

12-2013

# Assessment of the Human Factors Analysis and Classification System (HFACS): Intra-rater and Inter-rater Reliability

Awatef Ergai

Clemson University, aergai@g.clemson.edu

Follow this and additional works at: [https://tigerprints.clemson.edu/all\\_dissertations](https://tigerprints.clemson.edu/all_dissertations)



Part of the [Engineering Commons](#)

---

## Recommended Citation

Ergai, Awatef, "Assessment of the Human Factors Analysis and Classification System (HFACS): Intra-rater and Inter-rater Reliability" (2013). *All Dissertations*. 1231.

[https://tigerprints.clemson.edu/all\\_dissertations/1231](https://tigerprints.clemson.edu/all_dissertations/1231)

This Dissertation is brought to you for free and open access by the Dissertations at TigerPrints. It has been accepted for inclusion in All Dissertations by an authorized administrator of TigerPrints. For more information, please contact [kokeefe@clemson.edu](mailto:kokeefe@clemson.edu).

ASSESSMENT OF THE HUMAN FACTORS ANALYSIS AND  
CLASSIFICATION SYSTEM (HFACS): INTRA-RATER AND INTER-RATER  
RELIABILITY

---

A Dissertation  
Presented to  
The Graduate School of  
Clemson University

---

In Partial Fulfillment  
of the Requirements for the Degree  
Doctor of Philosophy  
Industrial Engineering

---

by  
Awatef Omar Ergai  
December 2013

---

Accepted by:  
Dr. Anand Gramopadhye, Committee Chair  
Dr. Scott Shappell  
Dr. Mary E. Kurz  
Dr. Julia Sharp

## ABSTRACT

Human error has been identified as the primary contributing cause for up to 80% of the accidents in complex, high risk systems such as aviation, oil and gas, mining and healthcare. Many models have been proposed to analyze these incidents and identify their causes, focusing on the human factor. One such safety model is the Human Factors Analysis and Classification System (HFACS), a comprehensive accident investigation and analysis tool which focuses not only on the act of the individual preceding the accident but on other contributing factors in the system as well.

Since its development, HFACS has received substantial research attention; however, the literature on its reliability is limited. This study adds to past research by investigating the overall intra-rater and inter-rater reliability of HFACS in addition to the intra-rater and inter-rater reliability for each tier and category. For this investigation, 125 coders with similar HFACS training coded 95 causal factors extracted from actual incident/accident reports from several sectors. The overall intra-rater reliability was evaluated using percent agreement, Krippendorff's Alpha, and Cohen's Kappa, while the inter-rater was analyzed using percent agreement, Krippendorff's Alpha, and Fleiss' Kappa. Because of analytical limitations, only percent agreement and Krippendorff's Alpha were used for the intra-rater evaluation at the individual tier and category level and Fleiss' Kappa and Krippendorff's Alpha, for the corresponding inter-rater evaluation.

The overall intra-rater and inter-rater results for the tier level and the individual HFACS tiers achieved acceptable reliability levels with respect to all agreement

coefficients. Although the overall intra-rater and inter-rater reliability results at the category level were lower than the tier level, both types of reliabilities achieved acceptable levels with inter-rater reliability being lower than intra-rater. In addition, the intra-rater and inter-rater results for the individual HFACS categories varied from achieving low reliability levels to being acceptable.

Both the inter-rater and intra-rater results found that the same 5 categories among the 19 – Skill Based Error, Decision Error, Inadequate Supervision, Planned Inappropriate Operations, and Supervisory Violation – were lower than the required minimum reliability threshold. While the overall findings suggest that HFACS is reasonably reliable, the fact that there were 5 categories with low reliability levels requires further research on ways and methods to improve its reliability. One such method could be to focus on training by designing and developing a standard HFACS training program that improves its reliability, which will have the potential to enhance both the confidence in using it as an accident analysis tool and the effectiveness of the safety plans and strategies based on it.

## DEDICATION

This dissertation is dedicated to my parents whose infinite love and dedication have been my main sources of motivation throughout my life. In addition, this work is dedicated to my husband, Ahmed, who has also been an important source of support and inspiration.

## ACKNOWLEDGMENTS

First, I would like to express my sincerest thanks and gratitude to Allah, who is the source of my success for accomplishing this research.

