Summer 2024

# Exploring a New Conceptual Framework in Aviation Maintenance Incident Reporting Using Natural Language Processing

Christopher P. Braun
*Embry-Riddle Aeronautical University*, braunc1@my.erau.edu

# Exploring a New Conceptual Framework in Aviation Maintenance Incident Reporting Using Natural Language Processing

Christopher P. Braun

Dissertation Submitted to the College of Aviation in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Aviation

Embry-Riddle Aeronautical University

Daytona Beach, Florida

June, 2024

.

**Exploring a New Conceptual Framework in Aviation Maintenance Incident Reporting Using Natural Language Processing**

By

Christopher P. Braun

This dissertation was prepared under the direction of the candidate's Dissertation Committee Chair, Dr. Dothang Truong, and has been approved by the members of the dissertation committee. It was submitted to the College of Aviation and was accepted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in Aviation.

Dothang Truong Digitally signed by Dothang Truong Date: 2024.06.27 16:50:58 -04'00'

Dothang Truong, Ph.D.
Committee Chair

David A. Esser Digitally signed by David A. Esser Date: 2024.06.28 07:07:52 -04'00'

David A. Esser, Ph.D.
Committee Member

Jennifer E. Thropp Digitally signed by Jennifer E. Thropp Date: 2024.07.02 13:07:09 -04'00'

Jennifer Thropp, Ph.D.
Committee Member

Kenneth Kuhn Digitally signed by Kenneth Kuhn Date: 2024.07.03 06:54:48 -07'00'

Kenneth Kuhn, Ph.D.
Committee Member

Committee Member (External)

Steven Hampton Digitally signed by Steven Hampton Date: 2024.07.16 13:43:25 -04'00'

Steven Hampton, Ed.D.
Associate Dean, School of Graduate Studies, College of Aviation

Alan J. Stolzer Digitally signed by Alan J. Stolzer Date: 2024.07.17 11:13:44 -04'00'

Alan J. Stolzer, Ph.D.
Dean, College of Aviation

June 24, 2024

Signature Page Date

## Abstract

Researcher: Christopher P. Braun

Title: Exploring a New Conceptual Framework in Aviation Maintenance Incident Reporting Using Natural Language Processing

Institution: Embry-Riddle Aeronautical University

Degree: Doctor of Philosophy in Aviation

Year: 2024

In the analysis of human error incidents, human factors specialists predominantly follow two schools of thought in categorization strategies. One is to categorize the operator’s actions by the physical properties of the activity (phenotypes); the other focuses on the cognitive behaviors preceding the incident (genotypes). These categorization strategies are intended to foster an understanding of the events, identify risks, and assist in the implementation of risk reduction interventions. Commercial aviation maintenance presents unique challenges to human factor practitioners in terms of task requirements, working environments, communication modalities, and safety standards. Complex tasks are handed from one shift of personnel to the next, and technicians are under constant pressure to rapidly assess an aircraft’s status, perform the needed maintenance, and return the aircraft to revenue service.

This exploratory qualitative research analyzed the way in which aviation maintenance incidents are reported and compared the thematic focus of maintenance reporting to established human factor paradigms. As a result of this analysis, a novel human factor conceptual framework is introduced that improves the alignment of incident

investigation with maintenance incident reporting styles and improves the existing communication gap between incident reporter and incident investigator.

The research methodology detailed in this dissertation describes the analysis of a large corpus of aviation maintenance reports. This analysis is accomplished using natural language processing to implement Latent Dirichlet Allocation in a topic modeling strategy. The topic modeling process distilled these reports into a set of topic word groups that reflect the prevalent themes within the corpus of documents, allowing for a reasonable effort of evaluation by subject matter experts. Without the topic modeling process, the volume of data within the selected corpus of documents would be overwhelming and impractical for direct review and evaluation.

The topic modeling process and the topic word group thematic assessment were reinforced with subject matter expert evaluations by human factor and aviation maintenance specialists. The findings of this process are the inspiration for a novel human factor classification strategy focused on the way maintenance personnel describe maintenance events and the organizational responsibilities under whose authority and direction these maintenance activities occur.

This research contributes to aviation maintenance and human factors research by offering a previously unexplored approach combining natural language processing and qualitative evaluation to address the challenges encountered in the analysis of aviation maintenance incidents. The proposed human factor framework is a departure from established human factor paradigms and can, if accepted and effectively implemented, allow aviation maintenance organizations to develop effective risk mitigation strategies and improve technician performance and aviation safety.

*Keywords*: human factors, phenotypes, genotypes, aviation maintenance, aviation safety, incident categorization, communication, shift work, natural language processing, topic modeling, latent Dirichlet allocation, ASRS, qualitative evaluation, semantic analysis, organizational management, risk mitigation strategies

## Dedication

This work is dedicated to my family, both those still with me and those who have departed.

I could not have struggled through this journey without the love and support of my wife. If only for the support she provided, Iris Chauntelle Braun deserves a contributor credit for this research. Shaunie, you reassured me when I doubted myself. You made me laugh when I was down, and your confidence in me was contagious enough for me to have confidence in myself. I am forever grateful for your love and partnership over the last 225 years!

My children, Aaron, and Arianna have been a continual source of inspiration for me on this path and I thank them both for being a part of my life from the bottom of my heart.

My brothers, Doug, and Matt have added to my stability as I struggled through this academic journey and have always been willing to listen to my grumbling without judgment. I am lucky to have you!

Sharon, you gave me life. Without you I would not be. Thank you.

Paul and Pat, my parents, did not live to see me through this academic adventure, but I know that they are with me always. I would not be who I am if it were not for them. For that I am profoundly grateful.

**Acknowledgments**

The more I sort things out, the more they get distorted.
I sort of think I'm better off just leaving it unsorted.
--The Mighty Mighty Bosstones, *Someday I Suppose*

To err is human;
To understand the reasons why humans err is science.
--Erik Hollnagel, *International Journal of Man-Machine Studies*

I must acknowledge the contributions of the academic stewards who helped me on my doctoral journey.

Dr, Alan Stolzer, my academic advisor throughout my coursework continually reassured me that I did, in fact, belong in this program.

Dr. Ian McAndrews, Dr. Michael O'Toole, Dr. Haydee Cuevas, my dissertation chairpersons along the way, gave me input and guidance that led me to the completion of this work. Dr. Dothang Truong, my final and current chairperson, was able to help me untangle the mess of ideas I had formulated and point me in the direction of the finish line. This was no small task; I can assure you!

I would also like to acknowledge all of the subject matter experts who graciously donated their time and effort to provide crucial data that was instrumental in the completion of this work. I am still amazed at their willingness to assist me in this research effort.

## Table of Contents

Page

Dedication ........ vii

Acknowledgments ........ viii

List of Tables ........ xiii

List of Figures ........ xiv

Chapter I: Introduction ........ 1

Background ........ 2

Statement of the Problem ........ 7

Purpose Statement ........ 8

Significance of the Study ........ 9

Research Question and Hypotheses ........ 10

RQ1 ........ 10

RQ2 ........ 10

RQ3 ........ 10

Delimitations ........ 11

Limitations and Assumptions ........ 12

Summary ........ 14

Definitions of Terms ........ 14

List of Acronyms ........ 15

Chapter II: Review of the Relevant Literature ........ 19

Human Behavior and Cognition ........ 19

Humans in Complex Systems ........ 23

Aviation Maintenance ........ 25
Human Error Theory ........ 29
Human Factors Analysis and Classification System ........ 34
Cognitive Reliability and Error Analysis Method ........ 38
Semantic Bias ........ 46
Other Classification Methods ........ 47
Gaps in the Literature ........ 50
Theoretical Framework ........ 51
Summary ........ 56
Chapter III: Methodology ........ 57
Research Method Selection ........ 57
Population/Sample ........ 60
Population and Sampling Frame ........ 60
Sample Size ........ 61
Data Collection Process ........ 61
Design and Procedures ........ 62
Sources of the Data ........ 67
Ethical Consideration ........ 67
Data Analysis Approach ........ 68
Reliability Assessment Method ........ 73
Validity Assessment Method ........ 74
Data Analysis Process ........ 76
Qualitative Data Analysis Process ........ 78

Summary ........ 84

Chapter IV: Results ........ 85

ASRS Data Demographics Results ........ 85

Report Submissions by Year ........ 85

Report Submitters ........ 87

NASA SME Input ........ 88

Descriptive Statistics ........ 90

Refinement of the Dataset to Enhance Topic Relevance ........ 92

Validity of the Topic Refinement Process ........ 99

Topic Modeling with LDA ........ 99

Content Analysis of Topics ........ 106

Reliability and Validity Testing Results ........ 111

SME Consensus via Delphi Analysis ........ 111

Alternate SME Evaluation Methodology Results ........ 116

Development of a Novel Framework ........ 119

Summary ........ 122

Chapter V: Discussion, Conclusions, and Recommendations ........ 123

Discussion of Research Results ........ 123

Aviation Maintenance Topic Modeling ........ 123

Evaluating Modeled Topics to CREAM and HFACS ........ 127

A Novel Human Factor Framework for Aviation Maintenance ........ 128

Conclusions ........ 136

Theoretical Contributions ........ 137

Practical Contributions ... 139

Limitations of the Findings ... 139

Recommendations ... 142

Recommendations for Aviation Maintenance ... 142

Recommendations for Future Research Methodology ... 144

Recommendations for Future Research ... 145

References ... 148

Appendix A: Permission to Conduct Research ... 157

Appendix B: Data Collection Device ... 159

Appendix C: Tables ... 161

Appendix D: SME Qualifications ... 181

NASA ASRS Program SME ... 181

Human Factors SMEs ... 181

Aviation Maintenance SMEs ... 182

Appendix E: SME Training Module ... 185

## List of Tables


Table Page

1 HFACS Unsafe Acts of the Operator ........ 53

2 Hollnagel's Error Modes from the CREAM Classifications ........ 54

3 Assessment of ASRS Primary Problem by NASA Reviewer ........ 89

4 Baseline Topic Word Groupings ........ 93

5 Interim Topic Word Listings After Word Replacement ........ 94

6 Interim Word Topic Listing After Stop Word Modification ........ 96

7 Topic Word Lists Resulting from Relevance Refinement Process ........ 97

8 Topic Renaming by Gamma Prevalence ........ 103

9 Final Word Topic Groups ........ 105

10 Round 1 Delphi Analysis of Quantitative Responses ........ 113

11 Round 2 Delphi Analysis of Quantitative Responses ........ 115

12 A Comparison of HF Frameworks to Observed Categories. ........ 121

13 Potential Narratives Theorized for Topics Using Words Within Topic Word Groups ........ 125

## List of Figures


Figure Page

1 Rasmussen's SRK Processing Model ........ 22

2 The HFACS Framework ........ 36

3 The Contextual Control Model of Cognition ........ 41

4 Links Between Consequents and Antecedents for a Retrospective Analysis ........ 43

5 Overall Grouping of Phenotypes and Genotypes ........ 45

6 A Visualization of Theoretical and Conceptual Frameworks in Human Factors .... 52

7 Process Milestones ........ 63

8 LDA Plate Notation ........ 65

9 Data Treatment and Analysis Process Flow ........ 70

10 Generalized Delphi Analysis Process ........ 79

11 ASRS Maintenance Report Submissions from 2000 through 2022 ........ 86

12 Role of the Report Submitter ........ 88

13 Count of HF Contributors from NASA ASRS Evaluators ........ 90

14 Top 25 Words ASRS Dataset Before Preprocessing and Topic Enhancement ..... 91

15 Top 25 Words ASRS Dataset After Preprocessing and Topic Enhancement ........ 98

16 Gibbs Analysis for the Selected Dataset of ASRS Maintenance Reports ........ 100

17 Average Gamma Values by Topic in Declining Order Before Topic Reduction 101

18 Final Topic Gamma Averages for 11 Topics ........ 102

19 Beta Values of the Top 15 Words in Each of the Top 11 Topics ........ 104

20 An Example of Theta Distribution for a Document Within the Selected Corpus 107

21 Categories Within Topic Word Groups ........ 109

22 Topic Words Grouped by Potential Categories .................................................. 110

23 Proposed Aviation Maintenance TECS Classification Framework..................... 135

# Chapter I: Introduction

The recording of human error incidents in aviation maintenance requires communication of both the physical context in which the incident occurred and the possible cognitive influences leading the technician to the error event. These two separate but associated descriptions have distinct vocabularies and semantic nuances originating in differing cognitive frameworks (Le Coze, 2015). The described research focused on the narrative descriptions of human error incidents in aviation maintenance and how these narratives align with existing human factors categorization frameworks. These narratives represent human involved incidents recorded by aviation maintenance technicians that may have caused issues in aircraft availability or serviceability and may have resulted in injury or accident. Because the research is focused on narrative descriptions of incidents occurring phenomenologically in natural settings, the varying nature of language and context necessitates a qualitative evaluation of the narrative content (Miles et al., 2014). However, it is impractical to qualitatively assess the full content of thousands of incident reports for evaluation to human factor frameworks. A reduction of the data is required if the relevant content is to become apparent (Miles et al., 2014). In order to overcome these challenges, the described research employed a combination of algorithms based on a Natural Language Processing (NLP) method called topic modeling, and a Subject Matter Expert (SME) evaluation method termed a modified Delphi analysis.

There are many different frameworks to describe the human factors that influence human error incidents, and many similarities exist among the various taxonomies. In terms of prevalence, two schools of thought are readily apparent in the extant literature regarding human error analysis (Le Coze, 2015). Le Coze (2017) defined these as the *cognitive systems engineering* viewpoint and the *cognitive psychological* viewpoint. From the cognitive systems engineering viewpoint, human error can be analyzed as a physical phenomenon in which its classification is dependent on phenotype or the physics of the action. This classification arises from a cognitive systems engineering viewpoint that considers how the operator interacts with their environment. Alternatively, classification strategies can bypass the physical form of the operator's actions and focus on the psychological factors, or genotypes, that influence or enable the error activity. This cognitive psychological point of view focuses on the operator's interpretation of their environment and the human factors that influence that environment (Le Coze, 2015).

While these two points of view are prevalent in the published literature, this brief list is not an exhaustive one. The published literature demonstrates that descriptions of human error exist in many forms and with many varying themes, topics, and points of view to be found in the incident descriptions (Dekker, 2017; Escalante, 1999; Hansen, 2006; Perrow, 1999; Reason & Mycielska, 1982). These descriptions of human error incidents and their alignment with one of these two schools of thought may have relevance to the incident investigation and the eventual determination of causal factors.

**Background**

In the commercial aviation industry, maintenance activities account for most of the effort in terms of work hours and human effort. Rashid et al. (2013) estimated that maintenance hours outpace flight hours by a factor of 12:1. Maintenance is determined to be a contributing factor in as many as 23% of aviation accidents and serious incidents (Rashid et al., 2013) with 60% to 70% of these occurrences involving human error (Reason & Hobbs, 2003). Maintenance activities often occur in a variety of suboptimal conditions over off shifts, in the dark of night, exposed to weather and other environmental challenges, and under extreme time pressure. These conditions create a work environment with a high opportunity for human error (Reason & Hobbs, 2003).

Reason and Hobbs (2003) described the conditions influencing human activities in aviation maintenance:

> If some evil genius were given the job of creating an activity guaranteed to produce an abundance of errors, he or she would probably come up with something that involved the frequent removal and replacement of large numbers of varied components, often carried out in cramped and poorly lit spaces with less-than-adequate tools, and usually under severe time pressure. There could also be some additional refinements. Thus, it could be arranged that the people who wrote the manuals and procedures rarely, if ever, carried out the activity under real-life conditions. It could also be decreed that those who started a job need not necessarily be the ones required to finish it. A further twist might be that a number of separate groups work on the same item of equipment either simultaneously or sequentially, or both together. (p.1)

The chaotic nature of aviation maintenance, as described by Reason and Hobbs (2003), presents challenges to incident reporting that exist to a greater degree than what is experienced in other more structured parts of commercial aviation (Kuhn, 2018; Mellema, 2018). The structure and methodology of the Aviation Safety Reporting System (ASRS) data collection tool have proven useful for exploring correlations in incident descriptions and in formulating hypotheses regarding causation. There are, however, systemic limitations in the reporting methodology that fail to capture, to some extent, the conditions leading up to an event and the processes that facilitated the event (Rouse & Rouse, 1983). Rouse and Rouse (1983) noted that in these archival data sets, the analyst is limited to what the reporter chose to include in, or omit from the recording, and furthermore, such narrative datasets capture very little of the process that led to the occurrence. Despite these limitations, the ASRS provides a large and well documented database of maintenance incident descriptions that can be used for exploring the frames of reference of the reporting parties and the prevalent themes.

Aviation maintenance has an obvious physical component requiring cognitive psychological influence in the decision to select an action. Although standardization in aviation maintenance exists in the form of maintenance and repair manuals, the variability from one aircraft maintenance operation to the next necessitates a high degree of human adaptability with a lesser degree of automation than what is available in other industrial disciplines (Rashid et al., 2013; Reason, & Hobbs, 2003). Considering the unique nature of maintenance within the aviation disciplines, and the diagnostic variability of human error taxonomies, the exploration of human factors within incident records is warranted (Rashid et al., 2013, 2014; Reason & Hobbs, 2003).

The study of human error in a theoretical context, and the factors which contribute to its occurrence lie along the intersection between psychological theories of human behavior and the empirical observations of engineered systems. Le Coze (2017) has noted that these two aspects of human error analysis represent the prevalent schools of thought regarding how human error occurrences can be viewed, analyzed, and categorized.

Empirical observations provide us with a view of human error incidents based on the physical aspects of the event that fall under direct observation. This frame of reference, termed *phenotypical*, is exemplified by the description of the event in terms of its physics. Common terms or categories include the speed and direction of the energies involved as well as the timing and sequence of the events taking place (Hollnagel, 1993). These phenotypical references give a description of events as directly observed without the interpretation of influences that may have been contributing factors. Le Coze (2015) characterized this type of description and the categorizations that develop from it as a *cognitive systems engineering* frame of reference. Hollnagel (1998) presented a very clear example of this classification methodology in the Cognitive Reliability and Error Analysis Method (CREAM). Hollnagel's classification of error modes includes observations of timing, duration, force, speed, and sequence. These categories provide the opportunity to begin the classification of the event based solely on the observations of the witnesses without inferring what psychological influences may have been present (Hollnagel, 1998).

In counterpoint to the phenotypical descriptions, there are other description frameworks based on the psychological theories of human behavior, what Le Coze (2015) described as the *cognitive psychological* school of thought, which focus on descriptions, analysis, and categorization of the actions of the human within the system, and the mental processes that influence these human actions. These *genotypical* categorization strategies, described this way for their association with the genesis or origin of the event, are descriptive of the levels of cognition that occur. Skill-based, rule-based, and knowledge-based thought processes are prevalent in genotypical classifications, as are the human, organizational, and environmental influences that can influence the skill-based, rule-based, and knowledge-based decisions leading to the event occurrence (Rasmussen, 1983). The Human Factors Analysis and Classification System (HFACS), appears, based on its prevalence within the published literature, to be one of the most common and ubiquitous examples of a genotypical human error classification methodology. This framework was intended to provide an organizationally layered causal taxonomy designed to assist incident investigators in determining the human factors contributors facilitating errant operator actions. HFACS is focused on Rasmussen's Skill-Rules-Knowledge (SRK) processing model of human cognition and incorporates aspects of Reason's Generic Error Management System (GEMS) to provide a comprehensive, broadly descriptive taxonomy for the human factors that influence operator activities (Cohen et al., 2015).

HFACS, as a purely categorical taxonomy, does not contain any reference to interventional strategies or risk mitigation techniques. Despite criticism on this point and the lack of specificity in its taxonomical structure, the HFACS has demonstrated broad applicability in high risk/ high reliability industries (Cohen et.al, 2015). Studies by Cohen et al. (2015), Berry et al. (2010), and Tvaryanas and Thompson (2008) have demonstrated that while HFACS provides comprehensive diagnostic ability, the distribution among categories of the human factor contributors varies widely from one analysis to the next, indicating that there may be confounding variables within the research that affect this distribution of contributing factors. Diller et al. (2014) suggested that some portion of this variability can be attributed to the form and content of the information used to conduct the incident investigation. Diller et al. (2014) hypothesized that if the information collected for the investigation is not aligned with the investigational methodology, then the diagnostic ability of the investigational methodology may be compromised.

The investigational process in maintenance applications has resource limitations that are not present in other aviation disciplines. In incidents involving pilots and air traffic controllers, the investigative teams have data telemetry and voice recordings that provide a descriptive picture of the conditions and the situation surrounding the incident at the time of the occurrence. Maintenance incident investigators must rely on reporter recollections of the incident and the latent failure conditions that may have manifested long after the erroneous actions took place (Hobbs & Kanki, 2008). Furthermore, the inherent latency of maintenance errors allows for a phenomenon of latent error detection and recovery by maintenance technicians that may drive reduced reporting and analysis (Saward & Stanton, 2015). As observed by Saward and Stanton (2015), it is not uncommon for a technician to spontaneously recollect an error occurrence, such as a failure to replace an oil filler cap, at some later time while performing other work, and to go back and correct the error, often without recording the occurrence in any logbook or other reporting modality. These phenomena highlight the uniqueness of aviation maintenance activities and signal the necessity for dedicated analysis separate from other aviation disciplines.

## Statement of the Problem

The unique challenges of aviation maintenance activities as indicated by Reason and Hobbs (2003), Rouse and Rouse (1983), Mellema (2018), Kuhn (2018), and Rashid et al. (2013) create barriers to effective reporting and subsequent analysis. A review of the published literature did not indicate the existence of a comprehensive and focused human factor framework dedicated to the unique challenges faced in aviation maintenance from the perspective of the maintenance technicians, inspectors, and supervisors who report incidents, nor what effect their reporting may have on the subsequent analysis of aviation maintenance incidents. Descriptive incident records, such as those collected within the ASRS database reflect the impressions, experiences, and observations of the individual reporting the incident. The terminology used to describe the incident is reflective of the experiences and semantic choices of the reporter, and therefore is likely to convey biases inherent to the reporter and their background. Thus, understanding underlying causes of the incident requires an understanding of the context and frame of reference as viewed from the perspective of the individual reporting the incident. Each of the theoretical frameworks described by Le Coze (2015) and evidenced by the HFACS and CREAM systems, excludes a portion of the influences that may surround an aviation maintenance human error event. Because of this, a novel conceptual framework may be warranted to provide a better categorization strategy for these aviation maintenance events.

**Purpose Statement**

The general purpose of the described exploratory research was to gain an understanding of the underlying human factor subjects prevalent in aviation maintenance incident reporting by obtaining expert opinions of the relationship between the discovered subjects and established human factor conceptual frameworks. There were three specific tasks associated with the stated general purpose. The first was to determine what human factor subjects are prevalent in the selected corpus of aviation maintenance reports by employing machine learning methodology and NLP techniques. The second task was to obtain qualitative expert evaluations of the alignment between the discovered subjects in the ASRS reports and the existing human factor conceptual frameworks. The HFACS *unsafe acts of the operator* and the CREAM *error modes* were used as conceptual framework examples of genotypical and phenotypical frameworks. This qualitative evaluation was performed by aviation maintenance SMEs with human factor familiarity and refined through a structured consensus-seeking process. The third task was to propose a novel human factors framework aligned specifically with the prevalent human factor related subjects from the selected corpus of aviation maintenance incident reports.

**Significance of the Study**

The described research has both practical and theoretical significance. As a demonstration of theory, the described research provides an analysis of the selected ASRS reports that uses the verbiage and semantic structure of aviation maintenance technicians as a foundation for a novel framework of incident contributors. This novel framework deviates from existing paradigms in that it is not based solely on the actions or psychological state of the operator but rather on the tasks and environments developed by the organizations that direct the operator's activities. The analysis described in this research demonstrates a communication gap between the way aviation maintenance incidents are reported and the way in which they are investigated. These contributions to the body of knowledge in human error theory, while limited in scope by the delimitations of the described research, can be leveraged to more generalizable applications of human factor analysis.

In practical application, the results of the described research may provide guidance for awareness training to both the incident reporting and incident investigation processes. This training can be used to enhance the value of minor incident reports by aligning reporting and investigative activities to a common conceptual framework tailored to the aviation maintenance process. The application of the proposed novel framework can, if embraced by aviation maintenance organizations, assist in developing risk mitigations targeted towards the issues as they are described by maintenance technicians. These management interventions and the associated enhanced training, based on the realizations in the described research, could therefore reduce the risk of future major maintenance related accidents.

## Research Question and Hypotheses

The described research was based on the following research questions realized in review of the extant literature:

### *RQ1*

What human factor subjects are prevalent in the selected corpus of ASRS maintenance reports?

### *RQ2*

How are these human factor subjects aligned with existing phenotypical and genotypical human factor theoretical frameworks, including CREAM and HFACS?

### *RQ3*

What novel conceptual framework can be proposed to align the human factor subjects within the ASRS maintenance reports to the prevalent phenotypical and genotypical theoretical frameworks?

The described research represents a journey into the unexplored complexity of human error reporting and analysis. The belief is that undiscovered recurring human factor references exist within the qualitative reports recorded by aviation maintainers, and that these subjects within the corpus of documents can be used to provide new classification strategies in incident investigation. Although there are many event classification frameworks documented within the extant literature, there is no indication of thematic analysis based on the recording style or frame of reference held by the aviation maintenance event observers. Therefore, no hypothesis is proffered in advance of this exploratory research.

**Delimitations**

Although there would be practical and theoretical benefit in comprehensive analysis that would encompass the complete sphere of human error and the aviation maintenance industry, certain delimitations were necessary to contain the practical scope of the described research. The current study utilized aviation maintenance incident reports submitted to the ASRS database only by maintenance personnel. Any incident not recorded in the ASRS database or not pre-screened by National Aeronautics and Space Administration (NASA) ASRS personnel was not included in the analysis. The timeframe of the research was limited to the twenty-three-year period from 2000 through 2022. The Federal Aviation Administration (FAA) made the HFACS taxonomy available to the public in February of 2000 through the publication of DOT/FAA/AM-00/7 by the FAA Civil Aerospace Medical Institute, Office of Aviation Medicine (Shappell & Wiegmann, 2000). This allowed for the acceptance of the HFACS taxonomy by the commercial aviation industry in the United States (Shappell & Wiegmann, 2000). The selection of the timeframe in which the HFACS taxonomy has been active in commercial aviation increases the likelihood that technicians reporting maintenance incidents in 14 CFR Part 121 organizations may associate the reported incident with terms contained within the HFACS taxonomy. Furthermore, to maintain applicability to the commercial aviation sector, the described research was delimited to reports submitted under the 14 CFR Part 121 category with incidents occurring domestically in the United States. These two delimitations ensured a level of cultural commonality and reporting rigor.

**Limitations and Assumptions**

The described research was focused on the qualitative analysis of narrative data. Event narratives, as with all language-based data, are subject to certain limitations. Narrative descriptions are subject to ambiguity in terminology, cultural nuances, memory bias, and contextual dependence (Reissman, 2007). Despite these limitations, these phenomenological narratives also provide rich contextual descriptions that can capture ordinary events in their natural settings (Miles et.al, 2014). The parameters selected for the chosen ASRS data set, such as reporting role and incident location were selected to minimize the effect of these limitations, and to enhance the validity of the data for the research purpose.

The ASRS database contains a large volume of data regarding aviation maintenance incidents. The potential value of the ASRS data set in terms of human reliability analysis (HRA) is largely unexplored regarding the human factor contributors in aviation maintenance. While the information documented within the ASRS system is proffered voluntarily, reports vary greatly in detail and comprehensive content, thus may not be representative of other incidents or operating conditions. The NASA staff at the

ASRS program has a rigorous process that ensures report confidentiality, validates report relevance, and ensures publication accuracy. The ASRS program staff is comprised of experienced aviation professionals including pilots, air traffic controllers, and maintenance technicians, alongside a management team with aviation and human factors experience (NASA, 2023). The reports were assumed to be honest and without intentional deception because they are collected in a non-compulsory, non-punitive and confidential manner. While not assumed to be intentionally dishonest, there remains the possibility that documented events are recounted or recorded inaccurately, even though no subversive intent exists. Similarly, the report screening and recording performed by the NASA ASRS personnel were assumed to have been performed accurately and with an honest best effort. The ASRS process employs aviation professionals and subject matter experts who are familiar with all aspects of aviation operations and human factors investigations to ensure effective screening and pre-coding of reports before publication (NASA, 2023). The research further assumed that all software utilities used for analysis performed as designed within the limitations described in their associated usage documentation. The limitations of anonymous archival reports added to the challenges of analysis and enhanced the importance of understanding the frame of reference and mindset of those submitting reports.

The panel of SME’s participating in the Delphi analysis were screened for their experience in both professional aviation maintenance and human factors training. They were assumed, based on these experiences, to hold a high level of professional and academic expertise in both fields. The capability of the Delphi qualitative analysis method is limited to producing a consensus of expert opinions. This consensus, once realized, cannot be considered as scientific fact. It is the agreed upon opinion of the SME panelists who have agreed to participate in the research.

**Summary**

Aviation maintenance plays a significant role within the larger aviation industry, and yet historically lacks the analysis required to determine an effective standard method for the diagnosis and categorization of human error contributors in the practice of incident investigation. Investigating the differences in categorization methodologies may facilitate improved investigation strategies.

**Definitions of Terms**

| | |
|---|---|
| 14 C.F.R. Part 121 | The Code of Federal Regulations that defines operating requirements for U.S. domestic air carriers (Operating Requirements: Domestic, Flag, and Supplemental Operations, 2023). |
| Error Mode | The CREAM categories proximal to an event occurrence (Hollnagel, 1998). |
| Genotype | Descriptive term associated with the psychological cause of a human error event (Hollnagel, 1993). |
| Latent Dirichlet Allocation | A statistical model used in topic modeling to organize and categorize a collection of documents into topics. (Blei et al., 2003). |
| Phenotype | Descriptive term associated with the physical manifestation of a human error event (Hollnagel, 1993). |

| | |
|---|---|
| Topic Modeling | A method of data reduction used to uncover thematic concepts within a group of documents (Blei et al., 2003). |
| Unsafe Acts of the Operator | The HFACS category group that is proximal to an event occurrence (Shappell & Wiegman, 2000). |
| stopwords | Uninformative words removed from analysis text based on context (Sarica & Luo, 2021). |
| add-back words | Words deleted from R native stopwords lists based on contextual relevance. |

**List of Acronyms**

| | |
|---|---|
| AI | Artificial Intelligence |
| AMM | Aircraft Maintenance Manual |
| ASAP | Aviation Safety Action Programs |
| ASRS | Aviation Safety Reporting System |
| ATS | Action Trigger Schema |
| CAN | Aviation Confidential Number |
| COCOM | Contextual Control Model |
| CRAN | Comprehensive R Archive Network |
| CREAM | Cognitive Reliability and Error Analysis Method |
| CSV | Comma Separated Variable |
| ERAU | Embry-Riddle Aeronautical University |
| FAA | Federal Aviation Administration |

| | |
|---|---|
| FAR | Federal Aviation Regulations |
| FMEA | Failure Mode Effects Analysis |
| FRAM | Functional Resonance Analysis Method |
| GEMS | Generic Error Management System |
| GUI | Graphical User Interface |
| HF | Human Factors |
| HFACS | Human Factors Analysis and Classification System |
| HFACS-ME | Human Factors Analysis and Classification System Maintenance Extension |
| HRA | Human Reliability Analysis |
| IQR | Inter-Quartile Ratio |
| LDA | Latent Dirichlet Allocation |
| IRB | Institutional Review Board |
| MEDA | Maintenance Error Decision Aid |
| MEL | Minimum Equipment List |
| MMI | Man-Machine Interface |
| MRO | Maintenance, Repair, and Overhaul |
| NASA | National Aeronautics and Space Administration |
| NLP | Natural Language Processing |
| PSA | Probabilistic Safety Assessment |
| PSF | Performance Shaping Factors |

| | |
|---|---|
| SCM | Swiss Cheese Model |
| SHEL | Software, Hardware, Environment, Liveware |
| SHELL | Software, Hardware, Environment, Liveware, Liveware |
| SME | Subject Matter Expert |
| SRK | Skill, Rules, Knowledge |
| SRM | Structural Repair Manual |
| STAMP | Systems-Theoretic Accident Model and Processes |
| TECS | Task, Environment, Communication, Safety |

# Chapter II: Review of the Relevant Literature

This chapter will provide a review of cognitive behavioral theory, the role of humans in complex systems, and divergent schools of thought regarding the nature of human error and its contributing factors. The discussion will continue into aviation maintenance, its sensitivity to human influence, the unique characteristics that set it apart from other sectors of the aviation industry, and its importance within the scope of the aviation industry. At the confluence of these topics, existing gaps in the published research will be highlighted to demonstrate the significance of this research.

## Human Behavior and Cognition

Research into human behavior and cognition, specifically in industrial settings, began as early as 1918 with analysis performed by Woodworth (1918) and with Mayo's (1930) Hawthorne studies. The research, at this point, differed from previous behavior analysis in that the focus shifted from a psychological approach to a stimulus-response model. This model established the connection between the actions of the individual in response to the inputs received from environmental conditions (Pan et al., 2017). Building from the stimulus-response point of view, Rasmussen (1983) developed a novel viewpoint by introducing a model of separate levels of cognitive processing and implementing the SRK framework to describe how human cognition processed inputs at different levels based on familiarity and complexity (Pan et al., 2017).

A cognitive analysis of human error depends on analysis of human behavior. Human behavior is more than the result of deterministic input processing with resultant reflexive outputs. Rasmussen (1983) theorized that human behavior was teleological, based on purpose and conation over mere processing of environmental inputs. Humans are goal-oriented and actively seek information to assist in the fulfillment of selected goals (Rasmussen, 1983). Therefore, human behavior is goal oriented and dependent on the accurate processing of informational inputs to achieve results that service our selected goals.

In aviation, maintenance activities occur within a complex system of inputs and controls intended to achieve specific recordable results. The design of these complex systems and their specific attributes drive the human activity within the systems. Because complex systems are designed and engineered for specific purposes, the human actions required within the system are purposeful by nature and intended to meet defined goals, rather than being purely responsive to environmental stimuli (Rasmussen, 1983).

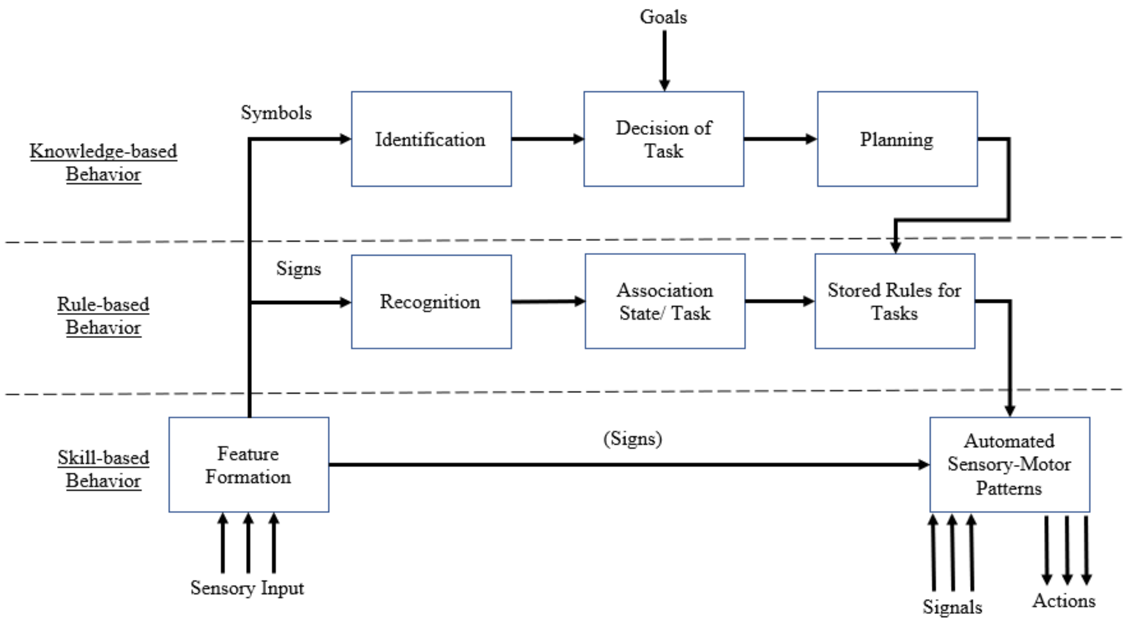
In describing his Theory of Reasoned Action, Ajzen (2005) noted, “Generally speaking, people intend to perform a behavior when they evaluate it positively, when they experience social pressure to perform it, and when they believe that they have the means and opportunity to do so” (Ajzen, 2005 p. 118). Ajzen discerned four influencing categories that determined how an individual would behave in each situation: the individual’s attitude towards the situation, beliefs regarding the situation, their behavioral intention or conation, and their previous experiences with observable behavior associated with the situation. These four factors play a key role in determining the action an individual will take to address the situation (Fishbein & Ajzen, 1980).

Norman (1981) offered a different perspective and proposed an Activation-Trigger-Schema (ATS) system to describe how human behaviors are enacted. Humans recognize situational patterns and activate behavioral sequences based on comparisons to previous experiential data. Norman discusses human behavior in terms of schema. A schema is an organized memory unit, a pattern which is used for comparison to current experiences. Schema are arranged in parent-child hierarchies based on complexity of behavior. High level parent schema are comprised of multiple child schema, which are, in turn parent to further lower-level child schema. At each level of complexity, comparison, recognition, and activation trigger patterned behavioral actions of comparable complexity which are also comprised of a series of lower-level action patterns (Norman, 1981).

Rasmussen's SRK model of cognition is likely the most ubiquitous view of human understanding in the human factors and safety research fields (Le Coze, 2015). Rasmussen (1983) established three levels of human performance based on the behavioral constraints the deterministic environment places on the human actor. Skill-based, rule-based, and knowledge-based behaviors are enacted based on the actor's responses to specific system inputs and expected goals of the actions. A simplified illustration of the SRK is presented in Figure 1.

**Figure 1**

*Rasmussen's SRK Processing Model*




*Note*. A simplified illustration of three levels of performance of skilled human operators. Adapted from "Skills, Rules, and Knowledge: Signals, Signs, and Symbols, and Other Distinctions in Human Performance Models" by J. Rasmussen, 1983, IEEE Transactions on Systems, Man, and Cybernetics, SMC-13(3), May/June 1983. Copyright 1983 IEEE.

At the most basic level, skill-based behaviors are largely autonomic and do not require analysis of complex mental models to enact the largely automated behavioral patterns. Environmental features signal autonomic behavioral patterns that do not require complex mental processing, what are commonly referred to as muscle memory behaviors. Skill-based behaviors are utilitarian activities that are recognized as single units in the mental process (Rasmussen, 1983).

Intermediary behaviors, termed rule-based, are composite behaviors triggered by environmental signs and controlled by stored patterns determined to be effective through previous personal experience or transferred knowledge. The operator recognizes the sign, associates it with a familiar task pattern and implements their personal set of stored rules for the task. Using rules based on past experiences, the actor matches a pattern of activities to a stored mental model to achieve an expected outcome (Rasmussen, 1983).

At the highest level of human cognition, associated with knowledge-based behavior, the actor can address unfamiliar situations where no known rules have been established, in order to achieve specific goals. The incoming information is symbolic and must be associated with known symbols based on pattern similarity. The actor accomplishes this by comparing a series of known mental models to current conditions, searching for commonalities and familiar patterns and identifying personal goals to formulate a high order response plan (Rasmussen, 1983).

