarms and result in higher mean square errors (accuracy). Task load significantly correlates positively with Recovery Time (p = 8.8714E-15), Consequence (p = 0.0001), Overall Performance (p = 0.0009), and Error (p = 1.6674E-07), suggesting increased task load leads to longer recovery times and higher errors, ultimately affecting overall performance.

The positive correlation between Alarm List Support and Alarm Priority Support is highly significant (p < 0.00001), indicating that effective alarm listing is closely tied to prioritisation. Similarly, it shows a notable positive correlation with Procedure Support (p = 0.0002). However, the correlation with Training Average (p = 0.9894) and Familiarity Average (p = 0.8766) is weak, indicating minimal impact on alarm list support. The correlation with Reaction Time (p = 0.0522) is low, while the correlation with Response Time (p = 0.9217) is weak. Alarm list support shows limited influence on the number of alarms (p = 0.1518) and Accuracy (p = 0.3690). It has a negative significant correlation with consequences (p = 0.0279) and error (p = 0.0013).

The correlation between Alarm Priority Support and Procedure Support is very strong (p < 0.00001), indicating that prioritising alarms significantly enhances the effectiveness of structured procedures. However, the correlation with the Training Average is weak (p = 0.9318), suggesting that training does not substantially impact alarm prioritisation. The relationship with the Familiarity Average is weak (p = 0.8710), indicating that familiarity does not notably influence this aspect of alarm management. In terms of performance metrics, the correlation with Reaction Time is weak (p = 0.0850), while it shows negligible impact on Response Time (p = 0.6671). Alarm prioritisation does not significantly affect the number of alarms (p = 0.0194) but has a moderate correlation with Accuracy (p = 0.5816) and a notable impact on Acknowledgements (p = 0.0250).

Additionally, it is linked to better alarm management (p = 0.7494) and influences the perception of consequences (p = 0.0063). The correlation with Overall Performance is weak (p = 0.5549), while it has a significant correlation with Error (p = 0.00009), highlighting its role in reducing errors during tasks.

The correlation between Procedure Support and Training Average is weak (p = 0.3180). On the contrary, the relationship with the Familiarity Average is strong (p = 0.0035). In terms of the objective metrics, Procedure Support shows no significant correlation with Reaction Time (p = 0.9795), Response Time (p = 0.2888), and Alarms Silenced (p = 0.2515). Acknowledgements show a negative significant correlation (p = 0.0222). The correlation with Mimics Opened is not significant (p = 0.6576), suggesting that procedures have little effect on the use of mimic displays. The negative relationship with the Number of Alarms is significant (p = 0.00003), indicating that effective procedures can help reduce no of alarms annunciated. Accuracy shows a negative correlation (p = 0.0116), suggesting procedure support influence on mean square error. There is a notable negative correlation with Recovery Time (p = 0.0043), indicating that effective procedures may shorten recovery durations. The negative correlation with Consequence is significant (p = 0.0016), suggesting that well-structured procedures influence consequences. Procedure support has no significant correlation with Overall Performance (p = 0.7949), but shows a strong negative correlation with Error (p = 0.000001) indicates that effective procedural support can significantly reduce error.

The correlation between Training Average and Familiarity Average is weak (p = 0.3057). Similarly, Reaction Time (p = 0.4386) and response time (p = 0.3867) indicate that the amount of training received does not notably affect reaction or response times. Training Average also showed no significant correlation with Alarms Silenced (p = 0.5288), Mimics Opened (p = 0.1100), Number of Alarms (p = 0.5081), Accuracy (p = 0.9811), Recovery Time (p = 0.4811), Consequence (p = 0.8723), Overall Performance (p = 0.2935) and Error (p = 0.4042). The negative correlation with Acknowledgements is significant (p = 0.0396), indicating that better training may improve alarm acknowledgement rates.

No significant correlation exists between Familiarity Average and Reaction Time (p = 0.3028), indicating that increased familiarity does not significantly enhance the speed of reactions. A similar non-significant correlation was observed with Response Time (p = 0.6546), Alarms Silenced (p = 0.7155), Mimics Opened (p = 0.0541) Acknowledgements (p = 0.2995), Alarms (p = 0.7022), Accuracy (p = 0.8901), Recovery Time (p = 0.9868),

Consequence (p = 0.2037), Overall Performance is weak (p = 0.4799) and Error (p = 0.1323). Generally, both training and familiarity averages do not seem to significantly with most of the behaviour and performance Variables and other subjective variables except for a few. This implies that these variables might not significantly impact the outcomes of the test.

Objective Measures:

Table 5.25 show the correlation between the objective measures and other variables, while Table 5.26 show if a significant difference exists.

Table 5.25 Correlation with objective measures.

	Reaction	Response	Silence	Ack.	mimics	alarms	Acc.	Rec. Time	Conseq.	Overall Perf.	Error
TLX_index	-0.170	0.429	0.363	0.400	0.095	0.527	0.425	0.550	0.220	0.255	0.323
Mental_demand	-0.116	0.502	0.447	0.494	0.231	0.580	0.337	0.625	0.305	0.312	0.403
Physical_demand	-0.014	0.302	0.268	0.172	-0.227	0.379	0.194	0.365	0.149	0.213	0.202
Temporal_demand	-0.168	0.286	0.274	0.408	0.146	0.424	0.448	0.406	0.187	0.166	0.340
Performance	-0.125	-0.296	-0.223	-0.277	-0.128	-0.484	-0.224	-0.442	-0.378	-0.299	-0.515
Effort	-0.066	0.431	0.270	0.373	0.042	0.479	0.375	0.573	0.213	0.313	0.273
Frustration	-0.021	0.291	0.288	0.298	0.041	0.474	0.325	0.444	0.245	0.238	0.360
SPAM Index	-0.092	-0.366	-0.344	-0.248	-0.068	-0.449	-0.290	-0.401	-0.146	-0.053	-0.263
SA1	0.019	-0.244	-0.209	-0.188	0.029	-0.224	-0.103	-0.279	-0.082	-0.064	-0.109
SA2	-0.113	-0.252	-0.155	-0.068	-0.026	-0.272	-0.094	-0.228	-0.091	-0.033	-0.166
SA3	-0.081	-0.234	-0.287	-0.270	-0.088	-0.365	-0.296	-0.306	-0.175	-0.049	-0.274
SART_Index	-0.036	-0.238	-0.247	-0.377	-0.075	-0.470	-0.229	-0.265	-0.191	-0.034	-0.353
SART_Demand	-0.136	0.564	0.380	0.545	0.186	0.724	0.463	0.705	0.368	0.311	0.516
SART_Supply	-0.142	0.368	0.254	0.284	0.153	0.387	0.364	0.480	0.210	0.175	0.292
SART_Understanding	-0.105	-0.015	-0.105	-0.230	0.010	-0.190	-0.052	-0.003	-0.094	0.024	-0.235
Task_load	-0.152	0.518	0.448	0.463	0.166	0.648	0.420	0.673	0.370	0.323	0.490
Alarm_list_support	0.193	-0.010	-0.006	-0.188	-0.044	-0.143	-0.090	-0.113	-0.218	0.020	-0.314
Alarm_priority_support	0.171	0.043	-0.032	-0.222	0.033	-0.231	-0.055	-0.115	-0.269	-0.059	-0.378
Procedure_support	-0.003	-0.106	-0.115	-0.226	-0.044	-0.398	-0.249	-0.280	-0.309	-0.026	-0.457
Training_average	-0.078	-0.087	-0.063	-0.204	-0.159	-0.066	0.002	-0.071	0.016	-0.105	-0.083
Familiarity_average	0.103	0.045	0.037	-0.104	-0.191	-0.038	-0.014	0.002	-0.127	-0.071	-0.150
Reaction_time	1.000	0.110	-0.064	-0.084	-0.119	-0.107	-0.176	0.092	-0.197	0.232	-0.243
Response_time	0.110	1.000	0.574	0.481	0.226	0.729	0.470	0.827	0.258	0.351	0.252
Alarms_silenced	-0.064	0.574	1.000	0.400	0.145	0.563	0.298	0.552	0.174	0.172	0.263
Acknowledgements	-0.084	0.481	0.400	1.000	0.172	0.645	0.344	0.514	0.170	0.110	0.259
mimics_opened	-0.119	0.226	0.145	0.172	1.000	0.190	0.337	0.205	0.112	0.017	0.106
No_of_alarms	-0.107	0.729	0.563	0.645	0.190	1.000	0.593	0.780	0.400	0.192	0.546
Accuracy	-0.176	0.470	0.298	0.344	0.337	0.593	1.000	0.471	0.080	-0.093	0.305
Recovery_time	0.092	0.827	0.552	0.514	0.205	0.780	0.471	1.000	0.450	0.592	0.474
Consequence	-0.197	0.258	0.174	0.170	0.112	0.400	0.080	0.450	1.000	0.417	0.817
Overall_Performance	0.232	0.351	0.172	0.110	0.017	0.192	-0.093	0.592	0.417	1.000	0.283
Error	-0.243	0.252	0.263	0.259	0.106	0.546	0.305	0.474	0.817	0.283	1.000

Table 5.26 Significance of objective measures.

	Reaction	Response	Silence	Ack.	mimics	alarms	Acc.	Rec. Time	Conseq.	Overall Perf.	Error
TLX_index	0.0870	6.93E-06	0.000173	3.17E-05	0.340	1.29E-08	8.41E-06	2.15E-09	0.0265	0.00975	0.000945
Mental_demand	0.244	7.67E-08	2.42E-06	1.29E-07	0.0193	1.63E-10	0.000536	2.22E-12	0.00182	0.00143	2.74E-05
Physical_demand	0.891	0.00202	0.00652	0.0831	0.0216	8.73E-05	0.0503	0.000163	0.136	0.0319	0.0415
Temporal_demand	0.0922	0.00357	0.00525	2.07E-05	0.144	9.19E-06	2.31E-06	2.28E-05	0.0596	0.0961	0.000479
Performance	0.210	0.00257	0.0246	0.00478	0.199	2.50E-07	0.0239	3.23E-06	8.78E-05	0.00231	3.01E-08
Effort	0.509	6.14E-06	0.00604	0.000111	0.678	3.63E-07	0.000102	3.12E-10	0.0319	0.00133	0.00554
Frustration	0.836	0.00298	0.00335	0.00232	0.684	4.94E-07	0.000852	2.93E-06	0.0131	0.0162	0.000199
SPAM-SAGAT	0.360	0.000155	0.000400	0.01197	0.495	2.19E-06	0.00312	2.98E-05	0.143	0.600	0.00756
SA1	0.849	0.0136	0.0349	0.0590	0.771	0.0234	0.304	0.00455	0.413	0.525	0.274
SA2	0.259	0.0106	0.119	0.498	0.797	0.00559	0.346	0.0213	0.361	0.743	0.0956
SA3	0.418	0.0181	0.00345	0.00604	0.381	0.000161	0.00250	0.00177	0.0790	0.628	0.00533
SART_Index	0.716	0.0161	0.0123	9.45E-05	0.456	6.37E-07	0.0205	0.00710	0.0544	0.736	0.000268
SART_Demand	0.174	6.89E-10	8.25E-05	3.28E-09	0.0620	7.83E-18	9.87E-07	1.28E-16	0.000141	0.00145	2.91E-08
SART_Supply	0.154	0.000144	0.00994	0.00382	0.123	5.88E-05	0.000168	3.26E-07	0.0338	0.0793	0.00288
SART_Understanding	0.293	0.883	0.293	0.0198	0.921	0.0559	0.602	0.976	0.347	0.811	0.0173
Task_load	0.128	2.49E-08	2.35E-06	9.84E-07	0.0946	1.84E-13	1.10E-05	8.87E-15	0.000128	0.000943	1.67E-07
Alarm_list_support	0.0522	0.922	0.954	0.0588	0.658	0.151	0.369	0.259	0.0279	0.839	0.00129
Alarm_priority_support	0.0850	0.667	0.749	0.0250	0.743	0.0194	0.582	0.251	0.00632	0.555	8.77E-05
Procedure_support	0.980	0.289	0.251	0.0222	0.658	3.36E-05	0.0116	0.00434	0.00156	0.795	1.37E-06
Training_average	0.439	0.387	0.529	0.0396	0.110	0.508	0.981	0.481	0.872	0.294	0.404
Familiarity_average	0.303	0.655	0.716	0.299	0.0541	0.702	0.890	0.987	0.204	0.480	0.132
Reaction_time		0.270	0.525	0.402	0.235	0.285	0.0763	0.356	0.0468	0.0188	0.0137
Response_time	0.270		2.85E-10	3.08E-07	0.0225	3.66E-18	6.09E-07	9.97E-27	0.00871	0.000301	0.0107
Alarms_silenced	0.525	2.85E-10		3.13E-05	0.145	7.11E-10	0.00235	1.80E-09	0.0799	0.0839	0.00768
Acknowledgements	0.402	3.08E-07	3.13E-05		0.0844	2.41E-13	0.000397	3.20E-08	0.0878	0.271	0.00851
mimics_opened	0.235	0.0225	0.145	0.0844		0.0559	0.000529	0.0385	0.263	0.867	0.290
No_of_alarms	0.285	3.66E-18	7.11E-10	2.41E-13	0.0559		5.18E-11	3.10E-05	0.0529	3.01E-09	
Accuracy	0.0763	6.09E-07	0.00235	0.000397	0.000529	5.18E-11		5.78E-07	0.423	0.355	0.00184
Recovery_time	0.356	9.97E-27	1.80E-09	3.20E-08	0.0385	4.42E-22	5.78E-07		2.05E-06	5.62E-11	
Consequence	0.0468	0.00871	0.0799	0.0878	0.263	3.10E-05	0.424	2.05E-06		1.33E-05	1.13E-25
Overall Performance	0.0188	0.000301	0.0840	0.271	0.867	0.0529	0.355	5.62E-11	1.33E-05		0.00390
Error	0.0137	0.0107	0.00768	0.00851	0.290	3.01E-09	0.00184	4.95E-07	1.13E-25	0.00390	

The analysis of p-values reveals that Reaction Time exhibits significant correlations with other behaviour and performance variables. It positively correlates with Response Time (p = 2.85E-10) and negatively with Alarms Silenced (p = 3.13E-05). As expected, longer reaction times will lead to longer response times and the likelihood of less alarm silencing. Similarly, there is a significant negative correlation with Acknowledgements (p = 0.000397), suggesting longer reactions lead to fewer acknowledgements, as seen with alarm silencing. Reaction Time negatively correlates significantly with Number of Alarms (p = 1.84E-18) and Accuracy (p = 0.000529). Furthermore, it has a very strong positive correlation with Recovery Time (p = 9.97E-27), just like with response time. Overall Performance (p = 0.0003) positively correlated significantly with reaction time. However, negative significant correlations were observed with Consequence (p = 0.0087) and Error (p = 0.0137); no significant correlation was observed with Mimics Opened (p = 0.235).

The analysis of Response Time shows very strong positive correlation with the other behavioural and performance variables, Mimics Opened (p = 0.0225), Alarms Silenced (p = 2.85E-10), Acknowledgements (p = 0.0004), Number of Alarms (p = 7.11E-10), Accuracy (p = 6.09E-07), Recovery Time (p = 9.97E-27), Consequence (p = 0.0087) and Overall Performance (p = 0.0003), with the exception of reaction time.

The analysis of Alarms Silenced indicates significant correlations with Acknowledgements (p = 0.00003), Number of Alarms (p = 1.79E-09), Accuracy (p = 0.0023), Recovery Time (p = 3.20E-08) and Error (p = 0.0077). There was no significant correlation with Mimics Opened (p = 0.1451), Consequence (p = 0.0799) and Overall Performance (p = 0.0839).

The analysis of acknowledgement showed no significant correlations with mimics opened (p = 0.0844), Consequence (p = 0.878) and Error (p = 0.271). However, there are positive significant correlations with the number of alarms, Accuracy, Recovery time and overall performance.

Mimics Opened showed a positive correlation with Accuracy (p = 0.0005) and recovery time (p = 0.0385). However, it did not significantly correlate with no alarms, consequences, overall performance and error.

There is a significant negative correlation between the number of alarms and Accuracy (p = 5.18E-11), Recovery Time (p = 4.42E-22), and Error (p = 3.01E-09). There was no significant correlation with Overall performance (p = 0.0529).

The analysis of Accuracy shows a significant positive correlation between Recovery Time (p = 0.000000577) and Error (p = 0.0018). There was no significant correlation between accuracy and Consequence and Overall performance.

The analysis of Recovery Time reveals a positive significant correlation with Consequence (p = 0.00002048), Overall Performance (p = 5.62E-11) and Error (p = 0.000000495).

The positive correlation between Consequence and Overall Performance is significant (p-value of 0.00001332). Additionally, there is a strong positive significant correlation between Consequence and Error (p = 1.13E-25), indicating that higher consequences are associated with higher error rates. The relationship between Overall Performance and Error shows a significant positive correlation with a p-value of 0.003895207.

Correlation of the eye-tracking measures:

Table 5.27 displays the correlation matrix for the eye-tracking measures alongside other variables for Scenario 3 based on the selected participants. Whereas Table 5.28 highlights the significant correlations. In Table 5.28, significant correlations are denoted with asterisks and marked in bold, while those nearing significance (0.05 - 0.09) are italicised.

Tables 5.27 and 5.28 demonstrate a few noteworthy correlations between eye-tracking metrics and various other variables. Particularly, SA2 (understanding) is significantly negatively correlated with the number of fixations (r = -0.812, p = 0.0143), suggesting that an increased number of fixations may reduce situation understanding. The number of fixations also shows a significant negative correlation with fixation duration (p = 0.0465). Additionally, fixation duration is nearly significantly negatively correlated with the training average (p = 0.0760). The average fixation duration appears to lack a significant correlation with non-eye-tracking variables.

Additionally, the SART Index shows a significant negative correlation with the number of visits (r = -0.729, p = 0.0403), and perhaps more from the SART Understanding, suggesting that the number of visits can describe *objective situational awareness*. Invariably, an increase in the number of visits to the procedures area of interest decreases objective situational awareness and would lead to increased objective workload (confirmed by loading with the mental demand. The number of acknowledgements significantly correlates with visits (r= 0.756, p = 0.0299) but positively. Also, 'the number of acknowledgements' has been established to correlate significantly with the SART Index but negatively, just like 'no of visits' with the 'SART

Index'. So the *no of visits* and *acknowledgement* both share a similar relationship and nature of the impact on objective situational awareness.

Furthermore, a positively significant relationship exists between mimics opened and average saccade amplitude (r = 0.764, p = 0.0274). A larger Saccade amplitude in literature is known to indicate increasing search difficulty. Therefore, its relationship/factor loading with a number of mimics opened categories both to describe the construct - *Search difficulty* when using the procedure. The average pupil diameter (r = 0.878, p = 0.00412) and average saccade amplitude (r = 0.873, p = 0.00466), suggesting that longer recovery times reflect increasing cognitive workload. This aligns with the factor analysis, which categorised recovery time to describe the construct *Workload*. Pupil diameter is also positively significantly correlated with average familiarity (r = 0.752, p = 0.031) and average saccade amplitude (p = 0.0243). Average saccade amplitude, likewise, positively correlates significantly with saccade duration (p = 0.0282).

In addition to these significant correlations, some of the observed close-to-significant correlations are noted with italicised p-values. The correlation between effort and visits (p = 0.0932) hints at a possible negative relationship between the amount of effort and no of visits to the procedure AOI. Similarly, the training average shows a trend with average fixation duration (p = 0.0760), suggesting that longer fixations may describe a learning process. The number of visits seems to closely negatively correlate with SART Understanding and Accuracy (p = 0.0885). This calls to mind the negative significant correlation between the number of visits and the SART Index describing the *Non-informed SA* construct. Perhaps, an increase in the number number of visits will reduce situational awareness but reduce mean square error (accuracy). Other near-significant correlations were observed between average pupil diameter and overall performance (p = 0.0634), TLX Index and saccade duration (p = 0.0936), mimics opened and saccade duration (p = 0.0781), and finally, saccade duration and error (0.0941).

Table 5.27 Correlation with Eye-tracking Measures.

	Avg. fix. due. (ms)	Visits (count)	Avg. pupil dia. (mm)	Avg. sac. amp. (°)	Sac. dur. (ms)	fixations (count)
TLX_index	-0.483	-0.547	0.053	-0.266	-0.631	0.410
Mental_demand	-0.514	0.323	0.186	-0.105	-0.427	0.200
Physical_demand	-0.146	-0.427	0.285	-0.005	-0.319	-0.025
Temporal_demand	-0.244	-0.291	-0.287	-0.239	-0.536	0.528
Performance	0.075	0.013	-0.050	0.100	0.551	-0.250
Effort	-0.221	-0.631	0.099	-0.098	-0.377	0.327
Frustration	0.119	-0.064	-0.463	-0.601	-0.269	0.161
SPAM-SAGAT	0.470	-0.137	-0.035	-0.176	-0.266	-0.470
SA1	0.421	-0.415	0.084	-0.013	-0.255	-0.396
SA2	0.482	0.088	0.274	-0.044	0.047	-0.812
SA3	0.026	-0.099	-0.283	-0.192	-0.268	0.051
SART_Index	-0.337	-0.729	0.230	0.162	-0.024	0.371
SART_Demand	0.541	-0.057	-0.099	-0.325	-0.379	-0.556
SART_Supply	-0.536	-0.562	0.498	0.235	-0.152	0.240
SART_Understanding	-0.007	-0.638	-0.021	0.004	-0.020	0.119
Task_load	0.031	-0.225	-0.149	-0.203	0.047	0.187
Alarm_list_support	0.300	-0.411	0.249	0.348	0.243	-0.442
Alarm_priority_support	0.254	-0.151	0.319	0.322	0.047	-0.488
Procedure_support	-0.031	0.225	0.149	0.203	-0.047	-0.187
Training_average	-0.658	0.155	-0.033	-0.131	-0.244	0.506
Familiarity_average	-0.412	-0.210	0.752	0.372	0.191	-0.284
Reaction_time	0.139	0.110	0.226	0.170	0.587	-0.500
Response_time	0.033	-0.160	0.138	-0.152	0.077	-0.214
Alarms_silenced	-0.279	0.268	0.108	-0.025	-0.025	0.038
Acknowledgements	-0.243	0.756	-0.005	-0.127	-0.121	0.061
mimics_opened	0.109	0.169	0.494	0.764	0.655	-0.327
No_of_alarms	0.101	-0.286	0.116	-0.328	-0.592	-0.181
Accuracy	0.214	-0.638	0.419	0.429	-0.071	-0.238
Recovery_time	-0.327	-0.113	0.878	0.873	0.546	-0.327
Consequence	0.206	0.213	-0.602	-0.495	-0.412	0.371
Overall Performance	0.169	0.290	-0.680	-0.507	-0.282	0.394
Error	0.252	0.000	-0.317	-0.378	-0.630	0.252
Average fixation duration(ms)	1.000	0.221	-0.471	-0.273	0.046	-0.714
No of visits (count)	0.221	1.000	-0.062	0.025	0.270	-0.381
Average pupil diameter (mm)	-0.471	-0.062	1.000	0.774	0.526	-0.299
Average saccade amplitude (°)	-0.273	0.025	0.774	1.000	0.761	-0.286
Saccade duration (ms)	0.046	0.270	0.526	0.761	1.000	-0.286
No of fixations	-0.714	-0.381	-0.299	-0.286	-0.286	1.000

Table 5.28 Significant differences of Eye-tracking Measures.

	Avg. fix. due. (ms)	Visits (count)	Avg. pupil dia. (mm)	Avg. sac. amp. (°)	Sac. dur. (ms)	fixations (count)
TLX_index	0.225	0.161	0.901	0.524	*0.0936*	0.313
Mental_demand	0.192	0.436	0.659	0.805	0.292	0.634
Physical_demand	0.731	0.292	0.493	0.991	0.441	0.954
Temporal_demand	0.561	0.484	0.491	0.568	0.171	0.179
Performance	0.860	0.976	0.906	0.814	0.157	0.550
Effort	0.598	*0.0932*	0.815	0.817	0.358	0.429
Frustration	0.779	0.881	0.248	0.115	0.519	0.704
SPAM-SAGAT	0.240	0.747	0.935	0.676	0.524	0.240
SA1	0.298	0.307	0.844	0.976	0.542	0.332
SA2	0.226	0.837	0.512	0.917	0.911	**0.0143***
SA3	0.952	0.816	0.498	0.650	0.521	0.904
SART_Index	0.414	**0.0403***	0.584	0.701	0.954	0.365
SART_Demand	0.166	0.893	0.815	0.432	0.354	0.152
SART_Supply	0.171	0.147	0.209	0.575	0.719	0.568
SART_Understanding	0.988	*0.0885*	0.960	0.993	0.962	0.779
Task_load	0.942	0.592	0.725	0.630	0.912	0.657
Alarm_list_support	0.470	0.311	0.553	0.398	0.562	0.273
Alarm_priority_support	0.544	0.721	0.441	0.437	0.913	0.220
Procedure_support	0.942	0.592	0.725	0.630	0.912	0.657
Training_average	*0.0760*	0.713	0.938	0.758	0.561	0.201
Familiarity_average	0.310	0.617	**0.0313***	0.364	0.650	0.495
Reaction_time	0.743	0.795	0.590	0.687	0.126	0.207
Response_time	0.938	0.706	0.745	0.720	0.857	0.610
Alarms_silenced	0.503	0.521	0.798	0.952	0.952	0.929
Acknowledgements	0.562	**0.0299***	0.991	0.764	0.775	0.887
mimics_opened	0.797	0.690	0.213	**0.0274***	*0.0781*	0.429
No_of_alarms	0.812	0.493	0.784	0.428	0.122	0.668
Accuracy	0.610	*0.0885*	0.301	0.289	0.867	0.570
Recovery_time	0.429	0.791	**0.00412***	**0.00466***	0.162	0.429
Consequence	0.624	0.613	0.115	0.212	0.310	0.365
Overall Performance	0.689	0.485	*0.0634*	0.200	0.499	0.334
Error	0.547	1.000	0.444	0.356	*0.0941*	0.547
Average fixation duration(ms)		0.599	0.239	0.513	0.913	**0.0465***
No of visits (count)	0.599		0.885	0.954	0.518	0.352
Average pupil diameter (mm)	0.239	0.885		**0.0243***	0.181	0.471
Average saccade amplitude (°)	0.513	0.954	**0.0243***		**0.0282***	0.493
Saccade duration (ms)	0.913	0.518	0.181	**0.0282***		0.493
No of fixations	**0.0465***	0.352	0.471	0.493	0.493	

5.3.3 Sensitivity Analysis

Tables 5.29, 5.30 and 5.31 presents the results of the statistical analysis, focussing on the sensitivity of various workload, situation awareness, and performance measures to increasing task complexity in three scenarios (S1, S2 and S3). The complexity of the scenario increased incrementally, and the objective of the analysis was to determine whether the measures were sensitive to these changes. Statistical tests, including Kruskal-Wallis and ANOVA, were used to detect significant differences between scenarios, followed by post hoc comparisons to identify specific pairwise differences.

Sensitivity to Task Complexity/Scenarios: NASA TLX, SART and SPAM

The TLX Index, which measures overall perceived workload, was normally distributed, and the ANOVA test demonstrated strong sensitivity to increasing scenario complexity as shown in Table 5.29. The significant result (p = 3.50e-07) of ANOVA indicated that the total workload increased significantly as the complexity of the task increased from S1 and S2 to S3. Post hoc tests showed that scenario 3 (S3), the most complex scenario, had significantly higher TLX scores compared to scenarios of lower complexity (S1 and S2). Following the post hoc test, S1 and 2 did not show a significant difference. The data for NASA TLX dimensions had a non-parametric distribution, and the Kruskal-Wallis test showed sensitivity to increasing task complexity. However, the post hoc test for S1 versus S2 did not show sensitivity between the two. The SPAM index, SA1, SA2 and SA3 were all non-parametric and their Kruskal-Wallis test showed sensitivity to task complexity. Except for SART Understanding, the SART measures showed sensitivity to increasing task complexity, and SART Understanding showed no sensitivity to the scenario.

Sensitivity to other subjective measures

The support of the procedure was sensitive to changes in complexity (p = 0.0006) for Scenario 1 versus 2, and the participants struggled more in more complex Scenario 3 as shown in Table 5.30. This aligns with the increased workload and cognitive demands reported in the TLX and SART measures, indicating that the complexity of the task impacts procedural support scores. The task load also showed sensitivity to task complexity except for S1 verus S2. In contrast, Alarm List Support, Alarm Priority

Table 5.29 Statistical Test Results for TLX and SART Measures Across Scenarios.

Measure	Test	P-Value	Post-hoc Test	S1 vs S2	S1 vs S3	S2 vs S3
TLX Index	ANOVA (F-test)	3.50E-07	Tukey HSD	0.9527	0	0
Mental demand	Kruskal-Wallis test	7.64E-09	Dunn	0.856198	2.72E-08	8.64E-06
Physical demand	Kruskal-Wallis test	0.003495	Dunn	1.000	0.005367	0.024675
Temporal demand	Kruskal-Wallis test	4.40E-05	Dunn	1.000	0.000845	0.000131
Performance	Kruskal-Wallis test	0.000396	Dunn	1.000	0.001003	0.003544
Effort	Kruskal-Wallis test	1.64E-06	Dunn	0.902745	0.000003	0.000337
Frustration	Kruskal-Wallis test	5.00E-05	Dunn	1.000	0.000618	0.000205
SPAM Index	Kruskal-Wallis test	2.65E-05	Dunn	1.000	0.000499	0.000096
SA1	Kruskal-Wallis test	0.009118	Dunn	1.000	0.009269	0.089389
SA2	Kruskal-Wallis test	0.033197	Dunn	1.000	0.162281	0.038535
SA3	Kruskal-Wallis test	0.000813	Dunn	1.000	0.005985	0.001892
SART Index	ANOVA (F-test)	0.000247	Tukey HSD	0.7476	0.0004	0.0043
SART Demand	Kruskal-Wallis test	3.28E-13	Dunn	0.33288	1.60E-12	5.64E-08
SART Supply	ANOVA (F-test)	0.000032	Tukey HSD	0.9719	0.0003	0.0001
SART Understanding	Kruskal-Wallis test	0.725858	No significant diff	NA	NA	NA

Support, Training Average, and Familiarity Average did not show sensitivity to increasing scenario complexity.

Table 5.30 Statistical Test Results for Other Subjective Measures.

Measure	Test	P-Value	Post-hoc Test	S1 vs S2	S1 vs S3	S2 vs S3
Task Load	Kruskal-Wallis	7.08E-11	Dunn	0.61908	3.34E-10	6.41E-07
Alarm List Support	Kruskal-Wallis	0.133989	No significant diff	NA	NA	NA
Alarm Priority Support	Kruskal-Wallis	0.136275	No significant diff	NA	NA	NA
Procedure Support	Kruskal-Wallis	0.000647	Dunn	1.000	0.008102	0.00109
Training Average	Kruskal-Wallis	0.981670	No significant diff	NA	NA	NA
Familiarity Average	Kruskal-Wallis	0.973340	No significant diff	NA	NA	NA

Sensitivity of behavioural and performance measures

In Table 5.31, behavioural metrics, including reaction time, response time, number of alarms silenced, number of acknowledgements, and mimics opened, exhibited non-parametric distributions. They were sensitive to increasing task complexity, with the exception of reaction time, which showed no sensitivity. Similar to most subjective variables, no sensitivity was observed between scenarios 1 and 2. Conversely, all performance measures, except for error, demonstrated sensitivity between scenarios 1 and 2, as well as between other scenarios. This observed sensitivity in performance measures between scenarios 1 and 2 suggests a significant need to discuss the outcomes of scenario 1 in comparison to the others at the performance level.