I acknowledge the insightful instruction and guidance of my advisors, Dr. Anand Gramopadhye and Dr. Scott Shappell, who have given me continuous support throughout this research. I also thank my committee members Dr. Kurz and Dr. Sharp, for their valuable suggestions for improving the quality of this work. I am especially grateful to Dr. Julia Sharp; her guidance, constructive feedback, and support significantly contributed to the accomplishment of this research. Special thanks also go to Barbara Ramirez, Director of the Class of 1941 Studio for Student Communication, for her technical help and support in editing this research.

Much appreciation goes to my family: my parents, Omar and Mohra; my siblings, Sassia, Abdel-Hakim, Eman, Najmeddien, Wafa, and Housameddien; and my mother in law, Aisha. I am blessed that you are my family and am especially grateful for your prayers, thoughtfulness and emotional support. Special gratitude is also extended to my other family members, aunts, uncles, nieces, nephews, and cousins. Finally, I am very grateful to my husband, Ahmed; I could not have done any of this without you.

And, of course, no dissertation is complete without a little help from friends and colleagues in Clemson's Industrial Engineering Department; I thank each and every one of you. Finally, I thank everybody who helped me in one way or another during my Ph.D. journey.

## TABLE OF CONTENTS

<b>TITLE PAGE .....</b>	<b>i</b>
<b>ABSTRACT.....</b>	<b>ii</b>
<b>DEDICATION.....</b>	<b>iv</b>
<b>ACKNOWLEDGMENTS .....</b>	<b>v</b>
<b>TABLE OF CONTENTS .....</b>	<b>vi</b>
<b>LIST OF TABLES .....</b>	<b>viii</b>
<b>LIST OF FIGURES .....</b>	<b>xiii</b>
<b>CHAPTER 1: INTRODUCTION.....</b>	<b>1</b>
1.1 The Impact of Accidents.....	1
1.2 Causes of Accidents.....	2
1.3 Managing Human Error .....	4
1.4 Research Problem and Contributions.....	7
<b>CHAPTER 2: LITERATURE REVIEW .....</b>	<b>9</b>
2.1 Human Error .....	9
2.2 Human Error Models .....	11
2.2.1 SHELL Model .....	15
2.2.2 Swiss Cheese Model (SCM).....	16
2.2.3 Wheel of Misfortune .....	19
2.2.4 Incident Cause Analysis Method .....	20
2.2.5 Human Factors Analysis and Classification System .....	22
2.2.5.1 HFACS Framework .....	23
2.2.5.2 Validation of HFACS .....	28

2. 2.5.2.1 Comprehensiveness.....	31
2. 2.5.2.2 Usability.....	37
2. 2.5.2.3 Diagnosticity.....	37
2. 2.5.2.4 Reliability.....	43
<b>CHAPTER 3: METHODOLOGY .....</b>	<b>48</b>
3.1 Participants.....	48
3.2 Instrument .....	49
3.3 Procedures.....	52
3.4 Data Management and Statistical Analysis.....	53
<b>CHAPTER 4: RESULTS .....</b>	<b>68</b>
4.1 Intra-rater Reliability Analysis .....	69
4.2 Inter-rater Reliability Analysis .....	83
<b>CHAPTER 5: DISCUSSION .....</b>	<b>109</b>
5.1 Intra-rater Reliability Discussion.....	109
5.2 Inter-rater Reliability Discussion.....	118
<b>CHAPTER 6: CONCLUSIONS AND FUTURE WORK.....</b>	<b>130</b>
<b>APPENDICES.....</b>	<b>133</b>
Appendix A: Google Form .....	134
Appendix B: Consent.....	141
Appendix C: Numerical Notations of HFACS Categories .....	143
Appendix D: Identification of Rogue Coders .....	144
<b>REFERENCES.....</b>	<b>146</b>