Rasmussen (1983) noted that the distinctions between adjacent categories can become blurred and the classifications are not absolute. Processing of mental data to enact behavior requires the actor to perform mental activities related to aggregation, abstraction, and analogy to refine, interpret, and apply system inputs to the mental modeling process (Rasmussen, 1983).

**Humans in Complex Systems**

The intricacy of large technical systems, and the level of expertise needed for operator specialization, drive a locally rationalized point of view, prohibiting any single participant from achieving authority over the complete system. Behaviorally, human errors are activities that create unintended and non-normal conditions within complex systems (Dekker, 2017).

The role of humans within complex systems is commonly described in terms of the interactions between system components. Hawkins' (2017) refinement of Edwards' SHEL (Software, Hardware, Environment, Liveware) model places the human actor at the center of the system resulting in an improved SHELL (Software, Hardware, Environment, Liveware, Liveware) model, whereby the interaction with software, hardware, the environment, and other humans is demonstrated to be central to systemic function. Hawkins noted that the human at the center of the system is the most valuable and the most flexible component within the system. This flexibility, however, also exposes the human subject to elevated levels of inconsistency and variability, affecting system predictability (Hawkins, 2017). Rasmussen (1982) agrees on this point, noting that although human variability can be seen as a source of system instability, it goes hand in hand with human adaptability, allowing for greater system stability as humans are able to adapt to unforeseen design variation within complex systems. Industrialization in most human experiences, from manufacturing and energy production to transportation, finance, and health care, has deepened the need for understanding of human-machine interaction. The complexity of these endeavors heightens the need for understanding the role played by liveware at the center of Hawkins' SHELL model. Catastrophic incidents like those that have occurred at the Three-Mile Island nuclear facility and the Bhopal chemical plant provide stark reminders that a full understanding of the permutations of human-machine interaction in complex systems is all but impossible (Perrow, 1999). Estimations of these interactions rely on analysis of human behavior, human intention, and human performance (Hollnagel, 1998).

**Aviation Maintenance**

The complexity of the modern aviation industry is well documented within the published literature (Rashid et al., 2013). While the potential for human error exists within all sectors of the aviation industry, aviation maintenance has inherent characteristics that provide unique risks and analysis opportunities. Maintenance is a significant activity in aviation operations. Maintenance hours exceed flight hours by a factor of 12:1 (Rashid et al., 2013). Complex maintenance operations can extend work over successive shifts where one crew will remove a large assembly, another will perform maintenance activities, and a third will re-install. Work progress, assembly documentation and other relevant communication must pass from one crew to the next to assure completion of work to specification and full compliance to regulatory requirements. This work happens under organizational pressure to maintain schedule and profitability (Warren et al., 2013). Mellema (2018) highlights the importance of a participational reporting culture in the evaluation of safety performance within the context of aviation maintenance. Errors attributable to inadequate technical communication modalities like incorrect documentation and insufficient procedures are the most common incident occurrences in the maintenance disciplines (Chatzi et al., 2019). Human error is cited as a causal factor in a large majority of aviation accidents and incidents (Hollnagel, 1998; Rasmussen, 1983; Reason & Hobbs, 2003; Saward & Stanton, 2015; Shappell & Wiegmann, 2000) and is an operational commonality in flight operations, ramp operations, air traffic control operations and in aviation maintenance.

Maintenance activities differ from operational activities performed by flight crew. Errors in flight crew performance, though subject to influence from latent contributors, are active, and the impact of the error is readily noticeable, in most cases. In the investigative process, this facilitates diagnosis. Conversely, although the work environment in which maintenance errors are created is active, the defects often lie dormant and undiscovered until uncovered by a mishap or an unsuspecting flight crew. The identification of aviation maintenance errors relies largely on inspection and self-detection by mechanics who are subject to schedule and performance pressures from within their organizations (Saward & Stanton, 2015). These factors discourage discovery of errors and diagnosis of contributing factors, prohibiting effective analysis (Chiu & Hsieh, 2016). In order to learn from previous incidents, aviation maintenance reports are focused on post-incident learning and analysis. The latent nature of the errors in aviation maintenance highlights the importance of the inclusion of detailed information including processes, procedures, and technical competencies in the error reporting process. These activities facilitate the development of corrective mitigations and increase the body of knowledge used to mitigate future risks. This creation of new knowledge is, given the retroactive nature of incident communication in maintenance activities, as dependent on technician insights as it is on mechanistic recording of statistical data (Clare & Kourousis, 2021).

In addition to possessing technical proficiencies directly related to maintenance activities, maintenance personnel are required to be proficient in non-technical skills related to leadership, communication, decision making, and task management (Irwin et al., 2016). Organizational pressures, supervisory inputs, and environmental challenges provide aviation mechanics with unique physical and mental challenges. Human factors contributors strongly influence maintenance activities, suggesting that in-depth analysis is warranted (Irwin et al., 2016; Rashid et al., 2013).

Saward and Stanton (2015) highlight the latent nature of human error in aviation maintenance in their observance of instances where aviation mechanics recover self-created error conditions after operations have been completed. Hobbs and Williamson (2003) note that as many as 60% of maintenance personnel surveyed reported that they had corrected an error made by another mechanic without reporting the incident. This occurs in maintenance activities despite the non-punitive environment fostered through aviation safety action programs (ASAP) dedicated to error reduction in maintenance and other aviation sectors (Hobbs & Kanki, 2008). Saward and Stanton (2015) note that the disparate nature of daily tasks, time pressure, and resource constraints contribute to a phenomenon of interrupted schema and the observed correction after the fact of previously undetected errors.

Perrow (1999) described the characteristics of high-risk technologies while making the case for the inevitability of accidents associated with these technologies. In high-risk technologies, the potential cost of catastrophic failure is extremely high in terms of loss of life and fiscal impact. The operational system of a high-risk technology contains many components, including machine components, organizational procedures, and human operators. These components are tightly coupled in that a failure of one component cannot be isolated from adjacent components quickly enough to prevent a cascade of failures within those adjacent components. Additionally, the number of components, the complexity of their arrangement within the system, and the tight coupling are sufficiently complex that the effects of failure of a given component may not be known despite any design analysis that may have been performed. Examples of these high-risk technologies include nuclear power generation, chemical processing plants, and commercial aviation (Perrow, 1999). Noting examples of high-risk technologies that operated in nearly error free environments, Roberts and Rousseau (1989) identified organizational characteristics common to nearly error free, or high-reliability organizations. These characteristics are a high degree of accountability to the point where deviation from procedure or substandard performance will have severe adverse consequences, hyper complexity, and a tight coupling of tasks, as also noted by Perrow (1999). Also common to high reliability organizations are extreme hierarchical differentiation within the organization, large numbers of decision makers arranged in complex communication networks, high frequency of immediate feedback about decisions, compressed time factors where operational cycles are measured in seconds rather than hours or days, and that more than one critical outcome must happen simultaneously (Roberts & Rousseau, 1989). Taken as a single complex system, it seems evident that commercial aviation operations meet the requirement set by Perrow (1999) to be considered as a high-risk technology and that the elements of high reliability organizations are present as described by Roberts and Rousseau (1989). However, when considered as an isolated subsystem of the industry, aviation maintenance is missing key characteristics of high reliability organizations. Specifically, the operational cycles in aviation maintenance may stretch over multiple days comprised of many work shifts (Reason & Hobbs, 2003). Furthermore, there is a limited amount of immediate performance feedback. Although inspection operations are commonplace requirements within any maintenance operation, Reason and Hobbs (2003) noted that 60% of mechanics surveyed admitted to finding and correcting errors of previous mechanics without providing notification, and research by Saward and Stanton (2015) noted it was a common occurrence for a maintenance technician to self-realize a previous error and backtrack to make a correction without documenting either the error or the needed correction. The difference in timescale of operations, and the facilitation of latent error conditions created while shifting workloads and tasks, set aviation maintenance apart from other areas of the industry such as flight operations and air traffic control. This study provides insight regarding the dearth of research focused on the maintenance sector of the commercial aviation industry.

**Human Error Theory**

The study of human error lies along an intersection between psychological theories of human behavior and empirical observations of engineered systems. This intersection of diverse schools of research, cognitive psychology, and systems engineering, leads to diverse points of view regarding the nature of human error and even disagreement on the validity of the term *human error* itself (Le Coze, 2015).

Dekker (2017) provided clarification on the origin of human error by stating “Bad outcomes are not the result of human immoral choice, but the product of normal, locally rational interactions between people and systems through which control is often maintained, and sometimes lost” (Dekker, 2017, p.557). Accident prevention and risk mitigation require a comprehensive understanding of human behavior, human error, and human factors, particularly in complex systems. Dekker’s (2017) view of human error is reflective of the views on the role of human operators in complex systems that have developed over the last century. The inevitable nature of human error is well documented throughout the historical record. Socrates argued for it in Plato's Protagoras dialogue (Plato, ca. 380 B.C.E./1990), and the 17th century English poet Alexander Pope noted that "To err is human, to forgive divine" (Pope, 1903, p. 19). The birth of psychoanalysis, coincident with the advent of the industrial revolution, however, brought other points of view. Freud (1904/1989), in his 1904 publication *The Psychopathology of Everyday Life*, theorized that verbal and physical slips and lapses, which he termed parapraxes, were the expressions of subconscious desires of the id into the conscious mind. Freud's theorizations said that there were no accidents, or that they were extremely rare, and that the errors in speech and action were reflections of the true desire of the individual (Freud 1904/1989). Anecdotally, Freud's point of view appears to have been the prevailing wisdom until industrial psychologists and engineers began to seek novel methods to improve output in industrial production.

Reason and Mycielska (1982) continued with a similar view to Freud's early work albeit without the view that the mental slips and lapses which comprised the bulk of human errors were intentional on a subconscious level (Le Coze, 2015). The discussion of human error in an industrial setting accelerated after the 1979 Three Mile Island nuclear incident. Systems engineers with a concern for safety began to analyze human behavior as a component of the larger complex system in which the human operated. Le Coze (2015) characterized two schools of thought on the nature of human error: the cognitive psychological school pioneered by Reason (1990) which viewed human error solely as a result of mental processing, and the joint cognitive systems school which considered human error as a result of systemic influences that exploited variability in human performance. Reason's (1990) and Reason and Mycielska's (1982) views on human error considered the influences of human cognition that facilitated human error while cognitive systems engineers like Hollnagel (1993; 1998) viewed these occurrences as erroneous actions that did not produce the intended result due to the influences of the complex systems and inherent human variability (Le Coze, 2015).

Other industrial engineers viewed human influences as necessary inconveniences in complex systems like manufacturing. Shingo (1989) believed that the risk of human error, although inevitable in its nature, could be mitigated through mistake proofing methodologies. Deming (1992) noted that it was virtually impossible for a job to be performed correctly the first time thus requiring a continuous feedback loop, the Plan-Do-Check-Act cycle, to drive process improvement and reduce the risk of human error within manufacturing environments. These viewpoints recognized human variability as inevitable and believed that application of technology, such as Shingo's (1989) mistake proofing methodologies, could all but completely mitigate the risk of having humans as a systemic component (Escalante, 1999). Perrow (1999) minimized the importance of human errors within complex systems because, given the characteristics of complex systems, multiple and unexpected interactions of failures are inevitable, regardless of whether direct actions of operators are involved (Perrow, 1999).

Hansen (2006) analyzed human error as a theoretical concept, working from the etymology of the terms, through psychological theory, and into human-factors fields including focused research in transportation, accident investigation, nursing, and engineering. In searching for a commonality in the definitions of human error, several salient points regarding the defining attributes of an action that results in human error are noted:

a) The action is performed by a human being.

b) The action occurs at the interface between the human and another system (human, machine, environment).

c) The action is voluntary and deliberate.

d) The action exceeds tolerance limits (Hansen, 2006).

As a result of these findings, supported by case analysis, Hansen (2006) offered a novel definition of human error; “Human error is a voluntary and deliberate action by a human interacting with another system that exceeds established tolerances defined by that system” (Hansen, 2006, p.74).

Rasmussen (1982) pointed out an inherent difficulty in the identification of an action as human error. In cases where system performance fails to meet expectations due to a human act or a system disturbance that a human had an opportunity to correct, the cause of the performance shortcoming will likely be attributed to human error. This diagnosis, however, only occurs after the system effect has been observed and the potentially discrepant action has been analyzed. Human errors can only be identified after the fact making it difficult to point to an immediate comprehensive definition. Noting the difficulty in determining a universally satisfactory definition of human error, Rasmussen (1982) put forth the idea that the identification of a person’s action, or the result of that action, is dependent upon the context within which the action was taken, and the frequency with which the undesired result occurs. Rasmussen considers human errors to be mismatches at human-machine interface points that are present within complex systems. If the mismatch occurs frequently, it is more likely to be categorized as a design error in the system rather than an operator-induced condition. Errors that occur more infrequently, within the parameters of human variability, are generally considered to be human errors. These actions are only classified as human errors because they are performed in a complex and demanding work environment. Were the same action-result sequences to occur in an experimentally controlled test environment, they would be considered valuable data points that would be used to analyze and improve system stability. Thus, the identification of human error is dependent upon frequency and context (Rasmussen, 1982).

Hollnagel (1993) agreed with Rasmussen (1982) on the retrospective nature of human error identification. However, Hollnagel (1993) argued that the term human error as commonly used was, in fact, a misnomer. Hollnagel noted that the phenomenon that others called human error could not be directly observed and evaluated at the point of occurrence. This qualification of human error as a transient and, for any practical purpose, unobservable phenomenon indicated to Hollnagel that the occurrences fail to meet any traditional standard for scientific analysis. Hollnagel further clarified that human action can be observed and that the result of an action can subsequently be classified as either undesirable or unexpected, and assigned with human error as a causal factor. Following from the position that human error is not an observable phenomenon, Hollnagel posited a clarification of terminology to say that an incorrectly performed action that produces an unsatisfactory result should be termed as an erroneous action rather than a human error. He goes on to note that the distinction between human error and erroneous action is not trivial, as in common use, human error can refer to both the cause of an event and to a special class of action, driving inherent ambiguity regarding the intent. The proffered term--erroneous action--clarifies intent to refer to the action observed without assigning causal factors in advance of any desired analysis (Hollnagel, 1993).

Regardless of specific phraseology or linguistic delimitations employed to describe human actions with undesirable consequences, Dekker (2017) is very clear in making the point that identification of the human contribution to an undesirable outcome is not the endpoint of an incident investigation. Whether the human action is believed to be intentional, accidental, or misguided, the analysis of the error and its modality should be the starting point in identifying what systemic shortcoming influenced the behavior of the human within the system. Human behavior is, for the most part, consistent in that is locally rationalized. Therefore, the behavior that made sense to one person under a given set of systemic conditions will likely make sense to another person if the same set of conditions exist in the future, generating near identical behavioral patterns in response (Dekker, 2017).

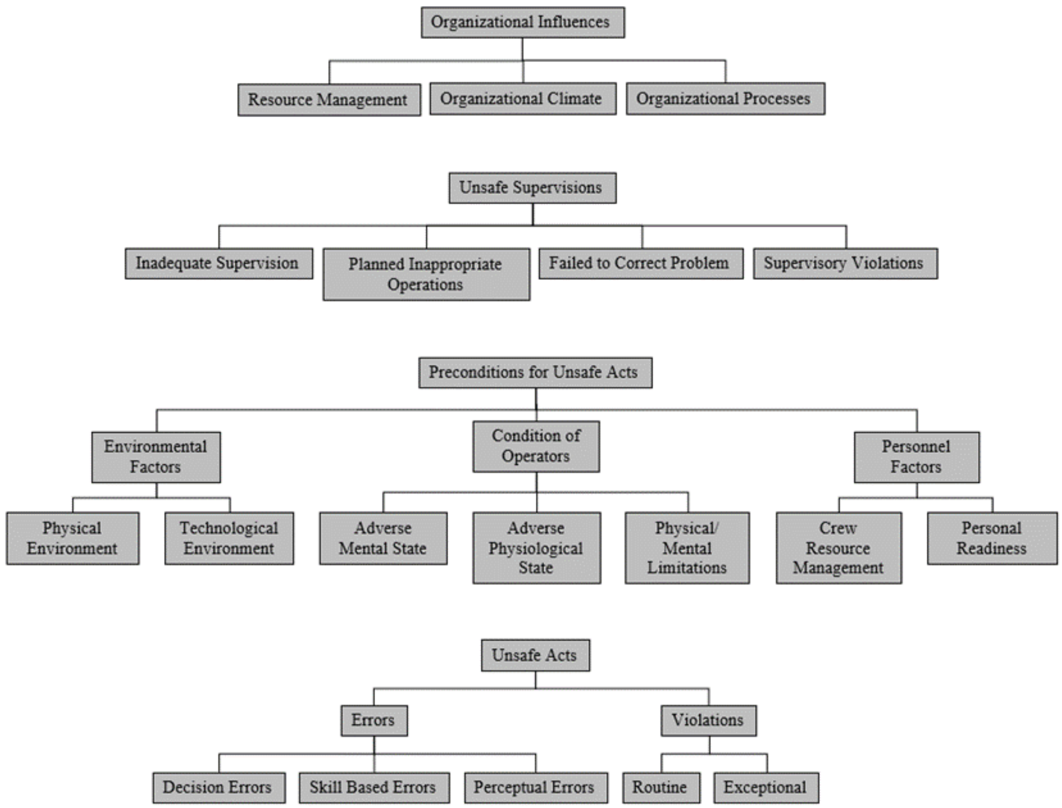
**Human Factors Analysis and Classification System**

Representative of the cognitive psychological school of thought (Le Coze, 2015) and an example of genotypical categorization, the Human Factors Analysis and Classification System (HFACS) provides an organizationally layered causal taxonomy designed to assist incident investigators in determining the human factors contributors facilitating errant operator actions (Cohen et al., 2015). HFACS is one of the most commonly used methods to connect human errors to accident causation (Lower et al., 2018). HFACS analysis has repeatedly demonstrated high inter-rater and intra-rater reliability. It is applicable to a wide range of industrial applications with high-risk, high-reliability, and human dependent operational factors (Ergai et al., 2016). Using a causal framework, HFACS provides investigators with a classification tool to categorize errant acts of human behavior that generate unexpected conditions within complex systems (Cohen et al., 2015). A representation of the HFACS categorization framework is shown in Figure 2.

Accident investigators employ the HFACS taxonomy to facilitate investigations in aviation, rail, maritime transportation, nuclear power generation and civil engineering (Lower et al., 2018). HFACS follows the structure of Reason’s Generic Error Modeling System (GEMS) and loosely defined layers of the Swiss Cheese Model (SCM) of accident causation (Reason, 1990) but enhances the effectiveness of these categories by adding 19 sub-categories detailing specific causal factors (Ergai et al., 2016). Thus, HFACS bridges the gap between theoretical and practical application of human factor concepts (Cohen et al., 2015).

**Figure 2**

*The HFACS Framework*




*Note*. Adapted from "The Human Factors Analysis and Classification System- HFACS" by S. Shappell and D. Wiegmann, 2000 (https://rosap.ntl.bts.gov/view/dot/21482). In the public domain.

HFACS categories are arranged in a hierarchical order loosely aligned with physical and temporal proximity to the incident occurrence. HFACS categorization begins with the active factors closest to the incident, termed "the unsafe acts of the operator" (Shappell & Wiegmann, 2000, p. 1). The unsafe acts of the operator include unintentional skill-based, decision-based, and perceptual errors by the operator. Unsafe acts also encompass intentional violations, which can be either routine or exceptional (Shappell & Wiegmann, 2000).

Preconditions for unsafe acts, the second layer in HFACS analysis, are the factors directly associated with the operator that affect the operator but are not active in accident causation. These factors include the operator's physical and mental state, the environment in which the operator operates, and the relationships between the operator and other individuals with whom the operator interacts. Although the preconditions are proximate to the accident, they are latent factors framing the conditions in which the incident occurs (Shappell & Wiegmann, 2000).

Unsafe supervision and organizational influences comprise the two upper layers of HFACS factors. These layers, both attributable to the organizational structure in command of the operator, influence the world view in which the operator exists. These layers provide the professional and regulatory guidance that determines an operator's work environment. The factors within these layers are latent contributors which create conditions that lie hidden, sometimes long before the accident occurs (Shappell & Wiegmann, 2000).

The extant literature suggests that HFACS is not intended to be a problem-solving strategy, but rather a taxonomical approach to human error categorization. The framework of HFACS analysis, as designed, lacks sufficient dimension to provide specific corrective action implementations. This has generated criticism within the safety and risk mitigation community (Harris & Li, 2011). This dimension of risk mitigation is, however, beyond the scope of HFACS' intent. Cohen et al. (2015) note that the purpose of HFACS is "to provide investigators with a tool for conducting a human factors analysis of accidents" (Cohen et al., 2015, p. 729). Despite this systemic shortcoming of the HFACS taxonomy, the connection that HFACS provides between behavioral theory and practical application makes it a benchmark tool for classification of contributing factors during accident investigation. The layered arrangement of HFACS categorizations allows for a *locally rationalized* viewpoint presented in an overall organizational context. This local rationalization aligns with Dekker's (2017) viewpoint that people do what makes sense to them at the time, given the available information. The extant literature suggests that this sets HFACS apart from other classification systems in common use, such as Dupont's Dirty Dozen (Cohen et al., 2015; Dupont, 1997; Mellema, 2018). While the Dirty Dozen methodology, adopted by American, Canadian, and other transportation safety agencies, provides clear common-sense guidelines for the prevention of the most common human error occurrences (Mellema, 2018), it lacks the resolution to establish how errors originate within an organizational structure or operational environment.

**Cognitive Reliability and Error Analysis Method**

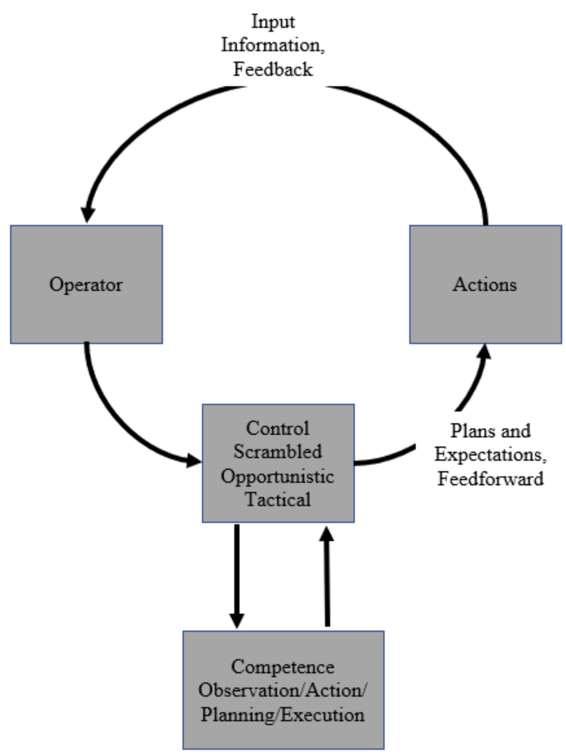
At the level most proximal to the incident occurrence, Hollnagel's (1998) Cognitive Reliability and Error Analysis Method (CREAM) is a distinctive example of a phenotypical classification system, and representative of the cognitive systems engineering school of thought (Le Coze, 2015). Citing the pervasiveness of human erroneous action and the focus upon human action failures in accident investigation, Hollnagel (1998) indicated the need for Human Reliability Analysis (HRA). Hollnagel (1998) provided an analysis of nine different first-generation HRA methodologies that expand upon previous Probabilistic Safety Assessment (PSA) conventions. Developed from a need to meet practical goals, the majority of first-generation HRA approaches lacked academic foundation, providing acceptable results in a practical sense but lacking the confidence of theoretical support. These first-generation methodologies improve on four recognized shortcomings of the previous PSA approaches. The PSA approaches were deficient in event tree specificity, limited error modes that could not account for cognitive errors, inadequate consideration of Performance Shaping Factors (PSFs), and insufficient operator models that lacked consideration for multi-stage information processing. Hollnagel (1998) noted that, while each of the methods analyzed addressed some of the shortcomings from the earlier PSA approaches, none spoke to all the cited shortcomings. This indicated a need for a second generation of HRA that would comprehensively enhance event tree analysis, provide error modes that would include consideration of cognitive errors, include PSFs early in the analysis, and enhance the operator model to consider multi-stage information processing (Hollnagel, 1998).

The Cognitive Reliability and Error Analysis Method (CREAM) is Hollnagel's enhancement of his previous Phenotypes of Erroneous Action Taxonomy intended to fulfill the need for second generation HRA. CREAM contains three key elements intended to address the shortcomings noted in the first-generation HRA analysis: an operator model of cognition, a bidirectional recursive classification method, and a non-hierarchical classification system. These three key elements support and define each other within the CREAM system, and it is apparent that they lose relevance when considered individually. Rather than adopting the common terms *cause* and *effect* to describe conditions that result in a particular action, Hollnagel (1998) establishes the designation of *antecedent* to describe conditions that are in existence when an erroneous action occurs. Also, the effects of antecedents are referred to as *consequents* rather than effects. Both antecedents and consequents can be designated as general (intermediary) or special (terminal) based on their systemic influence and a general consequent can be an intermediary antecedent of a more specific consequent. CREAM employs phenotypical consequents as error modes which are the effect of genotypical antecedent contributors to the erroneous action, thus creating a comprehensive error analysis system that can be used either predictively in system design or retrospectively in incident investigation. Hollnagel notes that his method is recursive rather than strictly sequential and contains an inherent clear stop rule to indicate when an analysis is sufficiently completed. The clear-stop rule ensures that use of the methodology is consistent across varied applications. Furthermore, the clear stop feature addresses the pitfall of endless analysis inherent in the recursive nature of the methodology (Hollnagel, 1998).

CREAM employs a Contextual Control Model (COCOM) derived from a Simple Model of Cognition employed in early iterations of the analysis method, to describe the mental processes associated with cognitive reliability. The COCOM model, as shown in Figure 3, centers on the assumption that “human performance is an outcome of the controlled use of competence adapted to the requirements of the situation, rather than the result of pre-determined sequences of responses to events” (Hollnagel, 1998, p. 154).

**Figure 3**

*The Contextual Control Model of Cognition*




*Note*. Adapted from “Cognitive Reliability and Error Analysis Method: CREAM” by E. Hollnagel, 1998. Copyright 1998 by Elsevier Publications.

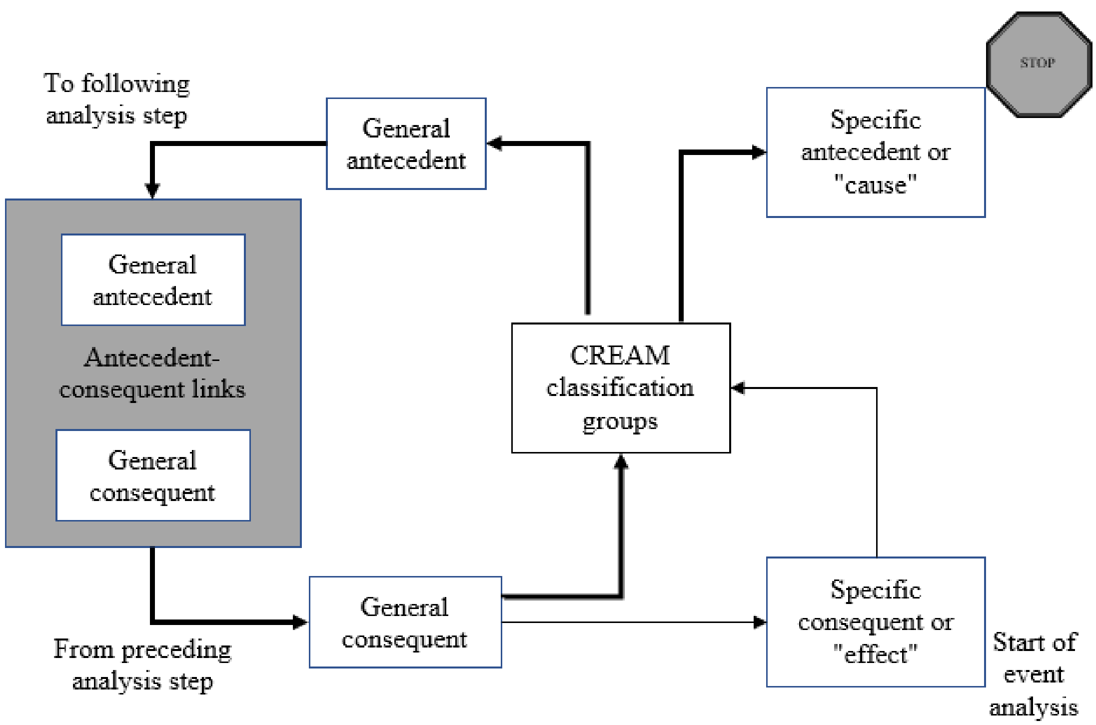
From this it can be inferred that human action is both intentional and reactive, with actions based on the operator’s contextual control of events and personal competence related to actions associated with the interpretation of the system state. Hollnagel (1998) describes four levels of control necessary to organize actions within the operator’s timeline. Ranging from a panic state to a long-term goal setting mode, the four control states are scrambled control, opportunistic control, tactical control, and strategic control. These four categories are not described as discrete states, but rather as waypoints along a continuum of an arc of control. Competence, in the COCOM model, is addressed in a more general manner. It is simply the operator's personal ability for observation, interpretation, planning, and execution of actions. Although a key characteristic of the CREAM method, the method is not derived from the COCOM model. The model gives structure to the classification system and serves as a framework for the relationship between some of the groups in the classification scheme. The COCOM model supports a distinction between observation of actions and an inference of their cause. The overt behavior of an individual, the action of task execution, can be observed. Further analysis of an action, determining the cognitive functions, is secondarily inferred from the observation. The model accounts for the cyclical nature of human cognition, recognizing that action events unfold in the context of an understanding of past events, and in anticipation of future effects. This process of understanding is recursive and cyclical, with development over time (Hollnagel, 1998).

Hollnagel (1998) notes that the CREAM methodology can be used both in a predictive manner and in retrospective investigation. The two methods are similar and share common principles. The retrospective method is germane to the current research and is discussed here, while the predictive method is not applicable and will not be described further.

Retrospective analysis in CREAM is intended to establish a probable path of cause and effect from a qualitative analysis of an erroneous action. Using a descriptive narrative of the incident, analysis begins with a determination of error modes present in the incident. The identification of error modes leads the user to the *genotypical antecedent* categorizations. Figure 4 illustrates the CREAM analysis method and demonstrates the consequent-antecedent relationships. The user then identifies either general or specific antecedents associated with the error mode. General antecedents then point to other antecedent categories, while the identification of specific antecedents signals an end point to that error mode branch of analysis. The analysis continues recursively until each identified error mode has a specific antecedent identified, if one exists (Hollnagel, 1998).

**Figure 4**

*Links Between Consequents and Antecedents for a Retrospective Analysis*




*Note*. Adapted from “Cognitive Reliability and Error Analysis Method: CREAM” by E. Hollnagel, 1998. Copyright 1998 by Elsevier Publications.

The CREAM classification system centers on the observable characteristics of the incident to be analyzed. These observable characteristics are designated as error modes. The observable error modes are connected to three general categories of antecedent conditions: man, technology, and organization. Error modes are observable occurrences while the antecedent conditions, both general and specific, must be inferred (Hollnagel, 1998).

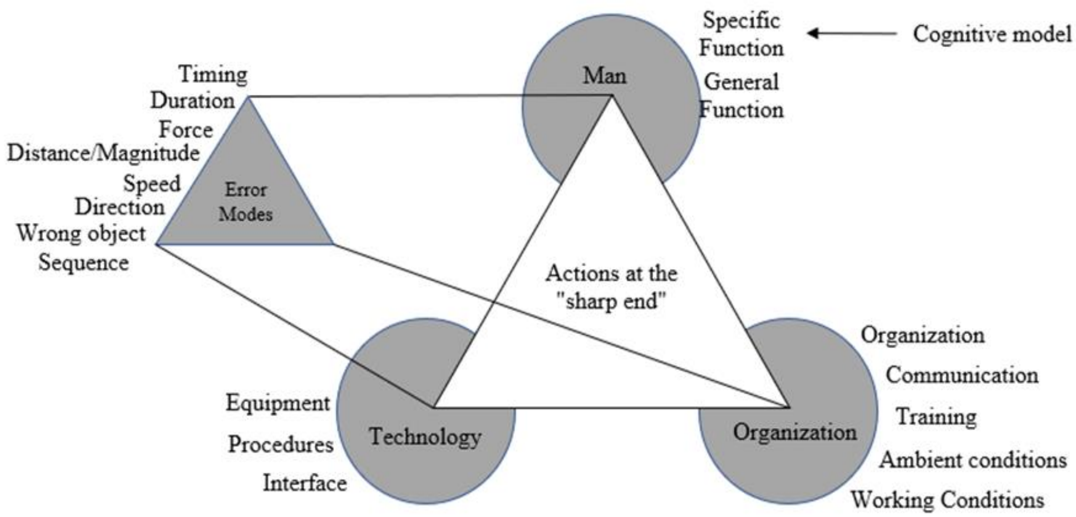
Hollnagel (1998) identifies four separate categories of error mode: action at wrong time, action of wrong type, action at wrong object, and action in wrong place. These four error modes are then divided into general consequent categories. Eight distinct phenotypes make up the general consequent class of error modes. The observable classifications are timing, duration, force, distance/magnitude, speed, direction, wrong object, and sequence. Each error mode references a subgrouping of two to five refinements termed specific consequents. In retrospective analysis, these classifications are intended to provide a basic starting point to answer the question “What happened?” (Hollnagel, 1998).

Antecedents are divided into three distinct groupings. The person related antecedent category contains five subcategories for observation, planning, interpretation, temporary and permanent person related causes. The technology related category contains four subcategories for equipment, procedures, and temporary and permanent human-machine related interface problems. Finally, the organization related genotype category contains antecedent subgroupings for communication, organization, training, ambient conditions and working conditions (Hollnagel, 1998). Figure 5 contains a graphical representation of the antecedent and consequent groupings.

The non-hierarchical nature of the CREAM methodology necessitates the need for designations of likely general and specific antecedents for each categorization. When listed as a general antecedent, a category becomes a consequent to another *downstream* antecedent. When identified as a specific antecedent, the category can be considered a contributing factor to the erroneous action, and therefore signals a stopping point for that specific branch of the investigation (Hollnagel, 1998).

**Figure 5**

*Overall Grouping of Phenotypes and Genotypes*




*Note.* Adapted from “Cognitive Reliability and Error Analysis Method: CREAM” by E. Hollnagel, 1998. Copyright 1998 by Elsevier Publications.

Hollnagel (1993) makes the point that the classifier should avoid mixing the observation of an activity with the interpretation of what caused the activity unless there is a strong supportable theory for the interpretation. In the noted case of erroneous action, no such strong theory exists, thus classification must begin with the pure observed facts of the action; “the systematic study of erroneous actions must clearly be based on what can be observed and verified” (Hollnagel, 1993 p. 3). This point of view provides the basis for the CREAM classification methodology and sets CREAM apart from other HRA techniques.

Hollnagel (2021) recently issued a disclaimer regarding CREAM on his personal website. Because CREAM focuses solely on the human in the system rather than adopting a complete system view, and due to the focus on how actions can go wrong over system variability, Hollnagel expressed the opinion that the CREAM methodology is obsolete and lacks relevance to more advanced views of system safety (Hollnagel, 2021). Although the author has declared the system to be obsolete in favor of more advanced system views such as the Functional Resonance Analysis Method (FRAM), the application of CREAM as a categorization and analysis method for erroneous human actions and their contributing factors remains relevant particularly as an exemplar of a phenotypical classification system within the cognitive systems engineering school of thought.

**Semantic Bias**

Preeminent work by Tversky and Kahneman (1981) established that decision framing, or how a choice is presented to a decision maker, has a marked influence on decisions made based on the narrative presented. The study of decision framing established that a rational decision maker will prefer the prospect that offers the highest level of utility based on the way in which the prospect is presented. This theory of decision framing further demonstrates that perceptions of an event narrative will vary based on the semantic and linguistic choices of the event reporter. Much of the published work on decision framing and interpretive bias has focused on media narratives and click-bait news reporting (Armstrong et al., 2020; van Hulst et al., 2014). Asher et al. (2021) use epistemic message exchange games to establish the foundations of a formal model for interpretive bias based on linguistic and cognitive mechanisms. The extant literature clearly indicates that the way an event is described will influence how it is evaluated, regardless of the intent (if any) of the semantic bias contained within the narrative.

**Other Classification Methods**

There is no shortage of methods for classification and analysis of human error. Each method has novel aspects of analysis and classification associated with its processes and level of focus. An overview of some of these investigational paradigms is presented here.

Dupont's (1997) Dirty Dozen provides a simple common-sense classification of human error types commonly seen in aviation maintenance. Originally presented as a series of workplace safety posters, each categorization is associated with an eye-catching cartoon drawing above a list of suggested mitigations to avoid the focus error type (Dupont, 1997). Although the Dirty Dozen lacked the veracity of academic research and peer reviewed publication, it became an industry standard for safety communication in aviation maintenance (Mellema, 2018).

Schmidt et al. (2000) developed the HFACS Maintenance Extension (HFACS-ME) from Shappell and Wiegmann's (2000) HFACS taxonomy based on an analysis of

450 Naval aviation maintenance incident reports. The HFACS-ME contains a series of subgroupings for each of the original HFACS categories. The subgroupings are intended to provide a finer resolution of the maintenance incident lending additional detail to the analysis (Schmidt et al., 2000).

The Boeing Maintenance Error Decision Aid (MEDA) provides a comprehensive system for aviation maintenance error investigation (Rankin, 2000). MEDA is arranged around a layered taxonomy similar to the HFACS model. Contributing factors can be associated with the maintenance personnel at the point of error commission, the immediate environment, supervision, or organizational philosophy. The MEDA system uses a multi-modal error model that recognizes the effect of multiple contributing factors that can increase incident probability due to serial or parallel co-occurrence. Error classifications within the taxonomy are phenotypical based on an omission-commission-violation model (Rankin, 2000).

Leveson (2015) describes the obsolescence of previous accident models due to increased system complexity, rapid technological change, increased reliance on automation, and the rapid evolution of socio-technical environments. While previous accident models depended heavily upon analysis of actions and operating environments, Leveson (2015) proposed the Systems Theoretical Accident Model and Process (STAMP) which focuses on the system control structures in determining the conditions that facilitate an accident occurrence. The occurrence itself is secondary to the operational control structures of the system. In this way, STAMP provides a model that accounts for the overall role human decisions and human behavior play in highly complex technical systems and the interactions between decisions made by multiple, interacting decision makers at varying system levels. STAMP focuses on system constraints and control structures as feedback mechanisms designed to enact system control. Thus, system failure is attributable to inadequate feedback or constraint and control design deficiency (Leveson, 2015).

Hollnagel (2012) cites a number of fallacies present in previous HRA methodologies in determining the need for a systems-based methodology for accident investigation. These include the inaccurate decomposition of system components and interactions, the predetermination of a linear order of events during system operation, and the absolute value of success or failure of system components. Hollnagel (2012) further noted that although these assumptions may be accurate in purely technical systems, their validity decreases in a blended socio-technical environment. The Functional Resonance Analysis Method (FRAM) relies on four basic principles of socio-technical systems to establish a method for system analysis:

- Success and failure result from the same behaviors.
- Work within complex socio-technical systems is underspecified and under resourced, requiring constant, approximated adjustment.
- It is not always possible to explain event causes from a set of known processes.
- Performance variability in complex systems occurs in a pattern of functional resonance.

FRAM relies on these four principles to identify and describe system functions, identify system variability, estimate the aggregation of system variability, and estimate the consequence of the analysis. FRAM is a highly theoretical method that establishes a holistic view of system performance intended to establish an estimate of system unknowns (Hollnagel, 2012).