Table 5.31 Statistical Test Results for Behavioural and Performance Measures Across Scenarios.

Measure	Test	P-Value	Post-hoc Test	S1 vs S2	S1 vs S3	S2 vs S3
Reaction Time	Kruskal-Wallis	0.271863	No significant diff	NA	NA	NA
Response Time	Kruskal-Wallis	3.99E-14	Dunn	0.68876	7.26E-13	2.77E-09
Alarms Silenced	Kruskal-Wallis	1.24E-07	Dunn	1.000	6.30E-07	1.97E-05
Acknowledgements	Kruskal-Wallis	6.37E-08	Dunn	1.000	0.000002	0.000002
Mimics Opened	Kruskal-Wallis	0.007598	Dunn	1.000	0.024785	0.017048
No. of Alarms	Kruskal-Wallis	1.24E-17	Dunn	1.000	2.51E-14	1.70E-13
Accuracy	Kruskal-Wallis	1.78E-14	Dunn	0.000306	1.41E-04	5.32E-15
Recovery Time	Kruskal-Wallis	4.37E-17	Dunn	0.01997	5.84E-17	2.20E-08
Consequence	Kruskal-Wallis	1.20E-06	Dunn	0.000042	0.000009	1.000000
Overall Performance	Kruskal-Wallis	1.27E-06	Dunn	0.000001	0.001179	0.373556
Error	Kruskal-Wallis	6.56E-06	Dunn	0.406948	0.000005	0.003083

5.4 Summary

The reliability of questionnaires is often assessed to ensure internal consistency. In this study, internal consistency was found to be high for the SART and Alarm Support questions and moderate for the NASA TLX questionnaires. SPAM displayed near-average consistency, as did the questions regarding familiarity with control rooms and process plants. However, the training questions showed very low internal consistency.

5.4.1 Correlations and Constructs

Factor and correlation analysis were used to validate the measures and correctly identify what construct they tend to measure in this study and how this relates to other literature. The loadings of the variables and the constructs derived from their loading pattern are summarised in Table 5.32.

Table 5.32 Constructs and Associated Variables.

Construct	Associated Variables
Workload	TLX Index, Mental, Temporal demand and Effort Task Load, SART Demand and Supply
Informed Situational Awareness	SPAM Index, SA1, SA2, SA3
Non-informed Situational Awareness	SART Index, SART Understanding and Familiarity
Behaviour	Response time, the number of alarms silence, the number of alarms annunciated, the number of acknowledgements
Performance	Consequence, error

The factor loadings of certain variables altered when examining scenario 3, indicating that the loadings are sensitive to the scenario. This was further validated by the sensitivity analysis. According to the sensitivity results, Scenarios 1 and 2 largely exhibited no significant differences; however, notable differences were observed between scenarios 2 and 3. Consequently, a new set of constructs is added for scenario 3 for the cases where eye-tracking measures are loaded with other variables. The summary is presented in Table 5.33. It is important to remember that the eye-tracking measures describe the operators' internal state and behaviour while looking at the procedures (paper or electronic) and not on any other areas of interest. This is particularly important when discussing the results of scenario 3 with the eye-tracking outcomes. The subjective rating of the support for the procedure did not correlate with any of the eye-tracking measures, but the results of Table 5.18 show that it loads positively with the SPAM index and negatively with the task load.

Table 5.33 Associating the eye-tracking measures obtained from when specifically reading through the procedure in scenario 3 with that of other variables in the same scenario.

Construct	Associated Variables
Objective SA	Mental demand (-ve), SART Index (+ve), SART Understanding (+ve) Alarm list support (+ve), Acknowledgement (-ve), No of Visits (-ve)
Processing capacity	Average pupil diameter (-ve), Saccade duration (-ve)
Search difficulty	Mimics opened (+ve), Average saccade amplitude (+ve)
Learning	SA2 (+ve), Average training (-ve), Average fixation duration (+ve)

The correlation between subjective measures and other variables is summarised in Table 5.34, while the correlation between objective measures (excluding eye-tracking) is summarised in Table 5.35 and with eye-tracking (only scenario 3) in Table 5.36.

The correlation results of Table 5.28 confirm that the number of visits correlates significantly with the SART index and the acknowledgements identified under the construct, *Objective SA*, as shown above in Table 5.33. The measure, number of fixations, correlated with SA2 and fixation duration, which aligns with the construct, *Learning*. The *Processing Capacity and Search Difficulty* constructs derivation for the pupil and saccade measures alongside variables like mimics opened are further confirmed by the correlation analysis. During discussions, it is important to remember the nature of the relationship between the variables and the constructs they describe.

Table 5.34 Summary of Significant correlations (positive and negative), $p < 0.05$.

Variable	Positive Correlation	Negative Correlation
TLX Index	Mental demand, Physical demand Temporal demand, SART Demand Effort, Frustration SART Supply, Task load Response time, Alarms silenced, No of alarms, Accuracy Familiarity average, Acknowledgements Recovery time, Consequence, Overall Performance, Error	SPAM Index, SA3 Alarm list support, Alarm priority support
SPAM Index	Performance SA1, SA2, SA3 SART Index, SART Understanding Procedure support	TLX Index, Mental Demand, Temporal demand, Effort Frustration, Task load SART Demand and Supply Response time, Alarm silenced, Acknowledgment No of alarms, Accuracy Recovery time, Error
SART Index	Performance, Familiarity average SPAM Index, SA1, SA3 SART supply, SART Understanding Procedure support, Training average Alarm list support, Alarm priority support	Frustration SART Demand Task load Response time, Alarm silenced, Acknowledgment No of alarms, Accuracy Recovery time, Error
Task load	TLX Index (5/6 dimensions) SART Demand No of alarms, Accuracy , Acknowledgment Response time, Alarm silenced, Overall performance Recovery time, Consequence, Error	Performance, SART Index, Procedure support SPAM Index, SA1, SA2, SA3 Alarm list support, Alarm priority support
Procedure	Performance SPAM Index, SA3, SART Index, SART Understanding Familiarity average, Alarm list support, Alarm priority support	Frustration, Task load SART Demand, Accuracy, Acknowledgment, Recovery time No of alarms Consequence, Error

Table 5.35 Summary of Significant correlations between Behaviour and Performance measures (positive and negative), $p < 0.05$.

Variable	Positive Correlation	Negative Correlation
Reaction time	Overall Performance	Consequence, Error
Response time	Alarms silenced, Acknowledgements, Error Mimics opened, No of alarms, Consequence Accuracy, Recovery time, Overall performance	
Alarms Silenced	Response time, Acknowledgements No of alarms, Accuracy Recovery time, Error	
Mimics Opened	Response time, Accuracy Recovery time	

Table 5.36 Summary of Significant correlations with eye-tracking (Scenario 3 only), $p < 0.05$.

Variable	Positive Correlation	Negative Correlation
Average fixation duration (ms.)		Number of fixations
No of Visits to AOI (Count)	Acknowledgements	SART Index
Average Pupil Diameter (mm)	Familiarity Saccade duration	
Average Saccade Amplitude °	Mimics opened	
Number of fixations (count)		SA2, Average fixation duration

5.4.2 Groupwise Comparison

There were no significant differences between the group with paper procedures and those with electronic procedures on their subjective workload using NASA TLX and task load. However, results showed that those with the paper procedures perceived their workloads to be higher. Additionally, there were no notable differences among the groups on the SPAM (informed situational awareness) and SART (non-informed situational awareness) metrics across the three scenarios. While participants using the paper-based method showed higher levels of non-informed and informed situational awareness in different scenarios, this was not the case for informed situational awareness in the most complex scenario (scenario 3).

Likewise, the perceived support from both the alarm and procedure showed no significant difference, especially regarding the alarm support. Nevertheless, in scenario 2, a notable difference was found between the two groups in their perceived procedure support. On average across all scenarios, participants utilising paper procedures felt better supported compared to those using electronic procedures. There were no significant discrepancies in their training and familiarity levels across the scenarios, which is consistent with the sensitivity analysis of these two variables.

Significant differences were not observed for the behaviour measures as well except for Response time in the first scenario. The response time result showed that those with the electronic procedure had significantly faster response time than those with the paper procedure in this less complex scenario. There was a close significant difference observed for response time in the most complex scenario (S3) where response time was expected to be longer. Those with the paper procedure seemed to perform better than their counterpart in this scenario, even if not significantly. Those with the paper procedure generally had better response times as scenario complexity increased. On the contrary, those with the electronic procedure had faster reaction times than those using the paper procedure. Additionally, there were other minor behavioural observations related to the opening of mimics, acknowledgement of alarms, the number of alarms, and silencing of alarms. As evidenced by the correlation and factor analysis, these variables may characterise one or two constructs or possibly provide unique interpretations as to the performance outcomes, which will be discussed in Chapter 7.

Substantial differences were noted between the two groups in terms of pupil diameter and average saccade amplitude. These measures already show a correlation with certain

subjective, behavioural, and performance measures, thereby aiding in the interpretation of the group-wise outcomes. With sensitivity and validity analysis as well as identified constructs in mind, the group-specific outcomes will be further discussed in Chapter 7.

Chapter 6

Results Methodology II

6.1 Introduction

This chapter focuses on presenting the results of the different phases of the R²HiTL performance assessment framework. The goal of the framework is to integrate both expert and real-time knowledge to provide risk-informed decision support to operators and improve their situational awareness, for example, by guiding their behaviour with an understanding of its contribution to risk during critical safety situations. Achieving this helps keep operators in the loop during complex scenarios or in case of prolonged out-of-the-loop situations. The research questions to be answered here are:

- RQ3: Which operator behavioural metrics are the best predictor of a successful recovery of a plant alarmed state?
- RQ4: How can live and legacy data be integrated for a real-time human-in-the-loop performance assessment framework, decision support, and procedure optimisation?
- RQ5: What are the benefits of integrating operator-system data and expert knowledge for risk-based real-time decision and situation awareness support?

The R2HiTL framework can be divided into three distinct phases: Perception, understanding and projection of risk. Phase 1, Section 6.2, focuses on Risk Perception, where the recognition of risky scenarios and data collection are prioritised. Phase 2, Section 6.3 involves a deeper dive into Risk Understanding, utilising techniques such as Step-wise Logistic Regression using live data, Hierarchical Task Analysis (HTA) and

SPAR-H using legacy data. The HTA method helps to break down complex tasks into fine levels for error mode identification and reliability assessment. Methods such as SPAR-H analysis allow for a comprehensive assessment by categorising failure groups, assigning and rating performance-shaping factors (PSFs), estimating human error probabilities (HEP), and accounting for dependencies. The results of these analyses, combined with the uncertainty analysis, provide a clearer picture of the likelihood of failure associated with key tasks/human failure events (HFE).

In Phase 3, Section 6.4, the risk projection comes into play. Here, Bayesian Networks (BN) SPAR-H are used to model and discretise the operator-system variables, in this case, selected behavioural variables assigned for each HFE. This process helps identify the behaviour tipping points and allows for the contextualisation of these behaviour pairs, which are then compared with error outcomes or predictions from legacy or live data, respectively. The likelihood of failures is refined through posterior estimations based on BN SPAR-H, which are validated through comparison with new participant data. Finally, the findings from this new participant data are consolidated with a consequence and risk analysis, offering a comprehensive view of potential risks of the combination of these behavioural variables and their impact in Section 6.4.

6.2 Phase 1: Perception (Risk Perception)

The risk perception phase involved collecting live data from the operational system interaction but also checking for legacy data that describe the current operators' configuration. For example, the type of procedure in use, the critical task and the associated prior probability from both expert records and past live data, the procedure for such task, and so on. In a case where legacy data, such as the prior probability, does not exist, an action is triggered for phase 2 of the framework.

However, the first step in risk perception is to collect live data on operators' behaviour, cognition, and plant process. In this research, the data source remains those obtained from the design of experiments as presented in Chapter 5.1. A summary of the data collected for the investigation is shown in Table 6.1. Despite the rich amount of data collected, the goal of the framework is to make predictions and guide behaviour in real time. Therefore, the type of data needed as input is those that can be easily retrieved in real time. This then excludes subjective data and those from eye-tracking, given that eye-tracking technology is not yet adaptable for real-time applications. Also, the

performance variables since they measure performance itself. The performance variable to be predicted in this study is the error [Success: 0, Failure: 1]. In subsequent phases, the variable 'consequence' will become handy for consequence analysis of the human failure event paths.

The selected variables, which are 1. defined to describe the behaviour construct in Table 6.1 are reaction time, response time, number of alarms silenced, the number of alarms, 2. objective situational awareness (SA) - number of alarms acknowledged, 3. Search difficulty - number of mimics opened. A summary of the selected variables, the subjective and performance measures with which they correlate, and the constructs they define are shown in Table 6.1. The variable *reaction time* did not correlate significantly with other variables except the performance measures, and neither did it describe any of the constructs identified. However, it is included for its significant negative correlation with error and consequence. Highlighting the constructs these selected variables describe and their associated significantly correlated variables helps with understanding how they impact final outcomes in terms of Error but also what other factors might have contributed to the final outcomes and would have been ideal for consideration even if subjective.

Table 6.1 Selected Input variables for the Framework.

Variable	**Construct**	**Significantly Correlated variables**
Number of alarms acknowledged (-ve)	Objective SA, Behaviour	
Mimics opened (+ve)	Search difficulty	
Response time	Behaviour	
Number of alarms silence	Behaviour	
Number of alarms	Behaviour	
Reaction time		

The variables are further clustered into 3 key alarm handling and process control activities: monitoring/alarm handling, planning, and intervention (troubleshooting, control, and evaluation) as shown in Table 6.2. It shows the type of variables that relate more to each of these tasks and will be used likewise in later phases of the framework.

Given that no priors or task analyses have been developed the next phase is to present the task analysis results and prior estimation for the safety-critical task performed by the operators during real-time data collection.

Table 6.2 Key process control room tasks.

Monitoring/Alarm handling	Planning	Intervention
No of alarms	Mimics opened	Reaction time
Alarms Acknowledged	Group/type of support	Group/type of support
	Scenario/Task complexity	Response time
Group/Type of support	*No of procedures opened*	Scenario/Task complexity
Scenario/Task complexity		Mimics opened

6.3 Phase 2: Understand (Risk Understanding)

6.3.1 Step-wise Logistic Regression

After introducing live data, an ensuing task involves forecasting errors utilising the chosen variables. This error forecast will be done for each scenario and encompass all merged groups (combining group PBP, which follows paper procedures, with EP, which utilises electronic procedures). In scenario 3's prediction, a single operator is excluded for validation later in this framework. Henceforth, beginning from the subsequent section, only scenario three (3) will be employed to explore other methods within the framework. Detailed reasons for this choice are provided in the next section. The results of the stepwise logistic regression analysis and prediction of the probability of failure [Error = 1] are presented in Table 6.3. Though there were best variables selected through feature selection, all variables contribute in some way to the prediction and performance of the prediction. Hence, the probability values in Table 6.3 are based on all variables defined for each of the tasks in Table 6.2.

The findings reveal that for the Monitoring alarm handling task, the *number of alarms* variable emerged as the most influential and was prioritised in the feature selection process. Meanwhile, the planning task identified *Group/type of support* and *Scenario/task complexity* as the key variables. For the intervention task, the significant factors were *Group/type of support*, *Scenario/task complexity*, and *response time*. The outcomes for the planning and intervention tasks were comparable to those of the monitoring task. Both groups demonstrated high probabilities of success in the initial two scenarios, with an increased likelihood of failure in the most challenging scenario, scenario 3. Overall, the probability of failure increased for both groups as task complexity escalated from scenario 1 to scenario 3.

Table 6.3 Failure prediction for each group per scenario and task /HFE (human failure event).

Monitoring/Alarm handling Task *Accuracy = 85.6%, Best variable = No of alarms*			
	S1	S2	S3
Group PBP	0.0504 (5%)	0.0971 (10%)	0.4230 (42%)
Group EP	0.0669 (7%)	0.1224 (12%)	0.5003 (50%)
Planning Task *Accuracy = 80.6%, Best variable = Group/type of support*			
	S1	S2	S3
Group PBP	0.0510 (5%)	0.1498 (15%)	0.3667 (37%)
Group EP	0.0664 (7%)	0.1891 (19%)	0.4339 (43%)
Intervention Task *Accuracy = 81.1%, Best variable(s) = Group, Scenario/task complexity, Response time*			
	S1	S2	S3
Group PBP	0.0526 (5%)	0.1682 (17%)	0.3388 (34%)
Group EP	0.0650 (7%)	0.1917 (19%)	0.4390 (44%)

Additionally, it can be noted that the probability of failure for both groups remained relatively constant in scenario 1 across all three tasks, but it increased as they moved from the planning phase to intervention. Conversely, in the most complex scenario, the failure probability decreased as they advanced from planning to intervention. This implies that in the first scenario, both groups were equally likely to succeed in all tasks without any variance in failure probability, whereas in the second scenario, the likelihood of failure increased for both groups as they progressed from monitoring to intervention. In contrast, during the most complex scenario, the monitoring task proved more challenging for both groups compared to the planning or intervention phases.

Validity using k-fold cross-validation Step-wise Logistic Regression Results:

Model Outputsvital presents key results, including coefficients, p-values, and performance metrics (accuracy, precision, recall, F1 score).

Validity and Reliability Assessment: Model Validation: Explain how the model's validity was assessed (e.g., goodness-of-fit tests, AIC/BIC criteria, cross-validation techniques).

Reliability Checks: Discuss the steps taken to ensure model reliability (e.g., stability of coefficients across different samples, sensitivity analysis).

6.3.2 Hierarchical Task Analysis

To recap the scenarios tested during the experiment, the operators performed three critical safety scenarios, which increased in complexity. Scenario 1 entailed following the procedures/task steps described for the alarm tag *PAL01* to manually control the nitrogen flow given the failure of the pressure indicator controller (PIC01), as shown in Table 6.4. The second scenario involved following the same procedure, but this time it involved extra steps to control the nitrogen flow manually, given the failure of the manual nitrogen control valve display element. The last scenario entailed following potentially more than one procedure and the supervisor's final instruction. The procedures before the supervisor's instruction are for the alarm tags TAH17 and FAL11. The tabular hierarchical task analysis (HTA) is shown for scenarios 1, 2 and 3 in Tables 6.4, 6.5, and 6.6.

Table 6.4 Task Steps and Subtasks for Scenario 1.

Task Steps	**Sub task**
Monitoring	Identify Critical Alarm
	Acknowledge critical alarm
Planning	Find procedure
	Check the Pressure value [ata] on the graph (see Graph on tank mimic). Cross check with nominal Pressure value [1 ata]. If pressure below or above 1 [ata], do step 2. Else, do nothing.
	Check the Nitrogen flow (see Primary system flow [Nm³/h] meter on tank mimic). Cross check with nominal Nitrogen flow value [4 Nm³/h]. If Nitrogen flow less than 3.5 [Nm³/h] or greater than 4.5 [Nm³/h], then continue to step 3. Else, go to step 1.
Intervention	Switch Nitrogen valve to manual
	Move and adjust Pointer on Nitrogen valve scale between 5.2 and 5.5 Nm³/h.
	Monitor, for a while, Tank Pressure with Plot on Tank mimic (nominal value = 1 ata). If Pressure increases, then continue to STEP 6. Else, go to STEP 9.
	Monitor alarm list, for a while, until PAL01 is recovered (turn off). If PAL01 is recovered, END (notify the supervisor in the room). Else, do STEP 8.

Table 6.5 Task Steps and Subtasks for Scenario 2.

Task Steps	**Sub task**
Monitoring	Identify Critical Alarm
	Acknowledge critical alarm
Planning	Find procedure
	Check the Pressure value [ata] on the graph (see Graph on tank mimic). Cross check with nominal Pressure value [1 ata]. If pressure below or above 1 [ata], do step 2. Else, do nothing.
	Check the Nitrogen flow (see Primary system flow [Nm³/h] meter on tank mimic). Cross check with nominal Nitrogen flow value [4 Nm³/h]. If Nitrogen flow less than 3.5 [Nm³/h] or greater than 4.5 [Nm³/h], then continue to step 3. Else, go to step 1.
Intervention	Switch Nitrogen valve to manual
	Move and adjust Pointer on Nitrogen valve scale between 5.2 and 5.5 Nm³/h.
	Monitor, for a while, Tank Pressure with Plot on Tank mimic (nominal value = 1 ata). If Pressure increases, then continue to STEP 6. Else, go to STEP 9.
	Monitor alarm list, for a while, until PAL01 is recovered (turn off). If PAL01 is recovered, END (notify the supervisor in the room). Else, do STEP 8.
	Switch Nitrogen System on Tank mimic to Backup system
	Check the Pressure value [ata] on the graph (see Plot on tank mimic). Cross check with nominal Pressure value [1 ata] IF, pressure below 0.988 [ata], continue with STEP 10. ELSE, continue observing trend
	Click Methanol mimic to open
	Locate Pump's Manual-Auto switch on Methanol mimic.
	Switch Pump to manual.
	Move and adjust Pump power below 0.85 kW but not less than 0.5 kW. $[nominal value = 0.85 kW]$.
	Monitor, for a while, Tank Pressure with Plot (nominal value = 1 ata).
	Monitor, for a while, Tank Pressure with Plot (nominal value = 1 ata). IF Pressure starts increasing, then continue STEP 15. ELSE call supervisor in the room.
	Monitor Tank Pressure with Plot until PAL01 is recovered (turns off). IF PAL01 is recovered, then continue STEP 16. IF Pressure starts increasing, then continue STEP 15 ELSE, call supervisor in the room.
	Move Pointer and set Pump power back to 0.85 kW.

Table 6.6 Task Steps and Sub-tasks for Scenario 3.

Task Steps	**Sub task**
1. Monitoring & Alarm Handling	Monitor the process (no alarm)
	Identify Critical Alarm (now multiple alarms)
	Acknowledge critical alarm
2. Planning	Find procedure
	Click to Open Heat Recovery mimic
	Check Cooling water system mass flow rate with scale [kg/s]. Nominal value is [5.2 kg/s]. IF, flow rate below 4.8 kg/s, continue STEP 3.1. ELSE, observe scale
3. Intervention 1	Switch Cooling water system mass flow to Manual
	Move and adjust Cooling water system mass flow (using scale) between 4.8 kg/s - 5.5 kg/s . [Nominal value = 5.2 kg/s]
	Monitor, for a while, Cooling water system mass flow rate with scale (nominal value = 5.2 kg/s). IF cooling water flow restored (FAL11 alarm cleared from alarm list), then do STEP 3.4; ELSE, do STEP 3.5
	CALL supervisor in the room to ask what to do
4. Alarm Handling	Follow Instructions given by Supervisor
	Identify First Up Reactor Alarm
	Acknowledge Critical alarm
5. Planning	Open reactor plant section button
	Check the cooling system temperature [°C] on the meter. Cross-check with nominal temperature value [110 °C]. IF cooling system temperature below 108 [°C] or above 109 [°C], continue 6.1; ELSE, do nothing
6. Intervention	Switch cooling system temperature to manual
	Monitor Max temperature graph, try to stabilise the temperature using the cooling system temperature scale. (BE VERY CAREFUL NOT TO MOVE A LOT OR TOO QUICKLY). NOTE: Temperature very sensitive to change by the cooling system! IF TAH13 is not recovered after 1 Minute, CALL Supervisor; ELSE CONTINUE STEP 6.2

6.3.3 SPAR-H Analysis Results

The HTA analysis underscores the critical function of the procedure as a primary data source for evaluating human reliability and performance. The manner in which the procedure is drafted affects not only operators regarding reading or search ease, etc., but also guides experts in valuing the importance of tasks and in assigning weights to specific variables that shape task outcomes. From now on, the analysis and result of the SPAR-H HRA method and outcomes of the analysis of the other phases in this Chapter will be based on Scenario 3. Scenario 3 has been selected to illustrate the phases of the R²HITL decision support framework for its level of complexity close enough to complex safety-critical control room scenarios and its significant difference from scenarios 1 and 2 as seen in the sensitivity analysis in Section 3.3.3 of Chapter 3. Scenario 3 was also the only scenario in which analysis of the eye-tracking result was presented, especially to compare paper versus electronic procedures. This eye-tracking result shaped the development of some of the constructs in this thesis, such as search difficulty, which is also influenced by the number of opened mimics (a variable relevant to the planning and intervention tasks).

The results of the SPAR-H steps are presented in the following subsections.

Failure groups

Following the HTA breakdown, the task or the so-called human failure events (HFE) and their subtasks are assigned a specific failure type: diagnosis or action failure, as shown for Scenario 3 in Table 6.7.

Assigning and rating PSFs

The performance shaping factors in the SPAR-H method include available time, stress, task complexity, experience/training, procedures, ergonomics / HMI, fitness for duty and work processes. Each subtask in table 6.7 is assigned 8 of the PSF, and estimates of the human error probability for that task are calculated using Equation 4.2 for diagnosis tasks and 4.3 for action task. Each PSF has multipliers with values that vary depending on the type of task (diagnosis or failure). The template for the PSFs and their multipliers are shown in Appendix B. Table 6.8 and 6.9 shows the PSF and multiples assigned for scenario three (3) for Group PBP and EP, respectively, by a human factor and process automation expert.

Table 6.7 Task Steps and Sub-tasks for Scenario 3.

Task Steps	**Sub task**	**Failure type**
1. Monitoring & Alarm Handling	Monitor the process (no alarm)	Diagnosis
	Identify Critical Alarm (now multiple alarms)	Diagnosis
	Acknowledge critical alarm	Action
2. Planning	Find procedure	Action
	Click to Open Heat Recovery mimic	Action
	Check Cooling water system mass flow rate with scale [kg/s]. Nominal value is [5.2 kg/s]. IF, flow rate below 4.8 kg/s, continue STEP 3.1. ELSE, observe scale	Diagnosis
3. Intervention 1	Switch Cooling water system mass flow to Manual	Action
	Move and adjust the Cooling water system mass flow (using scale) between 4.8 kg/s - 5.5 kg/s. [Nominal value = 5.2 kg/s]	Action
	Monitor, for a while, Cooling water system mass flow rate with scale (nominal value = 5.2 kg/s). IF cooling water flow restored (FAL11 alarm cleared from alarm list), then do STEP 3.4; ELSE, do STEP 3.5	Diagnosis
	CALL supervisor in the room to ask what to do	Action
4. Alarm Handling	Follow Instructions given by Supervisor	
	Identify First Up Reactor Alarm	Diagnosis
	Acknowledge Critical alarm	Action
5. Planning	Open reactor plant section button	Action
	Check the cooling system temperature [°C] on the meter. Cross-check with nominal temperature value [110 °C]. IF cooling system temperature below 108 [°C] or above 109 [°C], continue 6.1; ELSE, do nothing	Diagnosis
6. Intervention	Switch cooling system temperature to manual	Action
	Monitor Max temperature graph, try to stabilise the temperature using the cooling system temperature scale. (BE VERY CAREFUL NOT TO MOVE A LOT OR TOO QUICKLY). NOTE: Temperature very sensitive to change by the cooling system! IF TAH13 is not recovered after 1 Minute, CALL Supervisor; ELSE CONTINUE STEP 6.2	Action

Table 6.8 Experts PSF assignment of multipliers for Group PBP.

Task Steps	Sub task	Failure type	Ava. time	Stress	Compl.	Exp./Training	Procedures	Erg./HMI	Fit.	Work Proc.
1. Monitoring & Alarm Handling										
	Monitor the process (no alarm)	Diagnosis	0.1	1	1	1	1	1	1	1
	Identify Critical Alarm (now multiple alarms)	Diagnosis	10	2	2	0.5	5	10	1	2
	Acknowledge critical alarm	Action	1	2	2	0.5	5	1	1	5
2. Planning										
	Find procedure	Action	10	2	1	1	1	1	1	1
	Click to Open Heat Recovery mimic	Action	1	1	2	1	1	0.5	1	1
	Check Cooling water system mass flow rate with scale [kg/s]. Nominal value is [5.2 kg/s]. IF, flow rate below 4.8 kg/s, continue STEP 3.1. ELSE, observe scale	Diagnosis	10	2	5	0.5	5	10	1	2
3. Intervention 1										
3.1	Switch Cooling water system mass flow to Manual	Action	10	1	1	1	1	0.5	1	0.5
3.2	Move and adjust Cooling water system mass flow (using scale) between 4.8 kg/s - 5.5 kg/s. [nominal value = 5.2 kg/s]	Action	1	2	2	0.5	5	10	1	5
3.3	Monitor, for a while, Cooling water system mass flow rate with scale (nominal value = 5.2 kg/s). IF cooling water flow restored (FAL11 alarm cleared from alarm list), then do STEP 3.4 ELSE, do STEP 3.5	Diagnosis	10	2	2	0.5	5	10	1	2
3.4	CALL supervisor in the room to ask what to do	Action	1	2	2	0.5	1	10	1	5
4. Alarm Handling										
4.1	Follow Instructions given by Supervisor									
	Identify First Up Reactor Alarm	Diagnosis	1	2	2	0.5	0.5	1	1	1
	Acknowledge Critical alarm	Action	10	2	2	0.5	5	1	1	1
5. Planning										
	Open reactor plant section button	Action	10	1	1	1	1	1	1	0.5
	Check the cooling system temperature [°C] on the meter. Cross check with nominal temperature value [110 °C]. IF cooling system temperature below 108 [°C] or above 109 [°C], continue 6.1. ELSE, do nothing.	Diagnosis	10	5	2	1	1	1	1	0.8
6. Intervention										
6.1	Switch cooling system temperature to manual	Action	1	1	1	1	1	0.5	1	0.5
6.2	Monitor Max temperature graph, try to stabilise the temperature using the cooling system temperature scale. (BE VERY CAREFUL NOT TO MOVE A LOT OR TOO QUICKLY). NOTE. Temperature was very sensitive to change by the cooling system! IF TAH13 is not recovered after 1 minute, CALL Supervisor ELSE CONTINUE STEP 6.2.	Action	1	2	2	1	5	1	1	1

Table 6.9 Experts PSF assignment of multipliers for Group EP.