## LIST OF TABLES

Table		Page
1.1	Human Error Contribution to Accidents Across Industries .....	3
2.1	Perspectives of Human Error Models .....	12
2.2	Comparison Between Persons and Systems Approach to Human Error .....	14
2.3	Validation Criteria for Human Error Identification Techniques .....	30
3.1	Agreement Coefficients with Respect to Different Measures.....	54
3.2	List of Notations .....	55
3.3	Agreement Table of Two Coders' .....	63
3.4	Kappa Interpretations.....	64
3.5	Classification of Multiple Coders .....	65
3.6	Distribution of Coders by Item Number and Classification.....	66
4.1	Table Key.....	69
4.2	Intra-rater Reliability; Tier Level; PA, K, and $\alpha$ ; Individual Coders and Overall; Whole Data Set.....	70
4.3	Intra-rater Reliability; Tier Level; PA, K, and $\alpha$ ; Individual Coders and Overall; Excluding Compound Causal Factors .....	71
4.4	Intra-rater Reliability; Category Level; PA, K, and $\alpha$ ; Individual Coders and Overall; Whole Data Set .....	73

Table	Page
4.5 Intra-rater Reliability; Tier Category; PA, K, and $\alpha$ ; Individual Coders and Overall; Excluding Compound Causal Factors .....	75
4.6 Key Table for Figures 4.1 to 4.4 .....	78
4.7 Intra-rater Reliability; Individual HFACS Tiers; Overall Average PA, and $\alpha$ ; Whole Data Set .....	80
4.8 Intra-rater Reliability; Individual HFACS Tiers; Overall Average PA, and $\alpha$ ; Excluding Compound Causal Factors .....	80
4.9 Intra-rater Reliability; Individual HFACS Categories; Overall Average PA, and $\alpha$ ; Whole Data Set .....	81
4.10 Intra-rater Reliability; Individual HFACS Categories; Overall Average PA, and $\alpha$ ; Excluding Compound Causal Factors .....	82
4.11 Inter-rater Reliability; Tier Level; Overall PA, $K_F$ , and $\alpha$ ; Whole Data Set; Including Rogue Coders, First Session.....	85
4.12 Inter-rater Reliability; Tier Level; Overall PA, $K_F$ , and $\alpha$ ; Whole Data Set; Excluding Rogue Coders; First Session .....	85
4.13 Inter-rater Reliability; Tier Level; Overall PA, $K_F$ , and $\alpha$ ; Excluding Compound Causal Factors; Including Rogue Coders; First Session.....	85
4.14 Inter-rater Reliability; Category Level; Overall PA, $K_F$ , and $\alpha$ ; Whole Data Set; Including Rogue Coders, First Session.....	85
4.15 Inter-rater Reliability; Category Level; Overall PA, $K_F$ , and $\alpha$ ; Whole Data Set; Excluding Rogue Coders, First Session.....	85

Table	Page
4.16 Inter-rater Reliability; Category Level; Overall PA, $K_F$ , and $\alpha$ ; Excluding Compound Causal Factors; Including Rogue Coders, First Session .....	86
4.17 Inter-rater Reliability; Individual HFACS Tiers; Overall $K_F$ and $\alpha$ ; Whole Data Set; Including Rogue Coders, First Session .....	87
4.18 Inter-rater Reliability; Individual HFACS Tiers; Overall $K_F$ and $\alpha$ ; Whole Data Set; Excluding Rogue Coders, First Session .....	87
4.19 Inter-rater Reliability; Individual HFACS Tiers; Overall $K_F$ and $\alpha$ ; Excluding Compound Causal Factors; Including Rogue Coders, First Session .....	87
4.20 Inter-rater Reliability; Individual HFACS Categories; Overall $K_F$ and $\alpha$ ; Whole Data Set; Including Rogue Coders, First Session .....	88
4.21 Inter-rater Reliability; Individual HFACS Categories; Overall $K_F$ and $\alpha$ ; Whole Data Set; Excluding Rogue Coders, First Session .....	89
4.22 Inter-rater Reliability; Individual HFACS Categories; Overall $K_F$ and $\alpha$ ; Excluding Compound Causal Factors; Including Rogue Coders, First Session .....	90
4.23 Percentage of Coders Responses to Each Statement for First Session .....	94
4.24 Inter-rater Reliability; Tier Level; Overall PA, $K_F$ , and $\alpha$ ; Whole Data Set; Including Rogue Coders, Second Session .....	98
4.25 Inter-rater Reliability; Tier Level; Overall PA, $K_F$ , and $\alpha$ ; Whole Data Set; Excluding Rogue Coders; Second Session.....	98