Although some of these additional classification systems are tailored to the maintenance environment, there were additional considerations that omitted them as exemplars of the phenotypical and genotypical classification strategies. The lack of academic veracity in Dupont's (1997) Dirty Dozen, the over specificity of the Schmidt et al. (2000) HFACS-ME, and the inclusion of a fault tree analysis structure in Rankin's (2000) MEDA are examples of the factors that made them less desirable choices for the described research.

**Gaps in the Literature**

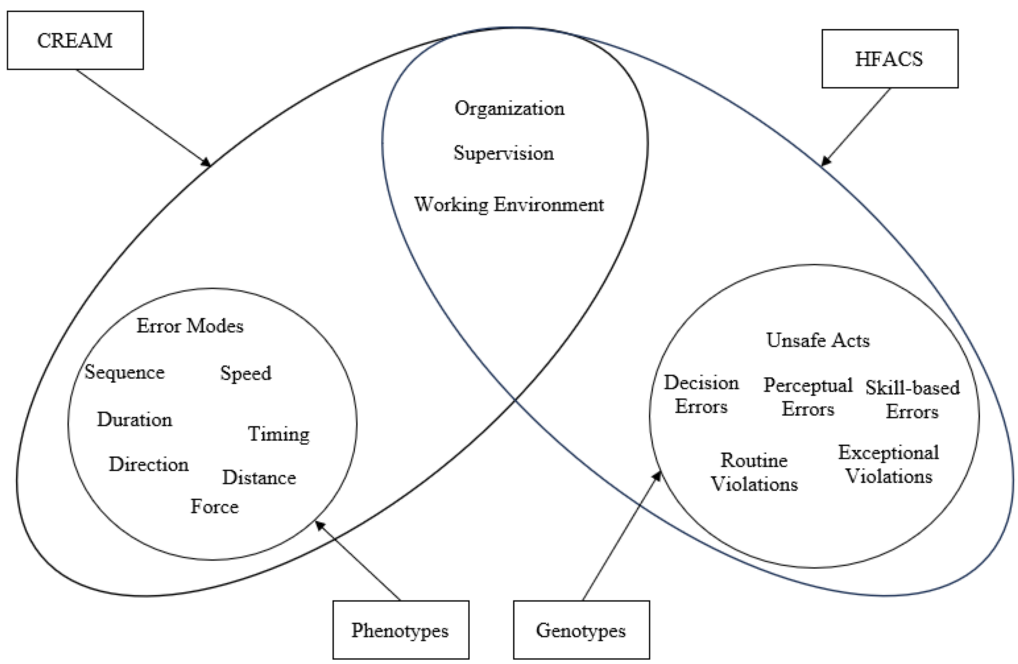
Categorization of human error has been explored in terms of both its genotypical origins and as a phenomenological occurrence. Each of these paradigms has received post initiation analysis in its own right (Cohen et al., 2015; Ergai et al., 2016; Phillips & Sagberg, 2014). A search of the extant literature, however, has not revealed any analysis focused on the linguistic style or semantic bias of the incident reporter in reference to these two divergent theories. Aviation maintenance has received some level of attention regarding human error analysis and novel approaches to mitigation efforts as evidenced by Rashid et al. (2014), Saward and Stanton (2015), Hobbs and Kanki (2008), and Mellema (2018). Although each of these works is dedicated to the specific set of challenges faced in aviation maintenance, there appears, at the technician level, a failure to comprehensively address all of the human factors that may influence the aviation maintainer. As a segment of the larger aviation industry, the maintenance discipline is underrepresented in the literature (Mellema, 2018) particularly in consideration of the ratio of flight hours to maintenance hours (Rashid et al., 2013). The described research addressed these opportunities by evaluating a corpus of aviation maintenance reports to determine the semantic themes and tendencies of the incident reporters for alignment with the aforementioned theoretical paradigms.

**Theoretical Framework**

Le Coze (2015) established that there are two very distinct schools of human error theory. One is based on psychological tradition, and the other on a cognitive view of systems engineering. These definitive schools of thought form a theoretical framework of human error with distinct concepts of phenotypical and genotypical categorizations overlapping at a higher cognitive level but distinctly separate at an operational level. Figure 6 presents a visualization of the theoretical frameworks and overlapping conceptual frameworks in human factor categorizations.

**Figure 6**

*A Visualization of Theoretical and Conceptual Frameworks in Human Factors*




*Note*. The phenotypical and genotypical theoretical frameworks are respectively central to the overlapping conceptual frameworks defined by the CREAM categorizations and the HFACS taxonomy.

Reason (1990) established a psychological framework of mental slips and lapses that can be attributed as causal factors in human error associated events. Le Coze (2015) noted that Reason and Mycielska (1982) initially developed this psychologically focused framework independent of Rasmussen's (1983) Skill-Rules-Knowledge cognitive framework, however, Rasmussen's influence is clearly evident in Reason's (1990) GEMS which would later provide a basis for Shappell and Wiegmann's (2000) HFACS taxonomy, representative of the genotype classification. The HFACS Unsafe Acts categories shown in Table 1 demonstrate the genotypical categorizations as described by Rasmussen (1983). The complete listing of HFACS categories is shown in Appendix C1.

**Table 1**

*HFACS Unsafe Acts of the Operator*

| Failure Level | Major Component | Causal Category | Failure Description |
|---|---|---|---|
| Unsafe Acts of the Operator | Errors | Skill Based Errors | Failure of automated sensory-motor patterns to meet the requirements of current conditions |
| | | Decision Errors | Failure of planned behavior to meet the requirements of current conditions |
| | | Perceptual Errors | Failure of the operator to correctly perceive current conditions |
| | Violations | Routine | Habitual disregard for recognized safe behaviors as routine behavioral patterns |
| | | Exceptional | Isolated departures from recognized safe behaviors not typical of routine behavioral patterns |

*Note.* Adapted from "The Human Factors Analysis and Classification System - HFACS" by S. Shappell and D. Wiegmann, 2000 (https://rosap.ntl.bts.gov/view/dot/21482). In the public domain.

In contrast, CREAM represents the phenotype perspective. Adopting a cognitive engineering view of humans in complex systems, Hollnagel (1998) established erroneous actions as a normal facet of human variability and hypothesized that, in order for the causal factors of the erroneous action to be understood, the physical attributes of the action must be established. Only then can an action be determined to be in error and further analyzed for causal influences (Hollnagel, 1998). Hollnagel's (1998) error modes from the CREAM system demonstrate phenotypical categorizations and are shown in Table 2. The complete listing of the CREAM categories is included in Appendix C2.

**Table 2**

*Hollnagel's Error Modes from the CREAM Classifications*

| Category | General Consequents | Specific Consequents | Definition/Explanation |
|---|---|---|---|
| Action at wrong time | Timing | Too early | An action started too early, before a signal was given or the required conditions had been established (Premature Action) |
| | | Too late | An action started too late (delayed Action) |
| | | Omission | An action that was not done at all (within the time interval allowed) |
| | Duration | Too long | An action that continued beyond the point where it should have stopped |
| | | Too Short | An action that was stopped before it should have |
| Action of wrong type | Force | Too little | Insufficient force |
| | | Too much | Surplus force |
| | Distance/ Magnitude | Too far | Movement taken too far |
| | | Too short | A movement not taken far enough |
| | Speed | Too fast | Action performed too quickly, with too much speed or finished too early |
| | | Too slow | Action performed too slowly, with too little speed or finished too late |
| | Direction | Wrong direction | Movement in the wrong direction, e.g., forwards instead of backwards or left instead of right |
| | | Wrong movement type (axis) | The wrong kind of movement, such as pulling a knob instead of turning it |
| Action at wrong object | Wrong Object | Neighbor | An object that is in physical proximity to the object that should have been used |
| | | Similar object | An object that is similar in appearance to the object that should have been used |
| | | Unrelated Object | An object that was used by mistake, even though it had no obvious relation to the object that should have been used |
| Action in wrong place | Sequence | Omission | An action that was not carried out This includes in particular the omission of the last action(s) of a series |
| | | Jump forwards | One or more actions in a sequence were skipped |
| | | Jump backwards | One or more earlier action that have been carried out is carried out again |
| | | Repetition | The previous action is repeated |
| | | Reversal | The order of two neighboring actions is reversed |
| | | Wrong Action | An extraneous or irrelevant action is carried out |

*Note*. Adapted from "Cognitive Reliability and Error Analysis Method: CREAM" by E. Hollnagel, 1998

The HFACS taxonomy and CREAM categories are, therefore, predominant representations of the two divergent viewpoints regarding the classification of human error. Each of the two conceptual frameworks, specifically at the category level closest to the described incident, exemplifies the theoretical principles that Le Coze (2015) describes.

Diagnostic analysis to establish potential causal factors is a critical step in reducing risk in complex systems. Although complexity and variability make catastrophic events normal in complex systems, the reduction of risk factors ensure that these events occur with less frequency (Perrow, 1999). As Perrow (1999) noted, death is, in fact, a normal part of human life, yet we rationally attempt to forestall it as long as is possible. Considering the importance of diagnostic analysis, it is therefore valuable to have an indication of the biases inherent in the reporting methodology. Although Kuhn (2018) demonstrated the effectiveness of machine learning and NLP methodology in identifying themes within ASRS reports, the published literature does not address the mindset or semantic styles of the incident reporter. Tversky and Kahneman (1981) demonstrated conclusively that the way in which a situation is described provides a decision frame for the subsequent decision maker's choices, thus the semantic themes within the ASRS reports may influence the impressions of subsequent evaluators and investigators. The use of aviation maintenance reports for this analysis provides service to an otherwise under-serviced segment of the aviation industry. As noted by Rashid et al. (2013) maintenance activities outpace flight hours by a factor of 12:1, while Mellema (2018) anecdotally notes that research on flight operations is ten-fold over research in aviation maintenance. While these figures do not constitute a direct comparison of opportunity and activity within the aviation industry sectors, they provide an indication of disparity in academic attention. The described research addressed that disparity by using aviation maintenance incident reports to evaluate the prevalent descriptive style of the reporting population in reference to phenotypical or genotypical error categorization methods.

**Summary**

The analysis of human error in accident investigation centers around two diverging themes. Some have viewed it as a psychological phenomenon (genotypes), while others consider it to be function of cognitive systems engineering (phenotypes). Shappell and Wiegmann’s (2000) HFACS taxonomy and Hollnagel’s (1998) CREAM are representative of these positions, genotypes, and phenotypes, respectively. Employing these classification methodologies in the described research provided insight into the way the described incidents are perceived by the individuals reporting the occurrences, and the applicability of the two schools of thought, as well. In using aviation maintenance reports for this analysis, the intent of the described research was to provide insight into aviation maintenance as an industry sector and to shed light on an underserviced portion of the aviation industry.

# Chapter III: Methodology

This chapter will detail the structure and methodology of the described research. Topics discussed in this chapter are research method selection, research population and sample, the data collection process, ethical considerations, the measurement instruments, and how the data was analyzed.

## Research Method Selection

The extant literature does not provide a basis for an analysis of aviation maintenance reports for thematic content. Furthermore, there is no support for the development of a human factors conceptual framework based on the way in which maintenance incidents are described by aviation maintenance technicians. The analysis of archival narratives to uncover previously unexplored themes and concepts necessitates a qualitative analysis performed in an exploratory manner.

The events recorded within the ASRS database represent human experiences recorded by individuals familiar with the events (NASA, 2023). In their program literature, NASA (2023) characterizes the primary concern of the ASRS system as "the quality of human performance in the National Airspace System" (p. 4). The impressions of the recording individuals are interpretative descriptions, and in order to address the research questions, these descriptions, as narrative passages of language, require analysis employing both qualitative and quantitative elements. These narrative reports contain vivid descriptions that have been nested in the real context of their occurrence (Miles et al., 2014). A qualitative analysis of these archival reports will allow the attitudes and impressions of the reporters to be considered in the analysis and is appropriate for the research purpose (Miles et al., 2014). The qualitative research method selected, a Delphi analysis, has been demonstrated as an effective method to establish a consensus of expert opinion on technical matters (Rowe & Wright, 1999).

Used as a method of data analysis, Artificial Intelligence (AI) algorithms and NLP methodologies provide a level of automation that allows for the examination of large corpora of documentation and the identification of recurrent topics, as defined by word groups, within the narrative texts. The volume of data contained within the ASRS database presents both opportunity and issue for any potential analysis. The large corpus allows for a broad sampling of error incident descriptions, but also makes direct human analysis impractical without a large research team of aviation maintenance SMEs. Blei et al. (2003) demonstrated the effectiveness of this NLP method of data reduction by using 100 years of scientific journal articles that had been scanned to a digital format and then analyzed using topic modeling algorithms. The methodology has also been employed in e-commerce and consumer entertainment (Blei et al., 2003). The described research employed a mixed-method strategy to reduce the selected large corpus of aviation maintenance reports to a much smaller listing of relevant thematic topics through the use of Latent Dirichlet Allocation (LDA) via machine learning algorithms. The resultant thematic topics were then qualitatively evaluated by human factors SMEs to determine if the topics were relevant to a phenotypical framework, a genotypical framework, or unrelated to either of the noted theoretical frameworks. This qualitative evaluation was intended to inform the level of need and potential content of a novel conceptual framework tailored to the prevalent themes within the selected ASRS incident report data set.

The exploratory review of the extant literature in the previous chapter revealed two theoretical schools of thought regarding the most effective way to categorize human action in the uncovering of contributing factors present in error incidents. The distinct differences between the two schools of thought identified by Le Coze (2015) and their associated categorization strategies indicated a need for a deeper understanding of the theoretical and practical applications of these points of view. The application of these methodologies to the practical activity of aviation maintenance requires a pragmatic world view and a nomothetic-deductive approach to evaluate these theoretical concepts in a physical environment.

The primary research focus observed within the published literature has been the analysis of the origins of the error event in terms of either phenotype or genotype. However, little attention has been given to the way in which the reporter has described the error event. The described qualitative research analyzed the bottom-up technician communication of human error events to establish whether these descriptions were aligned to the existing top-down analysis conceptual frameworks. The described research applied NLP to conduct an LDA analysis on a large corpus of aviation maintenance reports from the ASRS database. The LDA analysis then served as a data reduction method to identify recurrent semantic themes within the report narratives which, in turn, provided insight regarding the frame of reference under which the reporting technicians are operating. While there is no evidence in the extant literature documenting the combined use of topic modeling with a Delphi analysis method, the use of Delphi methodologies is well documented in business and healthcare (Keeney et al., 2002; Miles et al., 2014; Okoli & Pawlowski, 2004; Riessman, 2007).

Human error incidents are, by nature, unplanned, requiring an analysis of past occurrences for evaluation. In consideration of these factors, the described research was performed as a qualitative analysis of archival aviation maintenance incident reports recorded voluntarily in a narrative format by aviation maintenance personnel engaged in the support of 14 CFR Part 121 air carrier organizations. Thus, the selected research method was an archival study of preexisting data. This methodology allowed for the use of data collected from a broad span of the aviation industry in terms of both time and participating organizations with a minimal research investment. However, the nature of the secondary data collection also meant a loss of control in data collection methods and a lack of opportunity for follow-up or clarification (Vogt et al., 2012).

## Population/Sample

### *Population and Sampling Frame*

The relevant population for the described research was defined by the environment in which the research question applies. The research population encompassed all 14 CFR Part 121 air carrier maintenance incidents. Specifically, the selected research frame consisted of ASRS reports submitted by maintenance personnel engaged in the maintenance of aircraft operated by 14 CFR Part 121 air carriers in the United States from 2000 through 2022. The timeframe encompasses the most recent period of complete data and represents the most recent developments in aviation technology across a variety of industry operation conditions including the recent global pandemic. The complete set of ASRS maintenance reports from 14 CFR Part 121 organizations submitted during the noted timeframe was selected for the research sample.

This provided a comprehensive view of the maintenance environment of the commercial aviation industry throughout that time.

### *Sample Size*

An initial search of the ASRS database to the defined parameters yielded 5,929 reports from maintenance technicians, inspectors, and supervisors. Calculated at a 95% confidence interval with a 5% margin of error, this population required a minimum sample size of 361 reports in order to be adequately representative of the study population (Qualtrics, 2023). This calculation was generated using Qualtrics sample size calculator version 2023, Copyright© 2023, Qualtrics. Qualtrics and all other Qualtrics product or service names are registered trademarks or trademarks of Qualtrics, Provo, UT, USA (https://www.qualtrics.com).

The described research used the complete dataset as downloaded and filtered from the ASRS website. This is appropriate for the LDA topic modeling methodology and Gibbs sampling calculations as detailed in the Design and Procedures section of this chapter (Blei et al., 2003).

## Data Collection Process

ASRS reports are publicly available on the NASA website (https://asrs.arc.nasa.gov/search/database.html). The ASRS database includes reports across the complete spectrum of aviation activities. In order to ensure the data is reflective of the desired sample frame, the following data filters were employed when querying the ASRS database:

- Date of incident between January 2000 and December 2022

- Federal Aviation Regulations (FAR) Part 121 (i.e., Regularly scheduled air carriers)
- States Included: AK, AL, AR, AZ, CA, CO, CT, DC, DE, FL, GA, HI, IA, ID, IL, IN, KS, KY, LA, MA, MD, ME, MI, MN, MO, MS, MT, NC, ND, NE, NH, NJ, NM, NV, NY, OH, OK, OR, PA, RI, SC, SD, TN, TX, US, UT, VA, VT, WA, WI, WV, WY (i.e., all domestic United States locations)
- Reporter Function: Maintenance-Inspector, Instructor, Lead Technician, Other / Unknown, Parts / Stores Personnel, Quality Assurance / Audit, Technician, Trainee (i.e., all available maintenance functions)

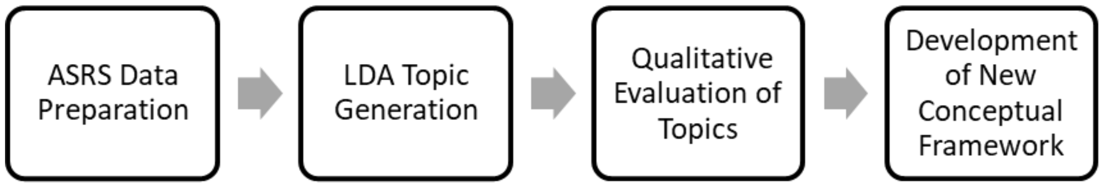
The generated report was downloaded and saved as a Microsoft Excel Macro-Enabled Worksheet.

### *Design and Procedures*

The described research required the data from the ASRS repository to be converted to a form suitable for use in LDA analysis. The LDA process was then used to reduce the corpus of ASRS reports to a set of word groupings, or *topics*, representing the prevalent themes in the ASRS report set. These topics were then qualitatively compared to the HFACS taxonomy and CREAM categories to evaluate any apparent similarities or differences that may be evident. These process milestones are shown in Figure 7.

**Figure 7**

*Process Milestones*




*Note*. These are the process milestones for topic generation and evaluation using LDA and Qualitative Analysis.

Achieving the objective of the described research required the evaluation of a large corpus of aviation maintenance reports for recurrent themes that indicate a tendency towards either a phenotypical or a genotypical frame of reference. It would be possible to accomplish this task through an SME review and referee process for categorization, but this endeavor would be both impractical and unreliable. Topic modeling uses a machine learning methodology and NLP to evaluate a large corpus of qualitative reports, identify recurrent topics, and establish distribution and association of these topics within the documents that comprise the corpus being evaluated. “Topic models are algorithms for discovering the main themes that pervade a large and otherwise unstructured collection of documents. Topic models can organize the collection according to the discovered themes” (Blei, 2012, p. 77). In essence, topic modeling employs mathematical methods to reduce large volumes of data down to a few vital sets of words that represent the themes within the documents. Thus, by using topic modeling techniques, the prevailing themes of a corpus of documents are uncovered. In the case of the described research, these sets of words represent the language, culture, and reporting environment of the recorded incidents. These prevailing themes, in the form of word groupings, were evaluated for their association to defined human factors categorization methods, specifically the HFACS taxonomy and CREAM categories.

LDA is a Bayesian analysis method useful in topic modeling. It was employed in this research to identify recurrent topics within the selected ASRS reports, to identify the distribution of words within the identified topics, and finally to identify the distribution of topics within the corpus of the ASRS reports. The output of the LDA process is a formal probabilistic model of the text corpus that describes a posterior probability based on prior frequency distributions of the words within topics and topics within documents (Blei et al., 2003).

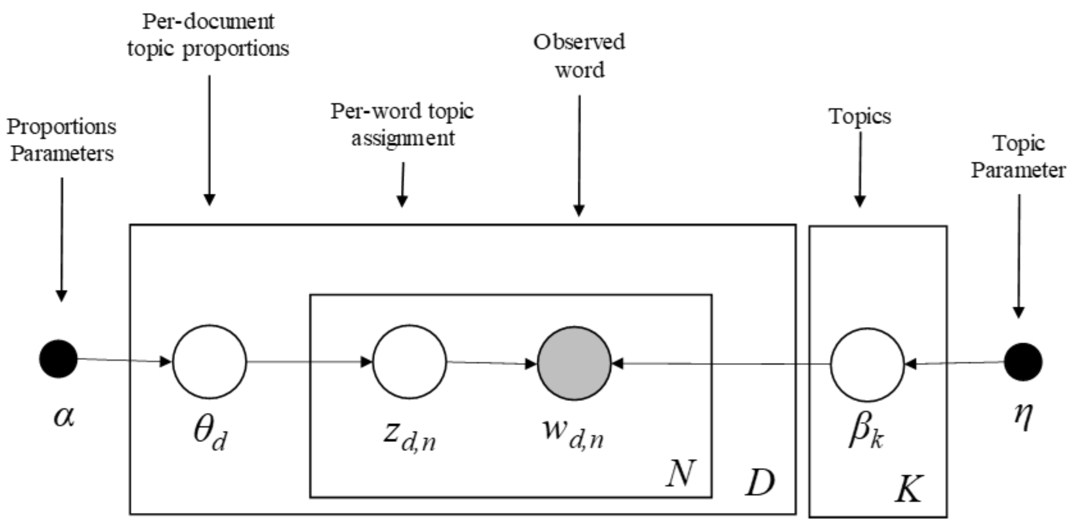
Blei et al. (2003) established a set of terms to define the elements required to describe the LDA process. The units defined are:

- Words: These are the basic units of discrete data contained indexed within a vocabulary as {1,…, V}. Words are represented using unit basis vectors that have a single component equal to one and all other components, equal to zero. Represented with superscript notation, the $\boldsymbol{v}$-th word within a defined vocabulary can be represented by a $\boldsymbol{V}$-vector $\boldsymbol{w}$ where $\boldsymbol{w}^{\boldsymbol{v}} = 1$ and $\boldsymbol{w}^{\boldsymbol{u}} = 0$ for $\boldsymbol{u} \neq \boldsymbol{v}$ (Blei et al., 2003).
- Document: A sequence of $\boldsymbol{N}$ number of words indicated by $\mathbf{w} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \boldsymbol{w}_N)$, where $\boldsymbol{w}_n$ is the $\boldsymbol{n}$-th word in the sequence (Blei et al., 2003).
- Corpus (pl. corpora): A collection of $\boldsymbol{M}$ documents indicated by $\boldsymbol{D} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_M\}$ (Blei et al., 2003).

Additionally, Blei et al. (2003) describe topics as groups of words associated by their probability of occurring together within a given document within the corpus. These notations are used in the expression of a probabilistic model that evaluates probabilities of words within topics and topics within documents for the selected corpus. Blei et al. (2003) express the LDA probabilities in a plate notation format as shown in Figure 8.

**Figure 8**

*LDA Plate Notation*




*Note.* Adapted from "Graphical Model Representation of the Smoothed LDA model," by D.M. Blei, A.Y. Ng, and M.I. Jordan, 2003, *Journal of Machine Learning Research 3* (https://jmlr.org/papers/v3/blei03a.html). Copyright 2003 David M. Blei, Andrew Y. Ng, and Michael I. Jordan.

The plate notation LDA model expresses collections in rectangles or plates and the circles represent node variables. In the plate notation of Figure 8, the collection of documents ***D*** within the corpus contains the collection of words ***N***. The only observed variable $\boldsymbol{w_{d,n}}$ is represented by a shaded node and indicates the occurrence of a word within the set ***N*** occurring within document group ***D***. Topics $\boldsymbol{\beta_{1:k}}$ on the ***K*** plate are distributed across the available vocabulary. The posterior parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$ are determined through analysis of the corpus and inform the topic distributions and the number of topics to be considered within the model. In summary, the LDA model considers the distribution of words within topics, the distribution of topics within documents, and the distribution of topics within the corpus of documents. The frequencies of the distributions are inferred to be an indicator of topic significance within the corpus of documents.

The research and analysis were performed in distinct phases. They were:

- Data collection and preparation: Data was collected from NASA ASRS public website using specified filtering parameters.
- Data preprocessing: Irrelevant data columns, words and punctuation were removed, and relevant words tokenized for LDA use.
- Data analysis: Topics were generated through LDA process and topic validation through SME review and concurrence.
- Data interpretation: Topics were evaluated for relevance to HFACS and CREAM categorizations and a novel conceptual framework was developed.

Each subsequent phase is dependent upon the treatment of the data in the previous phase. Data collection began with the download of maintenance incident reports from the ASRS website. The ASRS format does not require sanitization for identifying information as this is performed by the NASA SMEs prior to report publication in the ASRS database. The reports are also pre-coded by industry sector by the NASA SME reviewers.

### ***Sources of the Data***

The raw data used in the described research was in the form of aviation maintenance reports obtained from the NASA ASRS database. The ASRS database is a collection of voluntary and confidential safety reports provided by front line aviation personnel. The format established by NASA allows aviation personnel to describe, in rich contextual detail, events that might otherwise only be considered in terms of their statistical characteristics. The reports within the ASRS database are sanitized to remove identifying details and evaluated for descriptive event data by NASA SME personnel (NASA, 2023). ASRS reports utilized in the described research were descriptive of aviation maintenance events and submitted by aviation maintenance personnel. Although these types of event reports are "inherently limited by the fact that they capture very little of the process whereby the event occurred" (Rouse & Rouse, 1983 p. 542), the ASRS database provides a large and well documented database that can be useful for exploring correlations and developing hypotheses (Rouse & Rouse, 1983).

## **Ethical Consideration**

An application for Institutional Review Board (IRB) approval was submitted to the Embry-Riddle Aeronautical University (ERAU) IRB on the website (https://erau.edu/research/irb/application). The ERAU IRB determined that the research fell under the exempt category in accordance with 45 CFR 46.104, and noted it was determined that IRB review was not required. The detailed IRB approval is shown in Appendix A.

**Data Analysis Approach**

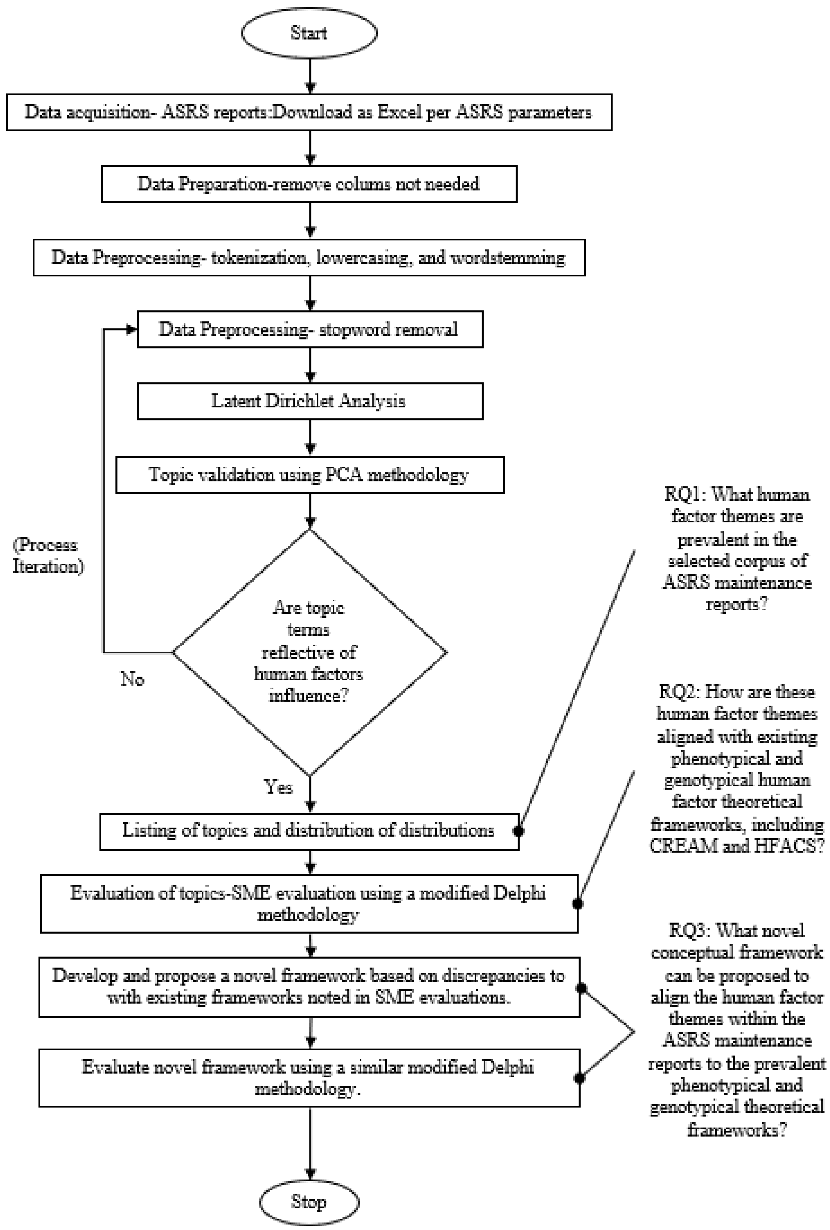
The data analysis in this described exploratory investigation required the reduction of a large corpus of text documents into a listing of relevant topic groupings to discern the prevalent themes of the corpus. This method was demonstrated by Blei et al. (2003) using 16,000 Associated Press documents from the 1992 Text Retrieval Conference to examine the prevalent themes in published news stories (Blei et al., 2003). Pereira et al. (2013) demonstrated the effectivity of topic modeling of traffic incident reports using two years of traffic accident records acquired from Singaporean official records. Kuhn (2018) also employed this LDA paradigm for a structured topic modeling methodology to identify latent topics and trends in aviation incident reports. Kuhn (2018) used quantitative ASRS data to provide context to the topic model outputs. (Kuhn, 2018). In a personal conversation, Kuhn (2021) remarked that he wished he had the time and resources to dig deeper into the relationship between the topics realized and the recognized human error modalities to gain a better understanding of the prevalent topics he had uncovered. The described methodology expanded the Blei et al. (2003) methodology and Kuhn’s (2018) work by investigating connections between the research corpus themes and existing recognized taxonomies and classification schemes, specifically Hollnagel’s (1998) CREAM error mode classifications and Shappell and Wiegmann’s (2000) HFACS taxonomy.

There are several computing platforms that have LDA algorithms available as verifiable code packages. The R for Windows (v4.3.2) is an open source publicly available platform (R Core Team, 2022) with an extensive library of available coding packages with referenceable traceability. This ensures that the selected research methodology will be able to be recreated for future validation or expansion. Downloads of the R base system and contributed coding packages are available through the Comprehensive R Archive Network (CRAN) at https://cran.r-project.org/ (R Core Team, 2022). A listing of employed coding packages for the described research is presented in Appendix C.

The data from the ASRS database used in this analysis was stored as plain text in discrete data fields. This formatting necessitated a conversion process that removed irrelevant information from the incident reports and allowed the R packages to consider each word individually, and place words within topics based on their association with other words and the probability that words will occur in groups within reports. The complete process from data download to topic grouping is represented as a process flow diagram in Figure 9.

**Figure 9**

*Data Treatment and Analysis Process Flow*




*Note*. The described research process flow begins with the ASRS data download and concludes after the proposal of a novel human factor conceptual framework. The research questions are addressed in the final three steps in the process. This flowchart is an expansion of the process milestones expressed in Figure 7.

The first two steps of the process acquired the data from the ASRS database. In the first step the data downloaded from the NASA ASRS website was filtered before acquisition. Pre-download filtering included four factors: reporting date, FAA classification of aircraft operational category, United States location by state, and maintenance role of the reporting individual. These filters were necessary because the ASRS database contains a wide variety of reports from all aviation disciplines and operational categories in a wide variety of US domestic and international locations. These additional categories range far beyond the delimitations of the described research. The data from NASA is provided in a CSV format compatible with Microsoft Excel. The data acquired in the first step of the process contained columns of information that, although useful in defining the scope of the research in accordance with the stated delimitations, are not useful in determining topics within the LDA process. In the second step of the defined process flow, the R package *tidyverse* was directed to ignore the irrelevant columns, leaving only the free text fields in the report narrative. This step ensured that the R packages used in the LDA processing are limited to the expressions and verbiage of the reporting personnel and the report synopsis provided by the NASA selected SMEs.

The next four steps of the described process were performed using the R for Windows GUI version 4.3.2 (R Core Team, 2022) to run a variety of preprocessing algorithms in the R computing language. The R GUI and the selected algorithms are open access, free software provided without warranty, but include specific referential instructions and use reference requirements. In this process, words were tokenized to individual items by removing irrelevant spacing and punctuation and all letters were lowercased for commonality. The third preprocessing step involves assigning common values to word forms associated with the same root. In this process, plurals, past tense, and other variations were assigned to their root word so that an accurate probability of the context can be established. In the final preprocessing step, *stopwords* were removed from the population dictionary. In the English language these include irrelevant common articles such as *a*, *an*, and *the*. This step is the point of iteration for the evaluation of topics because additional *stopwords* were added based on the specifics of the document corpus. For instance, the ASRS evaluators substitute the letters *zzzz* or *xxxx* for any identifying information contained within the document narrative to preserve the anonymity of the report. This is likely done with sufficient frequency to affect the topic content and distribution. Therefore, this occurrence of irrelevant letters was added to the stopword population to prevent them from being included in the topic distribution calculations.

The LDA process follows the data preprocessing in the R algorithm steps. The LDA process in R employs an iterative cycle to evaluate the probability of co-occurrence of the tokenized words within the documents comprising the entire corpus used for analysis. A Gibbs sampling algorithm uses a Markov chain Monte Carlo method to establish the posterior topic parameter $\boldsymbol{\eta}$, specifying the theoretical ideal number of topics to be generated. The process begins with the random assignment of words to topics. The tokenized words are treated in what Miyamoto et al. (2022) term a "Bag of Words" methodology. All the words in the dictionary of the corpus are considered equally with no weight or preference for associations to other words nor of the order in which they occur—it is a random bag of words. Once these arbitrary topics are assigned, the algorithm evaluates the probabilities of the words in the topics occurring together based on their occurrence within the documents in the corpus. The highest probabilities are maintained within the topics and the others returned to the bag of words. This begins the iterative process. Words are again randomly assigned to the existing topics, probabilities evaluated, and either maintained within the topic or returned to the bag of words based on the magnitude of the probability. This process continues until the point where no significant changes in the topic content occur. This is termed *convergence* and represents the end of the iterative process (Blei et al., 2003; Blei, 2012; Kuhn, 2018; Miyamoto et al., 2022).

The resultant topics, as identified by the NLP algorithms, were not associated with any meaning, as their identification occurs strictly as a matter of mathematical probability. Each topic needed to be qualitatively evaluated for relevance to either a phenotypical or a genotypical frame of reference. Although the ASRS reports used in the analysis were filtered based on their ASRS categorization for relevance to aviation maintenance, the possibility remained that many of the identified topics would fail to display an association to the desired error taxonomies. The evaluation of the generated topics was accomplished using an SME evaluation process as described in the *Qualitative Data Analysis Process* section in this chapter.

### ***Reliability Assessment Method***

The reliability of the described research is largely dependent upon the coding process of the maintenance reports being used. NASA employs SMEs to review and categorize reports upon receipt (NASA, 2023). The initial NASA SME coding of the reports determines the efficacy of the filtering when reports are downloaded for analysis. The report coding within the ASRS system was assumed to be consistent and accurate for the purposes of the described research. The integrity of the ASRS data was maintained in the condition in which it was obtained up to the point of interpretation of the analysis.

The use of machine learning algorithms for NLP increases reliability in data processing and analysis by reducing interpretive influence and human error opportunities throughout the research. Each algorithm package is well documented with clear usage instructions and well-defined user parameters. The described methodology can demonstrate repeatability through application of the selected code packages in the method described, using similar datasets from the ASRS database.

Topic coherence, or the degree to which a definitive concept can be derived from a set of words in a topic, can be an effective estimator of reliability in the LDA process. Coherence can be enhanced by the removal of words that occur either too frequently or too infrequently to provide meaning to the topics. Infrequent and unique words are unlikely to appear in modeled topics due to probability calculations (Röder et al., 2015). In subject-specific topic modeling, as with this described research, subject-specific words that appear too frequently to be meaningful across multiple topics can be added to the *stopwords* list to increase topic coherence (Sarica & Luo, 2021). Adding *stopwords* to the existing list of common words within the *tidytext* (Silge & Robinson, 2016) algorithm involved a process of preliminary analysis, SME consultation, and iterative modeling. SME's participating in this consultation were HF SMEs with a deep knowledge of HF concepts.

### *Validity Assessment Method*

The data obtained from the ASRS database were assumed to be valid for the described research and delimited in scope to avoid inaccurate generalization. The NASA dual SME review process and pre-download filtering ensure that the research data is representative of a common regulatory environment as well as general language and cultural norms (NASA, 2023). Tanguy et al. (2016) and Kuhn (2018) demonstrated the use of machine learning and NLP for the classification of ASRS incident reports using a topic modeling methodology. While the use of LDA through the application of NLP is not a novel concept, the described research further validates this methodology by employing a SME consensus process to examine the context of the words within the topics realized in the LDA process. This was instrumental in furthering the understanding of the underlying themes within the corpus, and providing validation to the LDA methodology, which also directly addresses the research purpose.

In describing the limitations of the Delphi methodology, both Riessman (2007) and Keeney et al. (2006) noted that the technique, as with other types of SME evaluation, is not capable of uncovering scientific facts. The Delphi methodology is an effective method of eliciting a concurrence of opinions from SMEs experienced in the field of research interest (Keeney et al. 2006; Riessman, 2007).

Validity of the SME opinions is dependent on the skills and experience of the SMEs who have voluntarily agreed to participate in the research. These individuals were required have fluency in the English language, as spoken in the United States, theoretical familiarity in Human Factors, and practical experience in aviation maintenance. Familiarity with Human Factors and the conceptual frameworks used in the described research were established by the volunteer’s professional and academic experiences and familiarity training issued before the evaluations took place. The SME volunteers were required to have at least two semesters of dedicated human factors courses at either a graduate or undergraduate level or commensurate professional exposure to the subject. Pre-evaluation training for HFACS and CREAM familiarity was drawn from the appropriate publications noted in the Literature Review section of this dissertation. The pre-evaluation training is included in Appendix E.

The FAA issues two different mechanic certifications. One certification is for airframe technicians and the other applies for power plant technicians. Individuals holding both certifications are referred to as A&P certified (FAA, 2023). There is no additional FAA designation for a master technician or other higher classification of qualification. SME evaluators who volunteered for this research were required to be current or former A&P certificated technicians or have practical experience commensurate with the stated FAA qualifications. This means they were at least eighteen years of age, able to read, write and understand English language, had at least 30 months of practical experience performing the duties appropriate to both the airframe and powerplant ratings, and previously demonstrated proficiency through a series of oral, written, and practical tests (FAA, 2023).