Subtasks	Failure type	Ava. Time	Stress	Compl.	Exp./Training	Procedures	Erg./HMI	Fitness	Work Proc.
1. Monitoring & Alarm Handling									
Monitor the process (no alarm)	Diagnoses	0.1	1	1	1	1	1	1	1
Identify Critical Alarm (multiple alarms)	Diagnoses	1	2	2	0.5	5	10	1	2
Acknowledge critical alarm	Action	1	2	2	0.5	5	1	1	5
2. Planning									
Find procedure	Action	1	2	1	1	1	1	1	1
Click to Open Heat Recovery mimic	Action	1	1	2	1	1	0.5	1	1
Check Cooling water system mass flow rate (kg/s)	Diagnosis	10	2	5	0.5	5	10	1	2
3. Intervention 1									
Switch Cooling water system mass flow to Manual	Action	10	1	1	1	1	0.5	1	0.5
Adjust Cooling water system mass flow (4.8-5.5 kg/s)	Action	1	2	2	0.5	5	10	1	5
Monitor Cooling water system mass flow rate	Diagnosis	1	2	2	0.5	5	10	1	2
CALL supervisor to ask what to do	Action	1	2	2	0.5	1	10	1	5
4. Alarm Handling									
Follow Instructions by Supervisor	Diagnosis	1	2	2	0.5	0.5	1	1	1
Acknowledge Critical alarm	Action	10	2	2	0.5	5	1	1	1
5. Planning									
Open reactor plant section button	Action	10	1	1	1	1	1	1	0.5
Check cooling system temperature (°C)	Diagnosis	10	5	2	1	1	1	1	0.8
6. Intervention									
Switch cooling system temperature to manual	Action	1	1	1	1	1	0.5	1	0.5
Monitor Max temperature graph	Action	1	2	2	1	5	1	1	1

The result of the expert assessment of Group PBP and EP shows that the subtasks 1.2, 1.3, 2.3, 3.2, 3.3, 3.4, 4.2, 5.2, and 6.2 had multipliers greater than 3, in both groups.

HEP estimation and Correction

In a case where there are three or more negative multipliers, multipliers greater than 1, equation 4.4 is applied to correct the HEP estimate. This correction ensures that HEP is, on no occasion, greater than 1. The average HEP estimates from two expert assessors after HEP correction are shown in Table 6.10 for Group PBP (those using the paper procedure) and Table 6.11 for Group EP (those using the electronic procedure).

Table 6.10 NHEP, HEP adjustment, dependency calculation and final HEP for Group PBP.

Task Steps	Sub Task	Type of Task	NHEP	HEP	Adj2	Crew	Time	Location	Cues	New_HEP
1. Monitoring & Alarm Handling										
	Monitor the process (no alarm)	Diagnoses	0.01	0.001	0.001	N/A				0.001
	Identify Critical Alarm (now multiple alarms)	Diagnoses	0.01	20.000	0.953	0	1	1	0	0.955
	Acknowledge critical alarm	Action	0.001	0.050	0.048	0	0	1	1	0.524
2. Planning										
	Find procedure	Action	0.001	0.020	0.020	0	0	1	1	0.510
	Click to Open Heat Recovery mimic	Action	0.001	0.001	0.001	0	0	1	1	0.501
	Check Cooling water system mass flow rate with scale [kg/s]....	Diagnosis	0.01	50.000	0.981	0	0	1	0	0.990
3. Intervention 1										
	3.1 Switch Cooling water system mass flow to Manual	Action	0.001	0.003	0.003	0	0	1	1	0.501
	3.2 Move and adjust Cooling water system mass flow (using scale).....	Action	0.001	0.500	0.334	0	0	1	1	0.667
	3.3 Monitor, for a while, Cooling water system mass flow rate with scale ...	Diagnosis	0.01	20.000	0.953	0	0	1	0	0.976
	3.4 CALL supervisor in the room to ask what to do	Action	0.001	0.100	0.091	0	1	1	1	0.221
4. Alarm Handling	Follow Instructions given by Supervisor									
	Identify First Up Reactor Alarm	Diagnosis	0.01	0.010	0.010	0	0	1	0	0.505
	Acknowledge Critical alarm	Action	0.001	0.100	0.091	0	0	1	1	0.545
5. Planning										
	Open reactor plant section button	Action	0.001	0.005	0.005	0	0	1	1	0.502
	Check the cooling system temperature [°C] on the meter...	Diagnosis	0.01	0.800	0.447	0	0	1	0	0.723
6. Intervention										
	6.1 Switch cooling system temperature to manual	Action	0.001	0.000	0.000	0	0	1	1	0.500
	6.2 Monitor Max temperature graph, try to stabilise the temperature ...	Action	0.001	0.020	0.020	0	0	1	0	0.510

Table 6.11 NHEP, HEP adjustment, dependency calculation and final HEP for Group EP.

Task Steps	Sub Task	Type of Task	NHEP	HEP	Adj2	Crew	Time	Location	Cues	New_HEP
1. Monitoring & Alarm Handling										
	Monitor the process (no alarm)	Diagnoses	0.01	0.001	0.001	N/A				0.001
	Identify Critical Alarm (now multiple alarms)	Diagnoses	0.01	2.000	0.669	0	1	1	0	0.685
	Acknowledge critical alarm	Action	0.001	0.050	0.048	0	0	1	1	0.524
2. Planning										
	Find procedure	Action	0.001	0.002	0.002	0	0	1	1	0.501
	Click to Open Heat Recovery mimic	Action	0.001	0.001	0.001	0	0	1	1	0.501
	Check Cooling water system mass flow rate with scale [kg/s]...e	Diagnosis	0.01	50.000	0.981	0	0	1	0	0.990
3. Intervention 1										
	3.1 Switch Cooling water system mass flow to Manual	Action	0.001	0.003	0.003	0	0	1	1	0.501
	3.2 Move and adjust Cooling water system mass flow ...	Action	0.001	0.500	0.334	0	0	1	1	0.667
	3.3 Monitor, for a while, Cooling water system mass flow rate with scale...	Diagnosis	0.01	2.000	0.669	0	0	1	0	0.834
	3.4 CALL supervisor in the room to ask what to do	Action	0.001	0.100	0.091	0	1	1	1	0.221
4. Alarm Handling	Follow Instructions given by Supervisor									
	Identify First Up Reactor Alarm	Diagnosis	0.01	0.010	0.010	0	0	1	0	0.505
	Acknowledge Critical alarm	Action	0.001	0.100	0.091	0	0	1	1	0.545
5. Planning										
	Open reactor plant section button	Action	0.001	0.005	0.005	0	0	1	1	0.502
	Check the cooling system temperature [°C] on the meter...	Diagnosis	0.01	0.800	0.447	0	0	1	0	0.723
6. Intervention										
	6.1 Switch cooling system temperature to manual	Action	0.001	0.000	0.000	0	0	1	1	0.500
	6.2 Monitor Max temperature graph, try to stabilise the temperature ...	Action	0.001	0.020	0.020	0	0	1	0	0.510

The results show that Sub-tasks are

Dependency and final priors for each key task

The dependency analysis remains consistent between both groups, using the conditions outlined in Figure 4.2. To recall, dependency is based on the following criteria: 1. Crew (same or different), Time(close or not close), Location (same or different) and cues (additional or no additional). However, the resulting HEP estimation after considering dependency differs for each group due to the initial HEP estimations. The results for Group PBP can be found in Table 6.10, while those for EP are presented in Table 6.11.

Subsequently, the New_HEPs are multiplied for each Human Failure Event (HFE): monitoring, planning, and intervention using Equation 4.7 to estimate their final prior probability. The results for each group are shown in Tables 6.12 and 6.13 for PBP and EP, respectively.

Uncertainty analysis

The final step in the SPAR-H analysis is the estimation of the α and β priors for each Human Failure Event as shown in Table 6.12 and 6.13. To recall the explanation of this estimation process as described in Subsection 4.2.2, alpha is estimated using the table provided by [170] while beta is calculated using Equation 4.7 given the prior probability $E(P0)$.

Table 6.12 Final HEP per Human Failure Event and the alpha and beta priors for Group PBP.

Task Steps	Success	Failure Probability, E(P0)	Alpha	Beta
Monitoring & Alarm Handling	0.005	0.995	0.500	0.002
Planning	0.000	1.000	0.500	0.000
Intervention	0.001	0.999	0.500	0.000

Table 6.13 Final HEP per Human Failure Event and the alpha and beta priors for Group EP

Task Steps	Success	Failure Probability, E(P0)	Alpha	Beta
Monitoring & Alarm Handling	0.034	0.966	0.500	0.017
Planning	0.000	1.000	0.500	0.000
Intervention	0.005	0.995	0.500	0.003

6.4 Phase 3: Project (Risk Projection)

6.4.1 BN SPARH

The aim of this Bayesian Inference (BI) is to covey the belief regarding the hypothesis that the HEP for the human failure events - monitoring, planning and intervention equals m, p and i, respectively. The data for this phase of the framework is obtained from the experiment record on the behavioural variables similarly used for error prediction of each HFE in Section 6.1. Also, data from the experts' evaluation of the HFE, that is, the prior probabilities and alpha and beta priors are inputs at this phase of the framework. Again, the scenario of interest following the SPARH expert analysis remains scenario 3. The results of the BN SPARH steps for scenario 3 are shown in the subsections below.

Discretisation

The discretisation was done for the variables that describe monitoring, planning and intervention HFEs, respectively. For the monitoring event, the optimal split for the number of alarms was 18 [≤ 18 & > 18], and that of the number of alarms acknowledged was 4.5 [≤ 4.5 & > 4.5]. The scenario/task complexity remained the same, that is, scenario 3 was labelled as 2, and the group/type of support was either Group PBP labelled 0 or EP labelled 1 as shown in Table 6.14.

For the planning HFE, the optimal split for the number of mimics opened was 3.5 [≤ 3.5 & > 3.5] as shown in Table 6.15.

For the Intervention HFE, the optimal split for the number of mimics opened was 3.5 [≤ 3.5 & > 3.5], for reaction time it was 266 seconds [≤ 266 & > 266], and response time it was 856.5 seconds [≤ 856.5 & > 856.5] as shown in Table 6.16.

Table 6.14 Monitoring HFE: discretisation of variables.

Group	Scenario	No_of_alarms_discretised	Acknowl._discretised	Error
1	2	≤ 18.0	≤ 4.5	1
1	2	> 18.0	> 4.5	1
1	2	> 18.0	≤ 4.5	1
0	2	≤ 18.0	> 4.5	0
0	2	≤ 18.0	≤ 4.5	0
1	2	> 18.0	≤ 4.5	1
0	2	> 18.0	> 4.5	1
0	2	> 18.0	> 4.5	1
1	2	≤ 18.0	≤ 4.5	0
0	2	> 18.0	> 4.5	0
0	2	> 18.0	> 4.5	0
1	2	≤ 18.0	≤ 4.5	0
1	2	≤ 18.0	≤ 4.5	0
0	2	> 18.0	≤ 4.5	1
1	2	> 18.0	> 4.5	1
1	2	> 18.0	≤ 4.5	1
1	2	> 18.0	≤ 4.5	0
0	2	≤ 18.0	> 4.5	0
0	2	≤ 18.0	> 4.5	0
0	2	≤ 18.0	≤ 4.5	0
0	2	> 18.0	≤ 4.5	1
0	2	≤ 18.0	> 4.5	0
1	2	≤ 18.0	> 4.5	0
0	2	≤ 18.0	≤ 4.5	0
0	2	≤ 18.0	≤ 4.5	0
1	2	≤ 18.0	≤ 4.5	0
1	2	≤ 18.0	> 4.5	1
0	2	> 18.0	≤ 4.5	1
1	2	> 18.0	≤ 4.5	1
1	2	> 18.0	> 4.5	1
1	2	≤ 18.0	≤ 4.5	0
0	2	≤ 18.0	> 4.5	1
1	2	> 18.0	> 4.5	1

Table 6.15 Planning HFE: discretisation of variables.

Group	Scenario	mimics_opened_discretised	Error
1	2	> 3.5	1
1	2	<= 3.5	1
1	2	> 3.5	1
0	2	<= 3.5	0
0	2	<= 3.5	0
1	2	<= 3.5	1
0	2	<= 3.5	1
0	2	<= 3.5	1
1	2	<= 3.5	0
0	2	<= 3.5	0
0	2	> 3.5	0
1	2	> 3.5	0
1	2	<= 3.5	0
0	2	> 3.5	1
1	2	> 3.5	1
1	2	> 3.5	1
1	2	<= 3.5	0
0	2	> 3.5	0
0	2	> 3.5	0
0	2	<= 3.5	0
0	2	<= 3.5	1
0	2	<= 3.5	0
1	2	> 3.5	0
0	2	<= 3.5	0
0	2	<= 3.5	0
1	2	> 3.5	0
1	2	<= 3.5	1
0	2	<= 3.5	1
1	2	<= 3.5	1
1	2	<= 3.5	1
1	2	<= 3.5	0
0	2	> 3.5	1
1	2	<= 3.5	1

Table 6.16 Intervention HFE: discretisation of variables.

Group	Scenario	mimics_opened_discretised	Reaction_time_discretised	Response_time_discretised	Error
1	2	> 3.5	<= 266.0	<= 856.5	1
1	2	<= 3.5	<= 266.0	<= 856.5	1
1	2	> 3.5	<= 266.0	<= 856.5	1
0	2	<= 3.5	> 266.0	> 856.5	0
0	2	<= 3.5	<= 266.0	> 856.5	0
1	2	<= 3.5	<= 266.0	<= 856.5	1
0	2	<= 3.5	<= 266.0	> 856.5	1
0	2	<= 3.5	<= 266.0	<= 856.5	1
1	2	<= 3.5	> 266.0	> 856.5	0
0	2	<= 3.5	> 266.0	<= 856.5	0
0	2	> 3.5	> 266.0	> 856.5	0
1	2	> 3.5	<= 266.0	> 856.5	0
1	2	<= 3.5	> 266.0	> 856.5	0
0	2	> 3.5	> 266.0	> 856.5	1
1	2	> 3.5	<= 266.0	<= 856.5	1
1	2	> 3.5	> 266.0	<= 856.5	1
1	2	<= 3.5	> 266.0	<= 856.5	0
0	2	> 3.5	> 266.0	> 856.5	0
0	2	> 3.5	> 266.0	> 856.5	0
0	2	<= 3.5	> 266.0	> 856.5	0
0	2	<= 3.5	> 266.0	<= 856.5	1
0	2	<= 3.5	> 266.0	> 856.5	0
1	2	> 3.5	<= 266.0	> 856.5	0
0	2	<= 3.5	<= 266.0	> 856.5	0
0	2	<= 3.5	> 266.0	> 856.5	0
1	2	> 3.5	<= 266.0	> 856.5	0
1	2	<= 3.5	<= 266.0	<= 856.5	1
0	2	<= 3.5	<= 266.0	<= 856.5	1
1	2	<= 3.5	> 266.0	<= 856.5	1
1	2	<= 3.5	<= 266.0	<= 856.5	1
1	2	<= 3.5	> 266.0	> 856.5	0
0	2	> 3.5	<= 266.0	<= 856.5	1
1	2	<= 3.5	<= 266.0	<= 856.5	1

Contextualisation

The variables for each of the HFEs were contextualised based on the unique combination of each of these variables for the HFEs they describe, as shown in Tables 6.17, 6.18 and 6.19, for the monitoring, planning and intervention HFEs, respectively.

Liklihood estimation

The contextualised data are then matched with the error [Success: 0, Failure: 1] to count/estimate the number of failures x and the number of independent opportunities for failure, n. The count result from which x and n will be derived for the monitoring, planning, and intervention HFEs is shown in Tables 6.20, 6.21, and 6.22. The context explains the unique combination of the variables at different levels. In the case of new live data, these operators' data is checked for which context it falls within.

Monitoring: Those in Group PBP fell under contexts D, E, F, and G, while those in the EP Group fell under contexts A, B, C, and H.

For those in context A, the number of failures, x, is 1 while n is 6. For those in context B, the number of failures, x, is 0 while n is 4. For those in context C, the number of failures, x, is 4 while n is 5. For those in context D, the number of failures, x, is 1 while n is 5. For those in context E, the number of failures, x, is 0 while n is 4. For those in context F, the number of failures, x, is 2 while n is 4. For those in context G, the number of failures, x, is 3 while n is 3. For those in context H, the number of failures, x, is 1 while n is 2.

Planning: Those in Group PBP fell under contexts C and D while those in the EP group fell under contexts A and B.

For those in context A, the number of failures, x, is 4 while n is 7. For those in context B, the number of failures, x, is 6 while n is 10. For those in context C, the number of failures, x, is 4 while n is 11. For those in context D, the number of failures, x, is 2 while n is 5.

Intervention Those in Group PBP fell under contexts C, D, E, G, H, and L, while those in the EP group fell under contexts A, B, F, I, J, and K.

For those in context A, the number of failures, x, is 1 while n is 3. For those in context B, the number of failures, x, is 1 while n is 5. For those in context C, the number of failures, x, is 0 while n is 4. For those in context D, the number of failures, x, is 1 while n is 3. For those in context E, the number of failures, x, is 1 while n is 2. For

Table 6.17 Monitoring HFE: contexualisation of variables.

Group	Scenario	No_of_alarms_discretised	Acknowl._discretised	Context	Error
1	2	≤ 18.0	≤ 4.5	A	1
1	2	> 18.0	> 4.5	B	1
1	2	> 18.0	≤ 4.5	C	1
0	2	≤ 18.0	> 4.5	D	0
0	2	≤ 18.0	≤ 4.5	E	0
1	2	> 18.0	≤ 4.5	C	1
0	2	> 18.0	> 4.5	F	1
0	2	> 18.0	> 4.5	F	1
1	2	≤ 18.0	≤ 4.5	A	0
0	2	> 18.0	> 4.5	F	0
0	2	> 18.0	> 4.5	F	0
1	2	≤ 18.0	≤ 4.5	A	0
1	2	≤ 18.0	≤ 4.5	A	0
0	2	> 18.0	≤ 4.5	G	1
1	2	> 18.0	> 4.5	B	1
1	2	> 18.0	≤ 4.5	C	1
1	2	> 18.0	≤ 4.5	C	0
0	2	≤ 18.0	> 4.5	D	0
0	2	≤ 18.0	> 4.5	D	0
0	2	≤ 18.0	≤ 4.5	E	0
0	2	> 18.0	≤ 4.5	G	1
0	2	≤ 18.0	> 4.5	D	0
1	2	≤ 18.0	> 4.5	H	0
0	2	≤ 18.0	≤ 4.5	E	0
0	2	≤ 18.0	≤ 4.5	E	0
1	2	≤ 18.0	≤ 4.5	A	0
1	2	≤ 18.0	> 4.5	H	1
0	2	> 18.0	≤ 4.5	G	1
1	2	> 18.0	≤ 4.5	C	1
1	2	> 18.0	> 4.5	B	1
1	2	≤ 18.0	≤ 4.5	A	0
0	2	≤ 18.0	> 4.5	D	1
1	2	> 18.0	> 4.5	B	1

Table 6.18 Planning HFE: contexualisation of variables.

Group	Scenario	Mimics Opened Discretised	Context	Error
1	2	> 3.5	A	1
1	2	<= 3.5	B	1
1	2	> 3.5	A	1
0	2	<= 3.5	C	0
0	2	<= 3.5	C	0
1	2	<= 3.5	B	1
0	2	<= 3.5	C	1
0	2	<= 3.5	C	1
1	2	<= 3.5	B	0
0	2	<= 3.5	C	0
0	2	> 3.5	D	0
1	2	> 3.5	A	0
1	2	<= 3.5	B	0
0	2	> 3.5	D	1
1	2	> 3.5	A	1
1	2	> 3.5	A	1
1	2	<= 3.5	B	0
0	2	> 3.5	D	0
0	2	> 3.5	D	0
0	2	<= 3.5	C	0
0	2	<= 3.5	C	1
0	2	<= 3.5	C	0
1	2	> 3.5	A	0
0	2	<= 3.5	C	0
0	2	<= 3.5	C	0
1	2	> 3.5	A	0
1	2	<= 3.5	B	1
0	2	<= 3.5	C	1
1	2	<= 3.5	B	1
1	2	<= 3.5	B	1
1	2	<= 3.5	B	0
0	2	> 3.5	D	1
1	2	<= 3.5	B	1

Table 6.19 Intervention HFE: contexualisation of variables.

Group	Scenario	mimics_opened_discretised	Reaction_time_discretised	Response_time_discretised	Context	Error
1	2	> 3.5	<= 266.0	<= 856.5	A	1
1	2	<= 3.5	<= 266.0	<= 856.5	B	1
1	2	> 3.5	<= 266.0	<= 856.5	A	1
0	2	<= 3.5	> 266.0	> 856.5	C	0
0	2	<= 3.5	<= 266.0	> 856.5	D	0
1	2	<= 3.5	<= 266.0	<= 856.5	B	1
0	2	<= 3.5	<= 266.0	> 856.5	D	1
0	2	<= 3.5	<= 266.0	<= 856.5	E	1
1	2	<= 3.5	> 266.0	> 856.5	F	0
0	2	<= 3.5	> 266.0	<= 856.5	G	0
0	2	> 3.5	> 266.0	> 856.5	H	0
1	2	> 3.5	<= 266.0	> 856.5	I	0
1	2	<= 3.5	> 266.0	> 856.5	F	0
0	2	> 3.5	> 266.0	> 856.5	H	1
1	2	> 3.5	<= 266.0	<= 856.5	A	1
1	2	> 3.5	> 266.0	<= 856.5	J	1
1	2	<= 3.5	> 266.0	<= 856.5	K	0
0	2	> 3.5	> 266.0	> 856.5	H	0
0	2	> 3.5	> 266.0	> 856.5	H	0
0	2	<= 3.5	> 266.0	> 856.5	C	0
0	2	<= 3.5	> 266.0	<= 856.5	G	1
0	2	<= 3.5	> 266.0	> 856.5	C	0
1	2	> 3.5	<= 266.0	> 856.5	I	0
0	2	<= 3.5	<= 266.0	> 856.5	D	0
0	2	<= 3.5	> 266.0	> 856.5	C	0
1	2	> 3.5	<= 266.0	> 856.5	I	0
1	2	<= 3.5	<= 266.0	<= 856.5	B	1
0	2	<= 3.5	<= 266.0	<= 856.5	E	1
1	2	<= 3.5	> 266.0	<= 856.5	K	1
1	2	<= 3.5	<= 266.0	<= 856.5	B	1
1	2	<= 3.5	> 266.0	> 856.5	F	0
0	2	> 3.5	<= 266.0	<= 856.5	L	1
1	2	<= 3.5	<= 266.0	<= 856.5	B	1

Table 6.20 Monitoring HFE: count of context - error pairing.

Context	Error	Count
A	0	5
A	1	1
B	1	4
C	0	1
C	1	4
D	0	4
D	1	1
E	0	4
F	0	2
F	1	2
G	1	3
H	0	1
H	1	1

Table 6.21 Planning HFE: count of context - error pairing.

Context	Error	Count
A	0	3
A	1	4
B	0	4
B	1	6
C	0	7
C	1	4
D	0	3
D	1	2

Table 6.22 Intervention HFE: count of context - error pairing.

Context	Error	Count
A	1	3
B	1	5
C	0	4
D	0	2
D	1	1
E	1	2
F	0	3
G	0	1
G	1	1
H	0	3
H	1	1
I	0	3
J	1	1
K	0	1
K	1	1
L	1	1

those in context F, the number of failures, x, is 0 while n is 3. For those in context G, the number of failures, x, is 1 while n is 2. For those in context H, the number of failures, x, is 1 while n is 4. For those in context I, the number of failures, x, is 0 while n is 3. For those in context J, the number of failures, x, is 1 while n is 1. For those in context K, the number of failures, x, is 1 while n is 2. For those in context L, the number of failures, x, is 1 while n is 1.

Recall SPAR-H posteriors and CNI

Following the likelihood estimation for each context per HFE for each group, the priors from the SPARH assessment are recalled. This would then be used for the calculation of posterior probabilities for each group and for each context within which they fall. The data recalled for the monitoring task for the PBP and EP group are shown in Tables 6.23 and 6.24, respectively.

The data recalled for the planning task for the PBP and EP group are shown in Tables 6.25 and 6.26, respectively.

Table 6.23 Monitoring HFE: recalling priors for group PBP.

Monitoring HFE Group PBP	x	n	EP0	Alpha	Beta
Context D	1	5	0.995	0.500	0.002
Context E	0	4	0.995	0.500	0.002
Context F	2	4	0.995	0.500	0.002
Context G	3	3	0.995	0.500	0.002

Table 6.24 Monitoring HFE: recalling priors for group EP.

Monitoring HFE Group EP	x	n	EP0	Alpha	Beta
Context A	1	6	0.966	0.500	0.017
Context B	0	4	0.966	0.500	0.017
Context C	4	5	0.966	0.500	0.017
Context H	1	2	0.966	0.500	0.017

Table 6.25 Planning HFE: recalling priors for group PBP.

Planning HFE Group	PBP	x	n	EP0	Alpha	Beta
Context C	PBP	4	11	1.000	0.500	0.000
Context D	PBP	2	5	1.000	0.500	0.000

Table 6.26 Planning HFE: recalling priors for group EP.

Planning HFE Group	EP	x	n	EP0	Alpha	Beta
Context A	EP	4	7	1.000	0.500	0.000
Context B	EP	6	10	1.000	0.500	0.000

The data recalled for the intervention task for the PBP and EP groups are shown in Tables 6.27 and 6.28, respectively.

Table 6.27 Intervention HFE: recalling priors for group PBP.

Context	x	n	EP0	Alpha	Beta
C	0	4	0.999	0.500	0.000
D	1	3	0.999	0.500	0.000
E	1	2	0.999	0.500	0.000
G	1	2	0.999	0.500	0.000
H	1	4	0.999	0.500	0.000
L	1	1	0.999	0.500	0.000

Table 6.28 Intervention HFE: recalling priors for group EP.

Context	x	n	EP0	Alpha	Beta
A	1	3	0.995	0.500	0.003
B	1	5	0.995	0.500	0.003
F	0	3	0.995	0.500	0.003
I	0	3	0.995	0.500	0.003
J	1	1	0.995	0.500	0.003
K	1	2	0.995	0.500	0.003

Estimate posteriors: probability, CNI

The last step of the BN SPARH was estimating the alpha and beta posteriors and the final posterior probabilities using equations 4.12, 4.13, and 4.14, respectively. The posteriors for the monitoring tasks are shown in Tables 6.29 and 6.30 for the PBP and EP groups, respectively.

Table 6.29 Monitoring HFE: Group PBP Data with Posterior-Parameters.

Context	x	n	EP0	Alpha	Beta	Alpha_post	Beta_post	E_m_1
Context D	1	5	0.995	0.500	0.002	1.500	4.000	0.273
Context E	0	4	0.995	0.500	0.002	0.500	4.000	0.111
Context F	2	4	0.995	0.500	0.002	2.500	2.000	0.555
Context G	3	3	**0.995**	0.500	0.002	3.500	0.002	**0.999**

Table 6.30 Monitoring HFE: Group EP Data with Posterior-Parameters.

Context	x	n	EP0	Alpha	Beta	Alpha_post	Beta_post	E_m_1
Context A	1	6	0.966	0.500	0.017	1.500	5.000	0.230
Context B	0	4	0.966	0.500	0.017	0.500	4.000	0.111
Context C	4	5	**0.966**	0.500	0.017	4.500	1.000	**0.816**
Context H	1	2	0.966	0.500	0.017	1.500	1.000	0.596

Table 6.31 Planning HFE: Group PBP Data with Posterior-Parameters.

Context	PBP	x	n	EP0	Alpha	Beta	Alpha_post	Beta_post	E_p_1
Context C	PBP	4	11	1.000	0.500	0.000	4.500	7.000	0.391
Context D	PBP	2	5	1.000	0.500	0.000	2.500	3.000	0.455

The posteriors for the planning HFE are shown in Tables 6.31 and 6.32 for group PBP and EP, respectively.

The posteriors for the Intervention HFE are shown in Tables 6.33 and 6.34 for the PBP and EP groups, respectively.

Validation by comparing with new participant data

Before the prediction of LR, one of the participants was randomly dropped from scenario 3. The participant had apparently passed the scenario and had the behavioural data as shown in Table 6.35. Following the data input of this operator, the next step is to check in which context he or she falls for each of the human failure events (HFEs) - Monitoring, Planning and Intervention.

Checking Table 6.17 for the monitoring task contexts, this operator falls under context D. Looking further in, Table 6.20 shows that those in D could either fail or pass the scenario. However, more than likely to succeed given that 4 out of 5 people have succeeded in that context. However, further confirmation from the logistic regression prediction would equally ascertain this assumption. Tables 6.36 confirm that the operator has a higher probability of success than failure (0.31) in the monitoring task ($\alpha = 0.4$ and

Table 6.32 Planning HFE: Group EP Data with Posterior-Parameters.

Context	EP	x	n	EP0	Alpha	Beta	Alpha_post	Beta_post	E_p_1
Context A	EP	4	7	1.000	0.500	0.000	4.500	3.000	0.600
Context B	EP	6	10	1.000	0.500	0.000	6.500	4.000	0.619

Table 6.33 Intervention HFE: Group PBP Data with Posterior-Parameters.

Context	x	n	EP0	Alpha	Beta	Alpha_post	Beta_post	E_i_1
C	0	4	0.999	0.500	0.000	0.500	4.000	0.111
D	1	3	0.999	0.500	0.000	1.500	2.000	0.429
E	1	2	0.999	0.500	0.000	1.500	1.000	0.600
G	1	2	0.999	0.500	0.000	1.500	1.000	0.600
H	1	4	0.999	0.500	0.000	1.500	3.000	0.333
L	1	1	**0.999**	0.500	0.000	1.500	0.000	**1.000**

Table 6.34 Intervention HFE: Group EP Data with Posterior-Parameters.

Context	x	n	EP0	Alpha	Beta	Alpha_post	Beta_post	E_i_1
A	1	3	0.995	0.500	0.003	1.500	2.000	0.428
B	1	5	0.995	0.500	0.003	1.500	4.000	0.273
F	0	3	0.995	0.500	0.003	0.500	3.000	0.143
I	0	3	0.995	0.500	0.003	0.500	3.000	0.143
J	1	1	**0.995**	0.500	0.003	1.500	0.003	**0.998**
K	1	2	0.995	0.500	0.003	1.500	1.003	0.599

Table 6.35 New Operator data for validation.