### *Data Analysis Process*

The data analysis process began with the output from the LDA calculations in the R packages. The first stated research question asked what human factor themes are prevalent in the selected corpus of documents. The output of the LDA process consisted of groups of words with a listing of their probabilities of occurrence.

The second research question asked if the topics generated through the LDA process could be associated with either a phenotypical or a genotypical frame of reference. A qualitative analysis of each topic, performed by the selected SMEs, was required in order to determine what human factor themes, if any, are prevalent within the LDA generated topics. In plain language, the SME evaluators were asked to provide their opinion on what subject was prevalent within the words of the topic in question. The SME responses were collected, communicated, and refined through a modified Delphi process. The employed modified Delphi process is discussed in detail in the *Qualitative Data Analysis Process* section.

The third research question concerned the proposal of a novel conceptual framework to describe the topics realized in the topic modeling process, and to provide HF categories aligned with the prevalent themes in these topic word groups, if evidence of need existed within the SME evaluations. This novel framework is intended to improve the relevance of a conceptual HF framework to the themes within the aviation maintenance reports.

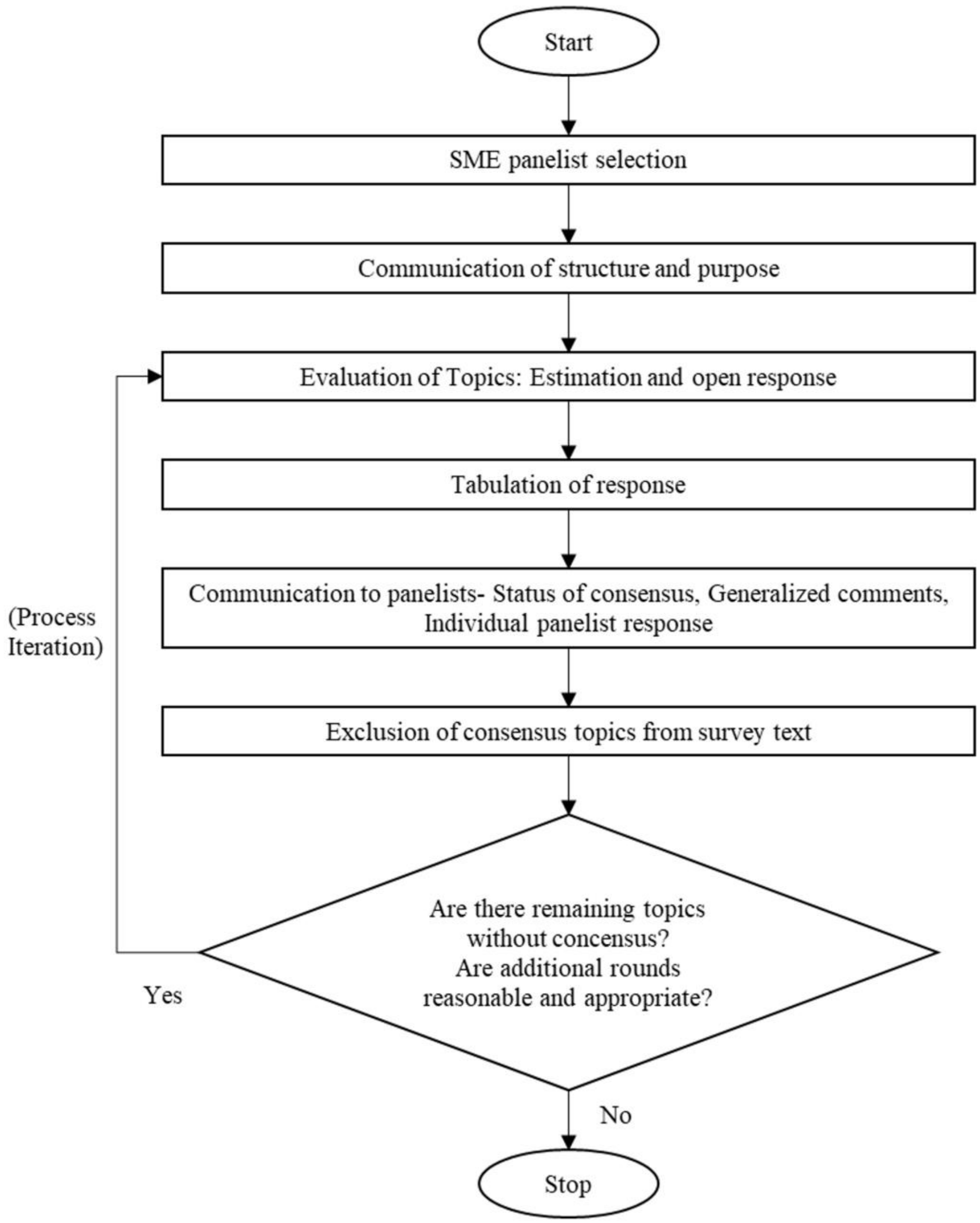
The LDA process of determining the prevalent topics within the document corpus based on probability of occurrence is quantitative, objective, and repeatable. The vast corpus of documents was distilled into a finite list of topics that were then evaluated to determine how well they align with the two selected conceptual frameworks. This process, as it involves interpretation of human communication and language, is subjective and qualitative. The fields of human factors and of aviation maintenance are specialized disciplines that require a deep familiarity with the subject matter, acquired through professional and academic experience. The evaluator or evaluators needed to be deeply conversant with both topics to adequately identify the associations between the topic word groups and the selected conceptual frameworks.

### ***Qualitative Data Analysis Process***

The described research employed a modified Delphi analysis methodology to develop a consensus of expert opinion regarding the alignment of the prevalent themes in the selected ASRS reports to the conceptual examples of phenotypical and genotypical human factor categorizations. Although expert opinion ranks low in the Strength of Evidence pyramid (Evans, 2003) it is deemed to be appropriate when other statistical analysis methods or case studies fail to address the purpose of the research (Rowe & Wright, 1999). The Delphi methodology is a recursive qualitative analysis process intended to elicit a consensus of expert opinion from a panel of subject matter experts (Evans, 2003). Rowe and Wright (1999) noted four characteristics that define a procedure as *Delphi*. In order to be considered as a true or modified Delphi process the methodology should ensure anonymity of response among panelists. There should be iterative survey rounds whereby panelists can refine their responses. The process should contain controlled feedback that will allow the panelists to consider the viewpoints and rationale of the other respondents while the aforenoted quasi-anonymity is maintained, and finally there should be a statistical aggregation of responses to determine consensus (Rowe & Wright, 1999). Keeney et al. (2006) noted that it is important to the validity and reliability of the study to establish the method and criteria for the index of consensus before the analysis begins. Figure 10 contains a generalized process flow that is applicable to both a true Delphi analysis and a modified Delphi analysis as used in this research. The steps detailed in this process flow were described by Miles et al. (2014), Riessman (2007), Okoli and Pawlowski (2006), Keeney et al. (2006), and Evans (2003), as well as Rowe and Wright (1999).

**Figure 10**

*Generalized Delphi Analysis Process*




*Note*. The Delphi qualitative analysis process is comprised of an iterative evaluation and communication flow. Communication of textual responses to panelists for each round are anonymized, except for the panelist's own response to the previous round. This process flow is a generalized representation of the full and modified Delphi processes.

As shown in Figure 10, the process begins with the selection of the SME panelists. There are no universal criteria for the number of expert panelists in a Delphi study. Okoli and Pawlowski (2004) noted that many researchers prefer to empanel eight to twelve experts, while Keeney et al. (2006) determined that the number of experts to be empaneled depended largely on the commonsense decisions and the practical logistics of the researcher. The described research solicited the participation of eight SMEs in order to ensure a minimum of five complete response sets at a 70% participation level. As detailed in the earlier *Validity Assessment Method* section of this dissertation, panelists were experienced aviation maintenance technicians with a deep familiarity of human factor topics.

The selected potential SME panelists were sent an introductory letter to request their participation in the Delphi analysis. It was important that SME panelists were informed as to the purpose and structure of the study before agreeing to participate. The panelists were also informed of their role in the research, the approximate time commitment for each round of evaluation, and the number of rounds. This letter also informed the participants of the quasi-anonymous nature of the research process. Responses were not identified by respondent names and only the facilitator knew how each SME responded and what rank was given for each topic. Communicating these items informed the potential panelists of what they were signing up for and was intended to have a positive effect on the response percentages (Keeney et al., 2006).

Following the communication of structure and purpose, each SME who agreed to participate in the analysis received a link to a personalized survey in the Qualtrics online research tool for the first round of evaluations. In each round of evaluation, the panelists received prompts for quantitative and qualitative Delphi evaluation sheets for each topic to each conceptual framework for which an evaluation consensus had not been achieved.

The evaluation sheet contained:

1. Instructions to evaluate the noted topic in relation to the human factor terms shown.
2. A listing of words in the topic to be evaluated. Words were ordered by probability but without probability values included.
3. The statement “In your opinion, how accurately do these topic words represent the human factor terms below?” Please indicate a value based on your opinion.
4. A qualitative scale stating, “Does not Reflect,” “Moderately Reflects,” and “Strongly Reflects.” A 9-point quantitative scale was included below the qualitative scale with 1-3 corresponding to the first category, 1-6 corresponding to the middle category, and 7-9 corresponding to the “Strongly Reflects” category.
5. A space for an open response to the request “Please provide support for your evaluation. Why did you select this value?”

The panelists were given the list of words in the topic set to be evaluated and a list of terms from either the HFACS Unsafe Acts categories or the CREAM Error Modes. The panelists were asked to evaluate how well the words in the topic set reflected the stated human factor categories. An open response question was also included, asking the panelist to explain how they came to the rating decision for that topic. Research by Lange et al. (2020) showed that a 9-point scale was an effective evaluation tool for SME evaluation of assertions in a Delphi research study. Lange et al. (2020) found that the 9-point scalar evaluation, divided into three parent categories increased test-retest reliability versus 3- or 5-point methodologies in a Delphi analysis of the identification of global treatment goals for total knee arthroplasty. A similar scalar evaluation was developed for this research. An example of the Delphi Evaluation Sheet used in this study is presented in Appendix B as a qualitative data collection device.

Upon receipt of panelists' responses, the facilitator performed two tabulation tasks. The facilitator composed a summarization of the responses capturing salient commentary for ratings that positively impact the evaluation and for ratings that negatively impact the valuation. These opposing summaries gave the panelists an impression of the other panelists' points of view while providing panelist confidentiality, and possibly reducing response time by limiting the amount of information that each panelist would need to review before providing the next round of evaluation. The preservation of confidentiality mitigated the risk of groupthink, or the tendency for panelists to appeal to the authority of other panel experts (Okoli & Pawlowski, 2004). These techniques were recommended as best practices by Keeney et al. (2006) after a review and analysis of an unspecified number of nursing studies conducted over a period of 10 years. The second task of the facilitator was to establish an inter-quartile range (IQR) for each set of scalar evaluations. The IQR was used to evaluate the measure to which the panelists reached a consensus on each evaluation. Evaluations with an IQR $\leq 1$ were declared to have reached a consensus and were not included in further rounds of evaluation. Birko et al. (2015) evaluated nine separate indices of consensus in Delphi studies and determined that the use of an interquartile range to evaluate consensus was the least sensitive to changes in the number of survey questions and to the effect of the group conformity index. The group conformity index is described as the tendency of panelists to fall in line with group responses without rationale or justification (Birko et al., 2015). The IQR displayed sensitivity to the number of panelists within the survey when evaluated for surveys with between 6 and 50 panelists. However, the number of participating panelists for this research was at the extreme low end of that variation analysis.

After a tabulation of responses, the facilitator removed any topics that met a standard of consensus (IQR $\leq$ 1) from the survey text. This reduced the length of survey communication for the next round and allowed the panelists to focus on the remaining topics. Establishment of the consensus index and criteria for consensus apriori increases the reliability of the analysis by mitigating the risk that an index would be chosen tailored to the responses after the fact (Keeney et al., 2006).

Following the removal of topics where consensus had been achieved, the facilitator was required to decide to either communicate the results of the previous round to the panelists, or to end the survey. After the first round of evaluations in the described research, since the majority of topics had not been agreed upon by the SME panelists, the facilitator made the decision to continue the study. The research plan required the facilitator to continue the iterative process if there were topics that remained for which a consensus has not been reached and the predetermined limit for survey rounds had not been attained. Keeney et al. (2006) suggested three to four rounds as optimal for a modified Delphi analysis, noting that additional rounds could make the time scale for the process impractical and lead to a sense of survey fatigue among the panelists. If no topics remained or the maximum number of rounds had been reached, the facilitator was to end the survey. Following the second round of SME evaluations, an analysis of participant responses indicated an inability to achieve sufficient consensus among the panelists. Potential response bias and declarations of participant fatigue were evident in the quantitative and qualitative responses provided. The study facilitator made the decision to terminate the study after the second round due to a lack of effectiveness of the research method. These qualitative evaluations in the modified Delphi process were intended to indicate an alignment with the established schools of thought noted by Le Coze (2015) or a deviation from these recognized paradigms, requiring a novel viewpoint and associated categorization system.

## Summary

This chapter presents a detailed research plan. The described research plan utilized NLP and machine learning methodology to analyze a large corpus of aviation maintenance incident reports. This analysis was intended to uncover the prevalent themes in the corpus of documents which could then be evaluated for their association to recognized phenotypical and genotypical human error categorization methodologies.

# Chapter IV: Results

This chapter discusses the results of the exploratory research developed and discussed in the previous chapter. The research was conducted in distinct phases intended to provide insight to the three research questions. In the first phase, a series of machine learning algorithms were employed in a topic modeling methodology to reduce a set of ASRS database reports to a collection of topic word groups. The intent of developing the word groups is to develop a sense of the prevalent topics contained within the data set of aviation maintenance incident reports. Secondly, the word topic groups were presented to a group of aviation maintenance SME's, along with two sets of categories from Human Factor frameworks to evaluate how the concepts from the ASRS incident reports aligned with the concepts of the Human Factor categories. Finally, a novel Human Factor framework was developed based on the themes within the topic word groups and the aviation maintenance SME comments to improve upon the alignment of the concepts from the ASRS database to extant HF conceptual frameworks.

## ASRS Data Demographics Results

The NASA process used for ASRS report submissions facilitates demographic data points that can be useful in inferring or extrapolating characteristics about the underlying population of the dataset. This includes the acceptance of the tool by the maintenance community, the experience levels of the technicians using the system, and prevalent human factor contributors.

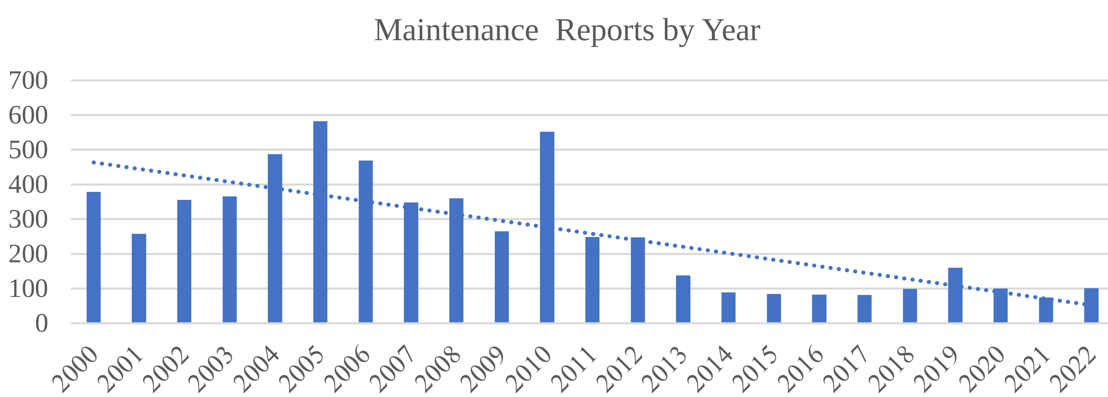
### *Report Submissions by Year*

The ASRS database search, as described in Chapter I, generated a set of 5,929 aviation maintenance incident reports. Each report is identified by a unique Aviation

Confidential Number (ACN) assigned by ASRS personnel and used to track the progress of the report while maintaining confidentiality of submission (NASA, 2023). Although the report filtering included a designation of all US states, the airport designation and state are anonymized to ensure reporter confidentiality. The number of reports submitted by year, as shown in Figure 11, ranges from 74 to 582 with an average of 258 reports per year.

**Figure 11**

*ASRS Maintenance Report Submissions from 2000 through 2022*




*Note*. The average number of maintenance reports submitted for the years 2000-2022 is 258. The trendline shows a general decline in maintenance reports submitted over the selected time period. The increase in reports during 2010 appears to be an anomaly. Data from ASRS.

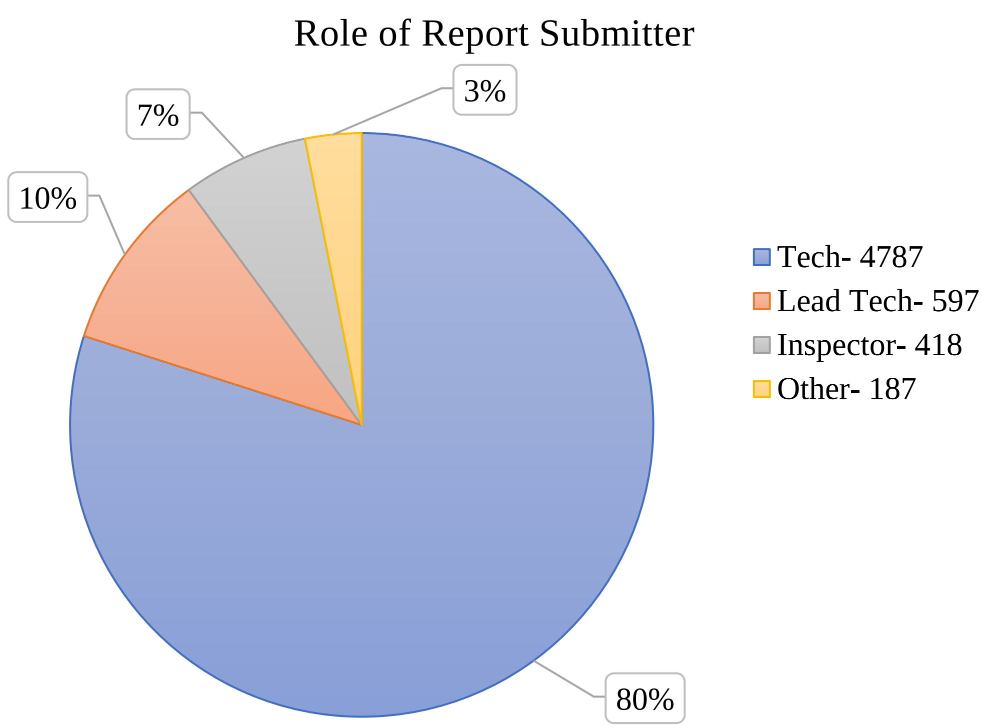
The declining trend in submission of maintenance reports is contrary to the overall report activity reported by NASA (2023) in the ASRS program briefing. The declining trend in maintenance report submittals could represent a number of factors not associated with the overall performance of the ASRS program. This trend could be attributable to a number of factors specifically associated with the maintenance segment of the commercial aviation industry. These factors could include organizational initiatives for internal reporting, shifts in safety culture within maintenance organizations, or general industry sentiment.

### ***Report Submitters***

The ASRS reports in the selected dataset were submitted by individuals in a variety of maintenance roles in FAR Part 121 organizations. Figure 12 details the role of the report submitter as identified by the individual submitting the report. ASRS maintenance reports are, in the great majority, submitted by maintenance technicians. In addition to the reports submitted by maintenance technicians, lead technicians, and inspectors, the reports submitted by the category identified as *other* include maintenance trainees, stores personnel, aircraft crew, and other aviation personnel.

**Figure 12**

*Role of the Report Submitter*




*Note*. The ASRS data collection device allows for multiple entries, so the total is greater than the number of reports in the data set. Data from ASRS.

### *NASA SME Input*

The NASA SMEs reviewing the reports add additional details about the reported incidents based on their experience and professional expertise (NASA, 2023). The NASA SME qualifications are listed in Appendix D. The added information provides insight into the problems described by the reporter and the potential human factor contributors that may have facilitated the occurrence. Table 3 shows the primary problem as identified by the ASRS reviewing SME for the selected data set.

**Table 3**

*Assessment of ASRS Primary Problem by NASA Reviewer*

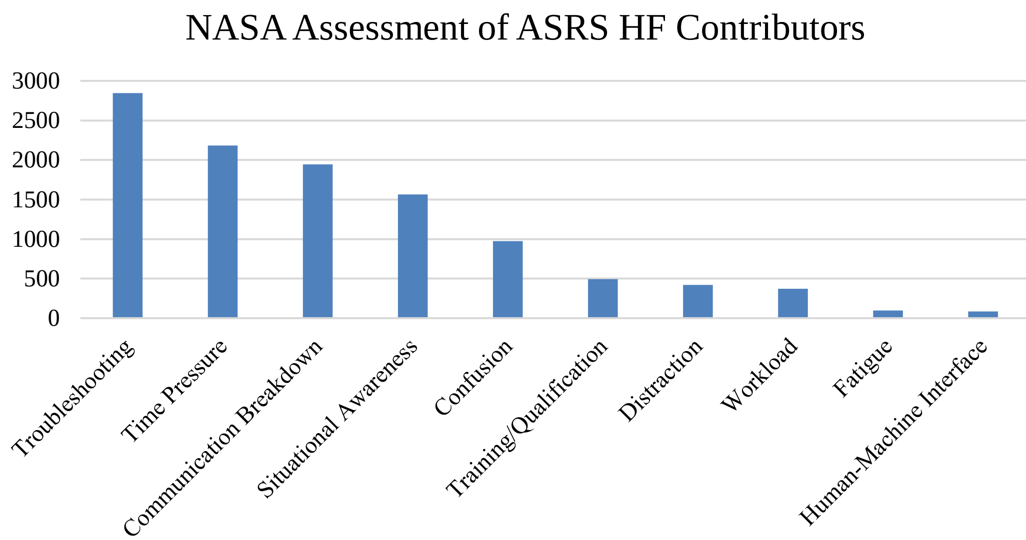
| Problem | Frequency |
|---|---|
| Human Factors | 3961 |
| Company Policy | 450 |
| Chart Or Publication | 450 |
| Procedure | 323 |
| Aircraft | 276 |
| Manuals | 129 |
| Ambiguous | 73 |
| Incorrect / Not Installed / Unavailable Part | 72 |
| Equipment / Tooling | 58 |
| MEL | 54 |
| Environment - Non-Weather Related | 19 |
| Airport | 15 |
| Logbook Entry | 10 |
| Weather | 5 |
| Staffing | 4 |
| Software and Automation | 3 |
| ATC Equipment / Nav Facility / Buildings | 1 |

Note. Human factor contributors are the most commonly identified primary problem. Subsequent problems consist of policies, procedures, and other directive types. Data from ASRS.

In addition to the assessment of primary problems, as shown in Table 3, the NASA SME evaluating each report also provides an assessment of human factor contributors that may have influenced the reported incident. These evaluations are shown in Figure 13. The evaluations of primary problems and human factor contributors provided by the NASA SME evaluators provide some insight regarding the content of the reports within the selected dataset. The overwhelming majority of the reports submitted by maintenance personnel in FAR Part 121 organizations involve human factor incidents related to troubleshooting, time pressure, and breakdowns in communication.

**Figure 13**

*Count of HF Contributors from NASA ASRS Evaluators*



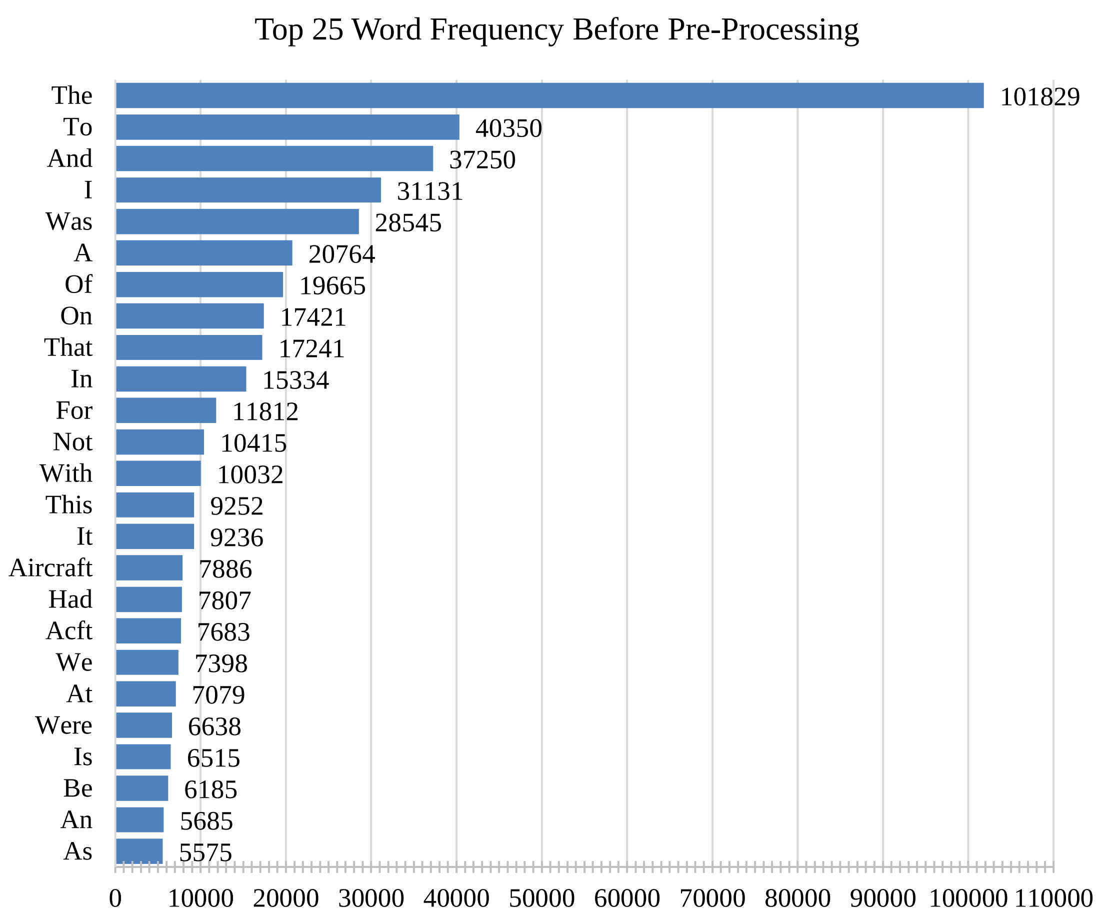

Note. In the majority of cases the ASRS evaluators listed more than one HF contributing factor. Data from ASRS.

## Descriptive Statistics

NLP methodologies rely on preprocessing of data to enhance relevance to the desired topic of analysis. The data preprocessing addresses the inconsistent use of abbreviations, removes words that lack contextual meaning, and allows the analysis algorithms to see the words as mathematical tokens (Miyamoto et al., 2022). Figure 14 shows the top 25 words within the selected ASRS dataset, as shown by their frequency of occurrence.

**Figure 14**

*Top 25 Words in the ASRS Dataset Before Preprocessing and Topic Enhancement*




*Note*. The top 25 words within the selected dataset contain a high percentage of article words lacking in relevance, particularly without any context.

As would be expected with any English language body of text, article words *a*, *an*, and *the* are heavily represented within the text of the reports. The only words represented that would indicate an association with aviation maintenance are *aircraft* and the short-hand abbreviation *acft*. However, finding the term aircraft within an aviation related report set would not be unexpected and without context contains no more relevance than the other words noted in Figure 14.

**Refinement of the Dataset to Enhance Topic Relevance**

The coding routines used to implement the R packages referenced in Appendix C were developed by a contracted programmer with experience in topic modeling. Using the RStudio interface, an iterative process was employed to generate topic word groups, analyze the terms, and establish stop-words, word substitutions, and additional word inclusions so that the relevance to human factor concepts could be enhanced. This process, including step-by-step topic word groups, was presented to three human factor SMEs to obtain their concurrence and to show that the strategy employed was reasonable and effective in enhancing the value of the topic word groupings. The first step in this iterative improvement process was to obtain an initial baseline of topic word groupings using the native word stemming and stop word sets within the published algorithms. This set of topic word groupings, consisting of 15 topics with the top 10 words included, is shown in Table 4.

**Table 4**

*Baseline Topic Word Groupings*

| Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 | Topic 6 | Topic 7 | Topic 8 |
|---|---|---|---|---|---|---|---|
| damag | sign | called | test | amm | start | aft | valv |
| document | night | mech | fault | step | hangar | log | main |
| engineer | inspector | informed | oper | wrong | move | page | lock |
| accomplish | block | approx | defer | incorrect | ground | writeup | assembl |
| report | question | arrived | indic | note | taxi | forward | land |
| carrier | veri | gnd | dispatch | data | stop | access | lower |
| refer | qualiti | ctlr | fail | error | power | bleed | torqu |
| limit | feel | proceeded | deferr | read | ramp | placard | jack |
| batteri | faa | capt | troubleshoot | amt | personnel | broken | mount |
| issu | didnt | engs | pilot | type | short | entri | posit |

| Topic 9 | Topic 10 | Topic 11 | Topic 12 | Topic 13 | Topic 14 | Topic 15 |
|---|---|---|---|---|---|---|
| installation | sys | engineering | fire | supervisor | leak | rptr |
| lndg | test | required | bottl | day | tool | stated |
| mech | pressure | damage | revers | manag | hydraul | revealed |
| assembli | pos | accomplished | verifi | chang | run | conversation |
| change | hyd | supvr | box | becaus | blade | callback |
| mechs | ops | signed | tag | assign | disconnect | reporter |
| missing | proc | ref | stock | befor | low | acn |
| changed | chked | items | posit | anoth | drain | required |
| signed | pwr | procs | rig | pressur | drive | pressure |
| assigned | performed | limits | cabl | arriv | clean | shop |

*Note*. The topic word groupings contain many word fragments as a result of the NLP algorithms. There are also a large number of word abbreviations resulting from the informal reporting conventions used by the incident reporters.

Increasing the relevance of the word groups required a change in word stemming strategy as well as a substitution for clarification of the words that appeared to be reporting shorthand terms within the reports within the data set. The stemming routine was removed from the algorithm and a word substitution routine was inserted to replace shorthand abbreviations with the appropriate words and to replace out of tense (past or plural) words that were previously stemmed with their complete root. The word replacement list is in Appendix C for reference. This strategy was refined through several iterations until there were no other replacements to be made. The resulting second set of interim topic lists is shown in Table 5.

**Table 5**

*Interim Topic Word Listings After Word Replacement*

| Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 | Topic 6 | Topic 7 | Topic 8 |
|---|---|---|---|---|---|---|---|
| amm | technician | reporter | hangar | leak | wrong | test | manager |
| training | step | revealed | ground | hydraulic | data | fault | management |
| amt | night | conversation | ramp | noted | tag | operations | call |
| issue | complete | callback | taxi | low | incorrect | operational | report |
| document | steps | lock | stop | drain | shop | message | situation |
| information | checks | closed | move | water | error | troubleshooting | controller |
| accomplish | occurred | acn | started | changed | received | warning | reported |
| personnel | event | access | moved | leaking | changed | normal | issue |
| issued | technicians | inboard | short | normal | stores | pulled | operations |
| proper | previous | properly | heard | run | serviceable | reset | chief |

| Topic 9 | Topic 10 | Topic 11 | Topic 12 | Topic 13 | Topic 14 | Topic 15 |
|---|---|---|---|---|---|---|
| didnt | power | landing | repair | aft | log | inspector |
| proceeded | bleed | main | limits | forward | page | routine |
| remove | closed | torque | damaged | fire | writeup | release |
| started | start | upper | srm | missing | discrepancy | sign |
| tool | run | install | shop | compartment | deferred | faa |
| notified | heat | lower | inch | pulled | deferral | quality |
| removal | trim | hardware | approximately | bay | action | signoff |
| box | engines | remove | repaired | emergency | entry | block |
| finished | duct | loose | days | floor | items | cards |
| blade | close | aileron | fuselage | access | inoperable | lhand |

*Note*. Some shorthand abbreviations could not be addressed with substitution because they represent multi-word terms. Substitution would alter the word probability of occurrence.

The next step in the process to increase the relevance of the topic word groups was to customize the stop word listing. *Stopwords* are eliminated from the analysis because they are determined to be irrelevant to the desired context of the analysis. New words were added to the existing *stopwords* list, which is native to the NLP algorithm. The added words consist primarily of aircraft part names, such as engine, nut, bolt, and so on, as well as terms inserted by the NASA SMEs to anonymize the reports by replacing any identifying reference within the narrative with null characters like a string of Z’s or

X's. The *stopwords* list was developed through repetitive iteration and analysis of the top 20 words in each of the 15 topic word groups. In each iteration, as word probability beta values changed and new irrelevant words appeared, they were added to the new stop word listing before the next iteration was performed. This analysis and adjustment were performed manually through the use of a data spreadsheet that was updated and reloaded into the algorithm for each iteration. The complete listing of *stopwords* includes the standard English set of *stopwords* from the NLP algorithm and the added *stopwords* from the iterative analysis. This *stopword* listing is included in Appendix C. The resulting word topic grouping after this iterative modification is shown in Table 6.

**Table 6**

*Interim Word Topic Listing After Stopword Modification*

| Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 | Topic 6 | Topic 7 | Topic 8 |
|---|---|---|---|---|---|---|---|
| power | night | technician | repair | call | landing | aft | issue |
| fault | didnt | inspector | limits | controller | main | forward | training |
| troubleshooting | proceeded | sign | damaged | operations | changed | emergency | manager |
| message | started | routine | authorized | captain | missing | closed | management |
| normal | morning | cards | document | received | incident | access | amt |
| bleed | tool | release | shop | pulled | discovered | secured | event |
| indication | finished | quality | srm | contacted | lhand | bay | personnel |
| closed | hours | technicians | inch | dispatch | rhand | missing | carrier |
| lock | happened | block | repaired | prior | broken | close | information |
| disconnected | dont | complete | written | applied | inspected | properly | report |

| Topic 9 | Topic 10 | Topic 11 | Topic 12 | Topic 13 | Topic 14 | Topic 15 |
|---|---|---|---|---|---|---|
| leak | reporter | torque | wrong | step | log | hangar |
| run | test | install | data | amm | page | ground |
| noted | revealed | remove | tag | replacement | writeup | ramp |
| low | conversation | upper | incorrect | proper | discrepancy | taxi |
| drain | callback | inboard | type | removal | items | stop |
| leaking | fire | lower | shop | performing | deferred | started |
| fluid | failed | hardware | stores | perform | action | move |
| level | category | outboard | serviceable | remove | deferral | short |
| disconnect | operations | trim | effectivity | checks | entry | set |
| normal | released | loose | error | hand | report | moved |

*Note*. Resulting topic word lists, after the addition of standard and customized stop word lists, appear to be focused on incident actions, contributor descriptives, locations, and the roles of those involved.

The standard English *stopwords* list from the NLP algorithm contains many terms that could, in certain contextual circumstances, be related to human factor concepts. Any words that could be broadly interpreted as having relevance to human factor concepts were removed from the *stopwords* list for the final iteration of topic word listings. The complete list of words meeting this criterion, described as *add-back* words, is shown in Appendix C. Extant literature does not universally recognize a defined term for these words, therefore the term *add-back* words is used in this research. The resulting list of topic words from this refinement process is shown in Table 7.

**Table 7**

*Topic Word Lists Resulting from Relevance Refinement Process*

| Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 | Topic 6 | Topic 7 | Topic 8 |
|---|---|---|---|---|---|---|---|
| different | issue | landing | technician | leak | wrong | repair | log |
| data | can | main | night | run | may | limits | page |
| tag | training | remove | morning | high | missing | damaged | writeup |
| incorrect | manager | torque | hours | bleed | changed | within | discrepancy |
| stock | management | install | approximately | power | brought | authorized | controller |
| shop | report | tool | until | low | mistake | srm | deferred |
| type | amt | removal | several | closed | notified | document | deferral |
| stores | carrier | installing | event | noted | can | repaired | action |
| received | personnel | proceeded | prior | quantity | proper | inch | entry |
| serviceable | many | removing | hangar | drain | mm | figure | call |

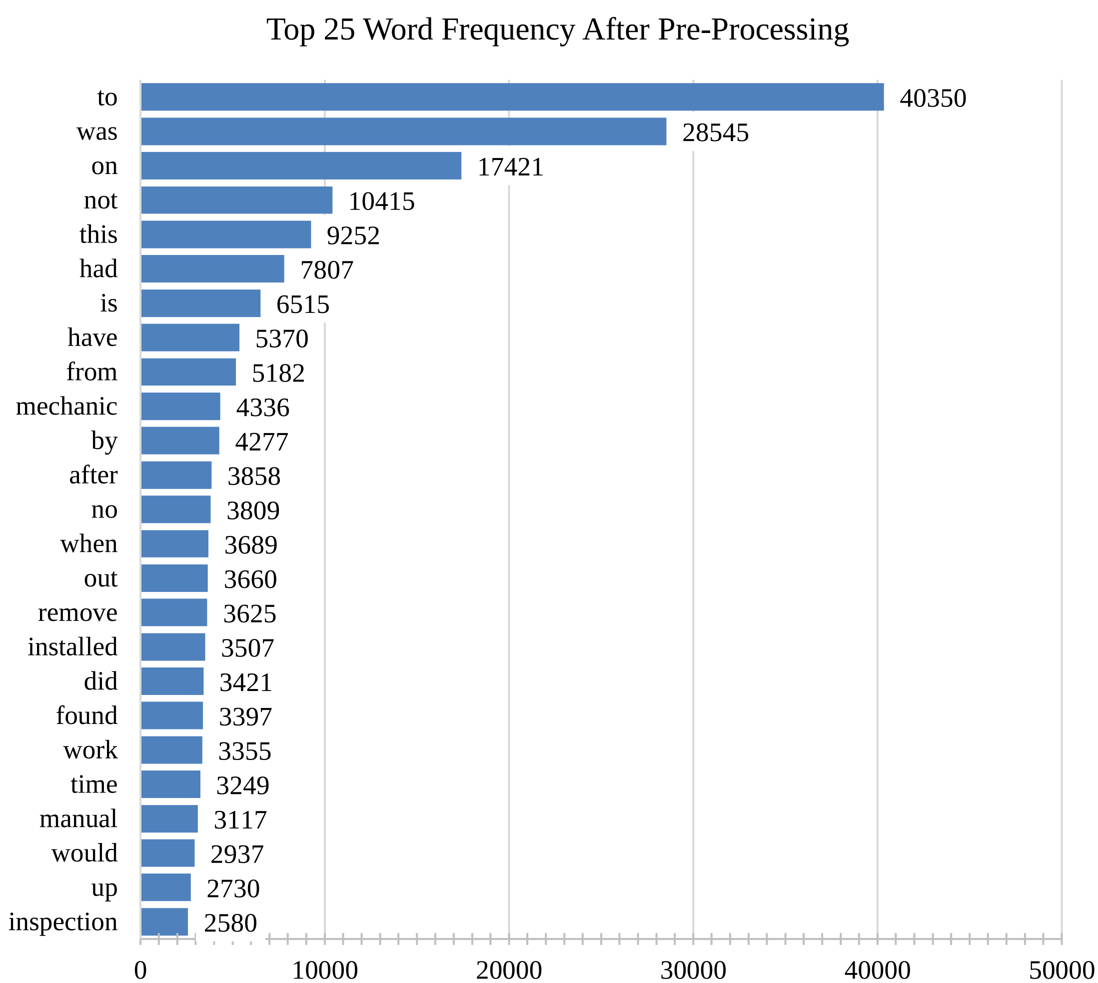
| Topic 9 | Topic 10 | Topic 11 | Topic 12 | Topic 13 | Topic 14 | Topic 15 |
|---|---|---|---|---|---|---|
| test | reporter | step | side | fault | going | hangar |
| amm | revealed | inspector | access | troubleshooting | put | ground |
| failed | conversation | routine | forward | message | go | power |
| checks | callback | items | aft | pulled | got | ramp |
| operations | fire | cards | inboard | normal | didnt | taxi |
| replacement | emergency | release | hardware | indication | know | stop |
| operational | aft | quality | lhand | warning | looked | started |
| category | lock | signoff | inspected | pilot | look | move |
| perform | closed | complete | outboard | start | never | short |
| note | properly | block | install | applied | thought | heard |

*Note*. The final iteration of the refinement process to increase relevance to human factor concepts contains 15 topic lists with 10 words in each topic.

After the data preprocessing steps were completed the data set consisted of 17,486 words distributed over 5,929 reports. The word-document frequency, or the number of documents that a word appears in, varied from 5,794 for the 40,350 occurrences of the word *to*, down to 7,510 words that only occurred within a single document. The resulting top 25 list changed considerably after the implementation of the preprocessing methodology. However, many of the words still lack context. Figure 15 shows the top 25 words by frequency after the data preprocessing steps.

**Figure 15**

*Top 25 Words in the ASRS Dataset After Preprocessing and Topic Enhancement*




*Note*. These word occurrences document the content of the selected dataset after data preprocessing using word elimination and word substitution.

The majority of the words on the top 25 list in Figure 15 were included in the *add-back* word list because they could theoretically add human factors context to the other words within the potential topic word lists. These include *to*, *from*, *up*, *by*, and *on*, which might imply a locational reference to an object. Other words such as *was*, *is*, *would*, *did*, *have*, and *had*, were included because they could provide a temporal reference to other words in the potential topic groupings. As noted in the description of the enhancement strategy, the *add-back* word decisions were based on broad possibilities of interpretation.

### *Validity of the Topic Refinement Process*

The process of refining the topic word lists was intended to increase the relevance of the topic word groups to human factor concepts. Ensuring the validity of this methodology required seeking the opinion of three human factor SMEs. These experts were contacted through emails to the Embry-Riddle School of Graduate Studies faculty and aviation Ph.D. program alumni. Three individuals agreed, by email, to review the methodology and provide input on its validity as well as suggestions for improvement.

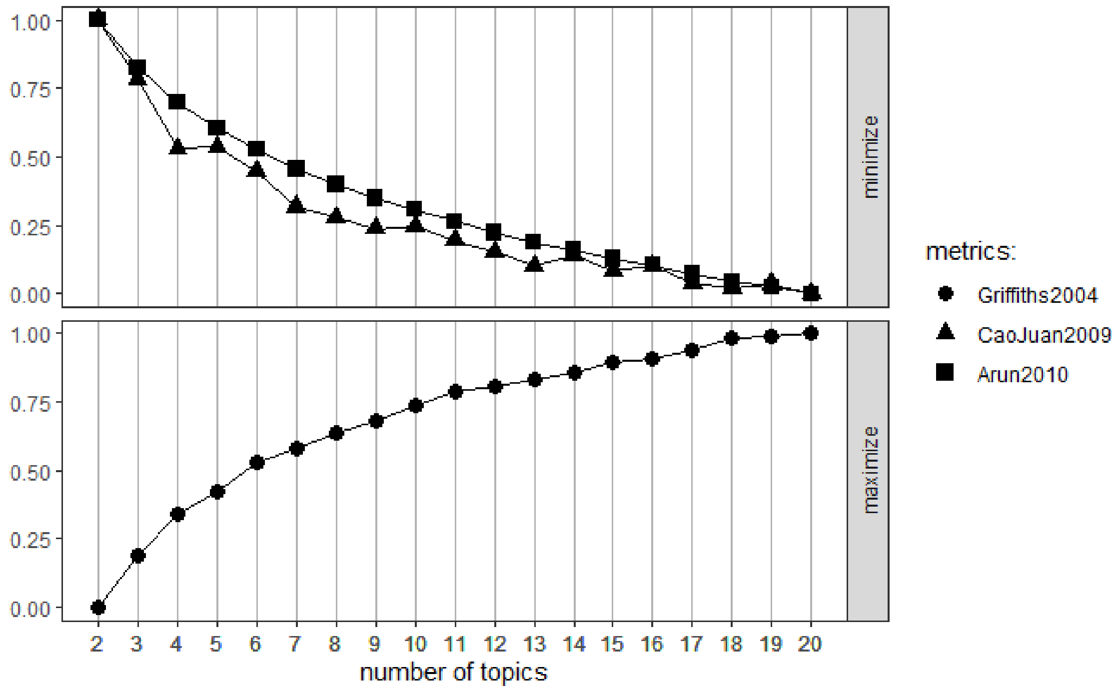
The replies from the three HF SMEs reflect a general agreement that the topic enhancement methodology constituted a sound approach and was needed to align the topics to human factor concepts, as demonstrated by the step-by-step process descriptions and the evolution of the topic lists through the four iterations. The qualifications of the HF SMEs are listed in Appendix D.

### *Topic Modeling with LDA*

The LDA process relies on a Gibbs analysis to establish a posterior parameter eta ($\eta$), the optimal number of topics for the dataset. The Gibbs analysis for the selected dataset indicated that the optimal number of topics was between 11 and 15. Blei et al. (2003) noted that a review of the generated topics for clarity and relevance should be used as a supplement to the Gibbs method. Figure 16 shows the results of the Gibbs analysis.

**Figure 16**

*Gibbs Analysis for the Selected Dataset of ASRS Maintenance Reports*



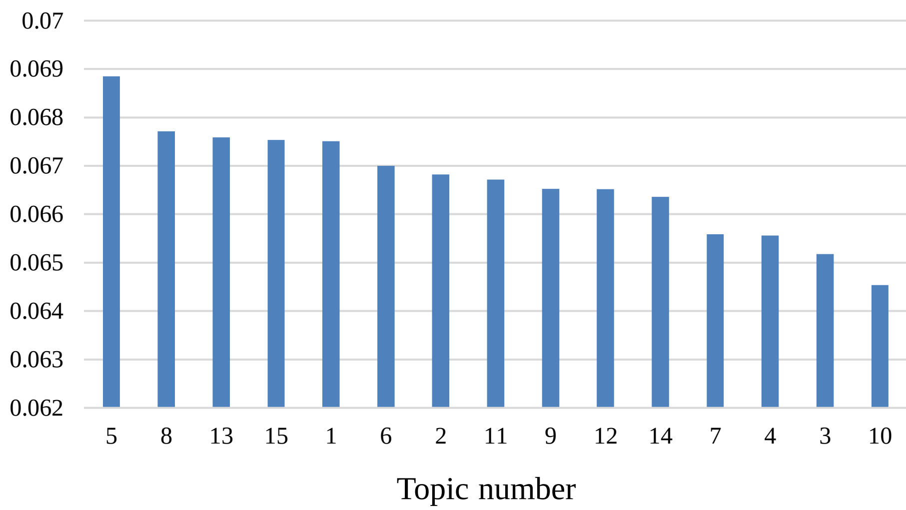

*Note*. The asymptotic curve of the Griffiths2004 plot indicates increased model coherence as the number of topics increases. The change in the slope of the CaoJuan2009 and Arun2010 plots indicates an approach to model coherence and the optimal value of **η,** while additional topics may demonstrate irrelevance and be an indicator of model noise (Röder et al., 2015).

The plots in Figure 16 represent three separate metrics used as coherence indicators in LDA modeling. *Griffiths2004* is a probability measure based on how words occur within topics (Griffiths & Steyvers, 2004). The *CaoJuan2009* metric evaluates the similarity between topics using a cosine distance between topic word distributions to evaluate coherence (Cao et al., 2009). Finally, *Arun2010* evaluates coherence between topics by measuring how distinct each topic is based on the topic's word distribution (Arun et al., 2010).

In the LDA model, the gamma statistic signifies the probability that a topic will occur in a given document. Each topic has a gamma value for each document within the corpus. A plot of the average of the gamma values as shown in Figure 17.

**Figure 17**

*Average Gamma Values by Topic in Declining Order Before Topic Reduction*

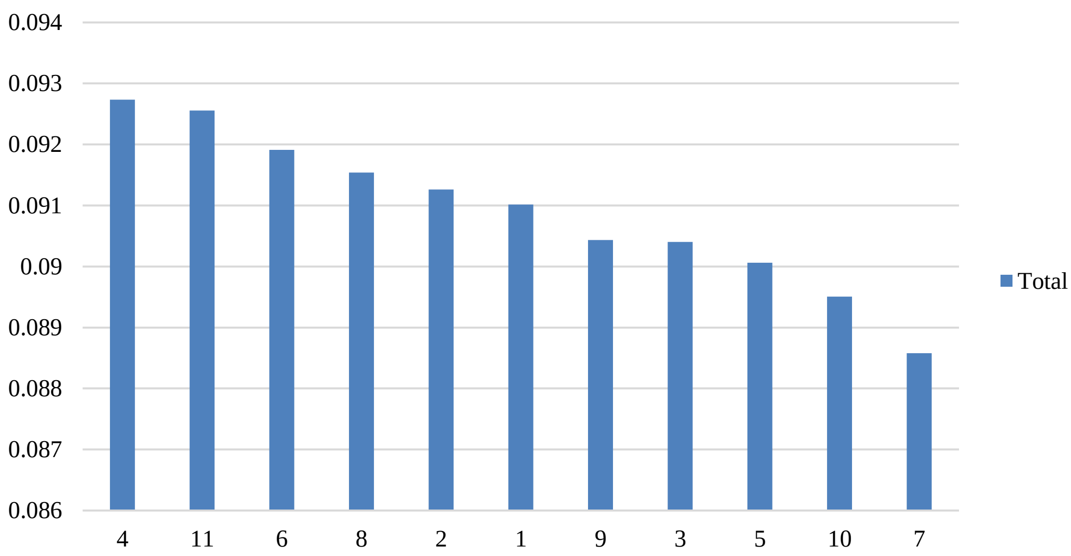



*Note*. The *topicmodels* package in R does not assign topic numbers in order of their prevalence. Topic numbers on the horizontal axis are the arbitrary designations assigned by the *topicmodels* package.

The average gamma values demonstrate groupings that indicate their relevance within the corpus of documents. The decline between the average gamma value between topic 14 (ranked 11th) and topic 7 (ranked 12th) supports the results of the Gibbs analysis that the optimal number of topics, or the likely point where model coherence is achieved, lies between 11 and 15 topics for the selected data set. Blei et al. (2003) and Röder et al. (2015) cautioned that estimations of model coherence, both through Gibbs analysis and the use of the gamma statistic, are subjective. The vagaries of language and semantic variations do not fit neatly into a mathematical model. Therefore, a review of actual topic word groupings was necessary to evaluate relevance to the selected corpus of documents. Because of the LDA formulation within the *topicmodels* algorithm, optimizing the number of topics alters the topic content. Reducing the number of topics to 11 yielded the gamma averages displayed in Figure 18.

**Figure 18**

*Final Topic Gamma Averages for 11 Topics*




*Note*. Topic numbers are arbitrarily assigned.

Each topic word group does not represent a particular report. The topic word group has a probability of occurring in a number of reports. A higher average gamma value for the topic word group indicates a probability that the group occurs in a larger number of reports. Additionally, each word within a group has a probability measure of beta (***β***) that expresses the likelihood that the word will occur concurrently with the other words in the group in a given report.

The topic modeling algorithm assigns numeric values for topic titles in a random manner that is not aligned with how words are discovered within the dataset. The prevalence of the topics by their gamma characteristic is shown in Figure 18. In order to avoid misinterpretation of the naming convention from the topic modeling output, the topic designations have been changed in further discussions to reflect their gamma prevalence. This conversion is shown in Table 8.

**Table 8**

*Topic Renaming by Gamma Prevalence*

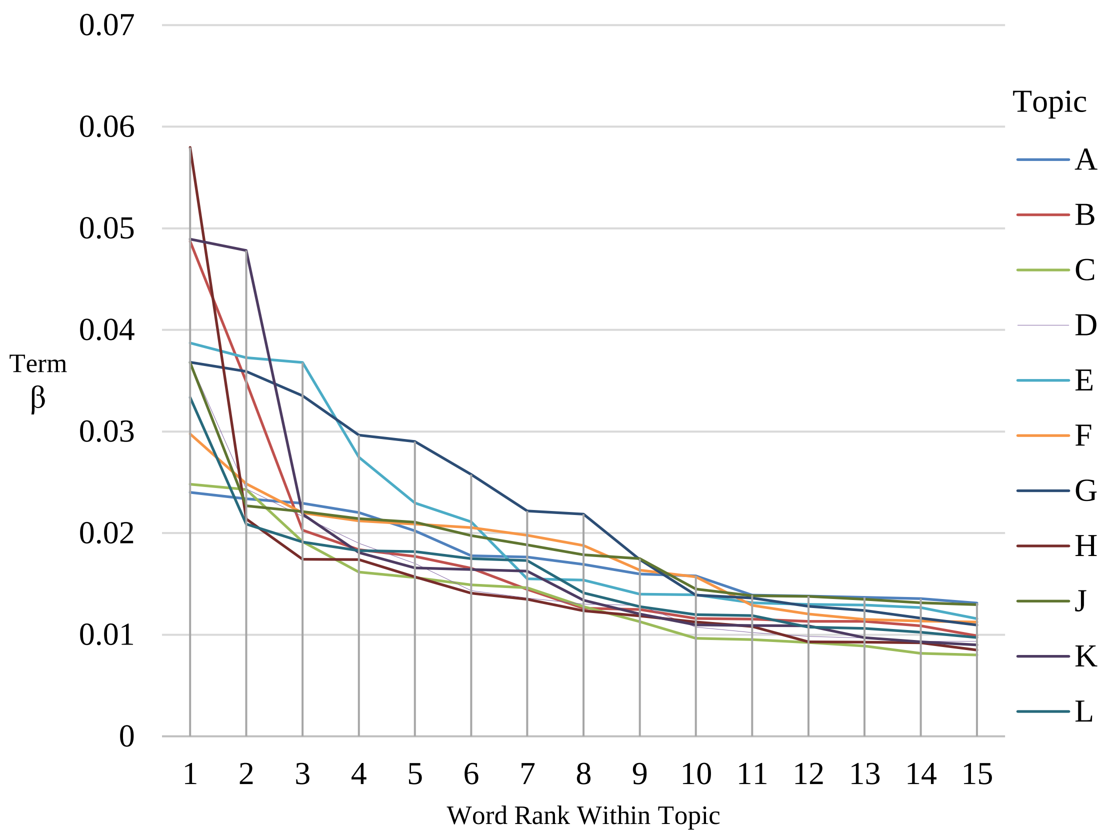
| Topic order rank | Generated topic number | Gamma designation | Gamma average |
|---|---|---|---|
| 1 | Topic 4 | Topic A | 0.09273448 |
| 2 | Topic 11 | Topic B | 0.09255508 |
| 3 | Topic 6 | Topic C | 0.09191001 |
| 4 | Topic 8 | Topic D | 0.09153899 |
| 5 | Topic 2 | Topic E | 0.09125965 |
| 6 | Topic 1 | Topic F | 0.09101408 |
| 7 | Topic 9 | Topic G | 0.09043527 |
| 8 | Topic 3 | Topic H | 0.09040409 |
| 9 | Topic 5 | Topic J | 0.09006363 |
| 10 | Topic 10 | Topic K | 0.08950406 |
| 11 | Topic 7 | Topic L | 0.08858066 |

*Note*. The renaming scheme simply substitutes a letter designation according to the topic gamma ranking. The letter *I* has been omitted from the naming convention to avoid confusion with the number one (1).

Determining the appropriate number of words within each topic requires evaluation of word beta values within the topic as well as an interpretation of word relevance (Blei et al., 2003). The beta statistic is a probability vector expressing the likelihood of a word occurring within the stated topic word grouping. Although the topic numbers are assigned arbitrarily, the output of the topic models package lists the words within topics in their order of prevalence by beta values. Figure 19 shows the beta values of the top 15 words in each of the top 11 topics. Table 9 shows the final topic word groupings after reduction to eleven topics containing ten words each.

**Figure 19**

*Beta Values of the Top 15 Words in Each of the Top 11 Topics*




*Note*. Each series plot represents the word grouping in a separate topic. Word relevance, as measured by the beta statistic, declines from one word to the next within each topic. The plot exhibits a static rate of change after word 10, indicating the prevalence of successive words does not significantly decrease compared to their predecessors in the topic.

**Table 9**

*Final Word Topic Groups*

| Topic A | Topic B | Topic C | Topic D |
|---|---|---|---|
| night | test | log | reporter |
| put | fault | page | leak |
| thought | operations | items | closed |
| going | controller | writeup | revealed |
| didn't | troubleshooting | entry | conversation |
| go | captain | routine | callback |
| know | pilot | data | fire |
| got | message | cards | high |
| morning | inoperable | release | bleed |
| looked | pulled | overnight | run |

| Topic E | Topic F | Topic G | Topic H |
|---|---|---|---|
| hangar | side | step | can |
| ground | aft | landing | several |
| power | missing | AMM | event |
| start | forward | main | issue |
| ramp | access | does | many |
| taxi | torque | under | may |
| started | lower | note | occurred |
| stop | install | states | times |
| set | inspector | accordance | possible |
| move | upper | read | involved |

| Topic J | Topic K | Topic L |
|---|---|---|
| report | wrong | repair |
| management | changed | limits |
| inspector | emergency | damaged |
| FAA | different | inch |
| information | tag | approximately |
| carrier | incorrect | within |
| manager | stock | hours |
| quality | type | hole |
| perform | old | shop |
| training | shop | SRM |

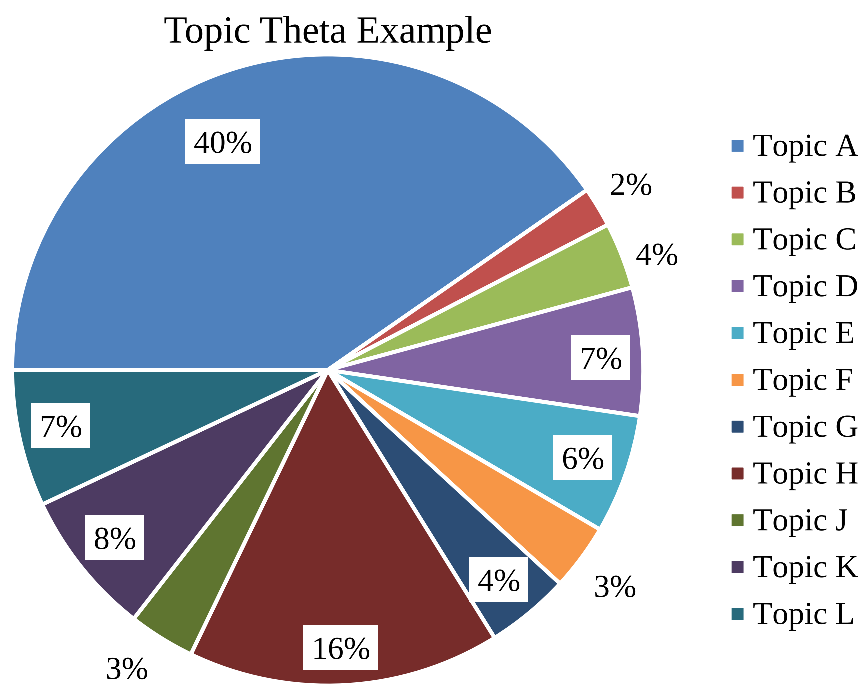
*Note*. Individual topic word lists are arranged in descending order of their beta values. The final word lists contain three abbreviations: FAA, AMM, and SRM. Because these abbreviations represent multiple words, it was impractical to address them through word substitutions.

***Content Analysis of Topics***

Each topic word group does not represent a particular report. The topic word group has a probability of occurring in a number of reports. A higher average gamma value for the topic word group indicates a probability that the group occurs in a number of reports. While the gamma value is a useful parameter for estimating posterior distributions of topics among documents, the theta vector is the true resulting distribution of topics among documents. It is a property of the LDA process that the sum of all theta ($\theta$) values for a given document must be equal to one. An example is shown in Figure 20.

**Figure 20**

*An Example of Theta Distribution for a Document Within the Selected Corpus*




*Note*. The values of the theta vectors are rounded for visibility in the chart. In the LDA formulations, the sum of the theta values is always equal to 1 by the definition of the parameter. The document selected as the example had the highest theta value of Topic A.

A simple evaluation of the topics listed in Table 9 would seem to indicate that the corpus of documents was clearly related to the aviation industry. This is evident by the terms *FAA*, *pilot*, and *hangar*. There is also a reference to an organizational structure indicated by the words *management*, *manager*, and *operations*. The theme of communications, both written and verbal, is prevalent in the word groupings as with the words *report*, *log*, *page*, *writeup*, *entry*, *data*, *cards*, *conversation*, *callback*, *note*, *states*, *read*, and *message*. Perhaps most relevant are the references to maintenance type activities demonstrated by the terms *AMM*, *SRM*, *inspector*, *damaged*, *repair*, *troubleshooting*, *test*, and *fault*. Taken as a whole, the words within the topic groupings clearly indicate an association with aviation maintenance activities accruing under an organized structure.

As a result of the NLP process, the words appear to have some distinctive thematic groupings. The majority of the terms are, unsurprisingly, associated with the requirements for tasks performed in aviation maintenance. The second most prevalent grouping is related to communication methods and aviation personnel which the incident reporter may have had cause or opportunity with which to communicate. The third group that can be observed within the generated topic word groups is connected to movement within an aviation maintenance working environment. There are several words that, depending on context, could be associated with one or more of these themes, as well as a potential classification for safety or regulatory concepts. Grouping the words by their potential categories demonstrates a prevalence of categories within some of the topic word groupings as shown in Figure 21.

**Figure 21**

*Categories Within Topic Word Groups*

| Topic A | Topic B | Topic C | Topic D | Topic E | Topic F |
|---|---|---|---|---|---|
| night | test | log | reporter | hangar | side |
| put | fault | page | leak | ground | aft |
| thought | operations | items | closed | power | missing |
| going | controller | writeup | revealed | start | forward |
| didnt | troubleshooting | entry | conversation | ramp | access |
| go | captain | routine | callback | taxi | torque |
| know | pilot | data | fire | started | lower |
| got | message | cards | high | stop | install |
| morning | inoperable | release | bleed | set | inspector |
| looked | pulled | overnight | run | move | upper |

| Topic G | Topic H | Topic J | Topic K | Topic L |
|---|---|---|---|---|
| step | can | report | wrong | repair |
| landing | several | management | changed | limits |
| amm | event | inspector | emergency | damaged |
| main | issue | faa | different | inch |
| does | many | information | tag | approximately |
| under | may | carrier | incorrect | within |
| note | occurred | manager | stock | hours |
| states | times | quality | type | hole |
| accordance | possible | perform | old | shop |
| read | involved | training | shop | srm |

| Legend |
|---|
| **Task** |
| **Environment** |
| **Communication** |
| **Safety** |
| **Ambiguous** |

*Note*. The color coding of the tentative categories within the word groupings demonstrates the prevalence of multiple categories within each topic word group.

The tentative task associated word category comprises 45% of the words in the final topic word groupings. In comparison, 25% of the words are associated with movement and the work environment, and 21% can be connected to communication modalities. Less than 2% of the terms can be positively associated with safety and regulatory structures. The nine words that were classified as ambiguous, or 8%, could fall into multiple categories based on the context within the specific report where they are observed. Figure 22 shows the distribution of the topic words within the possible classifications when considered out of context with the topic groups.

**Figure 22**

*Topic Words Grouped by Potential Categories*

| Task Associated | | Communication and Personnel | Environment, Location, and Movement | Safety and Regulatory | Ambiguous |
|---|---|---|---|---|---|
| access | missing | accordance | aft | emergency | **controller**- possibly task or personnel; |
| approximately | occurred | AMM | forward | FAA | **fire**- possibly task or safety; |
| bleed | old | callback | go | | **high**- possibly task or location; |
| can | operations | captain | going | | **issue**- possibly task or communication; |
| changed | performed | cards | ground | | **items**- possibly task or communication; |
| closed | possible | carrier | hangar | | **main**- possibly task or location; |
| damaged | power | conversation | hours | | **run**- possibly task or movement; |
| didn't | pulled | data | landing | | **step**- possibly task or movement; |
| different | put | entry | lower | | **type**- possibly task or communication |
| does | quality | information | morning | | |
| event | release | inspector | move | | |
| fault | repair | log | night | | |
| got | routine | management | ramp | | |
| hole | set | manager | shop | | |
| inch | several | message | side | | |
| incorrect | start | note | started | | |
| inoperable | stock | page | stop | | |
| install | test | pilot | taxi | | |
| involved | thought | read | times | | |
| know | torque | reporter | under | | |
| leak | training | revealed | upper | | |
| limits | troubleshooting | SRM | | | |
| looked | within | states | | | |
| many | wrong | tag | | | |
| may | | writeup | | | |

*Note*. Taken outside the context of the groupings generated in the topic modeling process described in Chapter III, these categories represent a potential categorization framework.

Figures 21 and 22 demonstrate that there are coherent themes within the corpus of documents from the ASRS database. While the listed categories are not exhaustive and are subject to interpretive methods and points of view, the observed categories demonstrate a focus on task associated ideas, movement within the working environment, and communication at discrete organizational levels. The generation of the final word topic groupings shown in Table 9 completes the first stage of the research plan and provides insight to the first research question.

## Reliability and Validity Testing Results

The reliability of the research method is evidenced by the structured methodology employed to obtain the final topic word groups. The topic modeling method was based on proven NLP algorithms and the relevance of the dataset was enhanced using methods based on sound logic and well reasoned practices, as described by the three-person panel of human factors SMEs. These HF SMEs were familiarized with the topic modeling process and reviewed the step by step progression of the data preprocessing and topic word group enhancement. Context in language can alter meaning and affect the validity of research methods that involve the use of application specific terms (Tracy, 2010). Therefore, the idiosyncracies of the language used in both human factors and aviation maintenance necessitated SME consultation to ensure the validity of the research at each step in the research method.

### *SME Consensus via Delphi Analysis*

As described in Chapter III, the research plan called for a modified Delphi analysis to validate the topics generated by the topic modeling using LDA. Seven of the eight solicited aviation maintenance SMEs agreed to participate in this activity, and opinions regarding the content and relevance of the topic word groups to the selected conceptual frameworks were solicited through an online survey platform. The group of SMEs represented a wide variety of specializations within the aviation maintenance community. These included academic educators with military and civilian maintenance experience, former A&P mechanics with doctoral degrees in aeronautics, and retired aviation incident investigators specializing in maintenance related incidents. Brief descriptions of the aviation maintenance SME qualifications are included in Appendix D.

In the first round of Delphi analysis, the SMEs were asked to provide 30 evaluations. Each of the 15 topics were evaluated for relevance to CREAM Error Modes and HFACS Unsafe Acts categories. Six of the seven panelists responded to the first round of evaluations. The SME comments varied widely. Some comments were overwhelmingly positive in terms of relevance to both selected HF frameworks, while others expressed a failure to see any relevance to either framework. One SME panelist repeatedly commented that the word groupings did not constitute topics in a traditional sense and felt strongly that the Delphi methodology was not practical. The IQR values were widely dispersed, and only two of the 30 evaluation sets achieved an IQR of $\leq 1$. The quantitative values for the first round of Delphi analysis are shown in Table 10.

**Table 10**

*Round 1 Delphi Analysis of Quantitative Responses*

| Topic | 9598 | 7144 | 5786 | 2663 | 2292 | 1922 | IQR | Range |
|---|---|---|---|---|---|---|---|---|
| 1-a | 1 | 6 | 2 | 6 | 2 | 5 | 3.75 | 5 |
| 1-b | 3 | 6 | 2 | 7 | 2 | 3 | 3 | 5 |
| 2-a | 1 | 8 | 3 | 6 | 3 | 1 | 3.75 | 7 |
| 2-b | 7 | 8 | 3 | 4 | 3 | 1 | 3.25 | 7 |
| 3-a | 7 | 8 | 1 | 2 | 4 | 1 | 5 | 7 |
| 3-b | 3 | 8 | 3 | 4 | 4 | 5 | 1.5 | 5 |
| 4-a | 5 | 8 | 3 | 4 | 4 | 4 | **0.75** | 5 |
| 4-b | 7 | 8 | 5 | 5 | 5 | 4 | 1.5 | 4 |
| 5-a | 7 | 8 | 4 | 4 | 4 | 1 | 2.25 | 7 |
| 5-b | 6 | 8 | 3 | 4 | 4 | 1 | 2.25 | 7 |
| 6-a | 5 | 8 | 3 | 4 | 4 | 1 | 1.5 | 7 |
| 6-b | 6 | 8 | 2 | 8 | 4 | 1 | 5 | 7 |
| 7-a | 7 | 7 | 2 | 5 | 4 | 3 | 3.25 | 5 |
| 7-b | 5 | 7 | 5 | 8 | 4 | 3 | 2.25 | 5 |
| 8-a | 6 | 8 | 1 | 5 | 4 | 1 | 4 | 7 |
| 8-b | 7 | 8 | 4 | 7 | 4 | 3 | 3 | 5 |
| 9-a | 4 | 8 | 1 | 3 | 4 | 1 | 2.5 | 7 |
| 9-b | 7 | 8 | 5 | 7 | 4 | 1 | 2.75 | 7 |
| 10-a | 5 | 8 | 1 | 4 | 4 | 1 | 3 | 7 |
| 10-b | 8 | 8 | 8 | 7 | 4 | 1 | 3.25 | 7 |
| 11-a | 4 | 8 | 5 | 3 | 4 | 4 | **0.75** | 5 |
| 11-b | 3 | 8 | 4 | 7 | 4 | 1 | 3 | 7 |
| 12-a | 4 | 7 | 1 | 4 | 4 | 1 | 2.25 | 6 |
| 12-b | 3 | 7 | 6 | 5 | 4 | 1 | 2.5 | 6 |
| 13-a | 1 | 8 | 1 | 3 | 4 | 1 | 2.75 | 7 |
| 13-b | 6 | 8 | 5 | 6 | 4 | 4 | 1.75 | 4 |
| 14-a | 4 | 8 | 2 | 4 | 4 | 2 | 1.5 | 6 |
| 14-b | 8 | 8 | 4 | 7 | 4 | 2 | 3.75 | 6 |
| 15-a | 4 | 8 | 7 | 4 | 4 | 1 | 2.25 | 7 |
| 15-b | 5 | 8 | 7 | 7 | 4 | 2 | 2.75 | 6 |

*Note*. The response values are grouped by the individual SME participant codes issued to maintain confidentiality of participants. The a and b designations for each topic number represent HFACS unsafe acts and CREAM error mode evaluations, respectively. The evaluation scale ranged from 1 to 9 with the description headings *Does not reflect* (1-3), *Moderately reflects* (4-6), and *Strongly reflects* (7-9).

In the second set of evaluations, the SMEs were given a brief synopsis of the comments from the first round, reminded of their individual evaluations and comments, and asked to reevaluate 29 of the original 30 evaluation sets. Six SMEs provided responses for the second round. SME 2663 participated in the first round but was unable to participate in the second round. Scheduling prevented SME 7512 from responding in the first round, but this SME was able to participate in the second round. Again, in the second round, the IQR values were widely dispersed and only two additional evaluations reached the consensus threshold. In addition to comments regarding how well the topics did or did not reflect the HF frameworks, several SMEs commented that the surveys were far too long and repetitive to be able to effectively evaluate the topic content to the selected HF frameworks. The IQR values for the second round are shown in Table 11.

**Table 11**

*Round 2 Delphi Analysis of Quantitative Responses*

| Topic | 9598 | 7144 | 5786 | 2292 | 1922 | 7512 | IQR | Range |
|---|---|---|---|---|---|---|---|---|
| 1-a | 3 | 6 | 2 | 2 | 3 | 5 | 2.25 | 4 |
| 1-b | 3 | 6 | 1 | 2 | 3 | 5 | 2.25 | 5 |
| 2-a | 2 | 8 | 1 | 3 | 1 | 4 | 2.5 | 7 |
| 2-b | 4 | 8 | 2 | 4 | 1 | 4 | 1.5 | 7 |
| 3-a | 5 | 7 | 1 | 1 | 1 | 6 | 4.75 | 6 |
| 3-b | 4 | 8 | 2 | 4 | 5 | 6 | 1.75 | 6 |
| 4-b | 7 | 8 | 7 | 5 | 4 | 5 | 2 | 4 |
| 5-a | 5 | 8 | 4 | 5 | 1 | 4 | 1 | 7 |
| 5-b | 5 | 8 | 4 | 5 | 1 | 4 | 1 | 7 |
| 6-a | 1 | 8 | 4 | 4 | 1 | 7 | 4.5 | 7 |
| 6-b | 6 | 8 | 2 | 4 | 1 | 7 | 4.25 | 7 |
| 7-a | 6 | 7 | 5 | 4 | 3 | 6 | 1.75 | 4 |
| 7-b | 6 | 8 | 6 | 5 | 3 | 7 | 1.5 | 5 |
| 8-a | 6 | 8 | 1 | 5 | 1 | 7 | 4.75 | 7 |
| 8-b | 7 | 8 | 5 | 5 | 3 | 6 | 1.75 | 5 |
| 9-a | 4 | 8 | 1 | 6 | 1 | 8 | 5.75 | 7 |
| 9-b | 6 | 7 | 7 | 4 | 1 | 7 | 2.5 | 6 |
| 10-a | 1 | 8 | 3 | 5 | 1 | 5 | 3.5 | 7 |
| 10-b | 7 | 8 | 9 | 6 | 1 | 5 | 2.5 | 8 |
| 11-b | 4 | 8 | 4 | 4 | 1 | 6 | 1.5 | 7 |
| 12-a | 1 | 7 | 1 | 4 | 1 | 4 | 3 | 6 |
| 12-b | 4 | 7 | 4 | 3 | 1 | 5 | 1.5 | 6 |
| 13-a | 2 | 8 | 1 | 3 | 1 | 3 | 1.75 | 7 |
| 13-b | 6 | 8 | 5 | 3 | 4 | 6 | 1.75 | 5 |
| 14-a | 1 | 8 | 3 | 3 | 2 | 6 | 3 | 7 |
| 14-b | 7 | 8 | 5 | 5 | 2 | 6 | 1.75 | 6 |
| 15-a | 4 | 8 | 6 | 4 | 4 | 7 | 2.75 | 4 |
| 15-b | 5 | 8 | 8 | 4 | 4 | 7 | 3.5 | 4 |

*Note*. The evaluation criteria were consistent between rounds of SME evaluation. In the second round of evaluations, two additional evaluation sets achieved an acceptable measure of consensus.

The qualitative responses of the SMEs contained repetitive statements. This, along with the narrow range intra-respondent quantitative responses, possibly demonstrating potential response set bias, led to an inability to achieve significant consensus. The lack of consensus, and the respondent comments regarding the lengthy and repetitive nature, therefore, led to the decision to discontinue the Delphi analysis without additional evaluation rounds.

### ***Alternate SME Evaluation Methodology Results***

Considering the ineffective results of the Delphi analysis process, an alternate method was required to obtain substantive aviation maintenance SME input. Two SMEs were selected from the original Delphi analysis group based on the quality of their comments and willingness to provide further input. Two additional SMEs agreed to participate in the scheduled virtual roundtable discussion. The SME qualifications are listed in Appendix D.

The intent of the roundtable discussion was to focus on three questions intended to address the research questions from the perspective of the SMEs' experience and body of knowledge. The questions and a sampling of the SME comments are:

- Can you theorize the operator level details of the incident from the words in the topic groupings?
    - The four SMEs discussed specific words within the topic word groups. One of the SMEs expressed surprise that the words *night* and *got* were listed. SME 4333 commented that he had never noted the word *got*, in his recollection, in any of the thousands of pages of incident reports he had reviewed in his career. SME 9598 pointed out the prevalence of shorthand in maintenance reports and the loss of context because of this. SME 7651 agreed and discussed the possibility of standardizing shorthand terminologies to improve the quality of reports. The SMEs discussed several incidents they had

encountered involving shift handovers, communication, and the use of work cards to convey information. SME 7651 said that he could put together a storyline using the words from the topic groups for each group although some, such as Topic E that they had previously discussed, were more apparent than others. SMEs 7512 and 4333 agreed with this. SME 4333 commented that he could develop several different scenarios for each of the topics and 9598 and 7651 expressed their agreements. SME 4333 said that hypothetical situations could be developed using a majority of words in the topics and 7512 agreed with this assertion.

- Are the topic word groupings more phenotypical or genotypical in their descriptions?
  - The four SMEs considered the question and asked for some clarification on the question. After a brief refresher on HFACS and CREAM, the SMEs reevaluated the topic word groups. SME 4333 pointed out that several words could belong to either category. For instance, *wrong* could be associated with phenotype if it were in reference to a part or a direction while it could be applied to a genotypical reference when referenced as a decision-making error and 7512 agreed. SME 4333 provided an example of an experienced mechanic who, when describing an incident, insisted on jumping over the physical cause because there was no sense in looking at how a bolt was tightened, or a safety wire was applied. SME 9598 noted that the only real representation of phenotypes in the topic word groupings was reference to wrong object or wrong sequence, and there were very few word sets that referenced magnitude as in too much of this or too little of that. SME 7651 said that although there

were some words that were clearly associated with the CREAM error modes, such as *times* and *forward*, and others with HFACS, like *know* and *thought*, neither system was significantly prevalent in the overall set of terms.

- Would HFACS or CREAM be a better choice to use as an investigation tool for an incident that had the topic groups in their descriptions?
    - To start the discussion over whether CREAM or HFACS would be a better choice for incident investigation, 9598 commented that the last three categories of CREAM Error Modes, direction, wrong object, and sequence, could really be fit into the HFACS taxonomy under the decision errors categories. SME 4333 agreed and further commented that when describing incidents, as he had mentioned before, mechanics seldom discuss the errors in terms of the physical aspects of the error; 7651 agreed and said that, although the CREAM error modes often show up during incident investigations, the specifics are rarely useful in terms of finding a root cause to the incident. SME 4333 agreed with this and stated that the HFACS taxonomy was more likely to be useful for incident investigation than CREAM. SME 9598 commented that HFACS is so general that it will be easily applicable to everything. The three other SMEs voiced their agreement on this.

Generally, the SME discussion participants found the methodology and its potential applications of particular interest. The discussion concentrated on the topic groups and the topic modeling methodology as a comprehensive process. All four participants agreed that, for the majority of the topics, they could theorize the details of incidents that contained the topics in their descriptions. The topic word groups were not thought to be more focused on either phenotypical or genotypical themes, but there was unanimous agreement among the participants that HFACS would be a better investigative tool than CREAM for the incidents that they theorized could have generated the descriptions from which the topic word groups were taken. Each of the participants identified particular terms that they found to be common within the industry and the incident descriptions with which they had experience. During extemporaneous discussions interspersed within the topical questions, the four SMEs discussed the lack of commonality in the investigation tools they had encountered and expressed the need for the industry to center itself around common tools and terms to improve incident investigation.

Although the discussion deviated from the intended parameters, the maintenance SMEs provided insightful comments and generally agreed that the topic word groups were easily relatable to specific aviation maintenance themes. Additionally, the two SMEs with investigative experience felt that the topic word groups were no more or less relatable to the selected HF frameworks than were the majority of the actual incident descriptions they had encountered in their experiences. These two SMEs noted that there is a lack of commonality in the way incidents are reported, and, specifically with maintenance incidents, this lack of commonality hinders the investigative process.