Group	Scenario	React._time	Resp._time	Ack.	mimics_opened	No_of_alarms
0	2	138	1020	9	3	14

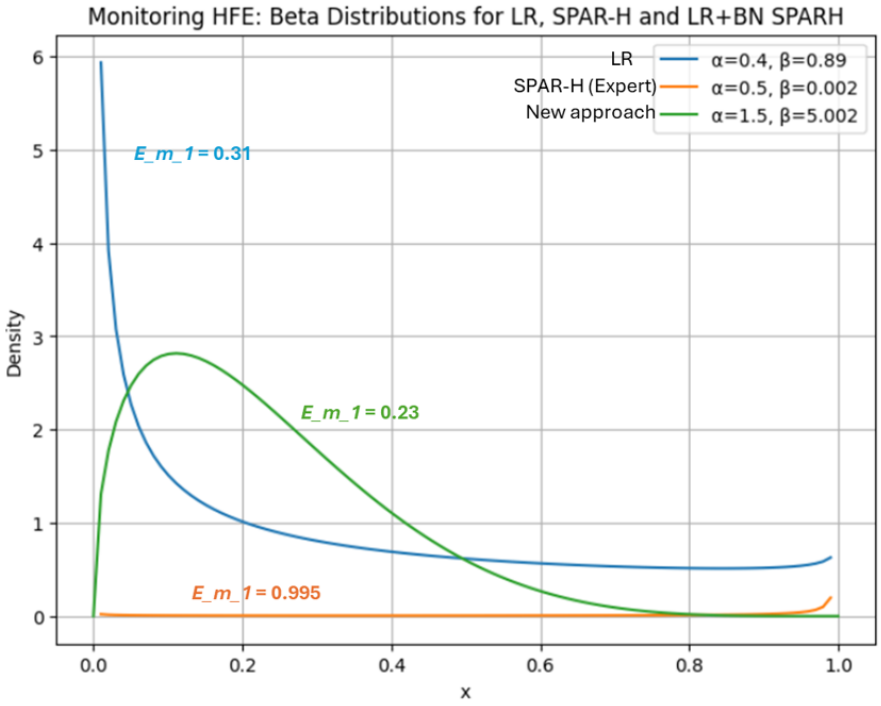


Fig. 6.1 Monitoring HFE: Beta Distributions for LR, SPAR-H and LR+BN SPARH. Participant outcome = Success.

$\beta = 0.89$). Adding these new data to Table 6.29, x still remains 1 and n becomes 6, given our assumption supported by the LR prediction that the operator is likely to succeed. With EP0 remaining the same, the new E_m_1 for context D becomes 0.231 where $\alpha_post = 1.5$ and $\beta_post = 5.002$. Figure 6.1 shows the distribution for the LR, SPARH and LR+BN SPARH results of the operators' failure probability in the monitoring task

Table 6.36 Monitoring Failure Probability from LR prediction.

Monitoring	Group	Scenario	Failure Probability
	PBP	3	0.310

For the planning task, Table 6.18 shows that the operator falls under context C, and a look into Table 6.21 shows that the operator is likely to succeed given that 7 out of 11 succeeded. This is further confirmed by the LR result in Table 6.37, which shows that the likelihood of failure is 0.33 ($\alpha = 0.37$ and $\beta = 0.75$). Adding these new data to Table 6.31, x still remains 4, and n becomes 12, given our assumption supported by the LR prediction that the operator is likely to succeed. With EP0 remaining the same (1.0)

while the new E_p_1 for context C becomes 0.36 where $\alpha_post = 4.5$ and $\beta_post = 8.0$. Figure 6.2 shows the distribution for the LR, SPARH and LR+BN SPARH results of the operators' failure probability in the planning task.

Table 6.37 Planning Failure Probability.

Planning	Group	Scenario	Failure Probability
	PBP	3	0.333

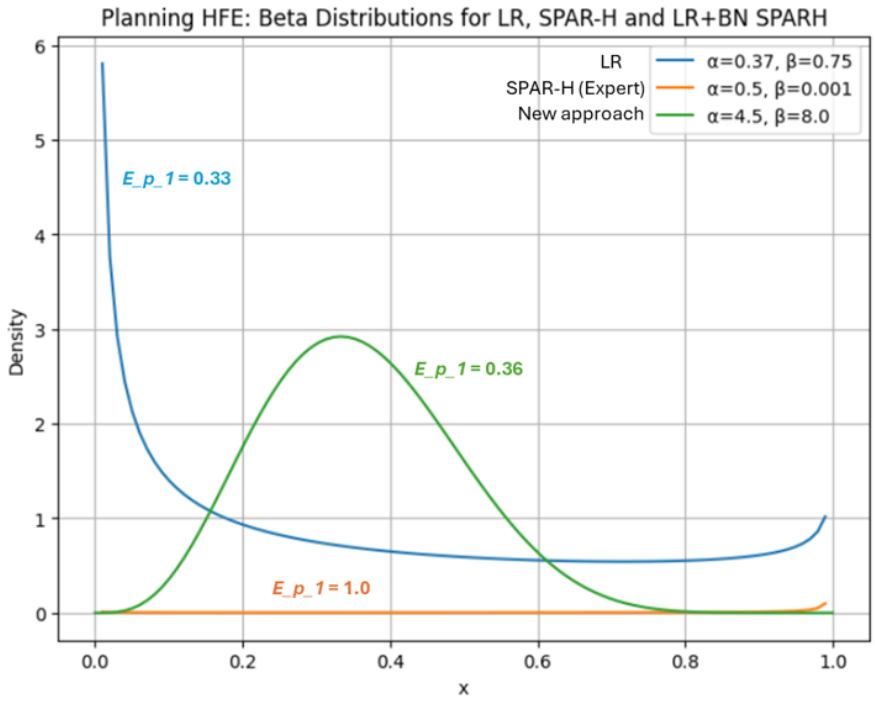


Fig. 6.2 Planning HFE: Beta Distributions for LR, SPAR-H and LR+BN SPARH. Participants' eventual outcome = Success.

For the intervention task, Table 6.19 shows that the operator falls under context D and a look at Table 6.22 shows that the operator is likely to succeed given that 2 of 3 succeeded. This is further confirmed by the LR result in Table 6.38, which shows that the likelihood of failure is 0.37 ($\alpha = 0.40$ and $\beta = 0.68$). Adding the new data to Table 6.33, x still remains 1 and n becomes 4, given our assumption supported by the LR prediction that the operator is likely to succeed. With EP0 remaining the same (0.999) while the new E_p_1 for context D becomes 0.33 where $\alpha_post = 1.5$ and $\beta_post = 3.0$. Figure 6.3

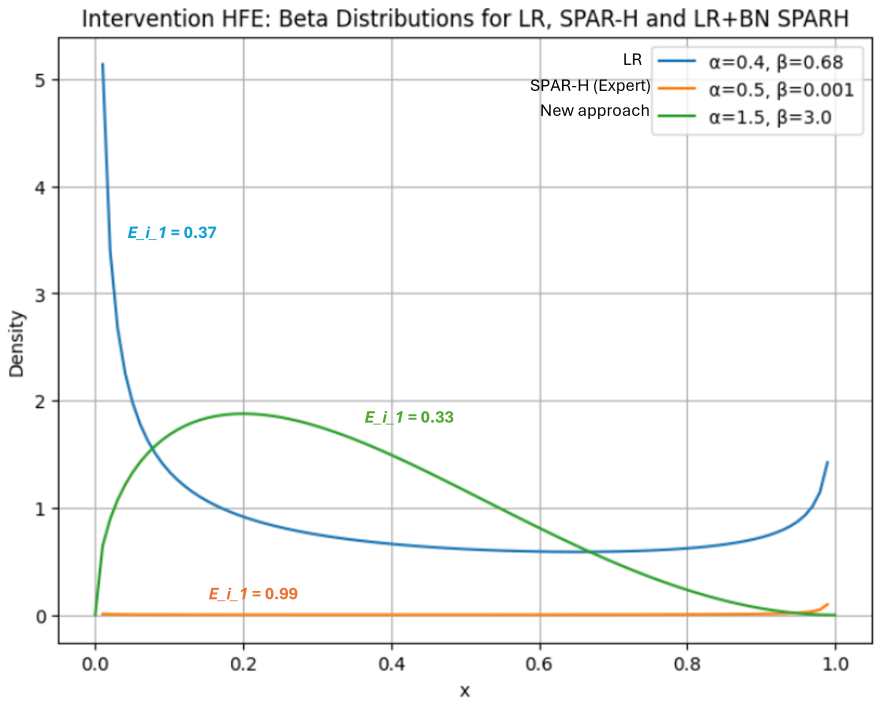


Fig. 6.3 Intervention HFE: Beta Distributions for LR, SPAR-H and LR+BN SPAR-H. Participants' eventual outcome = Success.

shows the distribution for the LR, SPARH and LR+BN SPARH results of the operators' failure probability in the intervention task.

Table 6.38 Intervention Failure Probability.

Intervention	Group	Scenario	Failure Probability
	PBP	3	0.367

Overall, the operator was likely to succeed in the monitoring, planning and intervention HFEs/tasks.

Gap: - Not considering the time of monitoring or planning etc., the model is based on end-of-study data and has not collected this data in real-time. - limited amount of data

6.4.2 Consequence and Risk Analysis

Figure 6.4 shows the error paths of the new operator with the likelihood of success and failure at each HFE node and the eventual consequence of each error path. The

emergency shutdown was optional for the participants to click in case they felt things were out of control and could no longer be managed. By clicking this button, they get a pop-up message that says the plant has been shut down. In Scenario 3 where they had to keep the plant from overheating for about 18 minutes, only two students clicked the emergency shutdown button (ESD). This action leads to the end of the test. The consequences in this scenario were: *safe*, *plant shutdown* in the case of a successful emergency shutdown, and *reactor overheating* (in the case of failed intervention and no ESD activation. The others either managed to survive for 18 minutes or performed actions that led to the reactor overheating. It is clear that the operator in Figure 6.4 managed to keep the plant safe with a 63% chance of succeeding.

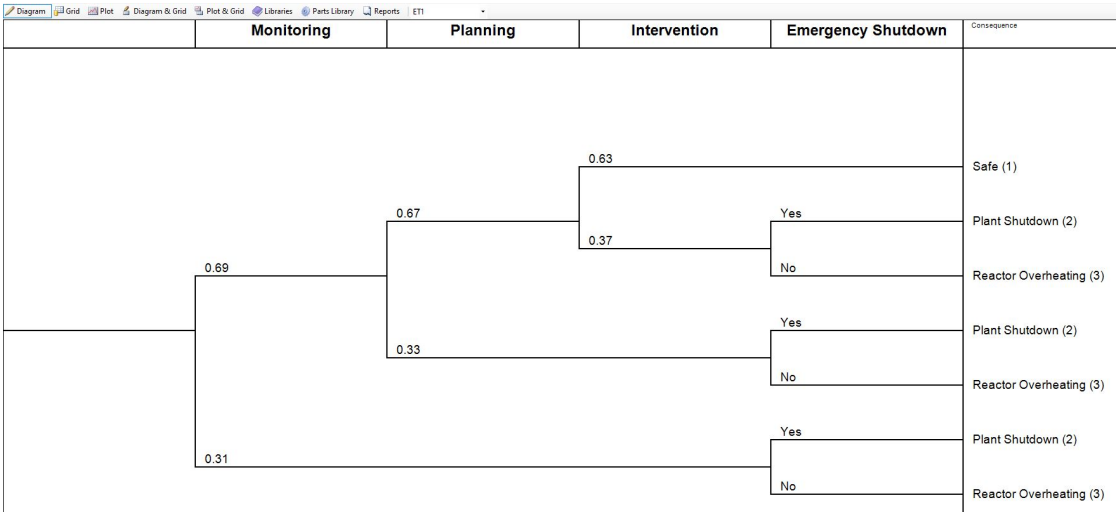


Fig. 6.4 Event tree for test operator outcome.

The risk matrix in 6.5 shows that the operator had a low-risk outcome for each of the human failure events (HFEs) with a failure probability of <0.4 and a severity value of 1 in each HFE.

6.5 Summary

The gap is that, though one scenario was needed for the sake of illustration, it is necessary to compare at least the outcomes of scenario 2 and scenario 3 at each phase of the framework. Scenario 1 and 2 are mostly not significantly different as seen from the

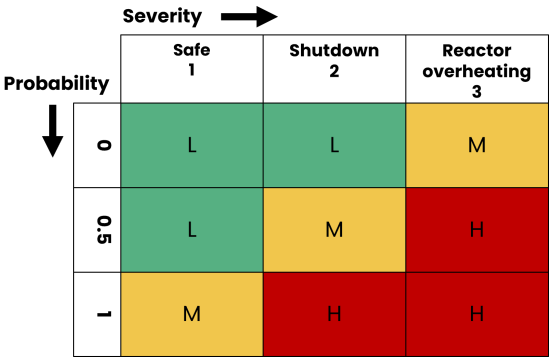


Fig. 6.5 Risk Matrix.

sensitivity analysis of almost all the measures, hence one can be used in place of the other.

Another gap is clustering the variables into key tasks. Ideally, all variables can be considered for each task. However, in this study, the analysis is done with data collected post-study. Hence, it is only fair to consider which variables would likely impact each task in order to estimate probabilities for each task. In future studies, the records can be kept for when operators move from one task step to the other which will entail a different design of the procedure and a synergy with the process.

Chapter 7

Discussion

7.1 Introduction

This chapter discusses the outcomes from Chapters 5 and 6. Section 7.2 delves into the findings of the metrics applied in the experimental study. These include subjective measurements such as the for workload assessment, and for evaluating situational awareness, and the perception of support from the procedures and alarm systems. Behavioural metrics covered were the count of acknowledged alarms, the number of mimics accessed, and reaction time, among others. Performance metrics focused on aspects such as errors, overall effectiveness, recovery duration, and impact. First, the subjective measures are discussed, and then the behaviour, performance and lastly, the eye-tracking measures. They are also discussed for how they relate to literature. Also, the results of the sensitivity and reliability analysis (correlation and factor analysis) are discussed in this section. The last part of Section 7.2 focuses on discussing the outcomes of the comparison of the study groups (those with paper versus those with electronic procedures) and the human factor issues that might have impacted the results.

Building on this, the chapter discusses the second phase of the thesis in Section 7.3, that is, the results from the methods applied in this phase. The methods included developed from the alarm response procedure and used as a basis by experts for the human reliability assessment using . Also, Logistic Regression was used for early prediction of the operators' probability of failure and to verify the error assumptions to be made while implementing the SPAR-H method. Section 7.3 concludes with a discussion on the validation outcomes of the real-time risk-based Human Performance Assessment

(HPA) and decision support framework. The discussion connects the findings from Phases 1, 2 and 3 of the R²HiTL framework.

7.2 Experiment: Human-in-the-loop in alarm handling and process control

This section aims to discuss the first phase of the thesis. It starts with a discussion of the measures introduced in this study, their interrelations with each other, the constructs they represent, and their implications regarding performance based on their positive or negative correlation with performance metrics. In addition, it provides an explanation for the potential reasons behind these relationships or effects on performance metrics. Further discussion includes sensitivity analysis and how results vary given different combinations of scenarios for factor analysis and correlation analysis. The goal is that at the end of this section, the second objective of this thesis which is to develop a dynamic, risk-based framework for human-in-the-loop assessment, addressing the following research questions:

- What factors are essential for HRA assessment in current human-in-the-loop configurations? And why? - How does the transition from paper to digital or computerised procedure options impact operators' general behaviour, cognition, and safety? Any difference between the two groups? If any, how significant are these differences?

The research question on the essential factors is addressed by examining the measures introduced in this study and the constructs derived from factor analysis. By identifying these constructs and their corresponding variables, the importance of these variables becomes apparent, emphasising the need to consider certain variables that are often overlooked in human factor assessments of operator-system interaction in process control rooms.

7.2.1 Workload, Situational Awareness, Perception of Support

The study investigated the subjective workload of the operators, the situational awareness and perception of the level of support from the alarm systems, and the post-scenario procedure. The measures used were checked for consistency and validity to determine

if they accurately measure, for this study, what they are known to measure in literature. The results are discussed in detail below for each measure.

The positive correlations of NASA TLX and most of its dimensions (mental, temporal, and physical demand, effort, and frustration) with SART's demand and supply, which are dimensions of situational awareness of SART, have been reported similarly in the literature by Braarud et al. [17]. This questions the interpretation of the SART Index to indicate situational awareness. As observed from the factor analysis, SART demand and supply also loaded similarly as NASA TLX under the construct, *Workload*, in Table 5.32. In contrast, the SART index itself and SART Understanding were positively correlated with the SPAM index and its dimensions, which assess situational awareness. The SART index's correlation with SPAM is largely driven by the SART understanding dimension. The reason for this observation could be explained by the SART demand and supply questions. SART demand is based on the questions that access demand on attentional resources, which are highly linked to the amount of load from the task. On the other hand, SART supply assesses the supply of attentional resources with questions on arousal, concentration, etc, that will be highly influenced by the task load and complexity. To this effect, SART demand and supply can be associated more with workload when discussing the group-wise results of these measures. Also, this textitworkload associated variables are observed to have a positive relationship with the behavioural variables, except reaction time and number of mimics opened, and all performance measures. Indicating that higher workload values increase the likelihood of failure in terms of error [Success: 0, Failure: 1], overall performance [optimal: 0, good: 1, poor: 3] and lead to more adverse consequences [Safe: 0, Impurity of air in tank atmosphere: 1, Plant Shutdown: 2, Reactor Overheated: 3]. Also, it shows that higher workload can reflect in terms of behaviour, for example, spurious acknowledgement of alarms as opposed to goal-specific acknowledgement, the number of alarms annunciated as a result of operators' actions, and consequently, the performance outcome.

Another interesting finding is the positive correlations between the perceived performance of the TLX measure and the situational awareness measures, SPAM and Index (particularly SART understanding). This lack of correlation between perceived performance and the TLX index, along with its dimensions, has similarly been noted in the literature by Braarud et al., [17]. This could entail that both informed and non-informed situational awareness influence one's perception of one's performance. However, subjective performance does not significantly load with these variables

following the factor analysis, and neither did it with the objective performance outcome - Overall performance, Error and Consequence. On the other hand, the SPAM index and its dimensions SA1, 2 and 3 correlated positively with the procedure support and the SART Index and Understanding. This confirms its worthiness for assessing situational awareness. However, factor analysis showed that they loaded separately from the SART index and understanding. The result of this loading is understandable given that SART and SPAM assessment assessed situational awareness in different ways. The SART questions were administered post-scenario, while the SPAM questions were asked following certain cues like the annunciation of an alarm a few minutes after the opening of a procedure and a few minutes/during the control task. In the former, the participant tries to recall their experience while responding to the questionnaire, while in the latter, the participant would sometimes cross-check with the alarm list or procedure to confirm their response. Braaud et. al., [17] in their study, referred to the former as a case of *non-informed situational awareness*. The latter approach applied through SPAM is somewhat similar to their SA3 situational awareness method was referred to as a case of *informed situational awareness*.

The significant positive loading of SPAM and SART Index with the perception of procedure support confirms the strong positive correlation/relationship between the procedure and situational awareness. In essence, better procedure support increases situational awareness (informed or non-informed). These two situational awareness constructs are worth distinguishing while discussing the situational awareness and performance of the two groups - those using paper versus those with electronic procedures. The idea of having good procedure support for better situational awareness is that operators can be kept in the loop regardless of task complexity or during the sudden annunciation of an alarm after prolonged monitoring, especially in present-day control rooms. This role of keeping operators in the loop is, however, not left to the procedures alone but all other human system interfaces, like the alarm system and the displays. Both SPAM and SART had an inverse relationship with the behaviour and performance variables they significantly correlated with. Specifically, response time, alarm silenced, acknowledgement, no of alarms, accuracy, recovery time, and error. This entails that, for example, as situational awareness increases, the response times become faster with a reduction in the number of alarms and those acknowledged or silenced. It also reflects on performance in that the mean square error, that is, accuracy, is reduced, recovery time is faster, and error tends towards success. These behaviour and

performance outcomes are typical with good situational awareness [33]. It is important to remember these relationships, especially when discussing the differences between the study groups. For example, looking at the correlations for scenario 3 in Table 5.36, the SART Index (non-informed situational awareness) negatively correlates with the number of visits to AOI, showing that they have opposite relationships and impact on performance results.

The alarm list and priority support were negatively correlated with the TLX index and its dimensions but positively correlated with the SART index and the SA1 and SA3 dimensions of SPAM. This shows that good alarm prioritisation would reduce workload while increasing non-informed situational awareness and situation perception (SA1) and projection (SA3). Although SART understanding was also accounted for in this positive correlation with the alarm support, there was no significant correlation with SPAM SA 2 - Situation understanding/comprehension. The alarm list and priority support also had a significant negative correlation with the acknowledgement of alarms and the consequences. This is an expected outcome for alarm design, as good alarm support can improve perception, which we see from the reduced acknowledgement of alarms. Perception is the first step in the information processing for a control room operator during alarm handling and process control and sets the floor for planning correctly, in this case, finding the correct procedure. Poor alarm design, with its associated issues such as alarm floods and so on, is known to overload the operators, with the tendency to reduce situational awareness. These situations have earlier on in Chapter 2 been mentioned as causes of some of the major accidents in the safety-critical industries, for example, in the case of Buncefield oil storage depot [87]. Interestingly, operator familiarity and training were also significant in this study, but more for non-informed situational awareness, such as in the case of alarm support. This shows the significance of operators' training and familiarity with non-informed situational awareness. In general, operators depend on their memory (a part of the Limbic System of the brain, see Chapter 2), and with effective attention filtering, they can recall memories from past experiences and training to accurately perceive, comprehend and project the possible results of a situation.

For most variables that showed sensitivity between scenarios 2 and 3, and following the factor loading for scenario 3, it becomes obvious that clear distinctions between workload and situational awareness cannot be made in a very complex scenario. Hence, there are variables for such situations that clearly describe cognitive workload, while some describe a combination of cognitive workload with either informed or non-informed

situational awareness as shown in Table 5.32. This must be taken into account when specifically discussing the comparison between the study groups.

7.2.2 Behaviour and Performance

The study also investigated behavioural and performance variables for their relationship with each other. The reaction time, which is based on the first control action, was observed to correlate positively with overall performance but negatively with consequence and error, indicating that a longer reaction time leads to poor overall performance, but does not translate to failure in terms of error or adverse consequences. Longer reaction times as earlier discussed in Chapter 2 can stem from increased cognitive flexibility and can result in a decrease in accuracy and sometimes lead to mental fatigue. Reaction time was not sensitive to task complexity as observed from the sensitivity analysis; therefore, the relationship between reaction time and performance measures can be considered the same in all scenarios. Furthermore, the reaction time did not show a significant correlation with subjective, eye-tracking and other behavioural variables or an influence from them, except for an almost significant positive correlation with the support of the alarm list, which is expected given the role of the alarm design in potentially guiding operators reaction times. Alarm support in this study included a combination of text, colour, and sound which in the literature causes visual cortex activation and fosters faster reactions [27]. However, in this thesis, this type of reaction is better observed from the variable *number of alarms acknowledged*. Hence, the almost close significant relationship between alarm support and reaction time, which in this study is a latter behaviour observable from successful or unsuccessful planning.

The response time, on the other hand, which is the time to correct action, is positively correlated with other behaviour and performance variables, with the exception of reaction time. The positive correlation of response time with consequence and error contrasts with the observation in the case of reaction time, which has an opposite correlation with these performance variables. The positive correlation of response time with other behavioural variables like the number of mimics opened or the number of alarms acknowledged, which are also positively correlated with workload, show that increasing workload leads to longer response times. Subsequently, longer response times would lead to failure outcomes [Error: 1], poor overall performance, and more adverse consequences. As shown in Table 5.34, response time shows a significant negative

correlation with both the SPAM and SART Index, which represent informed and non-informed situational awareness, respectively. Consequently, higher values in these situational awareness metrics typically result in faster response times. This effect aligns with expectations regarding situational awareness and procedure support (which also significantly negatively correlated with response time). Properly perceiving an alarm, understanding the situation, and troubleshooting allow operators to make timely control adjustments. Response time did not correlate significantly with the eye-tracking measures that describe operators' behaviour when using the procedure.

Furthermore, other behavioural variables, like the number of alarms silenced and acknowledged and mimics opened, did not significantly correlate with consequence and overall performance. However, alarms silenced and acknowledged positively correlated with error and accuracy. Therefore, these behavioural variables had a more significant impact on accuracy and error than they did on the overall performance and eventual consequence. Essentially, the more the silencing and acknowledgement of alarms by the operators, the more likely they are to have higher mean square errors and run into a failure outcome. This is much more evident in scenario 3 in that acknowledging many alarms indicates increasing workload and poor situational awareness since the operator is not specific as to the critical alarm to handle. Not being specific during the acknowledgement process is a sign of poor perception, a first step in situational awareness. Operators are expected to perceive which alarm is critical and resolve that before moving to the next critical alarm. Also, the impact of this behaviour of silencing alarms is more pronounced in the last scenario, given the alarm flood situation in this scenario. Given that the silencing of alarms impacts recovery time, accuracy and error, it is important to subsequently redesign the system to prevent silencing each alarm that pops up and provide an alternative means to remove the sounds without requiring operators' action on the alarm message list, especially in alarm flood situations. This could be done by a timely stimulation of the auditory senses for just a few seconds, enough to allow the recording of this alarm information in both the working and short-term memory. The alarm sound is then removed in this short time, leaving only the visual cues (colour and text) for visual attention support. Additionally, the number of mimics opened was found to significantly positively correlate with both accuracy and recovery time. Specifically, an increase in the number of mimics opened results in higher mean square error and prolonged recovery time. This effect is particularly evident in scenario 3, the most critical one. Further analysis of scenario 3 alongside eye-tracking measures indicates that both

the number of mimics opened and recovery time are significantly correlated with saccade amplitude (the distance the eye travels between two fixation points), which, when longer, may be associated with difficulty in shifting attention. Referring to the results of the factor analysis for this scenario in Table 5.17, it becomes clear that the number of mimics opened indicates cognitive workload, as an increase in mimics opened corresponds to a greater cognitive workload, leading to reduced accuracy and extended recovery times.

7.2.3 Eye-tracking measures

The eye-tracking measures considered in this study, that is, the average fixation duration, the number of visits, and so on, describe the behaviour of the operators and provide insight into the cognition of the operators while reading or navigating the alarm response procedure. These measures describe how the use of either a paper or electronic procedure impacts operators on a cognitive and behavioural level. The relationships between these measures and those of other variables in this study provide insight as to what they describe for this study and for potentially other human-in-the-loop process control setups where an alarm response procedure is used.

The average duration of fixations during the procedure was significantly inversely related to the number of fixations but seemed positively correlated with the SA understanding dimension of SPAM. This suggests that, although both fixation duration and fixation count are linked with attention, demonstrated by the SA understanding, they signal opposing trends in relation to SA understanding. Specifically, a higher number of fixations may indicate decreased situational understanding, whereas longer fixation durations might enhance it. The longer fixation and its positive relationship with situation understanding indicate a learning process as commonly observed when training operators [173]. Conversely, the SART index showed a significant negative correlation with the number of visits to the procedure, indicating that a higher frequency of visits may decrease objective situational awareness. In addition, Table 5.36 demonstrated that the frequency of visits to the procedure exhibited a positive correlation with the acknowledgement of alarms. This result is anticipated, as acknowledging an alarm is followed by seeking out the relevant procedure. Considering the association between these variables and their negative correlation with the SPAM index, it is suggested that they represent both objective workload and situational awareness, albeit with distinct relationships. However, the average pupil diameter and the average saccade amplitude

did not have significant correlations with the SPAM or SART measures. However, they correlated with some behavioural and performance variables.

Pupil diameter exhibited a notable positive correlation with the operators' familiarity with both the control room and the process, as well as with the duration of saccades. This suggests that increased pupil dilation, or larger pupil size, can indicate greater familiarity, even when there are longer saccade durations, which generally signify reduced information processing capacity or attention. In contrast, the average saccade amplitude showed a positive correlation with the number of mimics opened. As mentioned before, longer saccade amplitudes may be linked to difficulty in switching attention or search, as similarly mentioned by Shi et al., [129]. Consequently, an increased number of mimics and longer saccade amplitudes describe scenarios of *increasing search difficulty* for operators. It has been discussed that the number of opened mimics affects recovery, response time, and accuracy; more opened mimics typically result in longer recovery or response times. Therefore, increasing search difficulty will have a similar impact on these performance measures.

7.2.4 Discussion on Group-wise Comparison

This subsection delves into the results of the two study groups, focusing on constructs derived from factor analysis—such as Workload, informed and uninformed situational awareness, behaviour, performance, familiarity, and confusion. While most measures revealed no significant differences between the groups, certain measures distinctly differed. These particular measures and the constructs to which they relate will be examined in detail, together with an analysis of group performance despite the absence of significant differences. Additionally, it's crucial to identify factors that may have influenced these outcomes, whether significant or not and investigate other explanations based on observations made by the supervisors during the experiment.

Workload and Situational Awareness

Significant differences were not found between the two groups on the variables representing the construct *Subjective workload*. Specifically, TLX Index, Mental and Temporal demand, Effort, Task load, SART demand, and Supply showed no significant variation. However, on average, the group following the paper-based procedure (PBP) perceived a higher workload than the group using the electronic procedure across all

scenarios. For clarity, the PBP group sensed greater mental and temporal demand, particularly as tasks increased in complexity. They also reported exerting more effort during each scenario, which correlates with their assessments of the demand and supply components of SART, where the demand of SART is related to the demands for attentional resources and the supply of SART is related to the provision of attentional resources. The heightened perception of mental workload by the PBP group might be attributed to several factors. To explore these, a primary step would be to examine the participants' perception of support from the procedure or alarm design and the objective data on the participants' cognitive behaviour while reading and searching through the procedure. Following this, it would be beneficial to consider other experiment observations or design issues recorded post-study to understand how they might have contributed to these workload perceptions.

Participants in the PBP groups perceived that they had better support for the paper procedure and alarm design than those in the EP groups. Therefore, the procedure at a subjective level does not appear to have influenced their subjective workload ratings. Yet, they felt more frustrated and felt that they put more effort into the task. However, looking at the eye-tracking metrics from the use of the procedure and the constructs in Table 5.18 for Scenario 3, PBP had a significantly higher average saccade amplitude than EP with a slightly higher average number of mimics opened across scenarios 1 - 3. Higher saccade amplitude and mimics opened describe the construct search difficulty with a positive relationship. Therefore, the PBP group had higher search difficulty during their tasks, especially in Scenario 3. In addition, the PBP group had a significantly larger pupil diameter and a non-significant longer saccade duration than EP. Although this shows better familiarity for scenario 3, as seen in Table 5.36, it also shows that in scenario 3 and likely in the other scenarios, they had a decline in their average processing capacity. Another explanation for the high subjective rating for PBP could be their training performance. Although this was not significantly different from EP, PBP had a lower taking performance score than EP.

Though Group EP exhibited lower workload ratings across all tested scenarios, they also had reduced scores in subjective situational awareness as assessed by non-informed situational awareness (from SART) and informed situation understanding (from SPAM). They demonstrated superior informed situational awareness only in the more intricate scenarios, particularly regarding situation perception and projection. This enhanced situational awareness in Scenario 3 clarifies their low workload rating during this scenario.

The occurrence of low workload alongside diminished situational awareness, particularly in understanding, might stem from various factors. To investigate these factors, retracing the steps in procedures and alarm support and examining other study observations would be a logical starting point. Ideally, group EP's consistently lower workload values compared to PBP across scenarios would imply that they should exhibit higher SA than PBP; however, this was not the case. Part of the explanation is found in their reduced perception of support from the electronic procedure. The procedure's utilisation, as previously noted, shows a significant correlation with both SPAM and the SART index as seen in Table 5.34, indicating its influence on informed and non-informed situational awareness, respectively. Generally, the electronic procedure's menu window facilitated searching, but an inversion in the plant section and alarm tag ID arrangement on the alarm list was problematic [155]. To elucidate, maintaining the original sequence when scanning a list of titles is recommended best practice. For instance, Figure 7.1 (1) illustrates the alarm display with the tag ID on the left, followed by the section on the right, mirroring the typical reading order. Contrarily, the electronic procedure's layered menu reverses this order by displaying the section first (on the left) and the tag ID on the right, demanding extra cognitive processing. A proposed improvement involves altering this order, as depicted in Figure 7.1(2), so that the section precedes the tag ID in the alarm banner, matching the electronic procedure's format. Such a change is expected to further alleviate the cognitive load for EP users and enhance their situational awareness scores.