**Development of a Novel Framework**

The topic word groups generated using the topic modeling process indicate a focus on task details, working environment and communication deficiencies. The *Unsafe Acts of the Operator* in the HFACS taxonomy and the *Error Modes* consequents from the CREAM classifications used as exemplars of genotypical and phenotypical classification methods for the purposes of this research do not have a clear focus on the observed aspect of a maintenance technician's working environment. It must be noted, however, that the categorical levels of the two selected strategies adopted for this research do not make up the complete classification strategies of either system. As demonstrated in Chapter II, Figure 6, and Appendix C, both example frameworks contain multiple categorical levels that address more expansive groups of potential HF contributors. Table 12 presents a comparison of HFACS, CREAM, and a proposed novel framework to the observed categories of the words within the topic word groupings.

**Table 12**

*A Comparison of HF Frameworks to Observed Categories*

| Observed Categories | Selected Key Words | HFACS Category | CREAM Category | Novel Framework Category |
|---|---|---|---|---|
| Task Execution Requirements | access<br>changed<br>install<br>looked<br>missing<br>occurred<br>performed<br>routine<br>training<br>troubleshooting | Skill Based Errors<br>Decision Errors<br>Technological Environment<br>Physical/Mental Limitations<br>Inadequate Supervision<br>Failed to Correct Problem<br>Supervisory Violation<br>Resource Management<br>Organizational Management | Observation<br>Interpretation<br>Planning<br>Temporary person related functions<br>Equipment failure<br>Procedure<br>Permanent interface problems<br>Communication<br>Organization<br>Training | Task Requirements<br>Training<br>Task Definition |
| Action within the Working Environment | aft<br>forward<br>going<br>ramp<br>shop<br>started<br>stop<br>under | Skill Based Errors<br>Decision Errors<br>Perceptual Errors<br>Routine Violations<br>Exceptional Violations<br>Physical Environment<br>Adverse Mental State<br>Physical/Mental Limitations<br>Planned Inappropriate Operations | Observation<br>Planning<br>Temporary person related functions<br>Permanent person related functions<br>Equipment failure<br>Temporary interface problems<br>Organization<br>Training<br>Ambient Conditions | Time, Place, Duration<br>Situational Awareness |
| Communication | conversation<br>data<br>information<br>manager<br>captain<br>message<br>states<br>tag | Perceptual Errors<br>Routine Violations<br>Technological Environment<br>Crew Resource Management<br>Inadequate Supervision<br>Planned Inappropriate Operations<br>Failed to Correct Problem<br>Supervisory Violation<br>Organizational Processes | Observation<br>Planning<br>Temporary person related functions<br>Permanent person related functions<br>Equipment failure<br>Procedure<br>Temporary interface problems<br>Permanent interface problems<br>Communication | Personal Communication<br>Team Communication<br>Organizational/ Supervisory Communication |
| Regulatory and Safety Influences | emergency<br>FAA | Decision Errors<br>Perceptual Errors<br>Routine Violations<br>Exceptional Violations<br>Adverse Mental State<br>Adverse Physiological State<br>Personal Readiness<br>Failed to Correct Problem<br>Resource Management<br>Organizational Management<br>Organizational Processes | Interpretation<br>Temporary person related functions<br>Procedure<br>Organization<br>Training<br>Ambient Conditions<br>Working Conditions | Personal Readiness<br>Safety Culture<br>Policy Requirement |

*Note*. The HFACS and CREAM categories duplicate across multiple observed categories, while the novel framework categories provide a more concise evaluation.

The HFACS taxonomy focused on the cognitive-psychological point of view, and the extensive listing of CREAM consequents based in event phenomenology, intersect the observed categories at multiple duplicate points. However, this could lead to ambiguity during incident investigation and causal diagnosis. In counterpoint, the novel HF categories, designed specifically to address the topics extracted from the ASRS dataset, do not exhibit the same duplication and ambiguity found in the application of the two example HF frameworks.

**Summary**

This chapter presented the results of the research methodology described in Chapter III. The selected ASRS reports were reduced from a large corpus of documents to a set of topic word groupings that could be used to identify the general nature of the incidents described in the original reports. The topic word groups were not identified as being particularly relevant to either the HFACS unsafe acts or the CREAM error modes, but they did describe many aspects of aviation maintenance tasks. These results, and their potential impact on aviation maintenance will be further discussed in the next chapter.

# Chapter V: Discussion, Conclusions, and Recommendations

This chapter will discuss the results presented in Chapter IV as a result of the research method detailed in Chapter III. Following the discussion of the results presented in Chapter IV, this chapter contains a summarization of conclusions associated with the selected research method and recommendations for future research associated with the selected topics and methodologies.

## Discussion of Research Results

### *Aviation Maintenance Topic Modeling*

The first research question asks what human factors subjects are prevalent in the selected group of ASRS aviation maintenance reports. The volume of ASRS reports associated with maintenance activities would have made it impractical, if not impossible, to perform a comprehensive thematic analysis of these reports. The selected data set contained 5,929 reports submitted to the ASRS database over a span of 22 years. As described in Chapter IV, this large compendium of reports was reduced to a set of eleven topic word groups of ten words each using topic modeling and LDA.

As discussed in Chapter IV, the results of the topic modeling data reduction showed a focus on four potential categories. They are:

- Terms associated with the details and requirements of aviation maintenance tasks.
- Terms associated with locations and movement within an aviation maintenance working environment.
- Terms associated with communication and the personnel with which aviation maintenance technicians interact.

- Terms associated with the safety and regulatory structures of aviation maintenance.

Considering the report filters applied (i.e., reports submitted by aviation maintenance personnel for incidents occurring in Part 121 organizations) when acquiring the ASRS dataset from the NASA website, the observation of these four categories is not surprising. It can be reasonably expected that aviation maintenance personnel will submit reports detailing incidents relevant to their task assignments and work environments, and professional interactions.

Each topic word group displays some level of thematic coherence, and as noted by the aviation SMEs, a person with aviation maintenance familiarity could reasonably formulate a potential scenario of activities from the words in each group. Table 13 shows each topic word group, the observed categorical associations, selected key words and a potential narrative that may describe the occurrences described by the reports containing the selected topic.

**Table 13**

*Potential Narratives for Topics Using Words Within Topic Word Groups*

| Topic | Key Terms | Categorical Associations | Potential Narrative |
|---|---|---|---|
| A | looked, night, morning, didn't, know, thought, put | Task, Environment | An object was determined to be missing, and the reporter thought it should have been placed in a location the night before but it couldn't be located in the morning. |
| B | pulled, inoperable, troubleshooting, fault, message, captain | Task, Communication | The reporter was tasked to troubleshoot a fault reported by a captain that occurred when a control was pulled and a device or system was inoperable. |
| C | log, entry, writeup, routine, data | Communication | The reporter observed a shift log entry concerning data cards detailing routine items that should have been addressed overnight. |
| D | reporter, closed, leak, callback, revealed, fire, high | Communication, Task | The reporting technician had to get clarification from the initial fault reporter regarding a leak when a device was closed. The callback conversation revealed a high reading associated with a fire supression system during engine run. |
| E | ramp, taxi, started, stop, move, power | Environment | An incident occurred on the ramp during taxi movement when the the aircraft lost power and could not be started, requiring transfer to the hangar. |
| F | inspector, missing, access, torque, install | Environment, Task | An inspector found that fasteners were missing from an access panel, possibly due to improper torque during instalaltion. The panel may have been located forward or aft of an adjacent upper or lower feature. |
| G | note, AMM, main, states, accordance | Communication, Environment | A note in the Aircraft Maintenance Manual (AMM) states that a procedure on the main landing gear should be done in accordance with a particular specification. |
| H | event, occurred, issue, possible, | Task | An event occurred several times that may have involved an issue. Many times this event can include the stated issue. |
| J | report, management, FAA, inspector | Communication, Task | A report to the FAA by a quality inspector involved training information relevant to management and the air carrier. |
| K | wrong, changed, stock, tag, emergency | Task | A part was changed with the wrong type because of an incorrect tag when pulled from stock in shop, resulting in an emergency. |
| L | damaged, repair, limits, SRM | Task | A hole had to be repaired because the damage exceeded the limits stated in the Structural Repair Manual (SRM). |

*Note*. The potential narratives serve only to demonstrate that incident details can be theorized using only the words within the topic word groups.

The prevalence of task associated terminologies, which account for nearly half of the observed terms, indicates that the majority of the reports in the dataset are discussing an incident related to a deficiency of task requirements, or more simply put, cases where the resources provided to the maintenance technician were not adequate for the task assignment. In six of the eleven topics generated in the topic modeling process, a task associated word has the highest or second highest probability of occurrence within the topic indicating that, in incidents reported by maintenance personnel, task relevant themes were the most strongly emphasized in the reports.

From a standpoint of topic prevalence, and intra-topic word probability, the concepts of communication and movement within the working environment are close to equally represented. These themes highlight the importance, to aviation maintenance technicians, of interaction with work elements external to their assigned tasks, or possibly how their work environment and other aviation personnel affect the completion of their assigned task.

The lack of thematic representation observed for terms associated with safety and regulatory controls could indicate several factors. Aviation maintenance technicians may see these disciplines as beyond their areas of professional responsibility and therefore feel that reporting deficiencies in these systems would be pointless. The maintenance technicians may hold the opinion that incidents associated with these factors would be more appropriately addressed by regulatory or dedicated safety department personnel. The simplest explanation for the lack of representation of these themes within the topic word groups may be that fewer incidents occur that could be classified as being related to safety or regulatory environments.

***Evaluating Modeled Topics to CREAM and HFACS***

The second research question detailed in Chapter I concerns the way in which the subjects contained within the topic word groupings align or do not align with the phenotypical and genotypical concepts contained respectively in the CREAM and HFACS human factor classification frameworks.

Initially a Delphi analysis with a panel of seven aviation maintenance SMEs was arranged to evaluate the alignment of the topic word groupings to the selected HF frameworks. As described in Chapter IV, this method to obtain SME consensus was unsuccessful. A failure to garner a significant level of consensus, and an overwhelming presence of response set bias in the quantitative data required the termination of the study after the second round. These issues may have been caused by the length of the survey, the format of the online survey, or the diverse background of the aviation maintenance SMEs agreeing to participate in the research activity.

This question was addressed subsequently by a panel of four aviation maintenance SMEs in a virtual roundtable discussion as described in Chapter IV. The group of participants agreed unanimously that the topic word groupings did not significantly align with either of the HF frameworks, but HFACS would be a better choice for investigation of the incidents that could be theorized from the topic word groupings. The SME panel noted that the single word listing lacked the contextual richness that could be found in the full text reports. This statement was based on the past experiences of the SMEs in both incident investigation and academic research settings.

Additionally, the panel of SMEs hypothesized that the disconnection between the topic word groupings and the HF frameworks could have been exacerbated by a lack of commonality in reporting methodologies, data requirements, and linguistic styles. The ASRS reporting format places little limitation on what information must be included in the reports. The categorical entries are not limited to a single selection, and the open text responses are not limited in format or content. NASA requests only that reports are filled out as comprehensively as possible. Although this free formatting may not be conducive to the comprehensiveness or consistency of the reports submitted, it is likely essential for the success of the ASRS program. The submission of reports to the ASRS program is voluntary, confidential, and non-punitive (ASRS, 2023), and therefore strict control in the content of submissions could have a discouraging effect on those who would otherwise wish to submit reports.

The comments from the aviation maintenance SMEs, both regarding the ability to theorize events, and in the lack of alignment to the selected HF frameworks appear to indicate that a framework designed around the content of the reports solicited under the existing ASRS methodology could enhance the investigative process while maintaining the rich contextual format of the reports submitted by aviation maintenance personnel.

### *A Novel Human Factor Framework for Aviation Maintenance*

The final research question is about what improvements can be made to incident recorder/investigator communication in the form of a novel human factor conceptual framework. The topic modeling results of this research, and the opinions given by the maintenance SMEs indicate that, although the HFACS taxonomy is a better choice for incident investigation, the information recorded in the ASRS reports do not align with either a phenotypical or a genotypical theoretical framework. Both Shappell and Wiegmann’s (2000) HFACS, and Hollnagel’s (1998) CREAM touch on the influence that task-based resource allocation has on the erroneous actions of the operator, particularly at the organizational, supervisory, and environmental levels. As illustrated in Chapter II, Figure 6, both of these conceptual frameworks contain reference to organizational influences, environmental contributors, and supervisory oversight.

In terms of organizational responsibility, there is an agreement in both Taylor's (2004) bottom-up scientific management theory, and Fayol and Storrs' (2013) top-down leadership approach that it is the clear responsibility of the organization and its managers to provide employees with the policies, environment, and tools needed to complete the tasks assigned. Coupling this with McGregor's (2006) humanistic management theory, which says that given the needed resources, employees will strive to perform in a manner that benefits both themselves and the organization, it can be inferred that the risk of human error incidents can be minimized if the organization provides the resources required for the assigned task.

With consideration of the themes within the topic model groupings shown in Chapter IV, Table 9, the input from the aviation maintenance SME group, the nature of maintenance tasks described in the published literature, and the organizational theories of Taylor (2004), Fayol and Storrs (2013), and McGregor (2006), there is indication that the novel framework should be focused on the needs, requirements, and expectations of the task in which the operator was engaged or assigned to when the incident occurred.

As demonstrated in the review of the extant literature in Chapter II, there is "an already burgeoning list of human error taxonomies" (Shappell & Wiegmann, 2000, p. 13) and as further noted by Shappell and Wiegmann (2000), there is little purpose or practicality in contributing to this if the candidate classification system is not useful to the operator. Furthermore, the published literature makes it apparent that the candidate classification framework should be applicable and useful to both incident reporter and incident investigator. An incident report is a communication tool that must capture specific details of the occurrence, and, subsequently, be generalizable to a classification taxonomy if it is to be of use in preventing future occurrences in a similar category. Therefore, the selected classification system should exhibit a balance of comprehensiveness and ease of use (Shappell & Wiegmann, 2000).

The US Army training command developed a remarkably simple task-based framework for infantry training as a part of a force restructuring initiative in the late 1970s. Their framework consisted of operational training tasks required to ensure a soldier's ability to complete an assigned mission. These training tasks are arranged under four main group headings: shoot, move, communicate, survive. By grouping a soldier's training modules around these four main categories, the Army's training command could ensure that the needed training was delivered to the soldier in order to complete their battle assignments (Savage-Knepshield et al., 2011). Adaptation to aviation maintenance requires that this simple framework be restructured based on the basic nature of the categories, and organizational responsibilities rather than their derivations for a combat environment. Therefore, the Army categories are generalized for aviation maintenance organizations:

- *Shoot* translates to *Task Execution*. These are the things that the operator or technician needs in order to complete an assigned task.

- *Move* translates to *Control of Environment*. These requirements detail the aspects of the environment in which the subject operates and their behaviors within it.
- *Communicate* translates to *Communication*. These categories describe the communication that the operator or technician has with the organization, its supervisors, and the operator's coworkers and teammates.
- *Survive* translates to *Safety*. These items encompass factors related to personal readiness, safety culture, and regulatory policies that influence operator safety and the safety of those in contact with them.

In adapting this relatively simple list of military mission assurance categories, it became apparent that the specificity of the four main categories was insufficient to ensure the needed comprehensiveness and that sub-tier categories were required. These sub-tier categories are intended to add the needed umbrella of comprehensiveness of the developed framework without compromising the ease of use. The four parent classifications are enhanced, in terms relatable to both operator and organizational responsibilities, as follows:

- Task Execution
  - ➢ Task Requirements- This classification refers to the availability, functionality, and accuracy of the tools, technology, and team required for successful execution of the designated task.
  - ➢ Training- The training provided to the operator should be evaluated for comprehensive content specific to the task. Peer and supervisory oversight

are needed to ensure that the training was enacted and understood in an effective manner.

- Task Definition- The assigned task should include clear expectations for success, an appropriate level of detail in work instruction, and an unambiguous method of verification to validate successful completion.

- Control of Environment
    - Time, Place, and Duration- The operator should have a workspace that is appropriate for the completion of the assigned task. In addition to the physical characteristics of the workspace, this should include the time needed to complete the task and a workload appropriate to the complexity of the task.
    - Situational awareness- The operator should be able to perceive the conditions of the workspace, comprehend their own position within the workspace, and be able to reasonably project the situational occurrences in the immediate future. These requirements are similar to those described by Endsley & Jones (2012).
- Communication
    - Personal Communication- The operator should be required and reasonably able to communicate their own location, assignment, and activities to those in the immediate vicinity.
    - Team Communication- The operator should be readily able to send and receive communications, written, verbal, or multi-media, regarding their

awareness of their coworker's location and activity, the group's status, and the status of work assigned to the group.

  - Organizational and Supervisory Communication- The operator should be periodically made aware of the organization's progress on relevant issues, problems incurred by the organization that are relevant to the operator's task assignments, and changes in organizational policy and procedures that affect the operator's task assignments.

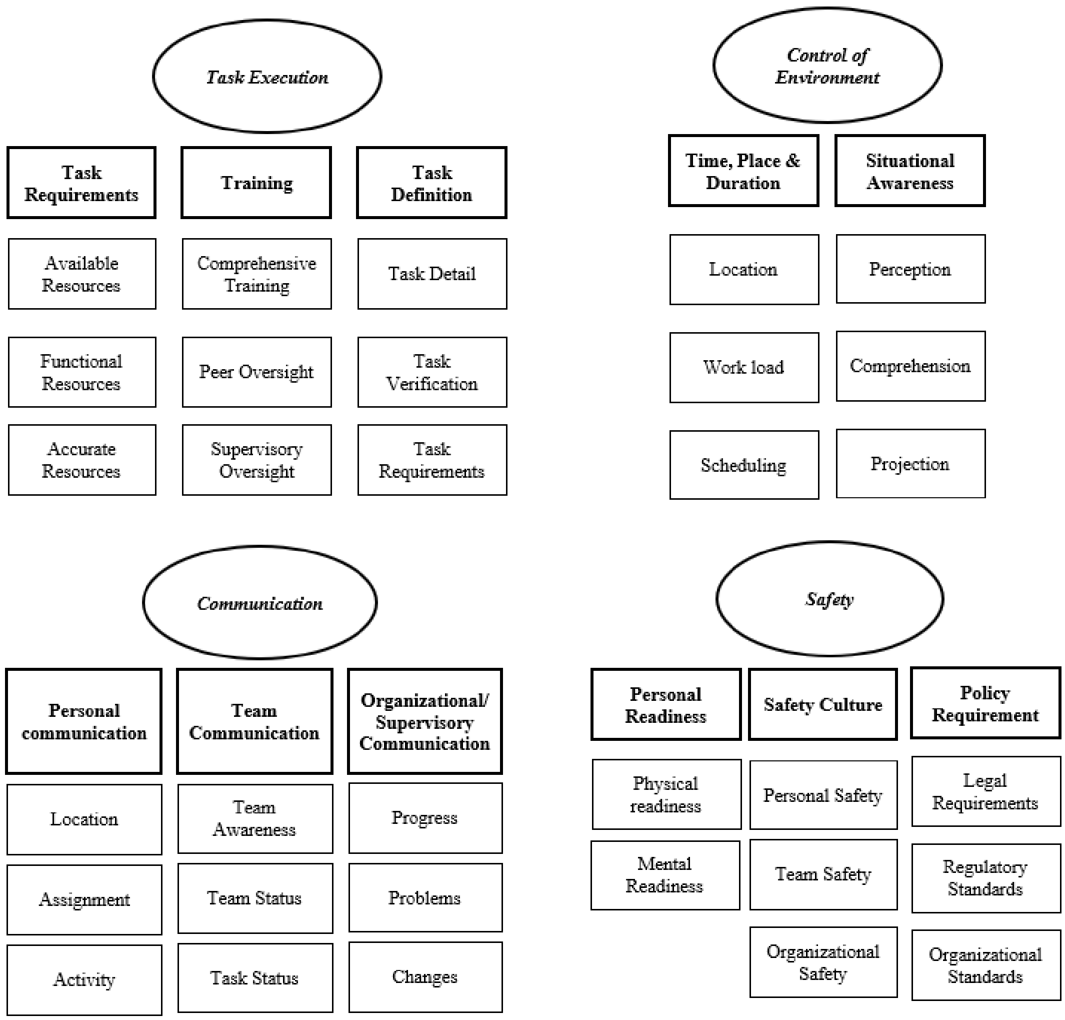
- Safety
  - Personal Readiness- The operator should, at all times, maintain a state of physical and mental readiness appropriate to the demands and complexity of the assigned task.
  - Safety Culture- The organization should develop and maintain a robust safety culture that fosters a comprehensive level of personal, team, and organizational safety. The operator should be aware of, and embrace, the organization's safety initiatives.
  - Policy Requirements- The organization shall analyze, interpret, and adhere to all legal requirements, regulatory standards, and organizational policies in such a way that will allow the operator to complete their task without violation of these requirements, standards, and policies.

This simple listing of task requirements, and therefore categories that could be associated with task failure, were developed through a review of the extant literature in human factors and industrial management theory. Each viewpoint in these fields of study contains references to how a work task should be supported, analyzed, and enacted.

While many theories overlapped and contained broad commonalities, there was no evidence that a framework specifically dedicated to task needs and requirements had been considered (Endsley & Jones, 2012; Hackman & Oldham, 1976; Hollnagel, 1998; McGregor, 2006; Rankin, 2000; Reason, 1990; Reason & Hobbs, 2003; Salas et al., 2005; Shappell & Wiegmann, 2000). The Task, Environment, Communication, and Safety (TECS) framework for aviation maintenance is shown in Figure 23.

**Figure 23**

*Proposed Aviation Maintenance TECS Classification Framework*




*Note*. This framework is presented in a box format to accommodate page dimensions. There is no hierarchical distinction between the four parent categories.

As noted by Reason (1990), failures in complex systems are seldom the result of a single fault, so it would be expected that application of this proposed framework would include the identification of more than one contributing factor. Each failure identification in a particular subcategory should be investigated as a separate causal factor with specific mitigation strategies developed to reduce the risk of reoccurrence. As noted in the literature review, investigation of aviation maintenance incidents relies heavily on review of archival reports (Rouse & Rouse, 1983). This is at least partially attributable to the latent nature of discovery of the recorded incidents (Saward & Stanton, 2015). Although the reporting methodology limits the investigator's ability to obtain clarifying details about the incident, these archival reports contain rich contextual data that illustrates the circumstances surrounding the event (Miles et al., 2014), therefore it is important that an analysis framework provides categorization based on the information included in the original report. Additionally, input from the aviation maintenance SMEs indicates there are frequent occurrences when the reporting individual or the operator involved in the incident cannot be made available to the investigator for clarifying questions or additional information about the incident. Therefore, it is essential that the records of the events in the incident report contain the appropriate information needed by incident investigators to discern the causal factors of the incident.

## Conclusions

This exploratory research study focused on human factors in aviation maintenance. This study asked what human factor subjects were present in the aviation maintenance incident descriptions stored in the ASRS database. Because a review of the published literature indicated a division in human factor schools of thought, this research further asked to which, if either, of the two prevalent schools of thought, the human factor subjects discovered in the maintenance incident reports were more applicable. Finally, this research asked if the alignment of the human factor subjects in the aviation maintenance reports to a human factor framework could be improved upon.

The size of the selected corpus of ASRS reports precluded a comprehensive review of each report. Therefore, a research methodology was developed using an NLP methodology to extract topic word groups from the ASRS dataset based on the mathematical probability of words within topic groups and topic groups within report documents to determine the content of the topic word groups. Although the topic word groups could be extracted based on mathematical probability, the topic word groups comprised of subjective contextual bits of language required SME evaluation to establish the degree of relevance to the selected human factor conceptual frameworks.

***Theoretical Contributions***

By focusing on the three research questions asked, the research method developed for this study was able to demonstrate the effective use of NLP techniques to reduce a large and indecipherable, at least in in terms of human comprehension, body of qualitative data related to aviation maintenance incidents to a focused set of topic word groups that could be readily analyzed for content by human subject matter experts.

Maintenance is an underserved sector of the aviation industry. And, although there is no shortage of human factor frameworks that have been developed through rigorous research and analysis, very few of these established frameworks deviate from the existing paradigms developed a quarter of a century ago. The thematic content of the topic word groups developed through analysis of the selected ASRS data showed limited relevance to the established human factor frameworks noted in the extant literature. The discovery, evaluation, and analysis of the ASRS maintenance topic word groups facilitated the proposal of a novel human factor framework focused on the requirements of aviation maintenance tasks, and the responsibility of the employing organization to facilitate the success of assigned tasks. As an analysis of the linguistic styles used by aviation maintenance incident reporters, this research fills a gap in the aviation maintenance body of knowledge. In terms of the human factor body of knowledge, this research provides a novel point of view that blends organizational management theory with human factor analysis, a paradigm that was not observed in the extant published literature. The purpose of this research did not include casting doubt on the efficacy of existing human factor frameworks. The proposal of a novel framework that bridges the communication gap between inspector and report enhances the existing body of knowledge and provides an alternative to the established theoretical paradigm. The connection of salient concepts in organizational management and human factors will allow future researchers to develop additional conceptual frameworks focused on the theoretical concepts realized in this research. This focus can enhance aviation safety by directing mitigations towards areas that were previously ignored or overlooked as minor contributors.

Although, due to the flexibility of NLP and customization options available to the algorithms, the research methodology is likely generalizable to virtually any large body of qualitative reports, the results of this research are specific to a narrow segment of the commercial aviation industry. The novel conceptual framework views the human factor contributors through a lens of organizational responsibility and therefore would not be generalizable to activities not under the supervision and direction of an organization with the scope and resources of a Part 121 air carrier.

Dissemination of this research and its realizations for aviation maintenance and human factors will facilitate additional studies on these and similar concepts, providing enlightenment on unique challenges faced by aviation maintenance organizations and the technicians they employ. The demonstration of the research methodology opens the possibility for other researchers to apply a similar methodology to diverse disciplines that have massive quantities of qualitative data with little or no awareness of topical content.

### *Practical Contributions*

The practical contributions of this research are, primarily and intentionally, focused on airline organizations and their practitioners of aviation maintenance. Analysis of the observed topic subjects can raise the awareness of the communication gaps that exist between maintenance incident reporters and incident investigators. Knowledge of the topical themes within maintenance incident reports will improve the quality of incident investigations through improved training to incident investigators and incident reporters.

The Aviation Maintenance TECS framework can be used retrospectively by maintenance organizations to develop targeted mitigations for risks associated with observed events, or proactively in Failure Mode Event Analysis (FMEA) or similar predictive methods to assess the risk of incidents before they occur. The TECS framework can be applied as a foundation for a standardized incident reporting and investigation tool to improve the quality and efficiency of these activities.

### *Limitations of the Findings*

The limitations of the described research are largely driven by the delimitations put in place to manage the scope of the study. Aviation maintenance activities occur in civil aviation settings, in military environments, Maintenance, Repair, and Overhaul (MRO) organizations, corporate organizations, and flight schools. While there are undoubtedly similarities that occur in these diverse maintenance disciplines, it was necessary to manage the scope of the data to ensure a level of commonality in environment and culture. As of this delimitation to Part 121 maintenance activities, the findings of the research are similarly limited specifically to maintenance activities performed under the supervision and direction of US based Part 121 airline organizations. Similarly, the application of the described Aviation Maintenance TECS framework, because of the focus on organizational influences, is only relevant to maintenance activities performed under the direction of a managed organizational structure.

The two conceptual human factor factor frameworks chosen as exemplars of the theoretical schools of thought were published and popularized in roughly the same timeframe. While it may be unlikely that these publications had significant influence on the verbiage and semantic styles of the incident reporters at the time of publication, the use of these two frameworks as reference categorizations required the dataset to be delimited to their time of active use. The selection of the reporting years from 2000 to 2022 reflects this delimitation, and therefore limits the generalizability of the research findings to the stated timeframe.

NASA maintains the ASRS database as a voluntary, confidential, and non-punitive incident reporting tool with wide public accessibility (NASA, 2023). These parameters are intended to encourage open reporting, but there are limitations. The reports within the ASRS database have the potential for numerous forms of reporter bias. Although the assumption must be made that the reports are submitted honestly and without ulterior motive, report submissions may be biased in favor of the reporter and the intent to submit a report may be based on a desire to insulate a reporter from regulatory enforcement or legal reprecussions. Reporter bias may also be attributabal to a lack of verification of report submissions. The reports are subjective and representative of the reporter's perceptions. The report formatting is loosely controlled and the level of detail within reports is incionsistent, depending only on the effort and recollection of the reporter. The NASA ASRS Program Briefing (2023) prohibits submission of incidents involving known criminal activities and major aircraft accidents. This restriction implies that any incident not included in the ASRS database has the potential to be sufficiently divergent in nature and scope from the reported incidents to a degree that the results of the described research would not apply. Because NASA prohibits the submission of reports associated with major aircraft accidents, association of the TECS framework to major accidents cannot be implied. The results of the described research are strictly limited to minor incidents. Therefore, the findings of the described research are limited to the reports submitted to NASA as part of the ASRS dataset with consideration given to the limitations driven by the program parameters and reporting format.

Due to the subjectivity of the decisions made in the data preprocessing and topic refinement processes applied in preparation of the LDA calculations, and the calculation of posterior parameters of the Gibbs analysis, future research findings could be affected by variations in these processes. Blei et al. (2003) and Miyamoto et al. (2022) provide guidance on the effects of parameter variation and were therefore instructive in the described research. Subjectivity was managed in the described research by seeking SME opinion on topic content, and by process variation through evaluation of expanded and reduced topic and word counts. Although SME opinion was valuable in assessing the effectivity of the topic refinement process, further limitations to the results arise from the subjective nature of the opinions given by the SME evaluators when establishing a consensus of the alignment of the topic word groups to the exemplar HF conceptual frameworks. This subjectivity limits the described research results to the topic word groups presented as the output of the LDA process. The selection of the SME roundtable participants based on their experience and familiarity with incident investigation and aviation maintenance, as well as the inclusion of multiple participants, were important factors in reducing subjectivity and opinion biases inherent in the SME evaluation process.

## Recommendations

The described exploratory research opens a number of possibilities to confirm the designed research methodology in similar aviation fields, to expand the application of the designed research methodology to other areas with similar corpora of archival documentation, and for the application of the Aviation Maintenance TECS framework for improvement in the process of minor incident investigation.

### *Recommendations for Aviation Maintenance*

The described research was intended specifically to gain a better understanding of how aviation maintenance incidents that occur under the supervision of airline maintenance facilities are described by the maintenance personnel who submit reports. The conclusions noted and the maintenance TECS framework may provide valuable guidance for maintenance organizations that want to have a better understanding of how incidents are described by their maintenance technicians. Specific recommendations are offered for these organizations based on the findings of the research.

**Recommendation #1.** A conceptual comparison of the extant literature and the thematic content of the topic word groups extracted from the ASRS data indicates a communication gap between aviation maintenance incident reporters and incident investigators. This was reinforced by SME assertions that the information required to effectively investigate an incident is seldom included in the archival narrative provided by the incident reporter. Therefore, aviation maintenance organizations should, preferably in a collaborative effort led by a regulatory body such as the FAA, develop an integrated system of awareness training and reporting tools for maintenance technicians and incident investigators. This common system of reporting and investigation will provide a common set of references for technicians and investigators, including best practices for minimum reporting informational standards and preferred investigational techniques.

**Recommendation #2.** Aviation maintenance organizations should validate the TECS framework using in-house incident data to determine how maintenance employees describe incidents and what verbiage or semantic styles they use when describing incidents within their work areas. This activity can be performed using the described research methodology as a model, or through direct evaluation of incident reports in lieu of the topic modeling data reduction method. This will give the organization the opportunity to determine the validity of the TECS framework for their own organization, and to determine what refinements may be needed to ensure relevance to local operations.

**Recommendation #3.** Airline maintenance organizations should implement a system of retrospective analysis for maintenance incidents using the maintenance TECS framework as a basis of categorization. The generalizability of the TECS framework under an organizational structure should facilitate a variety of investigational techniques commonly used in continuous improvement and risk management methods. This will allow the organization to discern the impact of the human and organizational factors noted in the TECS framework on safety performance. Subsequently, these impacts can be used to suggest targeted interventions aimed at preventing recurrence of the noted incidents due to the identified human factor and organizational contributors.

**Recommendation #4.** Airline maintenance organizations should implement a proactive risk analysis system focused on the human factor and organizational contributors detailed in the TECS framework. This can be accomplished by adapting current job hazard analysis of FMEA methodologies to highlight the concepts within the TECS framework. This proactive methodology will allow airline maintenance organizations to identify effective risk mitigations that leverage the organization's existing resources, safety culture, and policies, thus improving organizational safety performance.

### *Recommendations for Future Research Methodology*

The challenges encountered during the described research open several opportunities for improvements in the research methodology. The subjectivity of the SME agreement regarding data preprocessing and topic refinement should be addressed by the establishment of a standardized taxonomy of HF relevant terms. Future studies employing standardized taxonomy will have better reliability and generalizability to a larger population of HF associated research targets.

The LDA process employed for this described research treated the word tokens in each report as a *bag of words*, meaning the words were considered individually when establishing word within topic probabilities. As noted in the SME topic evaluation process, the single word methodology compromises a level of word context. Miyamoto et al. (2022) describe topic modeling alternatives to the bag of words methodology using *n-grams* of 1-4 words evaluated by the probability that words occur within the proximity of other words. Using the *n-gram* methodology while replicating the remainder of the research methodology could improve the contextual richness of the topic modeling process without sacrificing topic viability.

Finally, it is recommended that application of the Delphi analysis method using online survey tools be approached with caution. The lengthy and repetitious process of evaluating groups of word topics over multiple consensus seeking rounds led, in this research, to participant fatigue and a lack of objectivity. Use of a virtual roundtable method in an online meeting platform proved to be much more effective than the Delphi process, which was discontinued before an acceptable level of consensus could be established.

### *Recommendations for Future Research*

The described research combines applications in NLP, qualitative evaluation, and SME consensus to formulate a novel framework that describes task-based contributors to human error under the influence of large organizational structures. The TECS framework has potential generalizability to a wide range of disciplines. Although the described research focuses on the underserved field of aviation maintenance, the framework may be relevant to activities performed in any complex managerial organization that collects narrative data of a volume sufficient for thematic analysis.

The presentation of the Maintenance TECS framework in this described research is the result of exploratory, qualitative analysis. A confirmatory research study is warranted to evaluate the framework and its validity. The confirmatory analysis should consist of a content analysis and coding methodology using a set of aviation maintenance incident reports that were not included in this previous exploratory research. The confirmatory analysis should include a blind multiple SME coding methodology by at least two SME participants. Coding disagreements can be addressed by a third SME referee in consensus sessions. This methodology will allow for the evaluation of inter-rater reliability measured by the Cohen's Kappa statistic and evaluations of construct validity through SME feedback. The SME consensus meetings will also facilitate refinement of the proposed categorical definitions, thereby increasing the usefulness of the proposed TECS framework. This described methodology has proved to be effective in the initial development of the HFACS taxonomy (Shappell & Wiegmann, 2000), as well as evaluations of the HFACS taxonomy to additional practical applications (Tvaryanas & Thompson, 2008; Berry et al., 2010; Diller et al., 2014).

The qualitative nature of the described research will contribute to the complexity of a subsequent confirmatory analysis. Quantitative evaluations in LDA using measures of perplexity and topic coherence as described by Blei et al. (2003) may prove useful in confirming the model generated in the described research. Additionally, Shappell and Wiegmann (2000) and Cohen et al. (2015) provide guidance on using a variety of statistical measures to quantify framework validity and inter-rater reliability of the HFACS taxonomy. Additionally, quantitative evaluations in LDA using measures of perplexity and topic coherence as described by Blei et al. (2003) may prove useful in confirming the model generated in this research.

The proposed maintenance TECS framework developed for this study is directly applicable to aviation maintenance incident reports for activities occurring under the influence of Part 121 airline maintenance activities. There are opportunities to apply the described research methodology and the maintenance TECS framework to other sectors of the aviation industry. The TECS framework has potential generalizability to FAR Part 145 MRO operations, aerospace manufacturing organizations, and other high risk, high reliability disciplines in transportation and power generation.

Data reduction methods, using AI applications of NLP, as described in this research have been validated in a number of fields with large volumes of stored qualitative data. Blei et al. (2003) for instance, describe applications in publication, advertising, and entertainment. While this research focused on a limited sector of aviation maintenance, it is recommended that similar studies are initiated to analyze the large volumes of data collected by other sectors of the aviation industry.

**References**

Ajzen, I. (2005). *Attitudes, personality and behaviour*. McGraw-Hill Education.

Armstrong, G., Vijayakumar, L., Cherian, A. V., Krishnaswamy, K., & Gilligan, C. (2020). "It's a battle for eyeballs and suicide is clickbait": The media experience of suicide reporting in India. *PloS One*, *15*(9). https://doi.org/10.1371/journal.pone.0239280

Arun, R., Suresh, V., Veni Madhavan, C. E., Narasimha Murthy, M. N. (2010). On finding the natural number of topics with latent Dirichlet allocation: Some observations. In M. J. Zaki, J. X. Yu, B. Ravindran, & V. Pudi (Eds.) *Advances in Knowledge Discovery and Data Mining. PAKDD 2010. Lecture Notes in Computer Science, 6118* (pp. 391-402). Springer. https://doi.org/10.1007/978-3-642-13657-3_43

Asher, N., Hunter, J., & Paul, S. (2021). Bias in semantic and discourse interpretation. *Linguistics and Philosophy*, *45*(3), 393-429. https://doi.org/10.1007/s10988-021-09334-x

Berry, K. A., Stringfellow, P. F., & Shappell, S. A. (2010). Examining error pathways: An analysis of contributing factors using HFACS in non-aviation industries. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *54*(21), 1900-1904. https://doi.org/10.1177/154193121005402115

Benoit K, Watanabe K, Wang H, Nulty P, Obeng A, Müller S, Matsuo A (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software* *3*(30), 774. https://doi.org/10.21105/joss.00774

Benoit K, Watanabe K, Wang H, Nulty P, Obeng A, Müller S, Matsuo A (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software.* *3*(30), 774. https://doi.org/10.21105/joss.00774

Birko, S., Dove, E. S., & Özdemir, V. (2015). Evaluation of nine consensus indices in Delphi foresight research and their dependency on Delphi survey characteristics: A simulation study and debate on Delphi design and interpretation. *PloS One*, *10*(8). https://doi.org/10.1371/journal.pone.0135162

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993-1022. https://jmlr.org/papers/v3/blei03a.html

Blei, D. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84. https://doi.org/10.1145/2133806.2133826

Bouchet-Valat M (2023). *SnowballC: snowball stemmers based on the C 'libstemmer'*(Version 0.7.1) [Software package], https://CRAN.R-project.org/package=SnowballC

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing* (Amsterdam), *72*(7), 1775-1781. https://doi.org/10.1016/j.neucom.2008.06.011

Chang J (2015). *lda: Collapsed Gibbs sampling methods for topic models*. (Version 1.4.2), [Software package] https://CRAN.R-project.org/package=lda

Chatzi, A. V., Martin, W., Bates, P., & Murray, P. (2019). The unexplored link between communication and trust in aviation maintenance practice. *Aerospace*, *6*(6). https://doi.org/10.3390/aerospace6060066

Chiu, M., & Hsieh, M. (2016). Latent human error analysis and efficient improvement strategies by fuzzy TOPSIS in aviation maintenance tasks. *Applied Ergonomics*, *54*, 136-147. https://doi.org/10.1016/j.apergo.2015.11.017

Clare, J. & Kourousis, K. I. (2021). Learning from Incidents in Aircraft Maintenance and Continuing Airworthiness Management: A Systematic Review. *Journal of Advanced Transportation*, *2021*, 1-13. https://doi.org/10.1155/2021/8852932

Cohen, T. N., Wiegmann, D. A., & Shappell, S. A. (2015). Evaluating the reliability of the human factors analysis and classification system. *Aerospace Medicine and Human Performance*, *86*(8), 728-735. https://doi.org/10.3357/AMHP.4218.2015

Dekker, S. W. A. (2017). Rasmussen's legacy and the long arm of rational choice. *Applied Ergonomics*, *59*(Pt B), 554-557. https://doi.org/10.1016/j.apergo.2016.02.007

Deming, W. E. (1992). *Out of the crisis* (18th pr. ed.). Cambridge University Press.

Diller, T., Helmrich, G., Dunning, S., Cox, S., Buchanan, A., & Shappell, S. (2014). The Human Factors Analysis Classification System (HFACS) applied to health care. *American Journal of Medical Quality*, *29*(3), 181-190. https://doi.org/10.1177/1062860613491623

Dupont, G. (1997). The Dirty Dozen errors in aviation maintenance. In *The 11th Federal Aviation Administration meeting on human factors issues in aircraft maintenance and inspection: Human error in aviation maintenance* (pp. 45-49). Federal Aviation Administration.

Endsley, M. R., & Jones, D. G. (2012). *Designing for situation awareness* (2. ed. ed.). CRC Press.

Ergai, A., Cohen, T., Sharp, J., Wiegmann, D., Gramopadhye, A., & Shappell, S. (2016). Assessment of the Human Factors Analysis and Classification System (HFACS): Intra-rater and inter-rater reliability. *Safety Science*, *82*, 393-398. https://doi.org/10.1016/j.ssci.2015.09.028

Escalante, E. (1999). Quality and productivity improvement: A study of variation and defects in manufacturing. *Quality Engineering*, *11*(3), 427-442. https://doi.org/10.1080/08982119908919259

Evans, D. (2003). Hierarchy of evidence: A framework for ranking evidence evaluating healthcare interventions. *Journal of Clinical Nursing, 12*(1), 77-84. https://doi.org/10.1046/j.1365-2702.2003.00662.x

Fayol, H., & Storrs, C. (2013). *General and industrial management*. Martino Publishing

Federal Aviation Administration. (2023). Become an aviation mechanic. United States Department of Transportation. https://www.faa.gov/mechanics/become

Feinerer I, Hornik K (2023). *tm: Text Mining Package*. (Version 0.7-11), [Software package], https://CRAN.R-project.org/package=tm

Fishbein, M., & Ajzen, I. (1980). *Belief, attitude, intention, and behavior* (4th ed.). Addison-Wesley.

Freud, S. (1989). *The psychopathology of everyday life* (J. Strachey, Ed. & Trans.). WW Norton & Co. (Original work published 1904)

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, *101* (suppl. 1), 5228-5235. https://doi.org/10.1073/pnas.0307752101

Grolemund, G. & Wickham, H. (2011). Dates and Times Made Easy with lubridate. *Journal of Statistical Software*, 40(3), 1-25. https://doi.org/10.18637/jss.v040.i03

Grün B, Hornik K (2023). *topicmodels: Topic Models*. (Version 0.2-14), [Software package], https://CRAN.R-project.org/package=topicmodels

Hackman, J.R., Oldham, G,(1976) Motivation through the design of work: test of a theory, *Organizational Behavior and Human Performance,16*(2) 250-279, https://doi.org/10.1016/0030-5073(76)90016-7.

Hansen, F. D. (2006). Human error: A concept analysis. *Journal of Air Transportation*, *11*(3).

Harris, D., & Li, W.-C. (2011). An extension of the Human Factors Analysis and Classification System for use in open systems. *Theoretical Issues in Ergonomics Science*, *12*(2), 108-128. https://doi.org/10.1080/14639220903536559

Hawkins, F. H. (2017). *Human factors in flight* (2nd ed.). Taylor and Francis. https://doi.org/10.4324/9781351218580

Hobbs, A., & Kanki, B. G. (2008). Patterns of error in confidential maintenance incident reports. *The International Journal of Aviation Psychology*, *18*(1), 5-16. https://doi.org/10.1080/10508410701749365

Hobbs, A., & Williamson, A. (2003). Associations between errors and contributing factors in aircraft maintenance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *45*(2), 186-201. https://doi.org/10.1518/hfes.45.2.186.27244

Hollnagel, E. (1993). The phenotype of erroneous actions. *International Journal of Man-Machine Studies*, *39*(1), 1-32. https://doi.org/10.1006/imms.1993.1051

Hollnagel, E. (1998). *Cognitive reliability and error analysis method: CREAM*. Elsevier Science.

Hollnagel, E. (2012). *FRAM, the functional resonance analysis method: Modelling complex socio-technical systems*. Ashgate.

Hollnagel, E. (2021). *CREAM - Cognitive Reliability and Error Analysis Method*. erikhollnagel.com. https://erikhollnagel.com/ideas/cream.html

Hornik, K. (2020). *NLP: Natural Language Processing Infrastructure.* (Version 0.2-1), [Software package], https://CRAN.R-project.org/package=NLP.

Irwin, A., Taylor, S., Laugerud, E., & Roberts, D. (2016). Investigating non-technical skills in Scottish and English aircraft maintenance teams using a mixed methodology of interviews and a questionnaire. *The International Journal of Aviation Psychology*, *26*(3-4), 105-119. https://doi.org/10.1080/10508414.2017.1319734

Keeney, S., Hasson, F., & McKenna, H. (2006). Consulting the oracle: Ten lessons from using the Delphi technique in nursing research. *Journal of Advanced Nursing*, *53*(2), 205-212. https://doi.org/10.1111/j.1365-2648.2006.03716.x

Kuhn, K. D. (2018). Using structural topic modeling to identify latent topics and trends in aviation incident reports. *Transportation Research, Part C, Emerging Technologies*, *87*, 105-122. https://doi.org/10.1016/j.trc.2017.12.018

Lange, T., Kopkow, C., Lützner, J., Günther, K.-P., Gravius, S., Scharf, H.-P., Stöve, J., Wagner, R., & Schmitt, J. (2020). Comparison of different rating scales for the use in Delphi studies: Different scales lead to different consensus and show different test-retest reliability. *BMC Medical Research Methodology*, *20*(1), 28–28. https://doi.org/10.1186/s12874-020-0912-8

Le Coze, J. C. (2015). Reflecting on Jens Rasmussen's legacy. A strong program for a hard problem. *Safety Science*, *71*, 123-141. https://doi.org/10.1016/j.ssci.2014.03.015

Le Coze, J. (2017). Reflecting on Jens Rasmussen's legacy (2) behind and beyond, a ‘constructivist turn’. *Applied Ergonomics, 59*(Pt B), 558-569. https://doi.org/10.1016/j.apergo.2015.07.013

Leveson, N. (2015). A systems approach to risk management through leading safety indicators. *Reliability Engineering & System Safety*, *136*, 17-34. https://doi.org/10.1016/j.ress.2014.10.008

Lower, M., Magott, J., & Skorupski, J. (2018). A system-theoretic accident model and process with Human Factors Analysis and Classification System taxonomy. *Safety Science, 110*(Part A), 393-410. https://doi.org/10.1016/j.ssci.2018.04.015

Mayo, E. (1930). Changing methods in industry. *Personnel Journal*, *8*(5), 326.

McGregor, D. (2006). *The human side of enterprise*. McGraw-Hill Education

Mellema, G. M. (2018). *Application of Dupont's Dirty Dozen framework to commercial aviation maintenance incidents* [Doctoral dissertation, Embry-Riddle Aeronautical University].Scholarly Commons. https://commons.erau.edu/edt/477

Miles, M. B. Huberman, A. M., Saldana, J. (2014). *Qualitative data analysis: A methods sourcebook* (3rd ed.). Sage.

Miyamoto, A., Bendarkar, M. V., & Mavris, D. N. (2022). Natural Language Processing of aviation safety reports to identify inefficient operational patterns. *Aerospace*, *9*(8). https://doi.org/10.3390/aerospace9080450

Müller, K., Wickham, H. (2023). *tibble: Simple Data Frames*. (Version 3.2.1), [Software package], https://CRAN.R-project.org/package=tibble

National Aeronautics and Space Administration. (2023). *ASRS Program Briefing*. https://asrs.arc.nasa.gov/docs/ASRS_ProgramBriefing.pdf

Nikita, M. (2020). *ldatuning: Tuning of the Latent Dirichlet Allocation Models Parameters*. (Version 1.0.2), [Software package] https://CRAN.R-project.org/package=ldatuning

Norman, D. A. (1981). Categorization of action slips. *Psychological Review*, *88*(1), 1-15. https://doi.org/10.1037/0033-295X.88.1.1

Okoli, C., & Pawlowski, S. D. (2004). The Delphi method as a research tool: An example, design considerations and applications. *Information & Management*, *42*(1), 15-29. https://doi.org/10.1016/j.im.2003.11.002

Ooms, J. (2023). *writexl: Export Data Frames to Excel 'xlsx' Format*. (Version 1.4.2), [Software package] https://CRAN.R-project.org/package=writexl

Operating Requirements: Domestic, Flag, and Supplemental Operations, 14 C.F.R. § 121 (2023). https://www.ecfr.gov/current/title-14/chapter-I/subchapter-G/part-121

Pan, X., Lin, Y., & He, C. (2017). A Review of cognitive models in human reliability analysis. *Quality and Reliability Engineering International*, *33*(7), 1299-1316. https://doi.org/10.1002/qre.2111

Pereira, F. C., Rodrigues, F., & Ben-Akiva, M. (2013). Text analysis in incident duration prediction. *Transportation Research, Part C, Emerging Technologies*, *37*, 177-192. https://doi.org/10.1016/j.trc.2013.10.002

Perrow, C. (1999). *Normal accidents: Living with high-risk technologies* (2nd ed.). Princeton University Press. https://doi.org/10.1515/9781400828494

Phillips, R. O., & Sagberg, F. (2014). What did you expect? CREAM analysis of hazardous incidents occurring on approach to rail signals. *Safety Science*, *66*, 92-100. https://doi.org/10.1016/j.ssci.2014.02.003

Plato. (1990). *Protagoras* (B. Jowett, Trans.). C. Scribner's Sons. (Original work published ca. 380 B.C.E.)

Pope, A. (1903). Essay on criticism. In M. Wantilove (Ed.), *Dent's Temple Series of English Texts*. J. M. Dent & Co.

R Core Team. (2023). *R: Language and environment for statistical computing* (Version 4.3.2 [Computer software]. The R Foundation. https://cran.r-project.org/

Rankin, W. L. (2000). The Maintenance Error Decision Aid (MEDA) process. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *44*(22), 795-798. https://doi.org/10.1177/154193120004402278

Rashid, H. S. J., Place, C. S., & Braithwaite, G. R. (2013). Investigating the investigations: A retrospective study in the aviation maintenance error causation. *Cognition, Technology & Work*, *15*(2), 171-188. https://doi.org/10.1007/s10111-011-0210-7

Rashid, H. S. J., Place, C. S., & Braithwaite, G. R. (2014). Eradicating root causes of aviation maintenance errors: Introducing the AMMP. *Cognition, Technology & Work*, *16*(1), 71-90. https://doi.org/10.1007/s10111-012-0245-4

Rasmussen, J. (1982). Human errors: A taxonomy for describing human malfunction in industrial installations. *Journal of Occupational Accidents*, *4*(2), 311-333. https://doi.org/10.1016/0376-6349(82)90041-4

Rasmussen, J. (1983). Skills, rules, and knowledge; Signals, signs, and symbols, and other distinctions in human performance models. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-13*(3), 257-266. https://doi.org/10.1109/TSMC.1983.6313160

Reason, J. (1990). *Human error*. Cambridge University Press. https://doi.org/10.1017/CBO9781139062367

Reason, J., & Hobbs, A. (2003). *Managing maintenance error: A practical guide*. Ashgate.

Reason, J., & Mycielska, K. (1982). *Absent-minded?*. Prentice-Hall.

Riessman, C. K. (2007). *Narrative methods for the human sciences* (1st ed.). SAGE Publications.

Roberts, K., & Rousseau, D. (1989). Research in nearly failure-free, high-reliability organizations: Having the bubble. *IEEE Transactions on Engineering Management*, *36*(2), 132-139. https://doi.org/10.1109/17.18830

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 399-408. https://doi.org/10.1145/2684822.2685324

Rouse, W. B., & Rouse, S. H. (1983). Analysis and classification of human error. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-13*(4), 539-549. https://doi.org/10.1109/TSMC.1983.6313142

Rowe, G. & Wright, G. (1999). The Delphi technique as a forecasting tool: issues and analysis. *International Journal of Forecasting, 15*(4), 353-375. https://doi.org/10.1016/S0169-2070(99)00018-7

Sarica, S & Luo, J. (2021). Stopwords in technical language processing. *PloS One*, *16*(8). https://doi.org/10.1371/journal.pone.0254937

Salas, E., Sims, D. E., & Shawn Burke, C. (2005). Is there a “Big Five” in teamwork? *Small Group Research*, *36*(5), 555-599. https://doi.org/10.1177/1046496405277134

Savage-Knepshield, P., Davison, A., Harper, W., Morelli, F., Redden, E., & Vaughan, B. (2011). Human Performance Research in Support of Current Military Operations. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *55*(1), 1255–1259. https://doi.org/10.1177/1071181311551261

Saward, J. R. E., & Stanton, N. A. (2015). Individual latent error detection and recovery in naval aircraft maintenance: Introducing a proposal linking schema theory with a multi-process approach to human error research. *Theoretical Issues in Ergonomics Science*, *16*(3), 255-272. https://doi.org/10.1080/1463922X.2014.969360

Schauberger P, Walker A (2023). *openxlsx: Read, Write and Edit xlsx Files*. (Version 4.2.5.2), {Software package], https://CRAN.R-project.org/package=openxlsx

Schmidt, J., Schmorrow, D., & Figlock, R. (2000). Human factors analysis of naval aviation maintenance related mishaps. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *44*(22), 775-778. https://doi.org/10.1177/154193120004402273

Silge J, Robinson, D. (2016). tidytext: Text Mining and Analysis Using Tidy Data Principles in R. *Journal of Open Source Software, 1*(3). https://doi.org/10.21105/joss.00037

Shappell, S. A., & Wiegmann, D. A. (2000). *The Human Factors Analysis and Classification System-HFACS* (No. DOT/FAA/AM-00/7). United States Department of Transportation. https://rosap.ntl.bts.gov/view/dot/21482

Shingo, S. (1989). *A study of the Toyota production system from an industrial engineering viewpoint* (A. Dillon, Trans.). Taylor & Francis Group.

Tanguy, L., Tulechki, N., Urieli, A., Hermann, E., & Raynal, C. (2016). Natural language processing for aviation safety reports: From classification to interactive analysis. *Computers in Industry*, *78*, 80-95. https://doi.org/10.1016/j.compind.2015.09.005

Taylor, F. W. (2004). *The principles of scientific management*. Project Gutenberg. https://doi.org/10.4324/9780203498569

Tvaryanas, A. P., & Thompson, W. T. (2008). Recurrent error pathways in HFACS data: Analysis of 95 mishaps with remotely piloted aircraft. *Aviation, Space, and Environmental Medicine*, *79*(5), 525-532. https://doi.org/10.3357/ASEM.2002.2008

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453-458. https://doi.org/10.1126/science.7455683

van Hulst, M. J., Siesling, M., van Lieshout, M., & Dewulf, A. (2014). The riot, the people and the neighborhood: Narrative framing of social disorder in four cases. *Media, Culture & Society*, *36*(4), 456-472. https://doi.org/10.1177/0163443714523810

Vogt, W. P., Gardner, D. C., & Haeffele, L. M. (2012). *When to use what research design* (1st ed.). Guilford Press.

Warren, W. R., Blickensderfer, B., Cruit, J., & Boquet, A. (2013). Shift turnover strategy and time in aviation maintenance. *Proceedings of The Human Factors and Ergonomics Society 57th Annual Meeting*, *57*(1) 46-50. https://doi.org/10.1177/1541931213571013

Wickham, H. (2016) *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York, 2016 https://ggplot2.tidyverse.org

Wickham, H. (2023). *forcats: Tools for Working with Categorical Variables (Factors)*. (Version 1.0.0), [Software package] https://CRAN.R-project.org/package=forcats

Wickham, H. (2023). *stringr: Simple, Consistent Wrappers for Common String Operations*. (Version 1.5.1), [Software package] https://CRAN.R-project.org/package=stringr

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). Welcome to the tidyverse. *Journal of Open Source Software, 4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham H, François R, Henry L, Müller K, Vaughan D (2023). *dplyr: A Grammar of Data Manipulation*. (Version 1.1.2), [Software package] https://CRAN.R-project.org/package=dplyr

Wickham H, Henry L. (2023). *purrr: Functional Programming Tools*. (Version 1.0.2), [Software package] https://CRAN.R-project.org/package=purrr

Wickham H, Hester J, Bryan J (2023). *readr: Read Rectangular Text Data*. (Version 2.1.4), [Software package], https://CRAN.R-project.org/package=readr

Wickham H, Vaughan D, Girlich M (2023). *tidyr: Tidy Messy Data*. (Version 1.3.0), [Software package], https://CRAN.R-project.org/package=tidyr

Woodworth, R. S. (1918). *Dynamic psychology*. Columbia University Press. https://doi.org/10.7312/wood90908

Xie Y, Cheng J, Tan X (2023). *DT: A Wrapper of the JavaScript Library 'DataTables'*. (Version 0.30), [Software package] https://CRAN.R-project.org/package=DT

# Appendix A: Permission to Conduct Research

**Embry-Riddle Aeronautical University**
**Application for IRB Approval**
**EXEMPT Determination Form**

**Principal Investigator:** Christopher Braun

**Other Investigators:** Dothang Truong

**Role:** Student **Campus:** Daytona Beach **College:** Aviation/Aeronautics

**Project Title:** A Delphi analysis to obtain expert consensus on human factors topics from the ASRS database.

**Review Board Use Only**

---

**Initial Reviewer:** Teri Gabriel **Date:** 11/20/2023 **Approval #:** 24-062

**Determination:** Exempt

Dr. Shawn Doherty
IRB Chair Signature: Shawn M Doherty Digitally signed by Shawn M Doherty

**Brief Description:**

As part of the investigator's research, "Exploring a New Conceptual Framework in Aviation Maintenance Incident Reporting Using National Language Processing" two separate Delphi studies will be used to develop a consensus of expert opinion regarding topic word groupings taken from an Aviation Safety Reporting System (ASRS) data and human factor framework terminology.

The purpose of the first study will be to develop a consensus of Subject Matter Expert (SME) opinions on how well topic word groups from the ASRS data base align with existing human factor terminologies from the Cognitive Reliability and Error Analysis Method (CREAM) framework and from the Human Factors Analysis and Classification System (HFACS) taxonomy.

The purpose of the second study will be to use an identical methodology to develop an SME consensus of opinion on how well the same set of topic word lists align with a novel human factor framework to be developed based on the analysis of the first study round.

This research falls under the **EXEMPT** category as per 45 CFR 46.104:

☑ (2) Research that only includes interactions involving educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures, or observation of public behavior (including visual or auditory recording) if at least one of the following criteria is met: (Applies to Subpart B [Pregnant Women, Human Fetuses and Neonates] and does not apply for Subpart C [Prisoners] except for research aimed at involving a broader subject population that only incidentally includes prisoners.)

☑ (i) The information obtained is recorded by the investigator in such a manner that the identity of the human subjects cannot readily be ascertained, directly or through identifiers linked to the subjects; (May apply to Subpart D [Children] involving educational tests or the observation of public behavior when the investigator(s) do not participate in the activities being observed.)

☑ (ii) Any disclosure of the human subjects' responses outside the research would not reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, educational advancement, or reputation; or (May apply to Subpart D [Children] involving educational tests or the observation of public behavior when the investigator(s) do not participate in the activities being observed.)

☐ (iii) The information obtained is recorded by the investigator in such a manner that the identity of the human subjects can readily be ascertained, directly or through identifiers linked to the subjects, and an IRB conducts a **Limited IRB review** (use the Limited or Expedited Review form) to make the determination. (Does not apply to Subpart D [Children])

# Appendix B: Data Collection Device

**Figure B1**

*Sample Delphi Analysis Quantitative Data Sheet*

EMBRY-RIDDLE
Aeronautical University

Consider this list of topic words:

| | |
|---|---|
| valve | performed |
| system | inspect |
| oil | document |
| damage | task |
| compliance | shift |

In your opinion, how strongly does this list of topic words reflect these phenotypical error mode categories from the CREAM methodology:

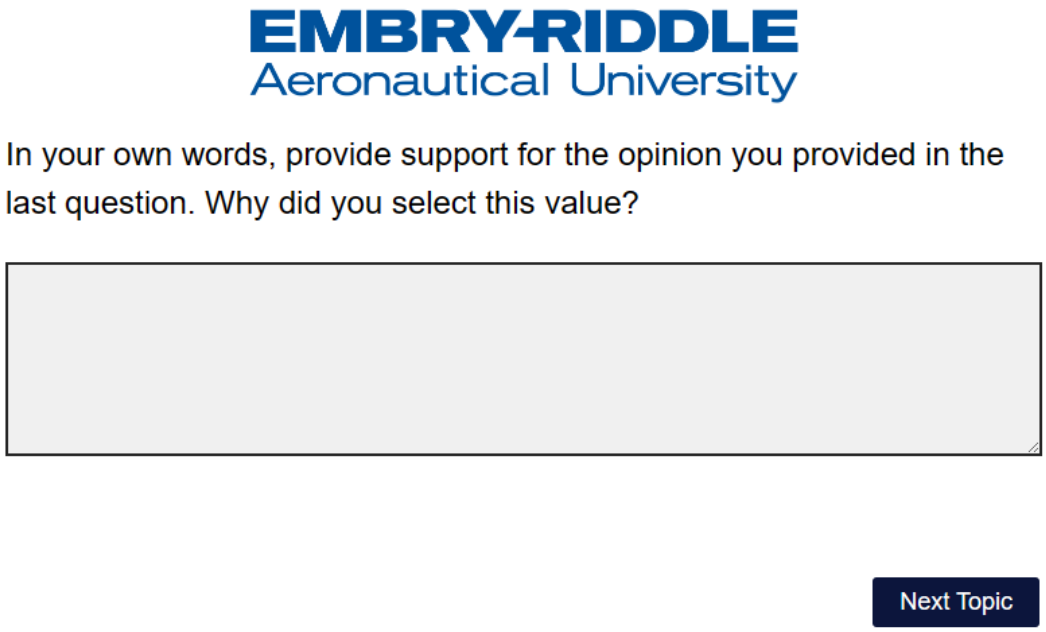
| | | | |
|---|---|---|---|
| Timing | Duration | Force | Distance/Magnitude |
| Speed | Direction | Wrong Object | Sequence |

| Does not Reflect these terms | | | | Moderately reflects these terms | | | | Strongly reflects these terms | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 3 | 4 | 5 | 6 | 7 | 7 | 8 | 9 |

*Note*. Evaluations were personalized for each of the Delphi analysis participants.

**Figure B1**

*Sample Delphi Analysis Qualitative Data Sheet*




*Note*. The qualitative data entry sheets in the second round also asked the participants if they had changed their response from the first round.

## Appendix C: Tables

**Table C1**

*The Complete HFACS Taxonomy*

| Failure Level | Major Component | Causal Category | Failure Description |
|---|---|---|---|
| Unsafe Acts of the Operator | Errors | Skill Based Errors | Failure of automated sensory-motor patterns to meet the requirements of current conditions |
| | | Decision Errors | Failure of planned behavior to meet the requirements of current conditions |
| | | Perceptual Errors | Failure of the operator to correctly perceive current conditions |
| | Violations | Routine | Habitual disregard for recognized safe behaviors as routine behavioral patterns |
| | | Exceptional | Isolated departures from recognized safe behaviors not typical of routine behavioral patterns |
| Preconditions for Unsafe Acts | Environmental Factors | Physical Environment | Characteristics of the operational or ambient environment that prevent safe operations |
| | | Technological Environment | Characteristics of operator and equipment interfaces that prevent safe operations |
| | Conditions of Operators | Adverse Mental State | Mental conditions that affect personal performance (i.e., situational awareness, distraction, mental fatigue) |
| | | Adverse Physiological State | Medical, pharmacological, or physiological conditions that preclude safe operations |
| | | Physical/Mental Limitations | Situations created when the operational requirements exceed the capabilities of the operator |
| | Personnel Factors | Crew Resource Management | Situations created by poor coordination or communication between personnel |
| | | Personal Readiness | Conditions created through an individual's failure to prepare physically or mentally for duty |
| Unsafe Supervision | N/A | Inadequate Supervision | Unsafe situational conditions generated by insufficient organizational training, leadership, oversight, or incentives |

| | | | |
|---|---|---|---|
| | | Planned Inappropriate Operations | Unsafe supervisory responses to resource shortage or risk conditions |
| | | Failed to Correct Problem | Unsafe condition created by a supervisory failure to correct a previously noted discrepancy |
| | | Supervisory Violation | Unsafe condition created by a supervisorial disregard for recognized safe behaviors |
| Organizational Influences | N/A | Resource Management | Unsafe condition created by a high-level organizational failure to appropriately allocate resources |
| | | Organizational Management | Unsafe condition created by a high-level organizational failure to establish and enhance a safe working culture |
| | | Organizational Processes | Unsafe condition created by a high-level organizational failure to establish safe operational policies and procedures |

*Note*. Adapted from “The Human Factors Analysis and Classification System- HFACS” by S. Shappell and D. Wiegmann, 2000 (https://rosap.ntl.bts.gov/view/dot/21482). In the public domain.

**Table C2**

*Listing of CREAM Antecedents*

| Main Category | Sub-Category | General Antecedent | Specific Antecedent | Definition/Explanation |
|---|---|---|---|---|
| Person Related Genotypes | Observation | Observation Missed | Overlook cue or Signal | A signal or an event that should have been the start of an action (sequence) is missed. |
| | | | Overlook measurement | A measurement or some information is missed, usually during a sequence of actions. |
| | | False Observation | False reaction | A response is given to an incorrect stimulus or event (e.g., starting to drive when the light changes to red). |
| | | | False recognition | An event or some information is incorrectly recognized or mistaken for something else. |
| | | Wrong Identification | Mistaken cue | A signal or a cue is misunderstood as something else. The difference from "false reaction" is that it does not immediately lead to an action. |
| | | | Partial Identification | The identification of an event or some information is incomplete (e.g., as in jumping to a conclusion). |
| | | | Incorrect Identification | The identification of an event or some information is incorrect. The difference from "false recognition" is that identification is a more deliberate process. |
| | Interpretation | Faulty Diagnosis | Wrong diagnosis | The diagnosis of the situation or system state is incorrect. |
| | | | Incomplete Diagnosis | The diagnosis of the situation or system state is incomplete. |
| | | Wrong reasoning | Induction error | Faulty reasoning involving inferences or generalizations (i.e., going from specific to general) leading to invalid results. |
| | | | Deduction Error | Faulty reasoning involving deduction (i.e., going from general to specific) leading to invalid results. |

| | | | |
|---|---|---|---|
| | | Wrong Priorities | The selection among alternatives (hypotheses, explanations, interpretations) using incorrect criteria, hence leading to invalid results. |
| | Decision Error | Decision paralysis | Inability to make a decision in a situation. |
| | | Wrong decision | Making the wrong decision (typically about action alternatives). |
| | | Partial decision | Making a decision that does not completely specify what to do, hence creates a need for further decisions to complete the course of action. |
| | Delayed Interpretation | No identification | An identification is not made in time (for appropriate action to be taken). |
| | | Increased time pressure | An identification is not made fast enough (e.g., because the reasoning involved is difficult) leading to a time pressure. |
| | Incorrect Prediction | Unexpected state change | A state change occurred which had not been anticipated. |
| | | Unexpected side effects | The event developed in the main as anticipated, but some side-effects had been overlooked. |
| | | Process speed misjudged | The speed of development (of the system) has been misjudged, so things happen either too slowly or too quickly. |
| Planning | Inadequate Plan | Incomplete Plan | The plan is not complete (i.e., it does not contain all the details needed when it is carried out). This can have serious consequences later in time. |
| | | Wrong Plan | The plan is wrong, in the sense that it will not achieve its purpose. |
| | Priority Error | Wrong Goal Selected | The goal has been wrongly selected, and the plan will therefore not be effective (the |

| | | | |
|---|---|---|---|
| | | | conventional definition of a mistake). |
| Temporary person related functions | Memory Failure | Forgotten | An item or some information cannot be recalled when needed. |
| | | Incorrect recall | Information is incorrectly recalled (e.g., the wrong name for something). |
| | | Incomplete recall | Information is only recalled partially (i.e., part of the information is missing). |
| | Fear | Random Actions | Actions do not seem to follow any plan or principle, but rather look like trial-and-error. |
| | | Action Freeze | The person is paralyzed (i.e., unable to move or act). |
| | Distraction | Task Suspended | The performance of a task is suspended because the person's attention was caught by something else. |
| | | Task not completed | The performance of a task is not completed because of a shift in attention. |
| | | Goal Forgotten | The person cannot remember why something is being done. This may cause a repetition of previous steps. |
| | | Loss of Orientation | The person cannot remember or think of what to do next or what happened before. |
| | Fatigue | Delayed Response | The person's response speed (physically or mentally) is reduced due to fatigue. |
| | Performance Variability | Lack of Precision | Reduced precision of actions (e.g., in reaching a target value). |
| | | Increasing Misses | An increasing number of actions fails to achieve their purpose. |

| | | | |
|---|---|---|---|
| | Inattention | Signal Missed | A signal or an event was missed due to inattention. This is similar to "observation missed"; the difference is whether it is seen as a random event or something that can be explained by a cognitive function. |
| | Physiological Stress | Many specific effects (physiological) | A general condition caused by physiological stress which may have many specific effects. |
| | Psychological Stress | Many specific effects (psychological) | A general condition caused by psychological stress which may have many specific effects. |
| Permanent person related functions | Functional Impairment | Deafness, Bad eyesight, Color blindness, Dyslexia/aphasia, Other Disability | These specific effects refer to well-defined functional impairments, mostly of a psycho-physical nature. They are therefore not defined further. Specific physiological disabilities may be added to this group if required by the analysis. |
| | Cognitive Style | Simultaneous Scanning | Search for data and information is accomplished by looking for several things at the same time. |
| | | Successive scanning | Search for data and information is accomplished by looking at one thing at a time. |
| | | Conservative focusing | Search for data and information starts from an assumption of which the various aspects are examined one by one. |
| | Cognitive Bias | Focus gambling | The search for data or information changes in an opportunistic way, rather than systematically. |

| | | | | |
|---|---|---|---|---|
| | | | Incorrect revision of probabilities | New information does not lead to a proper adjustment of probabilities - either a conservative or a too radical effect. |
| | | | Hindsight bias | Interpretation of past events is influenced by knowledge of the outcome. |
| | | | Attribution error | Events are (mistakenly) seen as being caused by specific phenomena or factors. |
| | | | Illusion of control | Person mistakenly believes that the chosen actions control the developments in the system. |
| | | | Confirmation Bias | Search for data or information is restricted to that which will confirm current assumptions. |
| | | | Hypothesis Fixation | Search for information and action alternatives is constrained by a strong hypothesis about what the current problem is. |
| Technology Related Genotypes | Equipment failure | Equipment Failure | Actuator stick/Slip | An actuator or a control either cannot be moved or moves too easily. |
| | | | Blocking | Something obstructs or is in the way of an action. |
| | | | Breakage | An actuator or a control or another piece of equipment breaks. |
| | | | Release | Uncontrolled release of matter or energy that causes other equipment to fail. |
| | | | Speed-up/ Slow down | The speed of the process (e.g., a flow) changes significantly. |
| | | | No Indications | An equipment failure occurs without a clear signature. |

| | | | |
|---|---|---|---|
| | Software Fault | Performance slow-down | The performance of the system slows down. This can in particular be critical for command and control. |
| | | Information Delays | There are delays in the transmission of information, hence in the efficiency of communication, both within the system and between systems. |
| | | Command Queues | Commands or actions are not being carried out because the system is unstable but are (presumably) stacked. |
| | | Information Not available | Information is not available due to software or other problems. |
| Procedure | Inadequate Procedure | Ambiguous text | The text of the procedure is ambiguous and open to interpretation. The logic of the procedure may be unclear. |
| | | Incomplete text | The descriptions given by the procedure are incomplete, and assume the user has specific additional knowledge. |
| | | Incorrect text | The descriptions of the procedure are factually incorrect. |
| | | Mismatch to actual equipment | The procedure text does not match the physical reality, due to, for example, equipment upgrades. |
| Temporary interface problems | Access Limitations | Item cannot be reached | An item is permanently out of reach (e.g., too high, too low, or too far away from the operator's working position). |
| | | Item cannot be found | An item is permanently difficult to find. Infrequently used items that are inappropriately |

| | | | |
|---|---|---|---|
| | | | labelled fall into this category. |
| | Ambiguous Information | Position Mismatch | There is a mismatch between the indicated positions of an item and the actual positions (e.g., controls have unusual movements). |
| | | Coding Mismatch | There is a mismatch in coding (e.g., in the use of color or shape). This may lead to difficulties in the use of equipment. |
| | Incomplete Information | Incomplete Information | The information provided by the interface is incomplete (e.g., error messages, directions, warnings, etc.). |
| Permanent interface problems | Access Problems | Item cannot be reached | An item (e.g., a control) cannot be reached, for instance because it is hidden by something or due to a change in the operator's working position. |
| | | Item cannot be found | An item, information, or a control, cannot be located when it is needed, or it is temporarily unavailable. |
| | Mislabeling | Incorrect Information | The labelling or identification of an item is not correct. |
| | | Ambiguous Identification | The labelling or identification of an item is open to interpretation. |
| | | Language error | The labelling or identification of an item is incorrectly formulated or is written in a foreign language. |

| | | | | |
|---|---|---|---|---|
| Organization Related Genotypes | Communication | Communication failure | Message not received | The message or the transmission of information did not reach the receiver. This could be due to incorrect address or failure of communication channels. |
| | | | Message misunderstood | The message was received, but it was misunderstood. The misunderstanding is, however, not deliberate. |
| | | Missing Information | No Information | Information is not being given when it was needed or requested (e.g., missing feedback). |
| | | | Incorrect Information | The information being given is incorrect or incomplete. |
| | | | Misunderstanding | There is a misunderstanding between sender and receiver about the purpose, form, or structure of the communication. |
| | Organization | Maintenance Failure | Equipment not operational | Equipment (controls, resources) does not function or is not available due to missing or inappropriate management. |
| | | | Indicators not working | Indications (lights, signals) do not work properly due to missing maintenance. |
| | | Inadequate Quality Control | Inadequate procedures | Equipment / functions are not adequate due to insufficient quality control. |
| | | | Inadequate reserves | Lack of resources or supplies (e.g., inventory, back-up equipment, etc.). |
| | | Management Problem | Unclear roles | People in the organization are not clear about their roles and duties. |

| | | | |
|---|---|---|---|
| | | Dilution of Responsibility | There is not clear distribution of responsibility; this is particularly important in abnormal situations. |
| | | Unclear Line of Command | The line of command is not well defined, and control of the situation may be lost. |
| | Design Failure | Anthropometric mismatch | The working environment is inadequate, and the cause is clearly a design failure. |
| | | Inadequate Man-Machine Interface (MMI) | The interface is inadequate, and the cause is clearly a design failure. |
| | Inadequate Task Allocation | Inadequate Managerial Rule | The organization of work is deficient due to the lack of clear rules or principles. |
| | | Inadequate Task Planning | Task planning / scheduling is deficient. |
| | | Inadequate work Procedure | Procedures for how work should be carried out are inadequate. |
| | Social Pressure | Group Think | The individual's situation understanding is guided or controlled by the group. |
| Training | Insufficient skills | Performance Failure | Lack of skills (practical experience) means that a task cannot be accomplished. |
| | | Equipment mishandling | Lack of skills (practical experience) means that equipment is incorrectly used. |
| | Insufficient Knowledge | Confusion | The person is not quite certain about what to do, due to lack of knowledge. |
| | | Loss of Situational awareness | The person has lost general situation awareness (understanding) due to lack of knowledge. |
| Ambient Conditions | Temperature | Too hot | It is uncomfortably warm. |
| | | Too Cold | It is uncomfortably cold. |
| | Sound | Too Loud | Noise level is too high. |

| | | | |
|---|---|---|---|
| | | Too soft | Signal level is too low. |
| | Humidity | Too dry | It is uncomfortably dry. |
| | | Too humid | It is uncomfortably humid. |
| | Illumination | Too bright | There is high luminosity, glare, or reflection. |
| | | Too dark | There is low luminosity, reduced color, and contrast. |
| | Other | Vibration | There may be other "dimensions" depending on the specific type of work. |
| | Adverse Ambient Conditions | None Defined | Highly context dependent and may coincide with some of the Common Performance Conditions. |
| Working Conditions | Excessive Demand | None defined | Excessive task demands or insufficient time / resources. |
| | Inadequate workplace layout | Narrow workspace | The available workspace is not large enough for the required activities. This is often the case for maintenance work. |
| | | Dangerous space | Work must be carried out in dangerous conditions (e.g., high voltage line work, radiation, unstable mass, or energy storage, etc.). |
| | | Elevated workspace | Work must be carried out where there is a risk of falling down. |
| | Inadequate Team Support | Unclear job description | The roles within the team are not well defined or well understood. |
| | | Inadequate communication | The distribution of work / responsibilities within the team is not mutually agreed upon. |
| | | Lack of team cohesiveness | There is little cohesiveness in the team, hence little collaboration. |

| | | |
|---|---|---|
| Irregular Working Hours | Circadian rhythm effects | Shift work leading to disturbances of physiological and psychological functions (jet lag, lack of sleep, etc.). |

*Note*. Adapted from “Cognitive Reliability and Error Analysis Method: CREAM” by E. Hollnagel, 1998.