Participants using the EP procedure potentially experienced reduced situational awareness due to the HMI configuration illustrated in Figure 3.2. These participants needed to locate and interpret the procedure on the diagnostic screen (4), then transition to the equipment details screen (3). This necessitated repeated head movements during the intervention period. Throughout this process, their focus may have been drawn away by elements on the central display (1) or other displays that become salient and detectable by their peripheral vision. Researchers have noted that such extensive head movements can impair attention and memory [33]. Additionally, the EP group rated their alarm list and priority system support lower compared to the PBP group. This negative perception of alarm support, coupled with their already low opinion of procedure support, diminished their view of situational awareness. Eye tracking data sheds further light on the procedure's impact on their situational awareness; those in the EP group exhibited longer fixation durations, as shown in Table 5.12 relative to the PBP group. Although

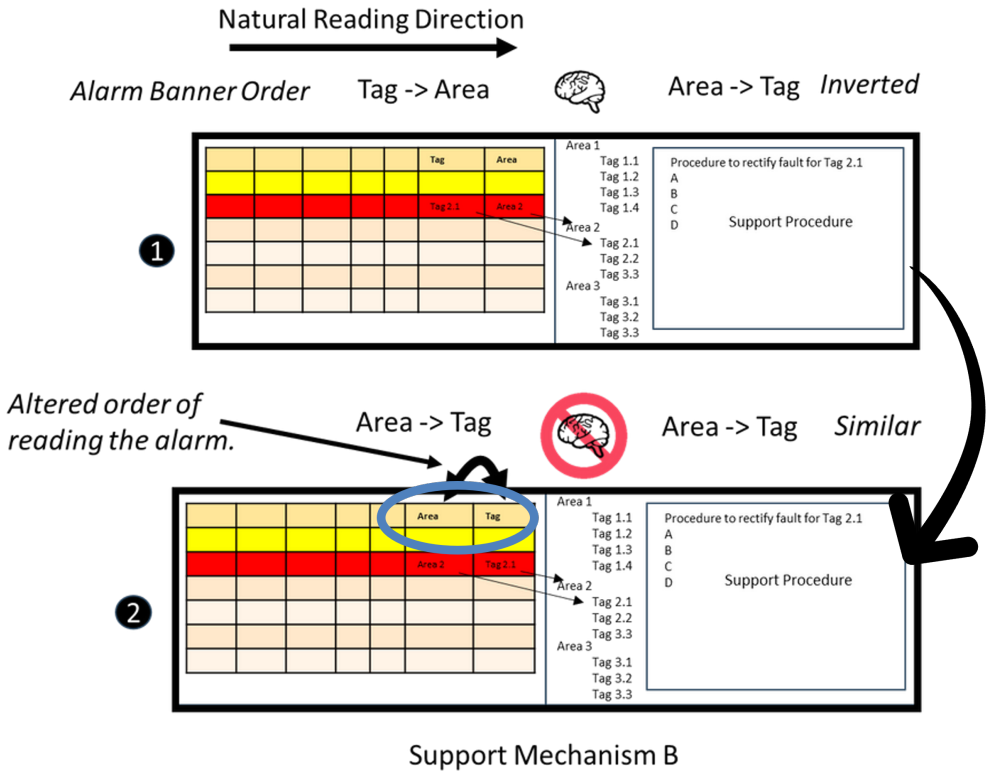


Fig. 7.1 Alarm list to Procedure Navigation human factor issues. 1. Incorrect layout 2. Proposed correct layout.

extended fixations typically suggest enhanced engagement and learning, the EP group demonstrated poorer situation comprehension compared to PBP in scenario 3. This indicates that while longer fixations might aid in perceiving and projecting situations, they fell short in improving situational understanding for EP participants. A possible explanation is that in critical, time-sensitive scenarios, extended fixations can result in Perseveration (difficulty shifting focus), Inattentional Blindness (visual fixation without cognitive processing), or reduced planning and evaluation abilities [33]. These factors, along with potential unidentified ones, might have contributed to the EP group's low situational awareness rating, despite their lower workload ratings compared to the PBP group.

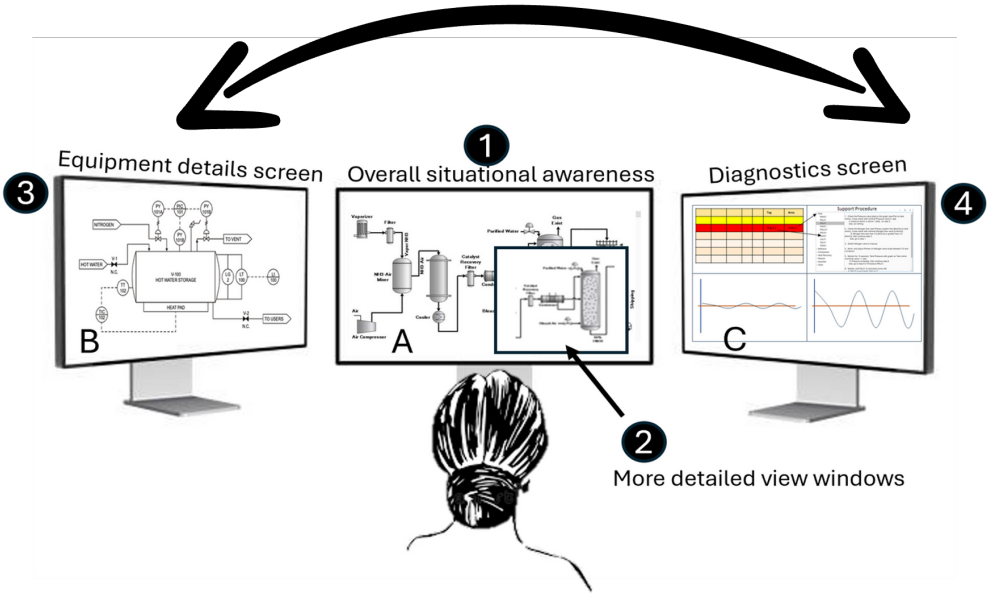


Fig. 7.2 Head movement for those using the electronic procedure.

Furthermore, those using the paper procedure despite their high workload scores might have had a better perception of the procedure support and situational awareness for the following reasons:

- Readability: Some individuals hold the view that reading is the same regardless of whether it is on paper or a digital screen. However, research suggests a more complex situation, especially when factors such as the reading purpose are considered. For instance, when the goal is learning, the paper proves

advantageous. Engaging with text through underlining, highlighting, and using a finger as a placeholder is beneficial to the learning process. Making summarising notes and visual aids helps in mentally organising the material, facilitating better memory encoding. This kind of deep reading enhances comprehension, memory retention, and critical analysis. Conversely, when the intent is to grasp the gist of a story, check texts, or browse social media, electronic formats are preferable. Yet, in this particular study, the reading material was not meant for learning or obtaining a gist but was focused on procedural following and operator assistance. Although learning was not the objective, those using paper formats gained the engagement benefits discussed from using prints and had fewer colour contrast issues that arise with screen use.

- Mixed text genres: Also, the activity of following a process-specific support mechanism is constrained by time and contains information that is a mix of information and narrative text. In such cases, printed documents are preferred for use as they support reading comprehension [137]. This supports the reported findings from the experiment showing that the PBP group had better situation understanding and overall situational awareness than EP.

Behaviour and Performance

The behavioural measure, number of mimics opened, loaded alongside the average saccade amplitude under the construct, Seach difficulty. The mimics opened positively and significantly correlated with accuracy, response time and recovery. Indicating that increasing search difficulty will lead to longer response and recovery time and larger mean square error. The difference between group PBP and EP was insignificant in each of the scenarios. However, from the use of the procedure in scenario 3, there was a significant difference in the saccade amplitude of the two groups. Those using the PBP had significantly larger saccade amplitude than EP. On the other hand, the number of alarms acknowledged loaded negatively under Factor 2 in table 5.33 with similar relationship as the number of visits and mental demand. These had an opposite relationship with the SART Index, SART understanding and alarm list support which had a positive loading on this factor. This shows that as the number of acknowledgements increases, objective situational awareness decreases. Participants in group PBP, on average, had a similar number of alarms acknowledged as those in the EP group in the

first scenario. However, as the scenarios progressed, the number of alarms increased for both groups but more so for those in PBP, especially in scenario 3. Confirming, as earlier discussed, that they had a higher workload and progressively lower situational awareness, especially in scenario 3. The number of alarms acknowledged has a significant positive relationship with the silencing of alarms and the number of alarms. These latter variables might thus have a similar relationship as the number of alarms acknowledged on objective situational awareness and workload.

Compared to the EP group, participants in the PBP group exhibited longer reaction and response times. Factors such as observed behaviours, initiation of processes, mimicry, and acknowledgement of alarms may contribute to increased reaction and response durations. These factors also inform workload and situational awareness, which subsequently affect performance metrics like recovery time, accuracy, overall performance, error rates, and resulting consequences. While both groups excelled in scenario 1, performance diminished as scenario complexity increased. In scenario 2, the EP group slightly outperformed PBP, securing one additional successful outcome. However, in scenario 3, the PBP group surpassed the EP group in minimising errors and consequences. Specifically, there were 6 failures with reactor overheating out of 17 participants in the PBP group, while 9 in the EP group encountered similar issues. Though the EP group showed an edge over search difficulty, processing capability, workload, and situational perception and projection in scenario 3, the PBP group excelled in the performance outcomes. This suboptimal performance might result from learning activities associated with EP procedures in intricate, time-critical scenarios and the inadequate situational comprehension noted within this group. Group EP's extended fixation duration, combined with frequent fixations, suggests possible reading challenges, potentially stemming from difficulties in identifying or correlating procedural rules [47]. For instance, in electronic formats, rules such as 'if,' 'else,' 'go,' and 'then' weren't bolded or italicised, while paper-based procedures did emphasise these terms. The suboptimal results could also relate to the HMI configuration in Figure 7.2, necessitating continuous head movement or other navigation challenges mentioned in this Chapter.

Additional Observation

A common observation from both groups highlighted the low accessibility and error caused by the contrast in the colour of the alarm list, specifically the use of black text on

a red alarm banner, as shown in Figure 7.3. The WCAG Criterion 1.4.3 (AA) deals with the contrast in visual presentations of text and images, specifying a minimum acceptable contrast ratio of 4.5:1. The contrast ratio measured during the experiment was 4.91:1, which slightly surpasses the WCAG AA requirement. Nevertheless, the AAA WCAG Criterion 1.4.6 for Enhanced Contrast elevates the minimum ratio to 7:1 [1]. This finding is crucial, considering that red banners are frequently utilised in HMI design to capture an operator's attention.



Fig. 7.3 Alarm list showing two alarms: FAL01 and PAL01.

Gap

The experimental research was not free of limitations that may impact the findings. One such constraint is the number of participants in the study. As discussed in Chapter 3, the experimental design should potentially involve over 30 students per group. Furthermore, the limited sample chosen for eye-tracking analysis in scenario 3 restricts the insights necessary for comparing the two groups. Also, solely relying on the results from scenario 3 for eye-tracking constraints discussions of scenarios 1 and 2, as the findings from scenario 3 may not fully generalise to the other scenarios. In conclusion, the eye-tracking data was exclusively collected for the procedure area of interest (AOI) and even more specifically during the intervention time window. Future analyses can investigate how interactions with other areas of interest, such as display elements (including values, graphs, and alarm lists), influence the constructs established in this study or affect workload, situational awareness, behaviour, and performance.

[1] W3C,2023.Web Content Accessibility Guidelines (WCAG) 2.2. [Online]. Available at: https://www.w3.org/TR/WCAG12/

7.3 Phase 2: Real-time Risk-Based HPA and Decision Support

This section discusses the result of Chapter 6 of this thesis. The data used came from behavioural and performance metrics collected during the experimental research detailed in Chapters 3 and 4, along with information from expert evaluation of the task steps as outlined in the procedure and presented in Chapters 5 and 6. The behavioural variables were clustered under the three main tasks or so-called human failure events of process control room operators. That is monitoring and alarm handling, planning, and intervention tasks. The monitoring task comprised the variables: Number of alarms annunciated, the number of alarms acknowledged, group/type of support, and scenario/task complexity. Planning comprised the number of mimics opened, the Group/type of support and the scenario/task complexity, while Intervention comprised the reaction time, response time, number of mimics opened, scenario and group. These were input to predicting the likelihood of failure for each of the HFEs using the logistic regression model and are key to answering the research question 3 - Which operator behavioural metrics are the best predictor of a successful recovery of a plant alarmed state?

7.3.1 Discussion on Phase 1 and 2 of the R²HiTL framework

The result of the stepwise logistics regression for the monitoring task showed that the best predictor of the probability of failure for this task was the number of alarms annunciated. However, all variables were combined for the final prediction, as they each contributed negatively or positively, at different weights, to the error prediction. The average cross-validation accuracy was approximately 86%. The results in Table6.3 showed that those in Group PBP (using the paper procedure) had a lower likelihood of failing the monitoring task in all three scenarios. A similar performance result was observed for the planning and intervention tasks. For the planning task, the best predictor variable was the Group/type of support with prediction accuracy from the combination of all variables at approximately 81%. The prediction accuracy for the intervention task with all variables was approximately 81%. However, the best variables for this task were the Group, scenario, and response time.

Hierarchical task analysis was helpful in structuring the tasks written in the procedure. It laid the foundation for the SPAR-H assessment of which tasks are diagnosis and action tasks. With a focus in scenario 3, the action tasks were those that required the operator to interact directly with the display elements, for example, acknowledging an alarm, clicking or moving a button, etc. Table 6.7 show how the steps are grouped into the HFEs - Monitoring, planning and Intervention. It also confirms the association of variables like the number of mimics opened to solely the planning and intervention tasks, as this is not required during monitoring or alarm handling. Clearly, from the table, each of these HFEs constitutes both action and diagnosis tasks. The next step involved the assessment of the subtasks by an expert on the likelihood of failure of members of either the Group PBP or EP on each of those tasks.

The expert assessment of Group PBP and EP resulted in different adjusted failure probabilities in only two tasks and similar outcomes in the others. Group PBP had a higher adjusted failure probability of 0.953 on the Monitoring/Alarm handling task, *identify the critical alarms*, as compared to group EP (Adjusted HEP = 0.669). Group PBP had a similar probability outcome compared to group EP on the Intervention task, *Monitor for a while cooling water system mass flow rate*. Following this, task dependency was checked for as is required for tasks that are highly dependent on the completion of the other. This was done by assigning values 0 or 1 depending on the possibility of changes to the following variables - crew, time, location, and cues, as the tasks progressed. For example, 0 is assigned to the task following another if the crew handling that task remains the same. As can be seen for both group PBP and EP, this value remained sthe ame (same crew = 0) for all tasks. This was the opposite for the location as the areas of interest changed for each task (different location = 1). The tasks were, in a few cases, not close to time (time = 1), while cues were present in some of the tasks (cues = 0). Cues here refer to the activation of an alarm or the change of trend in a process graph.

The HEP following the dependency assessment increased significantly for most of the tasks. This shows that tasks that are highly dependent on each other tend to increase the likelihood of failure. The tasks estimated to have the likelihood of failure from the operators in both groups were all diagnosis tasks. This shows that operators are likelier to fail a diagnosis task than an action task. This outcome can be due to the SPARHs assignment of a higher NHEP to diagnosis tasks (NHEP = IE-2) compared to the action task (1E-3). The outcome between the diagnosis and action tasks can be explained by understanding the effort exerted on memory for both error types. The diagnosis task

is more demanding on memory as it involves a lot of cognitive processing (perception, interpretation, comprehension and so on) that leads to a decision before an action occurs, as discussed in Chapter 2 under the subsection on *Memory*. On the contrary, the action task is merely semantic and procedural.

The final HEP estimation showed that both groups had the same high failure likelihoods in all the tasks/HFEs, especially in the planning task (E_p_0 = 1.000), while those in Group PBP had a high chance of failing the Monitoring and alarms handling (E_m_0 = 0.995) and the Intervention tasks (E_i_0 = 0.999). This estimation has its gaps in that it has been completed by only one expert and could exaggerate the error outcomes. It furthermore assumes that all participants in their groups have the same impact from the different performance shaping factors (PSFs), which is not the case in a realistic setting. For example, stress, fitness for duty, training and experience, and so on, are individual factors that will vary per participant. This gap in the SPAR-H method and thus the expert assessment, can be balanced by the introduction of the behavioural variables from the experiment and the analysis steps of the BN SPARH method to be discussed next.

7.3.2 Phase 3 Discussion

The last phase of the R²HiTL framework involved the BN SPARH method, a third-generation HRA assessment method. The BN SPARH method, as applied in this thesis, takes the variables under each HFE/task as input and discretises these variables using scenario 3 as an example. The result of the discretisation shows that the number of alarms is high when greater than 18 and low when less than or equal to 18. Likewise, the number of alarms acknowledged is high when greater than 5 and low when less than or equal to 4.5. The same operand pattern applies to the number of mimics opened with an optimal split of approximately 4, reaction time at 266 seconds and response time at 856.5 seconds. The group and scenario variables were not discredited, given that it was either group PBP or EP, and scenario/task complexity stayed the same.

The next results were those from the contextualisation of the variables in each HFE. This result shows that participants of the same support group can have different levels to the variables that describe their performance. For example, those in group PBP fell under contexts D, E, F and G in the monitoring task alone. Basically, there are four (4) outcomes from the combinations of the discretised variables for people in group PBP

under the monitoring HFE. It shows, too, that eventual performance outcomes might not be the same for all of them in the same group. The results of the pairing with the *error* in Table 6.20 further confirm this. The table shows that a participant in a context, say A, can either fail or succeed. So, if two operators have similar behaviour, it might not eventually lead to the same outcome. Predicting what this outcome would be is made possible by the Logistic regression model. This LR + BN SPARH method is in contrast to the SPAR-H method, which assumes that all participants are affected at the same level (PSF multipliers) by the PSFs because they fall under the same group/type of support [113].

Tables 6.29 to 6.34 show the expert estimate (EP0) of failure probability for each HFE and for each group, and the posterior probability estimation (E_m_1), (E_p_1) and (E_i_1) for each HFE - monitoring, planning, and intervention. As discussed above, the same expert estimate applies to all contexts since the SPAR-H method does not derive context during estimation. For the monitoring task/HFE, participants in contexts D, E and F in group PBP had similar probabilities as those in contexts A, B and H in group EP. Their probabilities of failing were far from the estimates of the expert. However, a context in each of the groups had high probability estimates close to those of the expert, as shown. In particular, context G of group PBP with a 99% probability of failing the monitoring task, which is close to the 99.5% expert estimate of the expert, and context C in group EP with an 82% chance of failing the same task close to the estimate of 96.6% experts. This shows that following the SPAR-H method alone without the inclusion of live data capturing *work-as-done* and the real impact on operators could lead to a wrong conclusion on or expectation of failure probabilities since it generalises for all cases.

Likewise, for the intervention task, only one context in each of the groups had a close estimate to the experts' estimation. For group PBP, this was context L, with only one participant in this category having a 100% chance of failing, which is close to the 99.9% experts estimate. Those in contexts E and G in this group had an average probability of failing this task with a probability of 60%. For group EP, context J had a 99.8% chance of failing, close to the 99.5% estimate of the expert. The number of contexts in the intervention HFE increased compared to the monitoring or planning tasks because of the number of variables discretised and included for the contextualisation.

On the contrary, the number of contexts in the planning tasks decreased because only two variables, *the number of mimics* and *the group/type of support* varied post discretisation since the scenario/task complexity stayed the same. Interestingly, the

expert's estimate on group PBP and EP's likelihood of failing the planning task was 100%, which was far off from the actual estimates in both groups. However, those in group EP had a higher likelihood of failing the planning task than those in group EP. This is in contrast to the monitoring and intervention tasks, which show that though some contexts in group PBP had lower failure probabilities than those in group EP, on average, those in EP had a lower chance of failing these two tasks. This outcome is in contrast to the logistic regression (LR) result in Table 6.3 when used alone, which shows that group PBP had a lower chance of failing all three tasks/HFE.

The LR prediction showed lower failure probability outcomes than when both expert and live data were combined. The monitoring result of the LR model is close to those in context F and context H in group PBP and EP, respectively. The LR planning outcomes were almost similar to the LR+BN SPAR-H for those in group PBP especially context C, while that of group EP was lower than the results from the latter approach. The results of the LR model for the intervention task were almost similar to those of context H and A in group PBP and EP, respectively. In general, the LR model failure predictions were lower than those of the LR+BN SPAR-H method and matched only one context from each group in each HFE. This was a similar outcome when SPAR-H was used alone, where, in contrast to the LR model, the HEP estimation by the experts was higher. Therefore, using the LR+BN SPARH method, which is a combination of expert and live data, compensates for the shortcomings of both approaches when applied as a standalone.

7.3.3 Validation outcome

Validating the framework is an essential step to understanding how it works with new operator data.

Chapter 8

Conclusion

8.1 Research Objective 1

The first objective of this thesis is to understand the state-of-the-art alarm response procedures and to compare the effectiveness of the transition from paper, typical in current process control rooms, versus electronic alarm response procedures for process control room operators handling abnormal situations. The first step to achieving this objective is understanding the variables that can explain the impact of using one procedure over the other on the operator and what operator states or characteristics, and performance measures can be impacted. Performance shaping factors such as training, experience, procedures, HMI etc., are common factors included for a human factors study of process control rooms. These variables are mostly organisational and technical, and when individual variables are considered, they lack in clarity and are assessed subjectively. It is not also considered what the interaction between the variables look like for each setup neither is there a consideration of which variables actually impact performance. There is also no consideration of the operators' behaviour in the overall human performance assessment picture. This raised the first research question for this objective - question.

Research Question (RQ 1)

The first research question was: what are the most critical factors or variables — both human, organisational, and technical — for decision and situational awareness support and human performance assessment of current dynamic human-in-the-loop process control configurations? Why are they important?

The variables considered in this thesis as input for the experimental study were training, familiarity, task complexity, human-machine interfaces (HMI) (alarms design, procedures, and displays), time. These variables are normally assessed individually or combined through human reliability methods to assess human performance. The combination of these factors determine operators' ability to respond correctly to an alarm but also plays a role is aiding recall or decision-making. They also play a role on the workload and situational awareness levels of operators besides performance. A typical measure to assess workload is the NASA-TLX while situational awareness measures vary from subjective to semi subjective measures. In experimental setting tools like eye-tracking and EEG have been introduced for objective workload, situation awareness and behaviour assessment. One missing piece however is the integration of direct operator-system interaction data for such assessment and observing how they can potentially describe workload, situational awareness or other operator characteristics.

These variables have been introduced in this thesis including some performance variables that can be derived form the operator-system data together with the typical variables earlier mentioned. The behavioural variables included: reaction time, response time, number of mimics opened, number of alarms acknowledged, and so on. While the performance measures derived included: consequence, recovery time, overall performance, error, and so on. Following the correlation and factor analysis of the variables, in addition to subjective and objective workload and situational awareness, three (3) additional constructs were derived: processing capacity, search difficulty and learning. These last three variables specifically describe behaviour while using the procedure. A comparison of how the operators outcome on these constructs and how it eventually impacts performance is addressed with the second research questions. However, it is clear that for current and future human in the loop configurations, especially when using alarm response procedures, typical or so called traditional performance shaping factors are not sufficient for human factor studies (of operator support tools or of the operators themselves) in these settings or for performance assessment.

Research Question (RQ2)

The second research question was: How does the transition from paper to digital or computerised procedure options impact operators' general behaviour, cognition, and

safety? Any difference? If any, how significant are these differences? This question is answered by comparing the outcomes of both groups on the different variables and constructs defined for the HiTL assessment to evaluate the actual impact of using either a paper or electronic procedure.

Although there were no significant differences in workload measures (e.g., TLX index, mental and temporal demand) between the paper-based procedure (PBP) and electronic procedure (EP) groups, participants using PBP consistently reported higher workloads, particularly in more complex scenarios. This higher workload perception was closely linked to increased search difficulty. Eye-tracking data revealed that PBP users exhibited larger saccade amplitudes and opened more mimics during task performance, indicating they faced greater challenges in locating and processing information. Despite their lower workload ratings, the EP group struggled with situational awareness (SA). They demonstrated reduced processing capacity in complex scenarios, which was evident in their extended fixation durations. This suggests that while they spent more time focusing on information, they faced difficulties in fully understanding or utilising it effectively, particularly in time-sensitive tasks.

Despite their higher perceived workload, the PBP group's superior performance in complex tasks points to better processing capacity and information retention. The ability to engage more deeply with the paper-based procedure, even with increased search difficulty, appeared to aid their cognitive processing and familiarity with the procedure, leading to better task execution. The EP group's lower processing capacity became evident in their learning difficulties with complex scenarios. Despite spending more time focusing on specific elements of the task (longer fixation durations), they struggled with key situational understanding and task progression. The design of the electronic procedure, including issues with head movement and navigation between screens, further contributed to this cognitive strain. In contrast, PBP users, although facing higher search difficulty, managed to process information more effectively, likely benefiting from the tactile and visual support provided by the paper. This facilitated their situational understanding and allowed them to outperform the EP group in high-stress, complex scenarios.

In terms of behaviour, both groups exhibited similar trends overall, but in Scenario 3, the PBP group's higher search difficulty correlated with longer response times. However, this did not hinder their performance; instead, they managed tasks more effectively than the EP group, which exhibited signs of cognitive overload and poorer situational

comprehension. Human factor issues observed in the design of the human system interfaces affected both groups. However, the EP group was particularly affected, as their extended fixations and increased head movements suggest they faced more challenges in navigating the interface, exacerbating their learning and processing difficulties.

In conclusion, it is clear that constructs like the search difficulty, processing capacity, and cognitive load played pivotal roles in shaping the performance of both groups.

8.2 Research Objective 2

The second goal of this thesis was to create a real-time risk-based framework aimed at assessing human performance, enhancing situational awareness, and aiding in decision-making within existing human-in-the-loop process control and safety systems. This framework aims to guide operators' behaviour by assessing its impact on risk. The R²HiTL framework integrates expert and real-time knowledge to provide operators with risk-informed decision support. It can potentially support the streamlining of alarm response procedures, mainly when there is an increased likelihood of subtask failures. By associating these failures with data from operators' behaviours or eye-tracking, one can determine whether issues stem from reading or search difficulties, among other factors. The research questions that guided the path to developing this framework are summarised below with an overview of how they have been addressed.

Research Question (RQ3)

The first research question of objective two was: Which operator behavioural metrics are the best predictors of a successful recovery of a plant-alarmed state? This question was answered by applying the logistic regression (LR) model to predict the probability that participants fail the monitoring, planning, and intervention human failure events (HFE) / tasks of each study group (group PBP and EP). The behavioural variables considered in the study were the following: the study group also known as type of support, the scenario also known as task complexity, number of alarms announced, number of alarms acknowledged, number of mimics opened, reaction time and response time. The scenario and group variables were associated with all three HFEs. The key variables associated with this HFE for the monitoring task were the number of alarms, and the alarms acknowledged. Following the prediction of LR failure, the most influential predictor

variable for the monitoring task was *the number of alarms*. However, other variables contributed to the eventual outcome. The key predictor variable for the planning task was the complexity of *the scenario/task complexity*, although the number of mimics and the group also contributed to the final outcome. For the intervention task, the key predictor variables were *the type of support, task complexity and response time*. However, the number of mimics opened contributed to the final prediction outcomes.

The contribution of the variables to failure prediction shows the importance of factoring in the operators' behaviour and potentially cognitive states when studying and planning to support the 'human-in-the-loop' in present-day human-in-the-loop configurations. This becomes even more paramount for future HITL configurations as one transitions from paper to electronic procedures, seeing that those within the group EP had a higher probability of failure from the LR model in all three HFEs. It is evident from the result, that the tools operators interact with impact their behaviour and performance. The Logistic regression model not only provides an understanding of the coefficient/weight of each variable used for prediction but also provides a backing for the assumption made for any new operator data on their likelihood to fail or succeed before final error probability estimations and risk assessment. The LR model provides the first step to having a real-time support method.

Research Question (RQ4)

The second research question under objective 2 was: How can live and legacy data be integrated for a risk-based real-time human-in-the-loop performance assessment framework, decision support, and procedure optimisation? Answering this research question covers the objective of having a real-time and risk-informed model to keep operators in the loop and guide their behaviour to reduce failure probabilities. Combining live and legacy data was possible by using a Bayesian Network Standardised Plant Analysis Risk-Human Reliability (BN SPARH) method in conjunction with the logistic regression predictions (LR + BN SPARH). Using the LR or SPARH methods as standalone comes with some challenges for a real-time model that should be tailored as much as possible to individual operators.

- The LR model does not account for task or scenario-specific nuances. For example, while it may identify the best predictor variables, it cannot capture the interaction effects between those variables or varying operator behaviour.

The SPARH, on the other hand, has the following shortcomings:

- Static Expert Estimates: The method relies on expert assessment and assumes that all participants are uniformly impacted by Performance Shaping Factors (PSFs). This leads to generalisation and lacks the ability to capture real-time data or differences in operator behaviour.

- Exaggerated Error Predictions: The expert-based approach often yields high probabilities of failure and, in some cases, extremely low probabilities that may not align with actual operator performance, leading to potential overestimation.

- Lack of Contextualisation: The SPAR-H method does not consider different contexts within a group, which means it misses the variation in how individuals may react under different conditions or subtasks.

Research Question (RQ5)

The third research question under objective 2 was: What are the benefits of integrating operator-system data and expert knowledge for risk-based real-time decision and situation awareness support? This question is asked to guide the development of the framework for clarity on what benefits come from the combination of legacy data from experts, past operator data, and live data. Combining Expert and Real-Time Data (LR + BN SPAR-H) thus provides a good combination of methods needed for real-time provides the following benefits as summarised:

- Contextualisation of Performance: The BN SPAR-H method allows for the discretisation of variables (e.g., number of alarms, reaction time) and recognises that even within the same group, different participants may exhibit different behaviours and have different probabilities of failure. This adds a layer of granularity to predictions.

- Improved Prediction Accuracy: By integrating expert assessments with live behavioural data, this approach compensates for the lower failure predictions from LR and the overestimation of failure by SPAR-H. It provides a more balanced, data-driven failure probability estimate.

- Task-Specific Precision: The combination enables a more precise estimation for each Human Failure Event (HFE), taking into account the variation in context across operators. This helps address both methods' limitations when used alone and aligns better with real-world outcomes.
- Adaptive to Individual Variability: Unlike SPAR-H's blanket estimates, this approach captures individual performance variability within the same group, avoiding the assumption that all participants are equally affected by each PSF.

The event trees delineating failure paths, along with the consequence analysis and risk matrix, provide a distinct visual representation of the operator's position within the risk landscape and reveal the impact of each human failure event or task on the ultimate consequence.

8.3 Potential for generalisation and Future Work

The first research objective and its research questions provided an understanding of the impact and interplay of the interaction of factors on operators' cognition, behaviour, performance, and ability to maintain a safe plant state when in different human-in-the-loop process control room configurations. It further provided a set of new factors that have been validated for assessing operators' cognitive process, for example, learning, and processing capacity, when using either paper or digital procedures. This set of factors, especially those derived specifically for evaluating the impact of using procedures, can be adapted by several industries where workers use job aids, procedures, charts and so on, to assess the impact of their procedural document or perhaps interface design on operators before approving them for deployment. The other validated factors impacting performance and safety, especially the behavioural factors that can be easily and non-intrusively recorded, can be applied for assessment in any human-in-the-loop configuration across industries and beyond the control room. Also, the effort from this thesis through the design of the experiment and the eventual validation of traditional and dynamic PSFs, not only provides a new set of factors but also provides a backing for some of the performance factors mentioned in standards and guidelines on the impact of alarms, procedures, and human-machine interfaces.

Through the second research objective, a couple of behavioural metrics stood out in their usefulness for predicting the capability of the operator to monitor a process,

handle alarms and intervene accordingly while following the task steps in the procedure. These metrics can be adapted in standards and guidelines for inclusion in the evaluation of human reliability. The thesis also described the methods with which these variables were derived and can likewise be adapted by industry and guidelines or standards. Furthermore, while trying to answer the last two research questions (RQ 4 and 5) concerning Objective 2, the researcher was able to demonstrate and validate the potential of combining live and legacy data using AI techniques and the traditional SPAR-H method. Live data easily collected can help the industry perform frequent updates on human reliability analysis following initial analysis performed during the deployment or set-up of the control configuration. The goal is to help the industry perform this analysis in real-time with a properly integrated and designed monitoring interface. Precisely, an interface where both legacy and live data can be recalled and fed as input to an in-built algorithm for a more realistic error probability estimation. Standards and guidelines in the process industry that can benefit from both the first and second objectives of this thesis include: IEC 63303, Human-machine interfaces for process automation systems, ANSI/ISA-18.2, Management of Alarm Systems for the Process Industries, HPOG's Best Practice in Procedure Formatting, and CCPS' Human Factors Handbook for Process Plant Operations: Improving Process Safety and System Performance, and so on.

The results of addressing objectives 1 and 2 of this thesis revealed some aspects that have yet to be examined and challenged thoughts on further exploitation opportunities. An aspect of interest for further research is the type and number of variables included in addressing objective 2. So far, only a few behavioural variables collectable in real-time were included because of the limited number of behavioural variables derived and validated. Hence, future research will explore and validate other potential variables for testing the framework. Secondly, the framework of objective 2 has only been applied to a simulated case study with a more controlled environment. Further work would include testing this framework in a live setting within the process domain and following final fine-tuning, as well as testing this out in other sectors and use cases. Testing this out in a live setting would require developing a collaborative system in the form of a support agent, which is also a future work of this thesis.

Bibliography

[1] CCPS. *Risked Based Process Safety Overview Risked Based Process Safety Overview*. 2014. ISBN: 9780816910809 (cit. on pp. 1, 32, 56).

[2] Chidera Winifred Amazu, Micaela Demichela, and Davide Fissore. "Human-in-the-Loop Configurations in Process and Energy Industries; A Systematic Review". In: 2023, pp. 3234–3241. DOI: 10.3850/978-981-18-5183-4{_}s33-04-572-cd (cit. on pp. 2, 11, 48).

[3] Omar Chamorro-Atalaya et al. "SCADA and Distributed Control System of a Chemical Products Dispatch Process". In: *International Journal of Advanced Computer Science and Applications* 12.12 (2021), pp. 710–717. ISSN: 21565570. DOI: 10.14569/IJACSA.2021.0121288 (cit. on p. 2).

[4] Neville A. (Neville Anthony) Stanton. *Human factors in the design and evaluation of central control room operations*. CRC Press, 2010, p. 416. ISBN: 9781439809914 (cit. on pp. 2, 33, 36).

[5] Steering Committee HPOG. *Best Practice in Procedure Formatting : REVISION 1*. Tech. rep. Human Performance Oil & Gas, 2021, p. 47 (cit. on pp. 2, 29, 36, 37, 40, 85).

[6] Harris Tarmimi Abu Bakar et al. "Analysis of main accident contributor according to process safety management elements failure". In: *Chemical Engineering Transactions* 56 (2017), pp. 991–996. ISSN: 22839216. DOI: 10.3303/CET1756166 (cit. on pp. 2, 39, 84).

[7] Zoltan R Rosztoczy. "Root causes of the Three Mile Island accident". In: March (2019), pp. 29–32 (cit. on p. 2).

[8] National Transportation Safety Board. *Aircraft accident report: Pan American World Airways, Inc., Boeing 707–321C, N799PA, Elmen- dorf Air Force Base, Anchorage, AK, December 26, 1968*. 1969 (cit. on p. 2).

[9] Qin Gao et al. "Mental workload measurement for emergency operating procedures in digital nuclear power plants". In: *Ergonomics* 56.7 (2013), pp. 1070–1085. ISSN: 00140139. DOI: 10.1080/00140139.2013.790483 (cit. on pp. 2, 4, 22, 35, 37, 52, 53, 64, 70, 72, 84, 96, 111).

[10] Punitkumar Bhavsar, Babji Srinivasan, and Rajagopalan Srinivasan. "Pupillometry Based Real-Time Monitoring of Operator's Cognitive Workload to Prevent Human Error during Abnormal Situations". In: *Industrial and Engineering Chemistry Research* 55.12 (2016), pp. 3372–3382. ISSN: 15205045. DOI: 10.1021/acs.iecr.5b03685 (cit. on pp. 3, 19, 22, 49, 71, 72, 88).

[11] Chao Shi et al. "Potential benefits of eye tracking within process control monitoring tasks". In: *Human Factors and Ergonomics In Manufacturing* 31.3 (2021), pp. 316–326. ISSN: 15206564. DOI: 10.1002/hfm.20892 (cit. on pp. 3, 48, 49, 51, 71, 88).

[12] G. Li et al. "Factors associated with pilot error in aviation crashes". In: *Aviation Space and Environmental Medicine* 72.1 (2001), pp. 52–58. ISSN: 00956562 (cit. on p. 3).

[13] Marcel F. Hinss, Anke M. Brock, and Raphaëlle N. Roy. "Cognitive effects of prolonged continuous human-machine interaction: The case for mental state-based adaptive interfaces". In: *Frontiers in Neuroergonomics* 3 (2022). DOI: 10.3389/fnrgo.2022.935092 (cit. on pp. 3, 12, 18, 19, 21–23, 53).

[14] Richard J. Simonson et al. "Impact of alarm management and automation on abnormal operations: A human-in-the-loop simulation study". In: *Applied Ergonomics* 100.May 2021 (2022), p. 103670. ISSN: 18729126. DOI: 10.1016/j.apergo.2021.103670. URL: https://doi.org/10.1016/j.apergo.2021.103670 (cit. on pp. 4, 49, 50, 71).

[15] Chih Wei Yang et al. "Assessing mental workload and situation awareness in the evaluation of computerized procedures in the main control room". In: *Nuclear Engineering and Design* 250 (Sept. 2012), pp. 713–719. ISSN: 00295493. DOI: 10.1016/j.nucengdes.2012.05.038 (cit. on pp. 4, 37, 38, 50, 71, 72, 84, 88).

[16] Vinh N. Dang, John A. Forester, and Ali Mosleh. "Developing a new HRA quantification approach from best methods and practices". In: *10th International Conference on Probabilistic Safety Assessment and Management 2010, PSAM 2010* 2 (2010), pp. 1336–1343 (cit. on pp. 4, 43).

[17] Per Øivind Braarud. "Investigating the validity of subjective workload rating (NASA TLX) and subjective situation awareness rating (SART) for cognitively complex human–machine work". In: *International Journal of Industrial Ergonomics* 86.October (2021). ISSN: 18728219. DOI: 10.1016/j.ergon.2021.103233 (cit. on pp. 4, 39, 50, 51, 53, 64–66, 71, 72, 88, 101, 108–111, 177, 248, 249).

[18] Chidera W. Amazu et al. "Exploring the Influence of Human System Interfaces: Introducing Support Tools and an Experimental Study". In: *International Journal of Human-Computer Interaction* 0.0 (2024), pp. 1–18. ISSN: 15327590. DOI: 10.1080/10447318.2024.2376354. URL: https://doi.org/10.1080/10447318.2024.2376354 (cit. on pp. 5, 32, 62, 67, 73, 85, 87, 111).

[19] Jenni Morrison. *100 largest losses in the hydrocarbon industry. 1974 – 2023*. 28th Edition. July 2, 2024. URL: https://www.marsh.com/en/industries/energy-and-power/insights/100-largest-losses-2024.html (cit. on pp. 11, 28, 32, 39, 42).

[20] David Howes. "Cross-talk between the Senses". In: *The Senses and Society* 1.3 (2006), pp. 381–390. ISSN: 1745-8927. DOI: 10.2752/174589206778476225 (cit. on p. 12).

[21] University of Utah. *Touch: Additional information*. URL: https://learn.genetics.utah.edu/content/senses/touch/ (cit. on p. 12).

[22] John Mcgrory. "Don ' t underestimate the power of the Pen". In: 10.1 (2022) (cit. on p. 12).

[23] A.T Beck and A. Miller. "The Instructor's Manual for Aaron Beck on Cognitive Therapy". In: (2012), pp. 1–46. URL: www.psychotherapy.net (cit. on p. 13).

[24] University of Utah. *Types of memory*. URL: https://learn.genetics.utah.edu/content/memory/types/ (cit. on p. 13).

[25] Westen Zubeyde Bayram. "The senses 4: touch – physiology of the sensation and perception of touch". In: 119.1 (2023), pp. 1–6. URL: www.nursingtimes.net (cit. on p. 13).

[26] Martin Schmelz. "Neuronal sensitivity of the skin." In: *European journal of dermatology : EJD* 21 Suppl 2 (2011), pp. 43–47. URL: https://api.semanticscholar.org/CorpusID:24454485 (cit. on p. 13).

[27] University of Utah. *Hearing: Additional information*. URL: https://learn.genetics.utah.edu/content/senses/hearing/ (cit. on pp. 13, 14, 251).

[28] John Knight, Zubeyde Bayram, and Maria Andrade. "The senses 1: hearing, balance and common pathologies of the ear". In: *Journal Nursing Times [online]* 118.10 (2022), pp. 1–5 (cit. on pp. 13, 14).

[29] University of Utah. *Hearing: Learn.Genetics*. URL: https://learn.genetics.utah.edu/content/senses/hearing/ (cit. on p. 14).

[30] Harvard Health. *Understanding the stress response*. 2024. URL: https://www.health.harvard.edu/staying-healthy/understanding-the-stress-response (cit. on p. 14).

[31] Oliver Carson. *How fast is human reaction time? Human Perception & Technology*. 2024. URL: https://www.pubnub.com/blog/how-fast-is-realtime-human-perception-and-technology/ (cit. on p. 15).

[32] Edmund T. Rolls. "The cingulate cortex and limbic systems for emotion, action, and memory". In: *Brain Structure and Function* 224.9 (2019), pp. 3001–3018. ISSN: 18632661. DOI: 10.1007/s00429-019-01945-2. URL: https://doi.org/10.1007/s00429-019-01945-2 (cit. on p. 15).

[33] Christopher D. Wickens et al. *Engineering psychology and human performance*. 2021, pp. 1–566. ISBN: 9781000401325. DOI: 10.4324/9781003177616 (cit. on pp. 15–20, 23, 24, 250, 256, 258).

[34] Margaret M. Hayes, Souvik Chatterjee, and Richard M. Schwartzstein. "Critical thinking in critical care: Five strategies to improve teaching and learning in the intensive care unit". In: *Annals of the American Thoracic Society* 14.4 (2017), pp. 569–575. ISSN: 23256621. DOI: 10.1513/AnnalsATS.201612-1009AS (cit. on pp. 15, 23).

[35] University of Utah. *Vision: Additional information*. URL: https://learn.genetics.utah.edu/content/senses/vision/ (cit. on pp. 16, 17).

[36] Yu Sun et al. "Functional cortical connectivity analysis of mental fatigue unmasks hemispheric asymmetry and changes in small-world networks". In: *Brain and Cognition* 85.1 (2014), pp. 220–230. ISSN: 02782626. DOI: 10.1016/j.bandc.2013.12.011. URL: http://dx.doi.org/10.1016/j.bandc.2013.12.011 (cit. on p. 19).

[37] Daniel T. Levin et al. "Change blindness blindness: The metacognitive error of overestimating change-detection ability". In: *Visual Cognition* 7.1-3 (2000), pp. 397–412. ISSN: 13506285. DOI: 10.1080/135062800394865 (cit. on pp. 19, 64).

[38] Ira E Hyman Jr et al. "Did you see the unicycling clown? Inattentional blindness while walking and talking on a cell phone". In: *Applied Cognitive Psychology* 24.5 (2010), pp. 597–607. DOI: https://doi.org/10.1002/acp.1638. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/acp.1638 (cit. on p. 19).

[39] Seema L. Clifasefi, Melanie K.T. Takarangi, and Jonah S. Bergman. "Blind drunk: The effects of alcohol on inattentional blindness". In: *Applied Cognitive Psychology* 20.5 (2006), pp. 697–704. ISSN: 08884080. DOI: 10.1002/acp.1222 (cit. on p. 19).

[40] David B. Kaber and Sang Hwan Kim. "Understanding Cognitive Strategy With Adaptive Automation in Dual-Task Performance Using Computational Cognitive Models". In: *Journal of Cognitive Engineering and Decision Making* 5.3 (2011), pp. 309–331. ISSN: 15553434. DOI: 10.1177/1555343411416442 (cit. on p. 19).

[41] Zizheng Guo et al. "The impairing effect of mental fatigue on visual sustained attention under monotonous multi-object visual attention task in long durations: An event-related potential based study". In: *PLoS ONE* 11.9 (2016), pp. 1–13. ISSN: 19326203. DOI: 10.1371/journal.pone.0163360 (cit. on p. 19).

[42] Roee Holtzer et al. "Cognitive Fatigue defined in the context of attention networks". In: *Aging, Neuropsychology, and Cognition* 18.1 (2011), pp. 108–128. ISSN: 13825585. DOI: 10.1080/13825585.2010.517826 (cit. on p. 19).

[43] Jens Rasmussen. *Information Processing and Human-Machine Interaction: An Approach to Cognitive Engineering*. USA: Elsevier Science Inc., 1986. ISBN: 0444009876 (cit. on p. 19).

[44] Barbara A Eriksen and Charles W Eriksen. "Effects of noise letters upon the identification of a target letter in a nonsearch task". In: *Perception & Psychophysics* 16.1 (1974), pp. 143–149. ISSN: 1532-5962. DOI: 10.3758/BF03203267. URL: https://doi.org/10.3758/BF03203267 (cit. on p. 19).

[45] Damon G. Lamb and Eric S. Porges. "Flanker Task". In: *Encyclopedia of Clinical Neuropsychology* 16.1 (2018), pp. 1444–1445. DOI: 10.1007/978-3-319-57111-9{_}9085 (cit. on p. 19).

[46] Nicholas J. Wade. "Pioneers of eye movement research". In: *i-Perception* 1.2 (2010), pp. 33–68. ISSN: 20416695. DOI: 10.1068/i0389 (cit. on p. 20).

[47] Keith Rayner et al. "Scientific Studies of Reading Eye Movements as Reflections of Comprehension Processes in Reading". In: *Scientific Studies of Reading* 10 (2006), pp. 241–255. DOI: 10.1207/s1532799xssr1003 (cit. on pp. 20, 260).

[48] Keith Rayner et al. "Eye movements and visual encoding during scene perception". In: *Psychological Science* 20.1 (2009), pp. 6–10. ISSN: 09567976. DOI: 10.1111/j.1467-9280.2008.02243.x (cit. on p. 20).

[49] Aared L Yarbs and Basil Haigh. "Eye Movements and Vision". In: *Problems of Information Transmission* (1967) (cit. on p. 20).

[50] Jonghyuk Park et al. "A bci based alerting system for attention recovery of UAV operators". In: *Sensors* 21.7 (2021), pp. 1–20. ISSN: 14248220. DOI: 10.3390/s21072447 (cit. on pp. 20, 34).

[51] M. R. Endsley. "Toward a theory of situation awareness in dynamic systems". In: *Human Factors* 37.1 (1995), pp. 32–64. ISSN: 00187208. DOI: 10.1518/001872095779049543 (cit. on pp. 20, 101).

[52] Linda Johnstone Sorensen, Neville A. Stanton, and Adrian P. Banks. "Back to SA school: Contrasting three approaches to situation awareness in the cockpit". In: *Theoretical Issues in Ergonomics Science* 12.6 (2011), pp. 451–471. ISSN: 1464536X. DOI: 10.1080/1463922X.2010.491874 (cit. on pp. 21, 22).

[53] K. S. O'Brien and D. O'Hare. "Situational awareness ability and cognitive skills training in a complex real-world task". In: *Ergonomics* 50.7 (2007), pp. 1064–1091. ISSN: 00140139. DOI: 10.1080/00140130701276640 (cit. on p. 21).

[54] Henry M. Holden. "The Boeing 247: the first modern airliner". In: *Choice Reviews Online* 30.02 (1992), pp. 30–1161. ISSN: 0009-4978. DOI: 10.5860/choice.30-1161 (cit. on p. 21).

[55] Vijaya K. Krishnan, D Dasari, and L Ding. "EEG Correlates of Fluctuation in Cognitive Performance in an Air Traffic Control Task". In: November (2014) (cit. on p. 21).

[56] Jessie Y.C. Chen and Michael J. Barnes. "Human - Agent teaming for multirobot control: A review of human factors issues". In: *IEEE Transactions on Human-Machine Systems* 44.1 (2014), pp. 13–29. ISSN: 21682291. DOI: 10.1109/THMS.2013.2293535 (cit. on pp. 21, 50, 52, 72).

[57] J J Marquez and M L Cummings. "Design and Evaluation of Path Planning Decision Support for Planetary Surface Exploration". In: *Journal of Aerospace Computing, Information, and Communication* 5.3 (2008), pp. 57–71. DOI: 10.2514/1.26248. URL: https://doi.org/10.2514/1.26248 (cit. on p. 21).

[58] Linda Onnasch, Stefan Ruff, and Dietrich Manzey. "Operators adaptation to imperfect automation - Impact of miss-prone alarm systems on attention allocation and performance". In: *International Journal of Human Computer Studies* 72.10-11 (2014), pp. 772–782. ISSN: 10959300. DOI: 10.1016/j.ijhcs.2014.05.001. URL: http://dx.doi.org/10.1016/j.ijhcs.2014.05.001 (cit. on pp. 21, 67).

[59] Young Woo Sohn and Stephanie M. Doane. "Memory processes of flight situation awareness: Interactive roles of working memory capacity, long-term working memory, and expertise". In: *Human Factors* 46.3 (2004), pp. 461–475. ISSN: 00187208. DOI: 10.1518/hfes.46.3.461.50392 (cit. on p. 22).

[60] Juan Jesús Roldán et al. "Multi-robot interfaces and operator situational awareness: Study of the impact of immersion and prediction". In: *Sensors (Switzerland)* 17.8 (2017), pp. 1–25. ISSN: 14248220. DOI: 10.3390/s17081720 (cit. on p. 22).

[61] Stacey D. Scott et al. "Assisting interruption recovery in supervisory control of multiple UAVS". In: *Proceedings of the Human Factors and Ergonomics Society* Figure 1 (2006), pp. 699–703. ISSN: 10711813. DOI: 10.1177/154193120605000518 (cit. on pp. 22, 66).

[62] Vsevolod Peysakhovich et al. "The neuroergonomics of aircraft cockpits: The four stages of eye-tracking integration to enhance flight safety". In: *Safety* 4.1 (2018). ISSN: 2313576X. DOI: 10.3390/safety4010008 (cit. on pp. 22, 48).

[63] Mica R. Endsley. "Direct measurement of situation awareness: Validity and use of SAGAT". In: *Situational Awareness* November (2017), pp. 129–155. DOI: 10.4324/9781315087924-9 (cit. on pp. 22, 51).

[64] Mica R. Endsley. "A Systematic Review and Meta-Analysis of Direct Objective Measures of Situation Awareness: A Comparison of SAGAT and SPAM". In: *Human Factors* 63.1 (2021), pp. 124–150. ISSN: 15478181. DOI: 10.1177/0018720819875376 (cit. on pp. 22, 51, 53, 70, 94).

[65] Ritske De Jong. "The Role of Preparation in Overlapping-task Performance". In: *The Quarterly Journal of Experimental Psychology Section A* 48.1 (1995), pp. 2–25. DOI: 10.1080/14640749508401372. URL: https://doi.org/10.1080/14640749508401372 (cit. on p. 23).

[66] Vanessa A. Petruo, Moritz Mückschel, and Christian Beste. "On the role of the prefrontal cortex in fatigue effects on cognitive flexibility - A system neurophysiological approach". In: *Scientific Reports* 8.1 (2018), pp. 1–13. ISSN: 20452322. DOI: 10.1038/s41598-018-24834-w (cit. on p. 23).

[67] Monicque M. Lorist et al. "The influence of mental fatigue and motivation on neural network dynamics; an EEG coherence study". In: *Brain Research* 1270 (2009), pp. 95–106. ISSN: 00068993. DOI: 10.1016/j.brainres.2009.03.015. URL: http://dx.doi.org/10.1016/j.brainres.2009.03.015 (cit. on p. 23).

[68] Ruston M Hunt and William B Rouse. "Problem-Solving Skills of Maintenance Trainees in Diagnosing Faults in Simulated Powerplants". In: *Human Factors* 23.3 (1981), pp. 317–328. DOI: 10.1177/001872088102300307. URL: https://doi.org/10.1177/001872088102300307 (cit. on p. 24).

[69] Christopher D. Wickens et al. "The anchoring heuristic in intelligence integration: A bias in need of de-biasing". In: *Proceedings of the Human Factors and Ergonomics Society* 3.2006 (2010), pp. 2324–2328. ISSN: 10711813. DOI: 10.1618/10711810X12829370266644 (cit. on p. 24).

[70] Max H Bazerman and Moore Don A. *Judgement in Decision Making*. 2009. ISBN: 9780470049457 (cit. on p. 24).

[71] Maia B. Cook and Harvey S. Smallman. "Human factors of the confirmation bias in intelligence analysis: Decision support from graphical evidence landscapes". In: *Human Factors* 50.5 (2008), pp. 745–754. ISSN: 00187208. DOI: 10.1518/001872008X354183 (cit. on pp. 24, 25).

[72] Eric R Stone, Jacques F Yates, and Andrew M Parker. "Effects of numerical and graphical displays on professed risk-taking behavior." In: *Journal of Experimental Psychology: Applied* 3 (1997), pp. 243–256. URL: https://api.semanticscholar.org/CorpusID:144210012 (cit. on pp. 24, 25).

[73] David A Schkade and Don N Kleinmuntz. "Information Displays and Choice Processes: Differential Effects of Organization, Form, and Sequence". In: *Organizational Behavior and Human Decision Processes* 57 (1994), pp. 319–337. URL: https://api.semanticscholar.org/CorpusID:145686633 (cit. on p. 24).

[74] Daniel J. Simons and Christopher F. Chabris. "Gorillas in our midst: Sustained inattentional blindness for dynamic events". In: *Perception* 28.9 (1999), pp. 1059–1074. ISSN: 03010066. DOI: 10.1068/p281059 (cit. on p. 25).

[75] Paul Swuste et al. "Process safety indicators, a review of literature". In: *Journal of Loss Prevention in the Process Industries* 40 (2016), pp. 162–173. ISSN: 09504230. DOI: 10.1016/j.jlp.2015.12.020. URL: http://dx.doi.org/10.1016/j.jlp.2015.12.020 (cit. on p. 26).

[76] Yong The. *Optimizing process safety: The Transformative Power of Integrated Software to Improve Process Safety Situational Awareness.* 2024. URL: https://dimensionsoftwarenz.com/balancing-safety-and-adaptability-understanding-overrides-and-bridges-in-process-plants/ (cit. on p. 26).

[77] Brian Nesbitt. *Handbook of valves and actuators*. 1st ed. Butterworth-Heinemann, Aug. 14, 2006. DOI: 10.1016/b978-1-85617-494-7.x5027-5. URL: https://doi.org/10.1016/b978-1-85617-494-7.x5027-5 (cit. on p. 26).

[78] Stuart Bennett. "History of Automatic Control to 1960: An Overview". In: *IFAC Proceedings Volumes* 29.1 (1996), pp. 3008–3013. ISSN: 14746670. DOI: 10.1016/s1474-6670(17)58135-5. URL: http://dx.doi.org/10.1016/S1474-6670(17)58135-5 (cit. on p. 26).

[79] C. J. Swan. "Automatic Control through Temperature or Pressure". In: *Industrial and Engineering Chemistry* 20.11 (1928), pp. 1152–1155. ISSN: 00197866. DOI: 10.1021/ie50227a011 (cit. on p. 26).

[80] Ismar Ginsberg. "Automatic Control in the Chemical Industries". In: *Industrial and Engineering Chemistry* 21.5 (1929), pp. 410–414. ISSN: 00197866. DOI: 10.1021/ie50233a004 (cit. on p. 26).

[81] Pierfrancesco Bellini et al. "High level control of chemical plant by industry 4.0 solutions". In: *Journal of Industrial Information Integration* 26 (2022), p. 100276. ISSN: 2452414X. DOI: 10.1016/j.jii.2021.100276. URL: https://doi.org/10.1016/j.jii.2021.100276 (cit. on p. 27).

[82] Sawyer Huth. *Evolution of control room consoles and process control operations | Tresco Consoles*. 2024. URL: https://www.trescoconsoles.com/blog/evolution/#::text=Control%20rooms%20trace%20their%20roots,of%20monitoring%20production%20became%20apparent (cit. on p. 28).

[83] Ravi Joshi and Ben Daum. "Human factors in design of control rooms for process industries". In: (2017), pp. 2–30. URL: https://engineering.purdue.edu/P2SAC/presentations/documents/Human%20Factors%20in%20Design%20of%20Control%20Rooms%20-%20grad%20student%20project.pdf (cit. on p. 29).

[84] ISO. "Ergonomic design of control centres (ISO 11064)". In: 3 (2000) (cit. on pp. 29, 80).

[85] X. Li et al. "The control room operator: The forgotten element in mineral process control". In: *Minerals Engineering* 24.8 (2011), pp. 894–902. ISSN: 08926875. DOI: 10.1016/j.mineng.2011.04.001 (cit. on p. 29).

[86] M L Bransby and J Jenkinson. "The management of alarm systems". In: *Health San Francisco* (1998), pp. 1–230 (cit. on pp. 31, 32).

[87] HSE. "Buncefield: Why did it happen?" In: *Control of Major Accident Hazards* December (2011), p. 36. URL: www.hse.gov.uk/comah/buncefield/buncefield-report.pdf (cit. on pp. 32, 34, 250).

[88] Bill Hollifield. "A High Performance HMI: Better Graphics for Operations Effectiveness". In: *ISA Water & Wastewater and Automatic Controls Symposium* 77062.281 (2012), p. 27. URL: www.isawwsymposium.com (cit. on p. 32).

[89] IEC 63303. "Human-Machine Interfaces for Process Automation Systems". In: *International Electrotechnical Commission* Cd (2021) (cit. on pp. 33, 80).

[90] Garth E Cummings. "Operator / Instrumentation Interactions During". In: 1 (1980), pp. 931–934 (cit. on p. 33).

[91] Chidera Winifred Amazu et al. "Decision Making for Process Control Management in Control Rooms: a Survey Methodology and Initial Findings". In: *Chemical Engineering Transactions* 99.May (2023), pp. 271–276. ISSN: 22839216. DOI: 10.3303/CET2399046 (cit. on pp. 33, 40).

[92] Julie Saint-Lot, Jean Paul Imbert, and Frédéric Dehais. "Red Alert: A Cognitive Countermeasure to Mitigate Attentional Tunneling". In: *Conference on Human Factors in Computing Systems - Proceedings* (2020), pp. 1–6. DOI: 10.1145/3313831.3376709 (cit. on p. 34).

[93] Caroline Morais et al. "Identification of human errors and influencing factors: A machine learning approach". In: *Safety Science* 146.October 2021 (2022), p. 105528. ISSN: 18791042. DOI: 10.1016/j.ssci.2021.105528. URL: https://doi.org/10.1016/j.ssci.2021.105528 (cit. on p. 34).

[94] Katrina M. Groth and Ali Mosleh. "Deriving causal Bayesian networks from human reliability analysis data: A methodology and example model". In: *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability* 226.4 (2012), pp. 361–379. ISSN: 1748006X. DOI: 10.1177/1748006X11428107 (cit. on pp. 34, 45).

[95] V. E. Barnes and L. R. Radford. "Evaluation of Nuclear Power Plant Operating Procedures Classifications and Interfaces". In: *Nureg/Cr-4613, Pnl-5852, Bharc-400/86/008* (1987) (cit. on pp. 35, 36, 39).

[96] Jinkyun Park and Wondea Jung. "Comparing the operators ' behavior in conducting emergency operating procedures with the complexity of procedural steps". In: May (2003), pp. 1–8 (cit. on pp. 35, 39).

[97] Song Xu et al. "An ergonomics study of computerized emergency operating procedures: Presentation style, task complexity, and training level". In: *Reliability Engineering and System Safety* 93.10 (2008), pp. 1500–1511. ISSN: 09518320. DOI: 10.1016/j.ress.2007.09.006 (cit. on pp. 35, 37, 65).

[98] J.M. O'hara et al. *Human-System Interface Design Review Guidelines*. 2002, p. 200. ISBN: 1800553684. URL: https://www.nrc.gov/docs/ML0217/ML021700337.pdf (cit. on pp. 35, 81, 83–85, 296).

[99] *Guidelines for the Preparation of Emergency Operating Procedures Resolution of Comments on NUREG-0799 U.S. Nuclear Regulatory Commission Office of Nuclear Reactor Regulation*. Tech. rep. (cit. on p. 36).

[100] Joseph Naser. "Computerized Procedures - Design and Implementation Guidance for Procedures, Associated Automation and Soft Controls". In: December (2007) (cit. on p. 37).

[101] Jinkyun Park and Yochan Kim. "Comparisons of human reliability data between analog and digital environments". In: *Safety Science* 149 (2022), p. 105701. ISSN: 18791042. DOI: 10.1016/j.ssci.2022.105701. URL: https://doi.org/10.1016/j.ssci.2022.105701 (cit. on p. 37).