**Table C3**

*Listing of R Coding Packages*

| Package Name | Code Title | Version | Date | Description |
|---|---|---|---|---|
| quanteda.textstats | An R package for the quantitative analysis of textual data | 0.96.4 | 2018 | Textual statistics functions formerly in the 'quanteda' package. Textual statistics for characterizing and comparing textual data. Includes functions for measuring term and document frequency, the co-occurrence of words, similarity and distance between features and documents, feature entropy, keyword occurrence, readability, and lexical diversity. These functions extend the 'quanteda' package and are specially designed for sparse textual data. |
| quanteda | Quantitative Analysis of Textual Data | 3.3.1 | 2023 | A fast, flexible, and comprehensive framework for quantitative text analysis in R. Provides functionality for corpus management, creating and manipulating tokens and n-grams, exploring keywords in context, forming and manipulating sparse matrices of documents by features and feature cooccurrences, analyzing keywords, computing feature similarities and distances, applying content dictionaries, applying supervised and unsupervised machine learning, visually representing text and text analyses, and more. |

| Package Name | Code Title | Version | Date | Description |
|---|---|---|---|---|
| SnowballC | Snowball Stemmers Based on the C 'libstemmer' UTF-8 Library | 0.7.1 | 2023 | An R interface to the C 'libstemmer' library that implements Porter's word stemming algorithm for collapsing words to a common root to aid comparison of vocabulary. Currently supported languages are Arabic, Basque, Catalan, Danish, Dutch, English, Finnish, French, German, Greek, Hindi, Hungarian, Indonesian, Irish, Italian, Lithuanian, Nepali, Norwegian, Portuguese, Romanian, Russian, Spanish, Swedish, Tamil and Turkish. |
| lda | Collapsed Gibbs Sampling Methods for Topic Models | 1.4.2 | 2015 | Implements latent Dirichlet allocation (LDA) and related models. This includes (but is not limited to) sLDA, corrLDA, and the mixed-membership stochastic blockmodel. Inference for all of these models is implemented via a fast collapsed Gibbs sampler written in C. Utility functions for reading/writing data typically used in topic models, as well as tools for examining posterior distributions are also included. |
| tm | Text Mining Package | 0.7-11 | 2023 | A framework for text mining applications within R. |
| lubridate | Dates and Times Made Easy with lubridate | 1.9.3 | 2011 | Functions to work with date-times and time-spans: fast and user friendly parsing of date-time data, extraction and updating of components of a date-time (years, months, days, hours, minutes, and seconds), algebraic manipulation on date-time and time-span objects. The 'lubridate' package has a consistent and memorable syntax that makes working with dates easy and fun. |
| topicmodels | Topic Models | 0.2-14 | 2023 | n Provides an interface to the C code for Latent Dirichlet Allocation (LDA) models and Correlated Topics Models (CTM) by David M. Blei and co-authors and the C++ code for fitting LDA models using Gibbs sampling by Xuan-Hieu Phan and co-authors. |
| ggplot2 | Elegant Graphics for Data Analysi | 3.4.4 | 2016 | A system for 'declaratively' creating graphics, based on ``The Grammar of Graphics". You provide the data, tell 'ggplot2' how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details. |
| NLP | Natural Language | 0.2-1 | 2020 | Basic classes and methods for Natural Language Processing. |

| Package Name | Code Title | Version | Date | Description |
|---|---|---|---|---|
| | Processing Infrastructure | | | |
| tibble | Simple Data Frames | 3.2.1 | 2023 | Provides a 'tbl_df' class (the 'tibble') with stricter checking and better formatting than the traditional data frame. |
| ldatuning | Tuning of the Latent Dirichlet Allocation Models Parameters | 1.0.2 | 2020 | For this first version only metrics to estimate the best fitting number of topics are implemented. |
| writexl | Export Data Frames to Excel 'xlsx' Format | 1.4.2 | 2023 | Zero-dependency data frame to xlsx exporter based on 'libxlsxwriter'. Fast and no Java or Excel required. |
| openxlsx | Read, Write and Edit xlsx Files | 4.2.5.2 | 2023 | Simplifies the creation of Excel .xlsx files by providing a high level interface to writing, styling and editing worksheets. Through the use of 'Rcpp', read/write times are comparable to the 'xlsx' and 'XLConnect' packages with the added benefit of removing the dependency on Java. |
| tidytext | Text Mining using 'dplyr', 'ggplot2', and Other Tidy Tools | 0.4.1 | 2016 | Using tidy data principles can make many text mining tasks easier, more effective, and consistent with tools already in wide use. Much of the infrastructure needed for text mining with tidy data frames already exists in packages like 'dplyr', 'broom', 'tidyr', and 'ggplot2'. In this package, we provide functions and supporting data sets to allow conversion of text to and from tidy formats, and to switch seamlessly between tidy tools and existing text mining packages. |
| forcats | Tools for Working with Categorical Variables (Factors) | 1.0.0 | 2023 | Helpers for reordering factor levels (including moving specified levels to front, ordering by first appearance, reversing, and randomly shuffling), and tools for modifying factor levels (including collapsing rare levels into other, 'anonymising', and manually 'recoding') |
| stringr | Simple, Consistent Wrappers for Common String Operations | 1.5.1 | 2023 | A consistent, simple and easy to use set of wrappers around the fantastic 'stringi' package. All function and argument names (and positions) are consistent, all functions deal with ``NA'''s and zero length vectors in the same way, and the output from one function is easy to feed into the input of another. |

| Package Name | Code Title | Version | Date | Description |
|---|---|---|---|---|
| tidyverse | Easily Install and Load the 'Tidyverse' | 2.0.0 | 2023 | The 'tidyverse' is a set of packages that work in harmony because they share common data representations and 'API' design. This package is designed to make it easy to install and load multiple 'tidyverse' packages in a single step. Learn more about the 'tidyverse' at <https://www.tidyverse.org>. |
| dplyr | A Grammar of Data Manipulation | 1.1.4 | 2023 | A fast, consistent tool for working with data frame like objects, both in memory and out of memory. |
| purrr | Functional Programming Tools | 1.0.2 | 2023 | A complete and consistent functional programming toolkit for R. |
| readr | Read Rectangular Text Data | 2.1.4 | 2023 | The goal of 'readr' is to provide a fast and friendly way to read rectangular data (like 'csv', 'tsv', and 'fwf'). It is designed to flexibly parse many types of data found in the wild, while still cleanly failing when data unexpectedly changes. |
| tidyr | Tidy Messy Data | 1.3.0 | 2023 | Tools to help to create tidy data, where each column is a variable, each row is an observation, and each cell contains a single value. 'tidyr' contains tools for changing the shape (pivoting) and hierarchy (nesting and 'unnesting') of a dataset, turning deeply nested lists into rectangular data frames ('rectangling'), and extracting values out of string columns. It also includes tools for working with missing values (both implicit and explicit). |
| DT | A Wrapper of the JavaScript Library 'DataTables' | 0.30 | 2023 | Data objects in R can be rendered as HTML tables using the JavaScript library 'DataTables' (typically via R Markdown or Shiny). The 'DataTables' library has been included in this R package. The package name 'DT' is an abbreviation of 'DataTables'. |

*Note*. All code packages and user manuals are available in the CRAN repository (https://cran.r-project.org/).

**Table C4**

*Replacement Words*

| Word in Report | Replaced with | Word in Report | Replaced with |
|---|---|---|---|
| approx | approximately | mech | mechanic |
| assemblies | assembly | mechs | mechanic |
| attn | attention | ops | operations |
| auth | authorized | plt | pilot |
| btwn | between | pos | position |
| capt | captain | pwr | power |
| cat | category | probs | problem |
| chked | checked | problems | problem |
| chk | checked | procedures | procedure |
| chks | checks | proc | procedure |
| ctlr | controller | procs | procedure |
| deferred | defer | ref | reference |
| deferral | defer | removal | remove |
| dep | departure | removed | remove |
| emer | emergency | rpt | report |
| eng | engine | rptd | reported |
| faults | fault | rpted | reported |
| gnd | ground | rptr | reporter |
| holes | hole | rwy | runway |
| hrs | hours | steps | step |
| hyd | hydraulic | supvr | supervisor |
| inches | inch | sys | system |
| inop | inoperable | tkof | takeoff |
| lndg | landing | tasks | task |
| leaks | leak | amt | technician |
| mgmnt | management | technicians | technician |
| mechanics | mechanic | amts | technician |

*Note*. Many of the words in this list represent colloquial abbreviations and shorthand terms used by aviation maintenance technicians in the ASRS reports.

**Table C5**

*Complete Stop Word Listing*

| Stop Words | | | | | | | |
|---|---|---|---|---|---|---|---|
| a | available | eight | howbeit | my | provides | than | use |
| a's | away | especially | however | myself | q | thank | used |
| able | awfully | et | i | n | que | thanks | useful |
| about | b | etc | i'd | name | quite | thanx | uses |
| according | be | ex | i'll | namely | qv | that | using |
| accordingly | becomes | example | i'm | nd | r | that's | usually |
| actually | becoming | except | i've | necessary | rather | thats | uucp |
| afterwards | been | f | ie | nevertheless | rd | the | v |
| again | beforehand | for | if | nobody | re | their | value |
| ain't | being | former | in | non | really | theirs | various |
| all | believe | formerly | inasmuch | none | reasonably | them | very |
| allow | beside | furthermore | inc | noone | regarding | themselves | via |
| allows | besides | g | indeed | nor | regardless | then | viz |
| alone | best | gets | insofar | novel | regards | thence | vs |
| also | better | getting | instead | now | respectively | there | w |
| although | brief | given | it | nowhere | right | there's | want |
| always | but | gives | it'd | o | s | thereafter | wants |
| am | c | gotten | it'll | of | self | thereby | way |
| among | c'mon | greetings | it's | off | selves | therefore | we |
| amongst | c's | h | its | oh | sensible | therein | we'd |
| an | came | happens | itself | ok | sent | theres | we'll |
| and | certain | hardly | j | okay | she | thereupon | we're |
| another | certainly | having | just | ones | since | these | we've |
| any | changes | he | k | only | so | they | welcome |
| anybody | co | he's | l | or | some | they'd | well |
| anyhow | com | hello | lately | other | somebody | they'll | went |
| anyone | come | help | latterly | others | somehow | they're | were |
| anything | comes | hence | lest | otherwise | someone | they've | what's |
| anyway | concerning | her | let | ought | something | those | whatever |
| anyways | consider | here's | let's | our | soon | though | whence |
| apart | considering | hereafter | like | ours | sorry | three | where's |
| appear | corresponding | hereby | liked | ourselves | specifying | throughout | whereafter |
| appreciate | course | herein | likely | overall | still | thus | whereas |
| appropriate | d | hereupon | looking | own | sub | together | whereby |
| are | definitely | hers | ltd | p | such | took | wherein |
| aren't | described | herself | m | particular | sup | truly | whereupon |
| around | despite | hi | mainly | particularly | sure | u | whither |
| as | doing | him | me | per | t | un | whoever |
| aside | done | himself | mean | perhaps | t's | unfortunately | whom |
| ask | during | his | meanwhile | please | take | unlikely | will |
| asking | e | hither | merely | plus | taken | unto | willing |
| associated | edu | hopefully | moreover | presumably | tends | upon | wish |
| at | eg | how | much | probably | th | us | with |

Stop Words

| | | | | | | |
|---|---|---|---|---|---|---|
| wonder | facts | ordering | assembly | engines | panels | wire |
| x | felt | orders | avionics | engs | pin | wires |
| y | finds | part | axle | fan | pins | wiring |
| you | furthered | parted | battery | fasteners | pitot | xa |
| you'd | furthering | parting | bay | filter | placard | xx |
| you'll | furthers | parts | bearing | flap | plane | xxx |
| you're | gave | place | blade | flaps | plate | xxxx |
| you've | general | places | blades | flight | plug | xyz |
| your | generally | point | bolt | floor | pump | zz |
| yours | give | pointed | bolts | flt | pylon | zzz |
| yourself | goods | pointing | bottle | fluid | reverser | zzzz |
| yourselves | great | presented | bottles | fuel | rii | zzzzz |
| z | greatest | presenting | box | fuselage | rod | zzzzzz |
| zero | group | presents | bracket | gate | rudder | zzzzzzz |
| she's | grouped | problem | brake | gauge | screw | zzzzzzzz |
| he'd | grouping | problems | brakes | gear | seal | zzzzzzzzz |
| she'd | groups | puts | breaker | generator | seat | zzzzzzzzzz |
| he'll | important | room | breakers | handle | seats | |
| she'll | interest | rooms | bushing | hardware | sensor | |
| mustn't | interested | sides | cabin | harness | skin | |
| when's | interesting | thing | cable | hyd | slat | |
| why's | interests | things | cables | hydraulic | slide | |
| how's | kind | thoughts | cannon | idg | spacer | |
| area | largely | today | cap | ipc | spoiler | |
| areas | lets | turned | cargo | latch | stabilizer | |
| asked | made | turning | chk | lavatory | strut | |
| asks | make | wanted | circuit | light | switch | |
| backed | making | wanting | cockpit | lights | tab | |
| backing | man | ways | compartment | lines | tail | |
| backs | member | wells | component | link | tank | |
| began | members | young | computer | maint | tape | |
| beings | men | younger | connector | maintenance | thrust | |
| big | mr | youngest | cover | mel | tire | |
| case | mrs | acft | cowl | mlg | tires | |
| cases | needing | acn | cowling | module | tube | |
| clear | number | acr | door | motor | unit | |
| differ | numbers | actuator | doors | nose | valve | |
| downed | open | aileron | duct | nut | valves | |
| downing | opened | aircraft | edge | nuts | washer | |
| downs | opening | airplane | electrical | oil | water | |
| face | opens | antiice | elevator | oxygen | wheel | |
| faces | order | apu | eng | pack | window | |
| fact | ordered | asrs | engine | panel | wing | |

*Note.* This list includes all the words that were removed in the data preconditioning and topic refinement process.

**Table C6**

*Add-Back Words*

| Add Back Words | | | | | | | |
|---|---|---|---|---|---|---|---|
| above | didn't | five | into | need | saw | somewhat | weren't |
| across | different | followed | inward | needed | say | somewhere | what |
| after | differently | following | is | needs | saying | specified | when |
| against | do | follows | isn't | neither | says | specify | whenever |
| almost | does | forth | keep | never | second | state | where |
| along | doesn't | four | keeps | new | secondly | states | wherever |
| already | don't | from | kept | newer | seconds | tell | whether |
| anywhere | down | full | knew | newest | see | think | which |
| back | downwards | fully | know | next | seeing | thinks | while |
| became | each | further | known | nine | seem | third | who |
| because | early | get | knows | no | seemed | this | whole |
| become | either | go | large | normally | seeming | thorough | who's |
| before | else | goes | last | not | seems | thoroughly | whose |
| behind | elsewhere | going | later | nothing | seen | thought | why |
| below | end | gone | latest | obviously | sees | through | within |
| between | ended | good | latter | often | serious | thru | without |
| beyond | ending | got | least | old | seriously | to | won't |
| both | ends | greater | less | older | seven | too | work |
| by | enough | had | little | oldest | several | toward | worked |
| can | entirely | hadn't | long | on | shall | towards | working |
| cannot | even | has | longer | once | shan't | tried | works |
| cant | evenly | hasn't | longest | one | should | tries | would |
| can't | ever | have | look | onto | shouldn't | try | wouldn't |
| cause | every | haven't | looks | out | show | trying | year |
| causes | everybody | here | many | outside | showed | turn | years |
| clearly | everyone | high | may | over | showing | turns | yes |
| consequently | everything | higher | maybe | placed | shows | twice | yet |
| contain | everywhere | highest | might | points | side | two | |
| containing | exactly | ignored | more | possible | six | under | |
| contains | far | immediate | most | present | small | unless | |
| could | few | indicate | mostly | put | smaller | until | |
| couldn't | fifth | indicated | must | relatively | smallest | up | |
| currently | find | indicates | near | said | sometime | was | |
| did | first | inner | nearly | same | sometimes | wasn't | |

*Note*. These words were removed from the *stopwords* list for inclusion to the topic modeling process.

# Appendix D: SME Qualifications

## *NASA ASRS Program SME*

NASA requires all ASRS Expert Analysts to have a minimum of 10 years of experience in their given field. Therefore, Analysts reviewing maintenance reports are required to have at least 10 years of experience in a maintenance technician role. The ASRS program director, Becky L. Hooey, ASRS Program Director, noted that additional requirements can be added by the hiring manager based on the staffing needs of the specific group (R. Hooey, personal communication, 2024).

## *Human Factors SMEs*

HF SME1 holds a Ph.D. in Business Administration from Northcentral University, an MBA from Park University, a Master of Science in Aeronautics from an ABET accredited university with an aviation program with specializations in unmanned systems and space studies, and a Bachelor of Science in Management from Park University. HF SME1 served in the United States Marine Corps for 14 years working as an avionics communication/navigation technician and supervisor for multiple aircraft platforms. A member of the faculty at an ABET accredited university with aviation programs since 2014, this SME has led both graduate and undergraduate degree programs. HF SME1 is currently serving as the Department Chair for the Graduate Studies Department in the College of Aeronautics for an ABET accredited university with aviation programs. HF SME1 has published research in aviation maintenance, human factors, organizational leadership, and unmanned systems, and furthermore, has a wide array of knowledge and experience with human factors, aircraft maintenance and inspections, unmanned aircraft systems, and business management.

HF SME2 holds a B.S. in Aerospace Engineering, M.S. in Project Management, and Ph.D. in Aviation Safety & Human Factors from an ABET accredited university with aviation programs. HF SME2 also holds a commercial pilot certificate and is currently the Flight Deck Chief Engineer at a major aircraft manufacturer with over 15 years of experience in the aviation and aerospace industry.

HF SME3 holds a Ph.D. in Human Factors and a Master of Science in Aeronautical Science-Aviation Safety from an ABET accredited university with aviation programs. HF SME3 has 22 years of experience in the US Airforce and the Federal Aviation Administration developing root cause assessments in human factors issues and identifying human factor contributors in aviation accidents. Additional areas of specialization include risk-based decision making, and educational outreach as a researcher, writer, editor, and producer of FAA safety briefings and FAA general aviation education and safety literature. HF SME3 is currently a Senior Human Performance Investigator with the National Transportation Safety Board.

***Aviation Maintenance SMEs***

The aviation maintenance SMEs were selected based on the following criteria:

- Current or former A&P license or commensurate Aviation Maintenance experience.
- Some academic or professional course work or training in Human Factors.
- English speaking.
- U.S. based.

These minimum requirements were established to ensure familiarity with the language in which the ASRS reports were written, the national culture of the working environments, the specific tasks of aviation maintenance, and the general concepts of human factors. The specific qualifications of the aviation maintenance SMEs are listed below.

AM SME 9598 participated in both rounds of Delphi analysis and was a contributor to the subsequent roundtable discussion. This SME is the Maintenance Manager and Chief Pilot for a corporate air carrier and holds a Ph.D. in Aeronautical Science from a regionally accredited university with aviation programs.

AM SME 2663 participated in the Delphi analysis studies. This SME is an Associate Professor of Aeronautics at an accredited university with aviation maintenance programs and has seventeen years of aviation maintenance experience with the US Navy and a Part 121 air carrier.

AM SME 5786 participated in the Delphi analysis studies. This SME holds a B.S. in Aeronautics from a state university with aviation programs and has eighteen years of experience as an aviation mechanic and flight instructor.

AM SME 7512 participated in both rounds of Delphi analysis and was a contributor to the subsequent roundtable discussion. This SME is the manager of quality assurance and chief inspector of the aviation department at a university flight school.

AM SME 7144 participated in the Delphi analysis studies. This SME holds a Master of Science in Aeronautics from an ABET accredited university with aviation programs, is an Associate Professor in the Aviation Maintenance Science department of an ABET accredited university and has seventeen years of professional aviation maintenance experience with a Part 121 air carrier.

AM SME 1922 participated in the Delphi analysis studies. This SME holds an A&P license and is a retired Quality Assurance Manager and FAA Liaison of an aerospace manufacturing and integration corporation. This SME holds a B.S. degree in Aviation Maintenance Management from a university specializing in aviation, and an M.B.A. from a university with prominent business programs.

AM SME 4333 participated in the aviation maintenance SME roundtable discussion. This SME has six years of experience with the FAA as a Principal Maintenance Inspector and a Program Manager. This SME also has seven years as an aviation industry consultant specializing in improving performance and efficiency in maintenance organizations.

AM SME 7651 participated in the aviation maintenance SME roundtable. This SME has 20 years of experience as an Air Safety Investigator with the National Transportation Safety Board. This SME has contributed a number of safety improvements to the aviation industry in the form of technological developments. This SME is a past Living Legends of Aviation honoree.

# Appendix E: SME Training Module

EMBRY-RIDDLE
Aeronautical University

## SME Topic Word Grouping Evaluation Training Module

**Developed in support of the dissertation research project:**

***Exploring a New Conceptual Framework in Aviation Maintenance Incident Reporting Using Natural Language Processing***

**Christopher Braun**
PhD Candidate
College of Aviation
Embry Riddle Aeronautical University

EMBRY-RIDDLE
Aeronautical University
2

## Introduction

- You have been invited to participate in two qualitative Delphi analysis studies that will be used to develop a consensus of expert opinion evaluating how well topic word lists taken from ASRS database, and human factor framework terminology align in concept and meaning. These studies are part of a research project titled: "Exploring a New Conceptual Framework in Aviation Maintenance Incident Reporting Using National Language Processing".
- A Delphi analysis is a structured research method similar to a subject matter expert focus group. It is designed to seek a consensus of opinion among a group of individual with expertise in a particular field. In this case you, and the others who have agreed to participate in the Delphi studies, have been invited to participate because you have expertise in aviation maintenance.

My background is in Aerospace Quality. I have worked in this field as an inspector, a supervisor, and an engineer and I have long been fascinated with the phenomenon of human error in its many forms. This research project is the final step in my journey to achieve a doctoral degree in aviation. I am sincerely grateful to you for agreeing to participate and for donating your time and expertise to my academic endeavor. I honestly cannot thank you enough.

Kindest Regards,

Chris Braun

EMBRY-RIDDLE
Aeronautical University
3

## Introduction

### Purpose

This is dissertation research that is intended to explore the way aviation maintenance incidents are described and how the descriptions given by technicians, inspectors and supervisors align with the way incident investigators classify the incidents in terms of possible human factor contributing causes. In doing research for this project, I did not find any evidence that this connection between reporters and investigators had been analyzed in the past. Although the research is being performed with strict academic rigor and structure, there is the potential for important practical significance along with the theoretical impacts.

### Significance

Practical

- Provide communication guidance for incident reporting and investigation.
- Facilitate enhanced training to improve incident investigation
- Improve efficacy of safety mitigations based on investigation of past incidents

Theoretical

- Insight to the psychological viewpoints in aviation maintenance reporting
- Evaluation of existing theoretical concepts
- Establish a novel framework for continued research

EMBRY-RIDDLE Aeronautical University 4

## Research Questions

This exploratory research is based on three research questions realized from a review of the published literature:

- ***RQ1***: What human factor themes are prevalent in the selected corpus of ASRS maintenance reports?
- ***RQ2***: How are these human factor themes aligned with existing phenotypical and genotypical conceptual frameworks? (i.e. CREAM and HFACS)
- ***RQ3***: What novel conceptual framework can be proposed to align the human factor themes within the ASRS maintenance reports to the prevalent human factor theoretical frameworks?

The topic word groupings that will be presented to you have been developed in response to research question #1. The Delphi analysis studies in which you have agreed to participate are intended to directly address research questions #2 and #3 and are an important part of the overall research project.

EMBRY-RIDDLE Aeronautical University 5

## Aviation Maintenance and Incident Investigation

Aviation Maintenance
- Activity Focused
  - Task based
  - Tactical- constantly changing
  - Variable environments
- Communication dependent
  - Multifunctional teams
  - Shift to shift pass-downs
  - Incident recording
  - "soft skills" required


Incident Investigation
- Theoretical Focus
  - Retrospective
  - Static- Historical records
  - Human Factor considerations
- Communication dependent
  - Archival data
  - Interpretation of reports
  - Incident Analysis

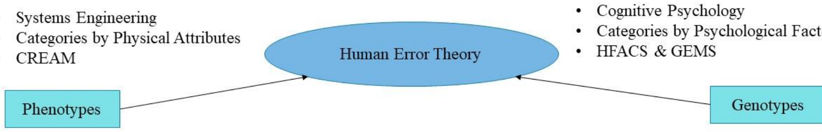
Archival incident records like the Aviation Safety Reporting System (ASRS) database are valuable tools for improving aviation safety but lack the feedback needed for effective communication. Without coordination between maintainers and investigators, we cannot be sure that the intended message from the maintenance shop is being clearly communicated to those responsible for determining what factors contributed to the incident.

EMBRY-RIDDLE Aeronautical University 6

## Theoretical Views of Human Error

- Systems Engineering
- Categories by Physical Attributes
- CREAM

- Cognitive Psychology
- Categories by Psychological Factors
- HFACS & GEMS




Most human error description strategies are centered around two theoretical concepts. One is that human error should be analyzed based on the physical characteristics of the event. These are called phenotypes. The second viewpoint is that events should be classified by the psychological influences that were present when the incident occurred. These are called genotypes.
The Cognitive Reliability Error Analysis Method (CREAM) developed by Erik Hollnagel is an example of a phenotypical classification method.
The Human Factors Analysis and Classification System (HFACS) from Scott Shappell and Douglas Wiegmann is an example of a genotypical classification system.
There are many different examples of both methods that have been used to varying degrees. These two were selected as examples because of their clarity of terminology and volume of support in terms of academic publication.

EMBRY-RIDDLE Aeronautical University 7

## HFACS- Human Factors Analysis and Classification System

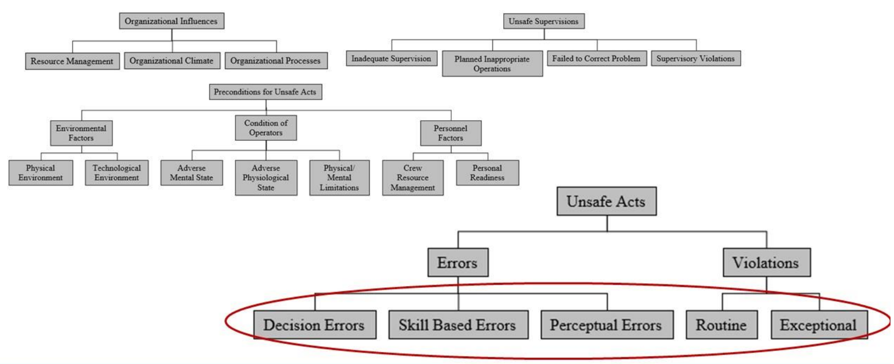
HFACS provides an organizationally layered causal taxonomy designed to assist incident investigators in determining the human factors contributors facilitating errant operator actions. HFACS is one of the most widely used methods to connect human errors to accident causation. It is applicable to a wide range of industrial applications with high-risk, high-reliability, and human dependent operational factors. Accident investigators employ the HFACS taxonomy to facilitate investigations in aviation, rail, maritime transportation, nuclear power generation and civil engineering.

Using a causal framework, HFACS provides investigators with a classification tool to categorize errant acts of human behavior that generate unexpected conditions within complex systems. HFACS categories are arranged in a hierarchical order loosely aligned with physical and temporal proximity to the incident occurrence.

HFACS categorization begins with the active factors closest to the incident, termed "the unsafe acts of the operator". The unsafe acts of the operator include unintentional skill-based, decision-based, and perceptual errors by the operator. Unsafe acts also encompass intentional violations, which can be either routine or exceptional. This research specifically uses the Unsafe Acts level of categorization as the example of a genotypical system

EMBRY-RIDDLE Aeronautical University 8

## HFACS- Human Factors Analysis and Classification System




EMBRY-RIDDLE Aeronautical University 9

## Clarification on HFACS meanings

- Skill-based Errors are failures of attention or memory. These errors occur without significant conscious thought and include:
  - Failure to observe an apparent defect
  - Poor use of tools or techniques
  - Omitted checklist items
  - Omitted steps in procedure
  - Task overload
  - Negative habits
  - Distraction
- Decision Errors are "honest mistakes". Intentional behaviors that proceed as planned but with inadequate or inappropriate planning These include:
  - Inappropriate procedures
  - Poor knowledge of system or procedure
  - Tasks beyond an operator's ability
  - Incorrect response to an emergency
- Perceptual Errors are a result of poor decision making due to faulty input information. These include:
  - Visual illusion or sensory perception (poor lighting, excessive noise)
  - Spatial disorientation/vertigo
  - Misjudged clearance, distance, speed, force
- Routine violations are habitual conscious decisions contrary to established rules or procedures. Examples include:
  - consistently driving slightly over the speed limit
  - Routinely performing operations contrary to procedure (i.e. installation or removal sequence)
- Exceptional Violations are unusual intentional acts performed against rules or procedures. They are not indicative of an individual's normal behavior.

EMBRY-RIDDLE Aeronautical University 10

# CREAM- Cognitive Reliability and Error Analysis Method

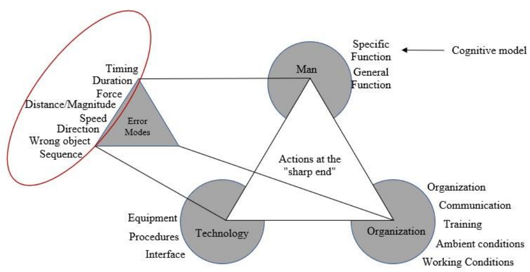
Erik Hollnagel theorized that traditional views on human error, and even the term *human error*, were inaccurate and misleading. Hollnagel believed that an event could not be determined to be a human error until its causal factors had been evaluated. Therefore, all of the existing frameworks that were rooted in the psychological and cognitive contributors were inaccurate because they did not consider the physical form of the action that was at the root of the investigation. His development of The Phenotypes of Erroneous Action was intended to provide categorization of the physical elements of the operator's action. His theory behind this categorization system also provided the term erroneous action as a substitute for human error, since it could not truly be called an error until the causal factors were known.

Hollnagel developed the CREAM methodology from his previous work to address the shortcomings he saw in other human reliability analysis methods. He noted that existing methods were missing event tree specificity, had limited error mode categories that could not account for cognitive errors , inadequate consideration of performance shaping factors, and insufficient operator models that lacked consideration for multi-stage information processing. Most importantly the existing models lacked any ability to analyze the actual action of the operator that was being classified as a human error.

The CREAM method contains a recursive model that begins with a phenotypical classification of the operator's actions and provides for subsequent antecedents (causes) and consequents (effects) of the operator's erroneous action (error). Although the CREAM system is not hierarchical or organizationally layered like HFACS, it makes a clear distinction of the error modes at the "sharp end" of the stick, where the operator's action occurs.

EMBRY-RIDDLE Aeronautical University 11

# CREAM- Cognitive Reliability and Error Analysis Method




EMBRY-RIDDLE Aeronautical University 12

# Clarification on CREAM Error Modes Meanings meanings

The CREAM error modes are clearly expressed as physical occurrences based on the recorded action of the operator.

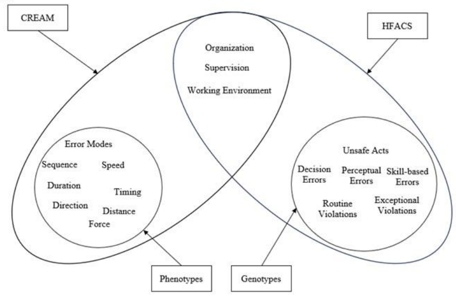
- Timing
  - Too Early
  - Too Late
  - Omitted (not completed within the allotted timeframe)
- Duration
  - Too Long
  - Too Short
- Force
  - Too Little
  - Too Much
- Distance/Magnitude
  - Too Far
  - Not Far Enough
- Speed
  - Too Fast
  - Too Slow
- Direction
  - Wrong Direction
  - Wrong Movement Type or Axis (i.e. pulling a knob instead of turning it)
- Wrong Object
  - Neighbor
  - Similar Object
  - Unrelated Object
- Sequence
  - Omission-skipped action
  - Jump forwards/ Jump Backwards/ Swapped order
  - Repetition- repeated action
  - Wrong Action- an unrelated action was substituted

EMBRY-RIDDLE Aeronautical University 13

## Overlapping Frameworks

The larger structures of the HFACS taxonomy and the CREAM analysis method overlap in the higher environmental and organizational categories.

The evaluations that you are being asked to perform use the HFACS Unsafe Acts and the CREAM Error Modes because they capture the differences between the two different ways to classify the actions of the operator.




The phenotypical and genotypical theoretical frameworks represent factors closest to and most influencing the incident occurrences.
These are respectively central to the overlapping conceptual frameworks defined by the CREAM categorizations and the HFACS taxonomy.

EMBRY-RIDDLE Aeronautical University 14

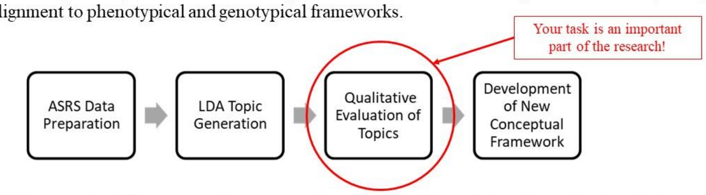
## Gaps in the Existing Research

- The published literature does not contain an analysis of how incident reports are associated with existing theoretical frameworks.
- Lack of research that analyzes the thematic content or semantic style of the incident reports in aviation maintenance.
- Lack of attention devoted to aviation maintenance, particularly at the maintainer level.

EMBRY-RIDDLE Aeronautical University 15

## Research Design

- The proposed research is a qualitative analysis of archival data.
- This methodology will employ quantitative methods to reduce the large corpus of aviation maintenance incident reports to a vital few sets of word groupings with probabilities of occurrence through natural language processing and Latent Dirichlet Allocation.
- A modified Delphi analysis will be used as a qualitative method to seek an expert consensus regarding topic alignment to phenotypical and genotypical frameworks.




- A novel framework will be proposed to capture the structure of the existing themes in the aviation maintenance incident reports.

EMBRY-RIDDLE Aeronautical University 16

## Data Source and Collection

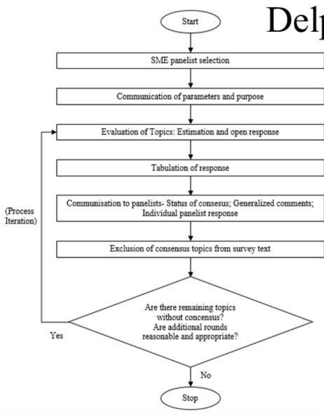
The topics that you will be asked to evaluate have been drawn out of a set of 6390 ASRS maintenance reports submitted from 2000 to 2023.

The LDA process uses machine learning and AI algorithms to reduce this vast set of incident narratives down to a vital few sets of words that represent the prevailing themes in the reports. The words are grouped into topics based on the probabilities that they will occur together in any given report.

- ASRS Database
  - Consistent collection methodology
  - Voluntary participation
  - Confidential reporting
  - Publicly accessible
  - NASA SME reviewed
  - Filtered downloads
  - Defined time period- 2000-2022


EMBRY-RIDDLE Aeronautical University 17

## Delphi Analysis Process

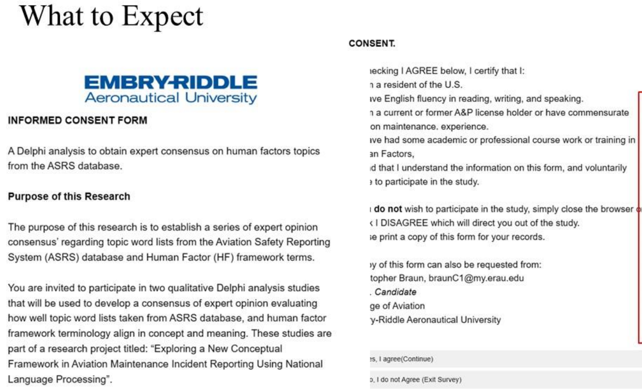



The Delphi Qualitative analysis process is comprised of an iterative evaluation and communication flow. Communication of textual responses to panelists for each round are anonymized, with the exception of the panelist's own response to the previous round. This process flow is a generalized representation of the full and modified Delphi processes.

The proposed modified process will empanel 5-8 SME's and be limited to four rounds of evaluation. These limitations are intended to abbreviate the Delphi process while maintaining reliability of results.

EMBRY-RIDDLE Aeronautical University 18

## What to Expect



EMBRY-RIDDLE Aeronautical University

**INFORMED CONSENT FORM**

A Delphi analysis to obtain expert consensus on human factors topics from the ASRS database.

**Purpose of this Research**

The purpose of this research is to establish a series of expert opinion consensus' regarding topic word lists from the Aviation Safety Reporting System (ASRS) database and Human Factor (HF) framework terms.

You are invited to participate in two qualitative Delphi analysis studies that will be used to develop a consensus of expert opinion evaluating how well topic word lists taken from ASRS database, and human factor framework terminology align in concept and meaning. These studies are part of a research project titled: "Exploring a New Conceptual Framework in Aviation Maintenance Incident Reporting Using National Language Processing".

**CONSENT.**

ıecking I AGREE below, I certify that I:
ı a resident of the U.S.
ıve English fluency in reading, writing, and speaking.
ı a current or former A&P license holder or have commensurate on maintenance. experience.
ıve had some academic or professional course work or training in ın Factors,
ıd that I understand the information on this form, and voluntarily ı to participate in the study.

ı **do not** wish to participate in the study, simply close the browser or ‹ I DISAGREE which will direct you out of the study.
ıe print a copy of this form for your records.

ıy of this form can also be requested from:
topher Braun, braunC1@my.erau.edu
. *Candidate*
ge of Aviation
y-Riddle Aeronautical University

ıs, I agree(Continue)

ɔ, I do not Agree (Exit Survey)

The Informed Consent Form will be the first thing you see at the beginning of each round of evaluations. This is a requirement for any research conducted under Embry-Riddle's umbrella of authority. Please read it carefully and then either Agree to move on to the evaluations or Do Not Agree to exit the survey.

EMBRY-RIDDLE
Aeronautical University
19

## What to Expect

You will receive an email from Qualtrics with a link for each survey round. The link will be personalized specifically for you, so please do not share it with others.
I will send an email requesting you to confirm that you received the Qualtrics link to ensure that I didn't make an incorrect data entry or otherwise mis-direct your survey questionnaire.

Since each survey round is dependent on the one before it, the sooner that you can complete your questionnaire, the sooner we can complete the project.

If for some reason you are unable to complete your questionnaire within a week, please call or email me to let me know if you will need more time or will withdraw from the project. My information is at the end of this presentation.

EMBRY-RIDDLE
Aeronautical University
20

## What to Expect

EMBRY-RIDDLE
Aeronautical University

Consider this list of topic words:

| | |
|---|---|
| valve | performed |
| system | inspect |
| oil | document |
| damage | task |
| compliance | shift |

In your opinion, how strongly does this list of topic words reflect these phenotypical error mode categories from the CREAM methodology:

| | | | |
|---|---|---|---|
| Timing | Duration | Force | Distance/Magnitude |
| Speed | Direction | Wrong Object | Sequence |

Does not Reflect these terms 1 2 3 3 Moderately reflects these terms 4 5 6 7 Strongly reflects these terms 7 8 9

For each topic word grouping, you will be asked to evaluate how closely the words in the topic reflect or are associated with the categories from either CREAM or HFACS on a 9-point scale. I would also like for you to explain why you provided the rating. Please be as descriptive as you can, using as many or as few statements that you think will be helpful to me as the researcher, and to the group for future evaluation rounds.

EMBRY-RIDDLE
Aeronautical University

In your own words, provide support for the opinion you provided in the last question. Why did you select this value?

Next Topic

EMBRY-RIDDLE
Aeronautical University
21

## What to Expect

In the second and any continuing rounds you will be asked to reevaluate the same topics while considering comments made by the group in the previous round. You will be told what the group mean average for the topic was and provided a summary of the group comments.

The comments you will be given will be a summary of the group opinions and will not be identifiable to any group member.

You can either change your responses or re-enter the same values based on your current evaluation of the topic group.

Each round will most likely be shorter than the one before it because topic evaluations will be removed from the survey as we come to a consensus on the numerical rating.

EMBRY-RIDDLE Aeronautical University 22

## What to Expect

This is what you will see at the end of each survey round…

EMBRY-RIDDLE
Aeronautical University

We thank you for your time spent taking this survey.
Your response has been recorded.

EMBRY-RIDDLE Aeronautical University 23

## The Second Study…

After 4 rounds of the first study have been completed (Or we have reached a consensus on all of the topics), I will use all of the information gathered to propose a new framework of error classification.

You will be asked to duplicate your efforts from the first study to evaluate the same topic word groups using my proposed categories.

This study will have the same format as the first study.

This will be a much shorter study because we will only be using one set of categories instead of two (HFACS and CREAM).

EMBRY-RIDDLE Aeronautical University 24

## Contact for Questions

- Contact information is also contained in the Informed Consent Form at the beginning of the Survey
  - Primary Researcher:
    - Chris Braun, BraunC1@my.erau.edu or (480) 235-7184
  - Faculty Advisor
    - Dothang Truong, truongd@erau.edu.
  - ERAU Institutional Review Board (IRB)
    - Teri Gabriel teri.gabriel@erau.edu or (386) 226-7179