[102] Esmaeil Zarei, Faisal Khan, and Rouzbeh Abbassi. "Importance of human reliability in process operation: A critical analysis". In: *Reliability Engineering and System Safety* 211.March (2021), p. 107607. ISSN: 09518320. DOI: 10.1016/j.ress.2021.107607. URL: https://doi.org/10.1016/j.ress.2021.107607 (cit. on pp. 39, 40, 43, 44, 47, 48, 54).

[103] Mohd Umair Iqbal and Rajagopalan Srinivasan. *Simulator based performance metrics to estimate reliability of control room operators*. 2018. DOI: 10.1016/j.jlp.2017.10.011 (cit. on pp. 39, 40, 48, 54, 72, 88).

[104] Jooyoung Park, Awwal Mohammed Arigi, and Jonghyun Kim. "A comparison of the quantification aspects of human reliability analysis methods in nuclear power plants". In: *Annals of Nuclear Energy* 133 (2019), pp. 297–312. ISSN: 18732100. DOI: 10.1016/j.anucene.2019.05.031 (cit. on pp. 41, 44–47).

[105] CCPS. *Human Factors Handbook For Process Plant Operations*. Ed. by The American Institute of Chemical Engineers. Wiley, Apr. 2022. ISBN: 9781119640493. DOI: 10.1002/9781119640516. URL: https://onlinelibrary.wiley.com/doi/book/10.1002/9781119640516 (cit. on pp. 41, 42).

[106] Katrina M. Groth and Laura P. Swiler. "Bridging the gap between HRA research and HRA practice: A Bayesian network version of SPAR-H". In: *Reliability Engineering and System Safety* 115 (2013), pp. 33–42. ISSN: 09518320. DOI: 10.1016/j.ress.2013.02.015. URL: http://dx.doi.org/10.1016/j.ress.2013.02.015 (cit. on pp. 43, 45–47, 52, 132–134).

[107] M. Abílio Ramos et al. "A human reliability analysis methodology for oil refineries and petrochemical plants operation: Phoenix-PRO qualitative framework". In: *Reliability Engineering and System Safety* 193.September 2019 (2020), p. 106672. ISSN: 09518320. DOI: 10.1016/j.ress.2019.106672. URL: https://doi.org/10.1016/j.ress.2019.106672 (cit. on p. 44).

[108] A Bye et al. "Petro-HRA, a new method for human reliability analysis in the petroleum industry". In: ... *and Management* ... October (2016). URL: https://www.researchgate.net/profile/Ronald_Boring/publication/308707744_Petro-HRA_A_New_Method_for_Human_Reliability_Analysis_in_the_Petroleum_Industry/links/57f41b2f08ae8da3ce538234.pdf (cit. on pp. 44, 45).

[109] REZIE BOROUN et al. "Fuzzy Logic in HEART and CREAM Methods to Assess Human Error and Find an Optimum Method Using a Hierarchical Fuzzy System: A Case Study in a Steel Factory". In: *International Journal of Occupational Hygiene* (2022). ISSN: 2008-5109. DOI: 10.18502/ijoh.v13i2.8372 (cit. on p. 44).

[110] Yundong Guo and Youchao Sun. *Flight safety assessment based on an integrated human reliability quantification approach*. 2020. DOI: 10.1371/journal.pone.0231391 (cit. on p. 44).

[111] Shokoufeh Abrishami, Nima Khakzad, and Seyed Mahmoud Hosseini. "A data-based comparison of BN-HRA models in assessing human error probability: An offshore evacuation case study". In: *Reliability Engineering and System Safety* 202.May (2020), p. 107043. ISSN: 09518320. DOI: 10.1016/j.ress.2020.107043. URL: https://doi.org/10.1016/j.ress.2020.107043 (cit. on p. 44).

[112] Junxiang Tu, Wei Lin, and Youxi Lin. "A Bayes-SLIM based methodology for human reliability analysis of lifting operations". In: *International Journal of Industrial Ergonomics* 45 (2015), pp. 48–54. ISSN: 18728219. DOI: 10.1016/j.ergon.2014.11.004. URL: http://dx.doi.org/10.1016/j.ergon.2014.11.004 (cit. on pp. 44–46).

[113] David I. Gertman et al. "The SPAR H human reliability analysis method". In: *American Nuclear Society 4th International Topical Meeting on Nuclear Plant Instrumentation, Control and Human Machine Interface Technology* (2004), pp. 17–24 (cit. on pp. 44, 126, 128–130, 132–134, 265, 309, 310).

[114] David Gertman et al. *Review of Findings for Human Performance Contribution to Risk in Operating Events*. August. 2002 (cit. on p. 44).

[115] Barry Kirwan. "The validation of three human reliability quantification techniques — THERP, HEART and JHEDI: Part 1 — technique descriptions and validation issues". In: *Applied Ergonomics* 27.6 (1996), pp. 359–373. ISSN: 0003-6870. DOI: https://doi.org/10.1016/S0003-6870(96)00044-0. URL: https://www.sciencedirect.com/science/article/pii/S0003687096000440 (cit. on p. 44).

[116] A. De Galizia et al. "Advanced investigation of HRA methods for probabilistic assessment of human barriers efficiency in complex systems for a given organisational and environmental context". In: *International Topical Meeting on Probabilistic Safety Assessment and Analysis, PSA 2015* 2 (2015), pp. 627–639 (cit. on p. 44).

[117] Myeong Soo Lee et al. "Development of human factors validation system for the advanced control room of APR1400". In: *Journal of Nuclear Science and Technology* 46.1 (2009), pp. 90–101. ISSN: 00223131. DOI: 10.1080/18811248.2007.9711511 (cit. on p. 44).

[118] Maria Chiara Leva et al. "Human factors issues and the risk of high voltage equipment: Are standards sufficient to ensure safety by design?" In: *Chemical Engineering Transactions* 26 (2012), pp. 273–278. ISSN: 22839216. DOI: 10.3303/CET1226046 (cit. on p. 44).

[119] Jie Geng et al. "Human error probability estimation in ATEX-HMI area classification: From THERP to FUZZY CREAM". In: vol. 43. May 2015. DOI: 10.3303/CET1543208 (cit. on pp. 44, 45).

[120] Jooyoung Park, Awwal Mohammed Arigi, and Jonghyun Kim. "Treatment of human and organizational factors for multi-unit HRA: Application of SPAR-H method". In: *Annals of Nuclear Energy* 132 (2019), pp. 656–678. ISSN: 18732100. DOI: 10.1016/j.anucene.2019.06.053. URL: https://doi.org/10.1016/j.anucene.2019.06.053 (cit. on p. 45).

[121] Ali Mosleh et al. "A model-based human reliability analysis methodology". In: *11th International Probabilistic Safety Assessment and Management Conference and the Annual European Safety and Reliability Conference 2012, PSAM11 ESREL 2012* 4.June (2012), pp. 3180–3189 (cit. on pp. 45, 129).

[122] Gh A. Shirali et al. “Modifying a method for human reliability assessment based on CREAM-BN: A case study in control room of a petrochemical plant”. In: *MethodsX* 6 (2019), pp. 300–315. ISSN: 22150161. DOI: 10.1016/j.mex.2019.02.008. URL: https://doi.org/10.1016/j.mex.2019.02.008 (cit. on pp. 45, 46).

[123] Katrina M. Groth, Curtis L. Smith, and Laura P. Swiler. “A Bayesian method for using simulator data to enhance human error probabilities assigned by existing HRA methods”. In: *Reliability Engineering and System Safety* 128 (2014), pp. 32–40. ISSN: 09518320. DOI: 10.1016/j.ress.2014.03.010. URL: http://dx.doi.org/10.1016/j.ress.2014.03.010 (cit. on pp. 46, 122, 134, 135, 137, 141).

[124] Per Braarud et al. “An Investigation of Speech Features, Plant System Alarms, and Operator–System Interaction for the Classification of Operator Cognitive Workload During Dynamic Work”. In: *Human Factors* 63.5 (2021), pp. 736–756. ISSN: 15478181. DOI: 10.1177/0018720820961730 (cit. on pp. 48, 50, 52, 106, 140).

[125] Chao Shi and Ling Rothrock. “Using eye movements to evaluate the effectiveness of the situation awareness rating technique scale in measuring situation awareness for smart manufacturing”. In: *Ergonomics* 66.8 (2022), pp. 1090–1098. ISSN: 13665847. DOI: 10.1080/00140139.2022.2132299. URL: https://doi.org/10.1080/00140139.2022.2132299 (cit. on pp. 48, 50, 52, 53, 72, 102, 111).

[126] Laya Das, Babji Srinivasan, and Rajagopalan Srinivasan. “Cognitive Behavior Based Framework for Operator Learning: Knowledge and Capability Assessment through Eye Tracking”. In: *Computer Aided Chemical Engineering* 40 (2017), pp. 2977–2982. ISSN: 15707946. DOI: 10.1016/B978-0-444-63965-3.50498-0 (cit. on pp. 48, 71).

[127] Mohd Umair Iqbal, Babji Srinivasan, and Rajagopalan Srinivasan. “Dynamic assessment of control room operator’s cognitive workload using Electroencephalography (EEG)”. In: *Computers and Chemical Engineering* 141 (2020), p. 106726. ISSN: 00981354. DOI: 10.1016/j.compchemeng.2020.106726. URL: https://doi.org/10.1016/j.compchemeng.2020.106726 (cit. on p. 49).

[128] Punitkumar Bhavsar, Babji Srinivasan, and Rajagopalan Srinivasan. *Quantifying situation awareness of control room operators using eye-gaze behavior*. 2017. DOI: 10.1016/j.compchemeng.2017.06.004 (cit. on p. 49).

[129] Chao Shi, Ling Rothrock, and Benjamin Noah. “Using Eye-Tracking to Predict Cognitive Workload in a Control Room Environment”. In: (2023), pp. 217–233. DOI: 10.1007/978-3-031-10780-1{_}12 (cit. on pp. 49, 67, 254).

[130] Jung Hyup Kim and Xiaonan Yang. “Applying fractal analysis to pupil dilation for measuring complexity in a process monitoring task”. In: *Applied Ergonomics* 65 (2017), pp. 61–69. ISSN: 18729126. DOI: 10.1016/j.apergo.2017.06.002. URL: http://dx.doi.org/10.1016/j.apergo.2017.06.002 (cit. on pp. 49, 51, 71, 88).

[131] Jung Hyup Kim, Ling Rothrock, and Jason Laberge. "Using signal detection theory and time window-based human-in-the-loop simulation as a tool for assessing the effectiveness of different qualitative shapes in continuous monitoring tasks". In: *Applied Ergonomics* 45.3 (2014), pp. 693–705. ISSN: 18729126. DOI: 10.1016/j.apergo.2013.09.008. URL: http://dx.doi.org/10.1016/j.apergo.2013.09.008 (cit. on pp. 49, 51, 64, 66, 67).

[132] Seunghyun Tina Lee, So Young Kim, and David Gilmore. "Human-in-the-loop evaluation of human-machine interface for power plant operators". In: *2017 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2017* 2017-Janua (2017), pp. 34–39. DOI: 10.1109/SMC.2017.8122574 (cit. on p. 49).

[133] Mostafa Mohammadian et al. "Cognitive demands and mental workload: A filed study of the mining control room operators". In: *Heliyon* 8.2 (2022), e08860. ISSN: 24058440. DOI: 10.1016/j.heliyon.2022.e08860. URL: https://doi.org/10.1016/j.heliyon.2022.e08860 (cit. on pp. 50, 64, 66, 110, 111).

[134] Bin Zheng et al. "Workload assessment of surgeons: Correlation between NASA TLX and blinks". In: *Surgical endoscopy* 26 (Apr. 2012), pp. 2746–50. DOI: 10.1007/s00464-012-2268-6 (cit. on p. 50).

[135] Laura H. Ikuma et al. *A guide for assessing control room operator performance using speed and accuracy, perceived workload, situation awareness, and eye tracking*. 2014. DOI: 10.1016/j.jlp.2014.11.001 (cit. on pp. 51, 53).

[136] Massimiliano De Ambroggi and Paolo Trucco. "Modelling and assessment of dependent performance shaping factors through Analytic Network Process". In: *Reliability Engineering and System Safety* 96.7 (2011), pp. 849–860. ISSN: 09518320. DOI: 10.1016/j.ress.2011.03.004. URL: http://dx.doi.org/10.1016/j.ress.2011.03.004 (cit. on pp. 64, 132–134).

[137] Pablo Delgado et al. "Don't throw away your printed books: A meta-analysis on the effects of reading media on reading comprehension". In: *Educational Research Review* 25.February 2019 (2018), pp. 23–38. ISSN: 1747938X. DOI: 10.1016/j.edurev.2018.09.003 (cit. on pp. 64, 259).

[138] Lee J. Cronbach. "Coefficient alpha and the internal structure of tests". In: *Psychometrika* 16.3 (1951), pp. 297–334. ISSN: 00333123. DOI: 10.1007/BF02310555 (cit. on pp. 64, 67, 108).

[139] A. Christmann and S. Van Aelst. "Robust estimation of Cronbach's alpha". In: *Journal of Multivariate Analysis* 97.7 (2006), pp. 1660–1674. ISSN: 0047259X. DOI: 10.1016/j.jmva.2005.05.012 (cit. on pp. 64, 65, 108, 173).

[140] Hamidur Rahman et al. "Non-contact-based driver's cognitive load classification using physiological and vehicular parameters". In: *Biomedical Signal Processing and Control* 55 (2020), p. 101634. ISSN: 17468108. DOI: 10.1016/j.bspc.2019.101634. URL: https://doi.org/10.1016/j.bspc.2019.101634 (cit. on pp. 52, 64, 66).

[141] Zahra Rahmatinejad et al. "A comparative study of explainable ensemble learning and logistic regression for predicting in-hospital mortality in the emergency department". In: *Scientific Reports* 14.1 (2024), pp. 1–17. ISSN: 20452322. DOI: 10.1038/s41598-024-54038-4. URL: https://doi.org/10.1038/s41598-024-54038-4 (cit. on pp. 64, 110, 140).

[142] Richard A. Armstrong. "Should Pearson's correlation coefficient be avoided?" In: *Ophthalmic and Physiological Optics* 39.5 (2019), pp. 316–327. ISSN: 14751313. DOI: 10.1111/opo.12636 (cit. on pp. 64, 67, 110).

[143] J. KEVIN FORD, ROBERT C. MacCALLUM, and MARIANNE TAIT. "the Application of Exploratory Factor Analysis in Applied Psychology: a Critical Review and Analysis". In: *Personnel Psychology* 39.2 (1986), pp. 291–314. ISSN: 17446570. DOI: 10.1111/j.1744-6570.1986.tb00583.x (cit. on pp. 65, 109).

[144] Robin K. Henson and J. Kyle Roberts. "Use of exploratory factor analysis in published research: Common errors and some comment on improved practice". In: *Educational and Psychological Measurement* 66.3 (2006), pp. 393–416. ISSN: 15523888. DOI: 10.1177/0013164405282485 (cit. on pp. 65, 109).

[145] Yu Chen, Shengyuan Yan, and Cong Chi Tran. "Comprehensive evaluation method for user interface design in nuclear power plant based on mental workload". In: *Nuclear Engineering and Technology* 51.2 (2019), pp. 453–462. ISSN: 2234358X. DOI: 10.1016/j.net.2018.10.010. URL: https://doi.org/10.1016/j.net.2018.10.010 (cit. on pp. 65, 72).

[146] Eric A. Sobie. "Parameter sensitivity analysis in electrophysiological models using multivariable regression". In: *Biophysical Journal* 96.4 (2009), pp. 1264–1274. ISSN: 15420086. DOI: 10.1016/j.bpj.2008.10.056. URL: http://dx.doi.org/10.1016/j.bpj.2008.10.056 (cit. on pp. 65, 66).

[147] Wen Si and Lixia Niu. "Enhancing Human Reliability Prediction in Smart Tower Crane Interfaces : A Refined Approach Using Simplified Plant Analysis Risk – Human Reliability Assessment and the Decision Making". In: *Buildings MDPI* 14.1083 (2024). DOI: 10.3390/buildings14041083 (cit. on pp. 65, 132, 133).

[148] Curtis L. Meinert. "Sample Size: Calculation". In: *Clinical Trials Handbook* (2012), pp. 259–260. DOI: 10.1002/9781118422878.ch86 (cit. on p. 66).

[149] Sebastiaan Mathôt et al. "Safe and sensible preprocessing and baseline correction of pupil-size data". In: *Behavior Research Methods* 50.1 (2018), pp. 94–106. ISSN: 15543528. DOI: 10.3758/s13428-017-1007-2 (cit. on pp. 66, 67, 113).

[150] Erin T. Solovey et al. "Classifying driver workload using physiological and driving performance data: Two field studies". In: *Conference on Human Factors in Computing Systems - Proceedings* (2014), pp. 4057–4066. DOI: 10.1145/2556288.2557068 (cit. on pp. 52, 67).

[151] Mohd Umair Iqbal, Babji Srinivasan, and Rajagopalan Srinivasan. "Multi-class classification of control room operators' cognitive workload using the fusion of eye-tracking and electroencephalography". In: *Computers and Chemical Engineering* 181.December 2023 (2024). ISSN: 00981354. DOI: 10.1016/j.compchemeng.2023.108526 (cit. on pp. 52, 95, 96, 140).

[152] Jing Yang et al. "Situation Awareness Classification Using Multi-modal Sensing in Automated Driving". In: *Proceedings of the Human Factors and Ergonomics Society* 65.1 (2021), p. 52. ISSN: 10711813. DOI: 10.1177/1071181321651095 (cit. on pp. 52, 70).

[153] Zuojin Li et al. "Online detection of driver fatigue using steering wheel angles for real driving conditions". In: *Sensors (Switzerland)* 17.3 (2017), pp. 1–12. ISSN: 14248220. DOI: 10.3390/s17030495 (cit. on p. 72).

[154] Micaela Demichela, Gabriele Baldissone, and Gianfranco Camuncoli. "Risk-Based Decision Making for the Management of Change in Process Plants: Benefits of Integrating Probabilistic and Phenomenological Analysis". In: *Industrial and Engineering Chemistry Research* 56.50 (2017), pp. 14873–14887. ISSN: 15205045. DOI: 10.1021/acs.iecr.7b03059 (cit. on pp. 73, 74).

[155] Chidera Amazu et al. "Human Factors in Alarm Response Procedures: an Empirical Analysis of Paper versus Digital support". In: *Human Interaction and Emerging Technologies (IHIET 2024)* 157.Ihiet (2024), pp. 116–125. DOI: 10.54941/ahfe1005470 (cit. on pp. 81, 256).

[156] Chidera W. Amazu et al. "Comparing Two Control Room Intervention Procedure Formats: Preliminary Insights from Eye Tracking Measures". In: *Chemical Engineering Transactions* 111.August (2024), pp. 223–228. ISSN: 22839216. DOI: 10.3303/CET24111038 (cit. on p. 86).

[157] K. K. Gupta et al. "Basic concepts for sample size calculation: Critical step for any clinical trials!" In: *Saudi Journal of Anaesthesia* 10.3 (2016), pp. 328–331. ISSN: 09753125. DOI: 10.4103/1658-354X.174918 (cit. on p. 88).

[158] Chidera Winifred Amazu et al. "Experiment data: Human-in-the-loop decision support in process control rooms". In: *Data in Brief* 53 (2024). ISSN: 23523409. DOI: 10.1016/j.dib.2024.110170 (cit. on pp. 97, 297–299).

[159] Bhanuka Mahanama et al. "Eye Movement and Pupil Measures: A Review". In: *Frontiers in Computer Science* 3.January (2022), pp. 1–22. ISSN: 26249898. DOI: 10.3389/fcomp.2021.733531 (cit. on p. 102).

[160] Keith Rayner and Martin H. Fischer. "Mindless reading revisited: Eye movements during reading and scanning are different". In: *Perception and Psychophysics* 58.5 (1996), pp. 734–747. ISSN: 00315117. DOI: 10.3758/BF03213106 (cit. on pp. 102, 103).

[161] Lei Fan et al. "Eye movement characteristics and visual fatigue assessment of virtual reality games with different interaction modes". In: *Frontiers in Neuroscience* 17.March (2023), pp. 1–8. ISSN: 1662453X. DOI: 10.3389/fnins.2023.1173127 (cit. on p. 103).

[162] Amr Gomaa et al. “What’s on your mind? A Mental and Perceptual Load Estimation Framework towards Adaptive In-vehicle Interaction while Driving”. In: *Main Proceedings - 14th International ACM Conference on Automotive User Interfaces and Interactive Vehicular Applications, AutomotiveUI 2022* (2022), pp. 215–225. DOI: 10.1145/3543174.3546840 (cit. on pp. 108, 109, 111).

[163] Winfred Arthur et al. *Team Task Analysis*. Vol. 54. 2. 2012, pp. 277–295. ISBN: 0415287006. DOI: 10.1177/0018720811435234 (cit. on pp. 109, 110).

[164] Philip Sedgwick. “Pearson’s correlation coefficient”. In: *BMJ (Online)* 345.7864 (2012), pp. 1–2. ISSN: 17561833. DOI: 10.1136/bmj.e4483 (cit. on p. 110).

[165] Student. “Biometrika Trust”. In: 6.1 (2024), pp. 1–25 (cit. on p. 111).

[166] David Embrey. *Task Analysis Techniques*. 2000. URL: https://www.humanreliability.com/wp-content/uploads/2023/07/Task-Analysis-Techniques.pdf (cit. on p. 127).

[167] T. Chandler et al. “Human Reliability Analysis Methods Selection Guidance for NASA”. In: *National Aeronautics and Space Administration* July (2006), p. 175. URL: http://www.hq.nasa.gov/office/codeq/rm/docs/HRA_Report.pdf (cit. on pp. 126, 138, 139).

[168] W. Galyean et al. “SPAR-H Step-by-Step Guidance”. In: (Jan. 2012) (cit. on p. 130).

[169] Mostafa Mirzaei Aliabadi et al. “Human Reliability Analysis (HRA) Using Standardized Plant Analysis Risk-Human (SPAR-H) and Bayesian Network (BN) for Pipeline Inspection Gauges (PIG) Operation: A Case Study in a Gas Transmission Plant”. In: *Health Scope* 8.3 (2019). ISSN: 2251-8959. DOI: 10.5812/jhealthscope.87148 (cit. on pp. 130, 132–134, 141).

[170] Corwin L. Atwood. “Constrained noninformative priors in risk assessment”. In: *Reliability Engineering and System Safety* 53.1 (1996), pp. 37–46. ISSN: 09518320. DOI: 10.1016/0951-8320(96)00026-9 (cit. on pp. 132, 226).

[171] Sung Il Ahn, Rafet Emek Kurt, and Emre Akyuz. “Application of a SPAR-H based framework to assess human reliability during emergency response drill for man overboard on ships”. In: *Ocean Engineering* 251.February (2022), p. 111089. ISSN: 00298018. DOI: 10.1016/j.oceaneng.2022.111089. URL: https://doi.org/10.1016/j.oceaneng.2022.111089 (cit. on pp. 132, 133).

[172] Gordon Sands, Enda F. Fallon, and Wilhelm J. van der Putten. “The Utilisation of Probabilistic Risk Assessment in Radiation Oncology”. In: *Procedia Manufacturing* 3.February 2016 (2015), pp. 250–257. ISSN: 23519789. DOI: 10.1016/j.promfg.2015.07.138. URL: http://dx.doi.org/10.1016/j.promfg.2015.07.138 (cit. on p. 132).

[173] Chandresh Sharma et al. “Eye Gaze Movement Studies of Control Room Operators: A Novel Approach to Improve Process Safety”. In: *Computers & Chemical Engineering* 85 (Oct. 2015). DOI: 10.1016/j.compchemeng.2015.09.012 (cit. on pp. 181, 253).

[174] Maria Eckstein et al. “Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development?” In: *Developmental Cognitive Neuroscience* in press (Nov. 2016). DOI: 10.1016/j.dcn.2016.11.001 (cit. on p. 181).

Appendix A

Tables

Table A.1 HMI Design Requirements

Requirement	**Source**
Avoid excessive contrast between ambient lighting and screen luminance to prevent eyestrain.	IEC 63303
Design should accommodate visual limitations of users.	"
Provide visual/audible signals with appropriate salience.	"
Reserve alarm colors exclusively to enhance cognitive recognition and speed.	IEC 62682
Use color to reflect information importance; most important info should be most salient.	IEC 63303
Avoid color gradients for static elements; use for dynamic highlights only.	"
Apply color coding conservatively and consistently across the HMI.	"
Ensure foreground and background colors provide sufficient contrast under expected lighting.	"
If displays show excessive info, redesign by grouping into sections or on-demand display. In this study the researcher grouped into sections for mimics, display on demand, and Grouped the procedures into sections too.	"
Use visual dynamic coding (e.g., motion, blinking) to highlight critical info. In this study, the researcher added a flashing acknowledgment button for the AI support; moving graphs to show pressure changes, etc.	IEC 63303
Provide redundant auditory signals with visual indications (e.g., alarms).	IEC 63303, IEC 62682
Consistent data grouping speeds cognitive processing and operator response. In this study, the researcher grouped procedures and mimics by plant sections (e.g., Tank, Methanol) for selective processing.	"

Table A.2 APR Requirement: Access and Location

Requirement	**Source**	**Implemented in the Thesis (Yes/No)**
Users should have immediate access to ARPs where alarm messages are read. Operators should not need to leave the display location to access ARP information.	NUREG/CR 0070	Yes
In a tile system, ARP identification should match the method of alarm identification.	"	Yes
A computerisd system may display the appropriate procedure for a given alarm when the operator selects the alarm message.	"	No
When alarm info is presented on VDU or printer displays, links to ARPs should be provided.	NUREG/CR 6105	Yes
Minimis alarms that refer users to more detailed displays outside the main operating area.	NUREG 0700	Yes

Table A.3 ARP: Content

Requirement/Recommendation	**As Implemented in this Thesis**
System or functional group to which the alarm belongs.	Sections of the plant are mentioned in-text.
Exact alarm text or legend.	Tags and texts visible in-text.
Alarm source (e.g., sensors, processors, actuating devices) with a reference to schematic diagrams.	-
Alarm setpoints.	Visible on some of the graphs.
Priority.	Complemented by alarm prioritisation design. Compared with and without priority, also evident on ARP display.
Potential underlying causes for the alarm.	Alarm tag nomenclature describes this within the procedure.
Required immediate actions, including confirmation steps.	Troubleshooting steps are available and include confirmation points.
Actions that occur automatically when the alarm triggers (should be verified).	-
Follow-up actions.	Evaluation steps available after control actions.
Explanations of relevant alarm processing (e.g., parameter comparisons).	-
Pertinent references.	Present in-text.
Text should not be abbreviated unless it improves space and remains clear. Underlining for emphasis shall not be used.	Used bold or italic text effects for emphasis where necessary.

Table A.4 Summary of Requirements for Alarm System Display [98]

Requirement	Status
Immediate Access	Implemented: Users can access ARPs directly from the alarm message location (O'Hara).
Display & Line of Sight	Implemented: Alarm indications are within 60 degrees of the user's normal line of sight (6105).
Alarm Displays & Controls	Implemented: Displays and controls allow for simultaneous viewing and operation (6105).
First-Out Alarms	Implemented: First-out displays are located at the main workstation or plant overview (0700).
Prompt Response	Implemented: Alarm displays and controls are arranged for timely access and response (6105).
Access to Controls & Displays	Implemented: Alarm panels are near relevant process controls and displays, or easy access is provided if on a VDU (6105, 6684, 0700).
Interference from Lights	Implemented: Non-alarm lights are positioned to avoid confusion with or obscuring alarm displays (6105).
Consistent Ordering	Implemented: Alarm display order matches the layout of related system controls and displays (6105).
Consistency with Main HSI	Implemented: Alarm system conventions align with the main HSI displays (e.g., red for critical alarms).
Guidelines Conformance	Implemented: Alarm system elements follow general HSI guidelines and specific alarm system guidelines (e.g., P&ID VDU displays).

Table A.5 Sample raw simulator log features by plant sections and alarms. Showing for Alarm, Tank, and Methanol [158].

Section/File	Acronym	Full Form	Y-Axis	Value
Allarm	All2	Allarm activation	Alarm label	0 = Not Run 1 = Run
	AckAll	Alarm Action	Status	0 = Not Run 1 = Run no action 2 = Run and Action
	sAll	Alarm Sound (no Master Silence)	Status	1 = Not Run 1 = Run Sound 2 = Run Single Sound Off
	proclis	Procedure visualisd	Procedure Number	0 = No Procedure 1001-1999 Tank Procedures 2001-2999 Methanol Procedures 3001-3999 Heat Recovery Procedures 4001-4999 Reactor Procedures 5001-5999 Assorber Procedures 6001-6999 Other Procedures
	t2	Duration		[s]
	intop	Open interface	1 = tank 2 = Methanol 3 = Compressor 4 = Heat recovery 5 = Reactor 6 = Assorber	0 = Close 1 = Open
	emmerO	Emergenci	Emergency status	0 = Off 1= On
	choice	Case study		
Tank	LSERB	Level	Level	[m]
	PSERB	Pressure	Pressure	[ata]
	FN2serb1O	Primary Nitrogen Flow	Flow	[Nm3/h]
	FN2serb2O	Back up Nitrogen Flow	Flow	[Nm3/h]
	FLO2	Air Flow	Flow	[Nm3/h]
	Nitsel	Nitrogen system selector	Status	0 = Off 1 = Primary 2 = Back up
	MNitsel	Manual Prymari nitrogen selector	Status	0 = Auto 1 = Manual
	x	Manual prymari flow value	Flow	0 = Off >0 = [Nm3/h]
Methanol	FMeps	Methanol Liqui pumped	Flow	[kg/h]
	WPOMP	Pump Power	Power	[kW]
	MpumpOld	Manual Pum selector	Status	0 = Auto 1 = Manual
	MWpopOld	Manual Pump power value	Power	0 = Off >0 = [kW]
	MliqbolO	Methanol Mass in the Boiler	Mass	[kg]
	FBOl	Methanol Boiled Flow	Flow	[kg/h]
	FBv	Boiler Steam flow	Flow	[kg/h]
	TBol	Methanol Boil Temperature		[C]
	MboilvapO	Manual Boiler Steam selector	Status	0 = Auto 1 = Manual
	FvapbolO	Manual Boiler Steam value	Flow	0 = Off >0 = [kg/h]
	TRis	Heater Methanol output Temperature	Temperature	[C]
	FRisv	Heater Steam	Flow	[kg/h]
	MhearvapO	Manual Heater Steam selector	Status	0 = Auto 1 = Manual
	MMvaprisO	Manual Heater Steam value	Flow	0 = Off >0 = [kg/h]

Table A.6 Sample raw simulator log features by plant sections and alarms. Showing for Compressor [158]

Section/File	Acronym	Full Form	Y-Axis	Value
Compressor	MaircompOld	Manual Air Compressor selector	Status	0 = Auto 1 = Manual
	MaircomppowOld	Manual Air Compressor value	Power	0 = Off >0 = [kW]
	WARglob	Air compressor Power	Power	[kW]
	FariaO	Air flow	Flow	[kg/h]
	GasselectOld	Gas Manual Selector	Status	0 = Auto 1 = Manual Vent Fraction 2 = Manual Gas Compressor
	frazventcOld	Vent Fractio	Fraction	[]
	FventO	Vent Flow	Flow	[kg/h]
	WRIglob	Gas Compressor Power	Power	[kW]
	FgaspomO	Gas Compressed Flow	Flow	[kg/h]
	MGascompowOld	Manual Gas Flow	Flow	0 = Off >0 = [kg/h]
	POutcomp	Gas Compressor Pressute	Pressure	[Pa]

Table A.7 Sample raw simulator log features by plant sections and alarms. Showing for Reactor [158]

Section/File	Acronym	Full Form	Y-Axis	Value
Reactor	Meold	Reactor methanol fraction In	Frac	[]
	Oold	Reactor Oxygen fraction In	Frac	[]
	Nold	Reactor Nitrogen fraction In	Frac	[]
	Wold	Reactor Water fraction In	Frac	[]
	Foold	Reactor Formaldehyde fraction In	Frac	[]
	Coold	Reactor Carbon Moxide fraction In	Frac	[]
	TinREaOld	Temperature Reactor In	Temperature	[K]
	PinREaOld	Pressure Reactor In	Pressure	[Pa]
	TMAXREATORE	Max Temperature in reator	Temperature	[C]
	TmreatO	Temperature long the reactor	Temperature	[C]
	MeOutRO	Reactor methanol fraction Out	Frac	[]
	OOutRO	Reactor Oxygen fraction Out	Frac	[]
	NOutRO	Reactor Nitrogen fraction Out	Frac	[]
	WOutRO	Reactor Water fraction Out	Frac	[]
	FoOutRO	Reactor Formaldehyde fraction Out	Frac	[]
	CoOutRO	Reactor Carbon Moxide fraction Out	Frac	[]
	TOutRO	Temperature Reactor Out	Temperature	[K]
	POutRO	Pressure Reactor Out	Pressure	[Pa]
	FRout	Reator Out Flow	Flow	[kg/h]
	FMW1	Steam Flow Cooling system	Flow	[kg/s]
	PEBre	Steam Pressure Cooling system	Pressure	[bara]
	TreatSteam	Steam Temperature Cooling system	Temperature	[C]
	MCoolreatOld	Manual Reactor Cooling selector	Status	0 = Auto 1 = Manual
	MCReatTempOld	Manual Reactor Cooling value	Temperature	0 = Off >0 = [C]

Table A.8 Sample raw simulator log features by plant sections and alarms. Showing for Heat Recovery [158]

Section/File	Acronym	Full Form	Y-Axis	Value
Heat recovery	PVREC1	Pressure Steam REC1	Pressure	[ata]
	TWrec1	Temperatura Steam REC1	Temperature	[C]
	MVAPrec1	Flow Steam REC1	Flow	[kg/h]
	Toutrec1	Temperatura Out REC1	Temperature	[C]
	Tinrec1	Temperatura In REC1	Temperature	[C]
	REC1SMO	Manual REC1 selector	Status	0 = Auto 1 = Manual
	REC1SMPO	Manual REC1 value	Pressure	0 = Off >0 = [ata]
	SezRec2O	Selector number of Heat Excanger in REC2	Number	1 = 1 heat exchanger used 2 = 2 heat exchangers used 3 = 3 heat exchangers used
	Toutrec2f	Reagent Out Temperature	Temperature	[C]
	Toutrec2c	Temperatura In REC3	Temperature	[C]
	Toutrec3	Temperatura Out REC3	Temperature	[C]
	TWrec3	Temperatura Steam REC3	Temperature	[C]
	MVAPrec3	Flow Steam REC3	Flow	[kg/h]
	REC3WMO	Manual REC3 selector	Status	0 = Auto 1 = Manual
	REC3WMFO	Manual REC1 value	Flow	0 = Off >0 = [kg/s]

Table A.9 Sample raw simulator log features by plant sections and alarms. Showing for Absorber [158]

Section/File	Acronym	Full Form	Y-Axis	Value
Absorber	LinasO	Water IN Flow	Flow	[kg/h]
	MAssWatO	Manual Assorber water selector	Status	0 = Auto 1 = Manual
	MAssWatFlowO	Manual Assorber water value	Flow	0 = Off >0 = [kg/h]
	TMEANCOL	Mean temperature in Assorber	Temperature	[K]
	TTOPCOL	Top temperature in Assorber	Temperature	[K]
	TLIQIN	Water IN temperature in Assorber	Temperature	[K]
	MfondassO	Liquid bottom Mass in the Assorber	Mass	[kg]
	yMeoutassO	Assorber Gas methanol fraction Out	Frac	[]
	yWoutassO	Assorber Gas Water fraction Out	Frac	[]
	yCOoutassO	Assorber Gas Carbon Moxide fraction Out	Frac	[]
	yFooutassO	Assorber Gas Formaldehyde fraction Out	Frac	[]
	yNoutassO	Assorber Gas Nitrogen fraction Out	Frac	[]
	yOoutassO	Assorber Gas Oxygen fraction Out	Frac	[]
	ToutassO	Temperature Assorber Gas Out	Temperature	[K]
	PoutassO	Pressure Assorber Gas Out	Pressure	[Pa]
	GAS	Assorber Gas Out Flow	Flow	[kg/h]
	xmasoutass	Assorber Liquid Formaldehyde fraction Out	Frac	[%]
	xmasoutassME	Assorber Liquid methanol fraction Out	Frac	[%]
	xmasoutassW	Assorber Liquid Water fraction Out	Frac	[%]
	ToutassOL	Temperature Assorber Liquid Out	Temperature	[C]
	LoutassO	Assorber Liquid Out Flow	Flow	[kg/h]
	MliqoutO	Manual Assorber Liquid Out selector	Status	0 = Auto 1 = Manual
	MliqoutflowO	Manual Assorber Liquid Out value	Flow	0 = Off >0 = [kg/h]

Appendix B

Figures

Next: Experiment Questionnaire showing personal data, demographics, training, experience, etc, SPAM etc



Fig. B.1 Recruitment Fyler: first part

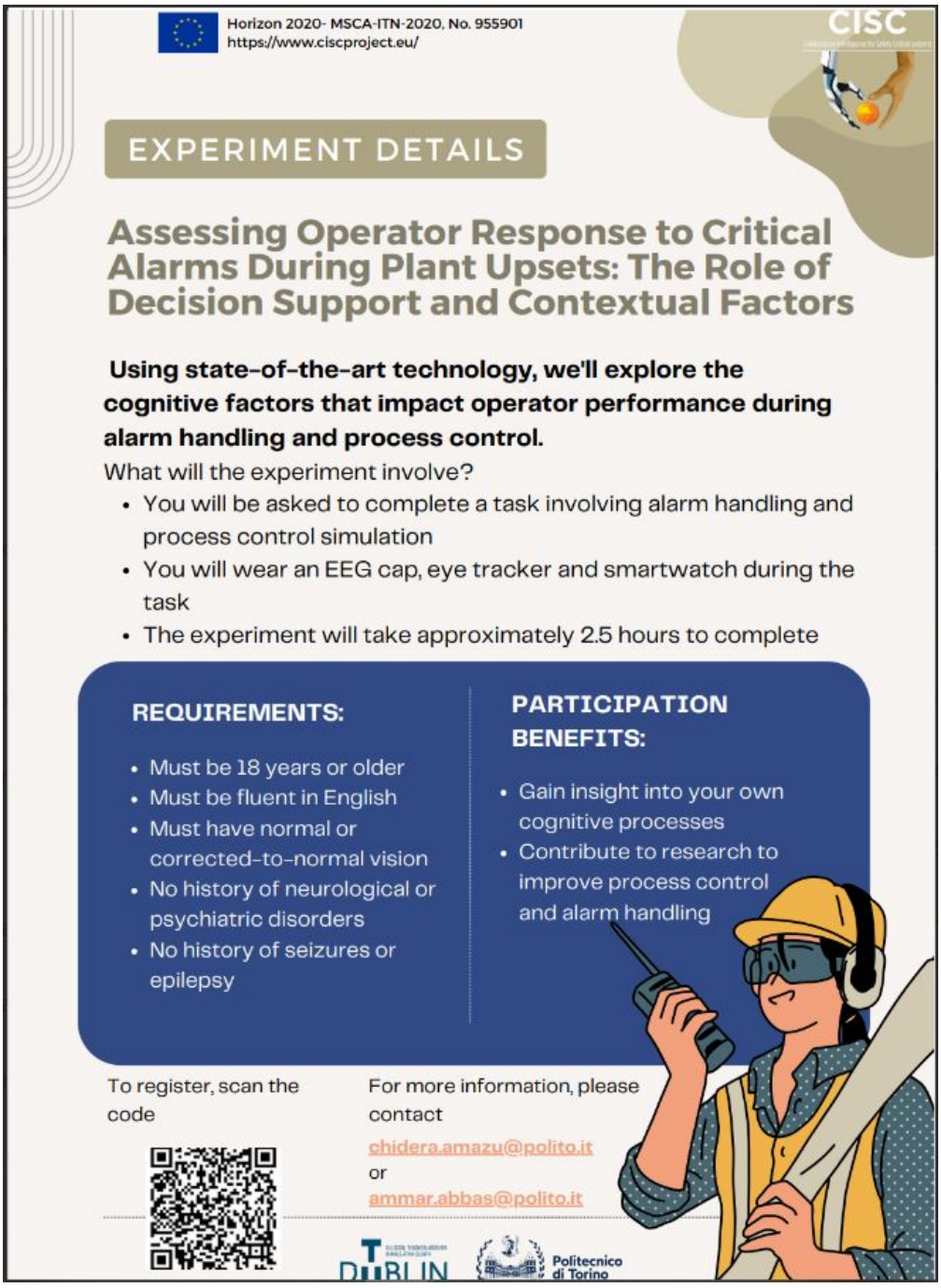


Fig. B.2 Recruitment Fyler: second part

Participation approval email

Subject: Confirmation: CISC LivLAb3 Research Experiment Participation

Dear [Participant's Name],

We are delighted to confirm your participation in our upcoming research experiment. Your involvement is invaluable to us, and we appreciate your willingness to contribute to our study.
Please take note of the following details:
Experiment Date: [Date]
Experiment Time: [Time]
Experiment Location: [Location]

We have carefully designed the experiment to be an engaging and informative experience. By wearing the EEG cap, eye tracker glasses, and smartwatch, you will have the opportunity to contribute to groundbreaking research in a fun and interactive manner.
To ensure the experiment runs smoothly and generates reliable data, we kindly request that you follow these **pre-experiment instructions**:

1. **Rest and Relaxation:** Prioritize getting a good night's sleep before the experiment to ensure you are well-rested and mentally alert during the session. Adequate rest contributes to optimal cognitive performance.
2. **Preparing for the EEG Cap:** Ensure that your hair is clean and free from styling products or excessive oils to facilitate proper electrode contact with your scalp. This will ensure accurate measurements.
3. **Eye Lenses:** If you wear contact lenses, please wear them during the experiment instead of eye tracker glasses. Make sure they are clean and comfortable for the duration of the session.
4. **Minimizing Movement:** During the experiment, try to minimize head and body movements to prevent artefacts in the EEG data and eye tracking measurements. This will help ensure accurate data collection.

We greatly appreciate your attention to these instructions and your commitment to the success of the experiment. If you have any further questions or require additional information, please do not hesitate to reach out to us at chidera.amazu@polito.it or ammar.abbas@polito.it
If, for any reason, you are unable to attend or need to reschedule, please notify us as soon as possible. Your participation is valuable to us, and we would appreciate prompt communication regarding any changes in your availability.

Once again, thank you for agreeing to participate in our research experiment. We look forward to your presence and the insights you will bring to our study.

Fig. B.3 Participation Approval Email

Participation denial email

Subject: Regretful Notification: Research Experiment Participation Denial

Dear [Applicant's Name],

Thank you for your interest in participating in our research experiment. We sincerely appreciate your enthusiasm and support for our study. However, after careful consideration, we regret to inform you that we are unable to include you in the experiment at this time due to specific requirements not being met.

We truly appreciate your understanding, and we encourage you to explore future research opportunities where you may have the opportunity to contribute. Your interest is highly valued, and we hope to have the chance to collaborate in the future.

Thank you once again for your interest, and we wish you all the best in your future endeavours.

Warm regards,

CISC LivLab3 Team
[Research Institution/Organization]

Fig. B.4 Participation Denial Email

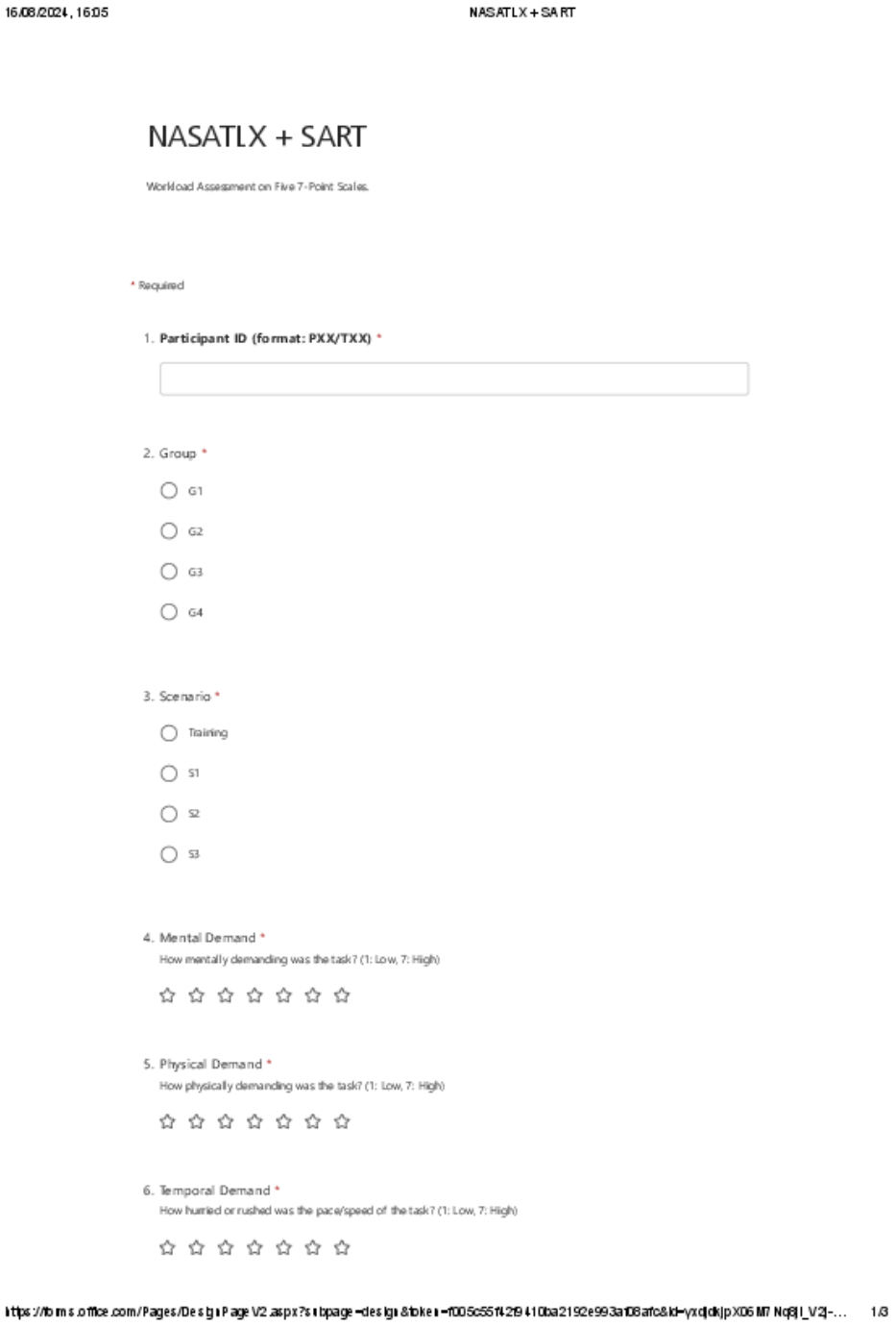


Fig. B.5 NASA-SART Questions with preliminary questions of participant ID, Group and Scenario, followed by NASA-TLX dimensions: Mental Demand, Physical Demand and Temporal Demand.

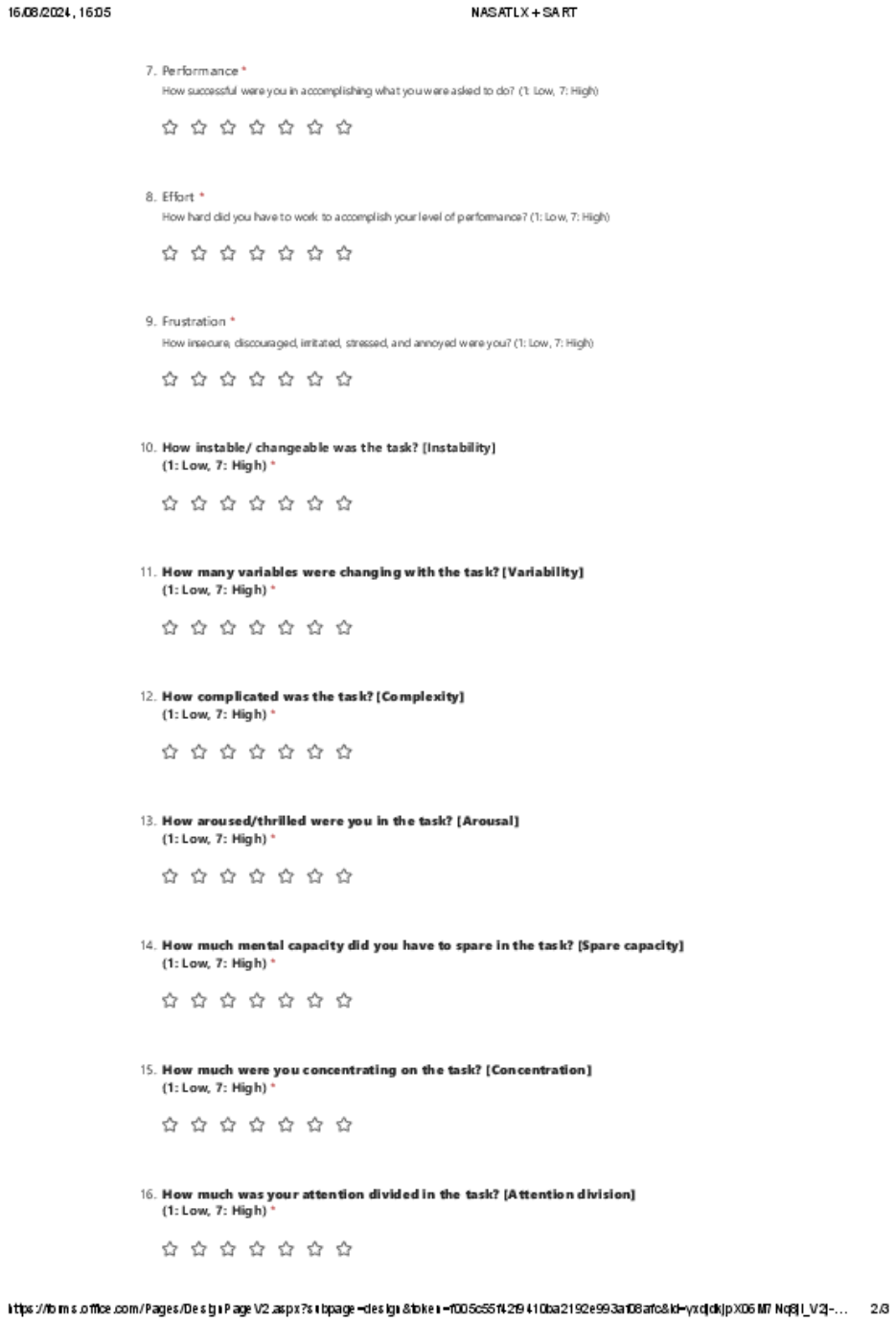


Fig. B.6 NASA-SART Questions showing NASA-TLX dimensions: Mental Demand, Physical Demand and Temporal Demand, and SART dimensions:

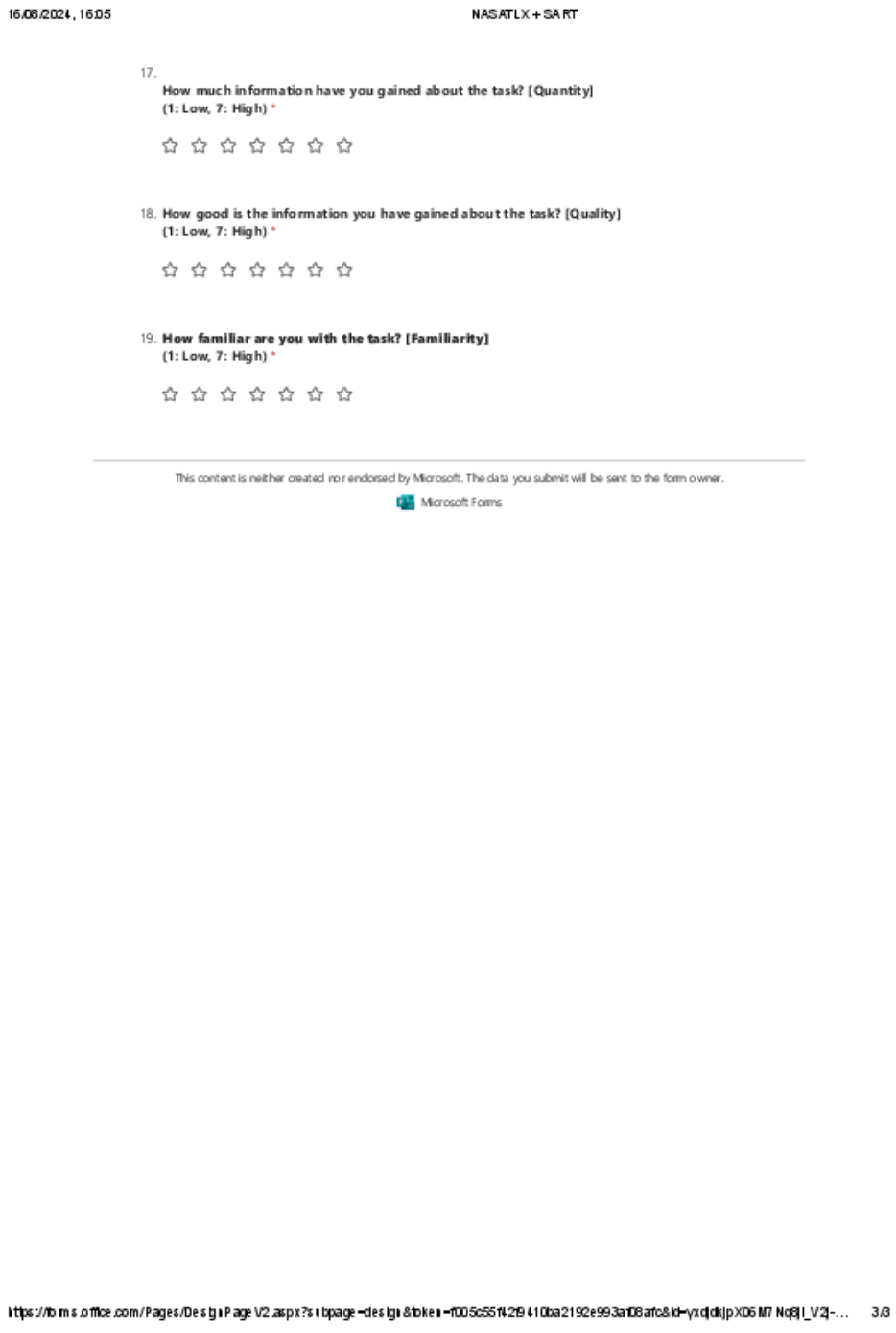


Fig. B.7 NASA-SART Questions showing SART dimensions:

This consent form was designed following the guidelines provided by the TU Dublin template

Researchers' Name	Houda Briwa, Chidera Winifred Amazu, Ammar Abbas, Joseph Mietkiewicz, Andres Alonso Perez
Academic Unit	School of Food Science and Environmental Health, Technological University of Dublin, Ireland
Title of Study	Assessing Operator Response to Critical Alarms During Plant Upsets: The Role of Decision Support and Contextual Factors

The following section should be completed by the research participant

	Yes	No
Have you been fully informed of the nature of this study by the researchers? *(Note that you should have received the participant information sheet)*	☐	☐
Have you had an opportunity to ask questions about this research?	☐	☐
Have you received satisfactory answers to all of your questions?	☐	☐
Have you been full informed of your ability to withdraw participation and/or data from the research at any time* without giving a reason for your withdrawal, or to decline to answer any particular questions? * *Since the data collected will not contain your identity, it can be withdrawn up until the end of the experiment, after which the researchers will not be able to distinguish your data from that of other participants.*	☐	☐
Have you been fully informed of what will happen to data generated by your participation in the study and how it will be kept safe?	☐	☐
Do you agree to provide information to the researchers under the conditions of confidentiality set out in the information sheet?	☐	☐
Do you wish to participate in the study under the conditions set out in the information sheet?	☐	☐
Do you consent to the use of anonymised / de-identified information (not containing your identity) collected in this research study to be used for scientific reports, books, conference proceedings, or journal publications?	☐	☐
Do you want to be contacted **before** the results of this research are published? (If so, please write your contact at the bottom of this page)	☐	☐
Do you want to be contacted **after** the results of this research are published? (If so, please write your contact at the bottom of this page)	☐	☐
Have you been informed that this consent form shall be kept securely and in confidence by the researchers?	☐	☐

Name of Participant			
Signature of Participant		**Date**	
Signature of Researchers		**Date**	

Fig. B.8 Consent Form

Plant: Plant X **Initiating Event:** Excess Water Criticality **Basic Event :** XHE **Event Coder:** bf

Basic Event Context: Cask Closure

Basic Event Description: Operators Fail to Perform Vacuum Drying System Connections/Setup

Part II. EVALUATE EACH PSF FOR ACTION

A. Evaluate PSFs for the Action Portion of the Task, If Any.

PSFs	PSF Levels	Multiplier for Action	Please note specific reasons for PSF level selection in this column.
Available Time	Inadequate time	P(failure) = 1.0 ☐	Adequate (nominal) time is assumed in this example.
	Time available is ≈ the time required	10 ☐	
	Nominal time	1 ☒	
	Time available ≥ 5x the time required	0.1 ☐	
	Time available is ≥ 50x the time required	0.01 ☐	
	Insufficient Information	1 ☐	
Stress/ Stressors	Extreme	5 ☐	There is no information available on stress in this example.
	High	2 ☐	
	Nominal	1 ☐	
	Insufficient Information	1 ☒	
Complexity	Highly complex	5 ☐	Observers of this task suggest moderate complexity for operators.
	Moderately complex	2 ☒	
	Nominal	1 ☐	
	Insufficient Information	1 ☐	
Experience/ Training	Low	3 ☐	Average (nominal) experience and training are assumed in this example.
	Nominal	1 ☒	
	High	0.5 ☐	
	Insufficient Information	1 ☐	
Procedures	Not available	50 ☐	Attachment refers to canister rather than cask, which employs different valve connections and contains inconsistent or missing symbols.
	Incomplete	20 ☐	
	Available, but poor	5 ☒	
	Nominal	1 ☐	
	Insufficient Information	1 ☐	
Ergonomics/ HMI	Missing/Misleading	50 ☐	There are no indications that ergonomics would significantly contribute to human error for this task.
	Poor	10 ☐	
	Nominal	1 ☒	
	Good	0.5 ☐	
	Insufficient Information	1 ☐	
Fitness for Duty	Unfit	P(failure) = 1.0 ☐	The fitness for duty of the crew is plant and crew specific. This analysis was determined from a generic model and therefore nominal is an appropriate choice.
	Degraded Fitness	5 ☐	
	Nominal	1 ☒	
	Insufficient Information	1 ☐	
Work Processes	Poor	5 ☐	The work processes are plant and crew specific. This analysis was determined from a generic model and therefore nominal is an appropriate choice.
	Nominal	1 ☒	
	Good	0.5 ☐	
	Insufficient Information	1 ☐	

Fig. B.9 PSF Evaluation Sheet for Action Task [113].

HRA Worksheets for LP/SD

SPAR Human Error Worksheet

Plant: Plant X Initiating Event: Excess Water Criticality Basic Event : XHE__ Event Coder: bf__

Basic Event Context: Cask Closure

Basic Event Description: Operators Fail to Perform Vacuum Drying System Connections/Setup

Does this task contain a significant amount of diagnosis activity? YES ☒ (start with Part I–Diagnosis) NO ☐ (skip Part I – Diagnosis; start with Part II – Action) Why? **Multiple steps, components, connections, and manipulations.**

PART I. EVALUATE EACH PSF FOR DIAGNOSIS

A. Evaluate PSFs for the Diagnosis Portion of the Task.

PSFs	PSF Levels	Multiplier for Diagnosis		Please note specific reasons for PSF level selection in this column.
Available Time	Inadequate time	P(failure) = 1.0	☐	Adequate (nominal) time is assumed in this example.
	Barely adequate time (≈ 2/3 x nominal)	10	☐	
	Nominal time	1	☒	
	Extra time (between 1and 2 x nominal and > 30 min)	0.1	☐	
	Expansive time > 2 x nominal & > 30 min	0.1 to 0.01	☐	
	Insufficient Information	1	☐	
Stress/ Stressors	Extreme	5	☐	There is no information available on stress in this example.
	High	2	☐	
	Nominal	1	☐	
	Insufficient Information	1	☒	
Complexity	Highly complex	5	☐	Multiple steps, components, connections, and manipulations.
	Moderately complex	2	☒	
	Nominal	1	☐	
	Obvious diagnosis	0.1	☐	
	Insufficient Information	1	☐	
Experience/ Training	Low	10	☐	Average (nominal) experience and training are assumed in this example.
	Nominal	1	☒	
	High	0.5	☐	
	Insufficient Information	1	☐	
Procedures	Not available	50	☐	Attachment refers to canister rather than cask, which employs different valve connections and contains inconsistent or missing symbols.
	Incomplete	20	☐	
	Available, but poor	5	☒	
	Nominal	1	☐	
	Diagnostic/symptom oriented	0.5	☐	
	Insufficient Information	1	☐	
Ergonomics/ HMI	Missing/Misleading	50	☐	There are no indications that ergonomics would significantly contribute to human error for this task.
	Poor	10	☐	
	Nominal	1	☒	
	Good	0.5	☐	
	Insufficient Information	1	☐	
Fitness for Duty	Unfit	P(failure) = 1.0	☐	The fitness for duty of the crew is plant and crew specific. This analysis was determined from a generic model and therefore nominal is an appropriate choice.
	Degraded Fitness	5	☐	
	Nominal	1	☒	
	Insufficient Information	1	☐	
Work Processes	Poor	2	☐	The work processes are plant and crew specific. This analysis was determined from a generic model and therefore nominal is an appropriate choice.
	Nominal	1	☒	
	Good	0.8	☐	
	Insufficient Information	1	☐	

Rev 1 (1/20/04)

Fig. B.10 PSF Evaluation Sheet for Action Task [113].