Table	Page
4.26 Inter-rater Reliability; Tier Level; Overall PA, $K_F$ , and $\alpha$ ; Excluding Compound Causal Factors; Including Rogue Coders; Second Session .....	99
4.27 Inter-rater Reliability; Category Level; Overall PA, $K_F$ , and $\alpha$ ; Whole Data Set; Including Rogue Coders, Second Session .....	99
4.28 Inter-rater Reliability; Category Level; Overall PA, $K_F$ , and $\alpha$ ; Whole Data Set; Excluding Rogue Coders, Second Session .....	99
4.29 Inter-rater Reliability; Category Level; Overall PA, $K_F$ , and $\alpha$ ; Excluding Compound Causal Factors; Including Rogue Coders, Second Session.....	99
4.30 Inter-rater Reliability; Individual HFACS Tiers; Overall $K_F$ and $\alpha$ ; Whole Data Set; Including Rogue Coders, Second Session.....	100
4.31 Inter-rater Reliability; Individual HFACS Tiers; Overall $K_F$ and $\alpha$ ; Whole Data Set; Excluding Rogue Coders, Second Session.....	100
4.32 Inter-rater Reliability; Individual HFACS Tiers; Overall $K_F$ and $\alpha$ ; Excluding Compound Causal Factors; Including Rogue Coders, Second Session.....	101
4.33 Inter-rater Reliability; Individual HFACS Categories; Overall $K_F$ and $\alpha$ ; Whole Data Set; Including Rogue Coders, Second Session.....	101
4.34 Inter-rater Reliability; Individual HFACS Categories; Overall $K_F$ and $\alpha$ ; Whole Data Set; Excluding Rogue Coders, Second Session.....	102
4.35 Inter-rater Reliability; Individual HFACS Categories; Overall $K_F$ and $\alpha$ ; Excluding Compound Causal Factors; Including Rogue Coders, Second Session.....	103

Table	Page
4.36 Percentage of Coders Responses to Each Statement for Second Session.....	106
5.1 Comparison of HFACS Reliability Studies Reported in the Literature with the Current Study .....	113
5.2 Comparison of Overall Inter-rater Reliability Results Between First and Second Session for Tier Level .....	126
5.3 Comparison of Overall Inter-rater Reliability Results Between First and Second Session for Category Level .....	127
5.4 Comparison of Inter-rater Reliability Results for Each HFACS Tier Between First and Second Session.....	127
5.5 Comparison of Inter-rater Reliability Results for Each HFACS Category Between First and Second Session .....	128

## LIST OF FIGURES

Figure	Page
1.1 HFACS Application Areas in the Human Error Process Loop.....	5
1.2 Percentage of Nonfatal US GA Accidents Associated with Unsafe Acts .....	6
2.1 SHEL Model .....	15
2.2 Reason’s SCM for Human Error Causation.....	16
2.3 Mark II Version of the SCM.....	17
2.4 Mark 3 Version of the SCM.....	18
2.5 The Wheel of Misfortune.....	20
2.6 The ICAM Model of Incident Causation .....	21
2.7 HFACS Framework .....	23
2.8 ASRM User Interface, Evaluation Form .....	34
2.9 Human Factors Analysis and Classification System-Mining Industry .....	36
2.10 Department of Defense - DoD-HFACS Framework .....	36
2.11 Significant Paths Between Categories in the HFACS Framework .....	39
2.12 Failure Paths Between HFACS Categories.....	40
2.13 Pictorial Associations Between HFACS Categories.....	72
4.1 Frequency and Distribution of PA, K, and $\alpha$ ; Intra-rater Reliability; Tier Level; Individual Coders; Whole Data Set.....	78

Figure	Page
4.2 Frequency and Distribution of PA, K, and $\alpha$ ; Intra-rater Reliability; Tier Level; Individual Coders; Excluding Compound Causal Factors .....	78
4.3 Frequency and Distribution of PA, K, and $\alpha$ ; Intra-rater Reliability; Category Level; Individual Coders; Whole Data Set.....	79
4.4 Frequency and Distribution of PA, K, and $\alpha$ ; Intra-rater Reliability; Category Level; Individual Coders; Excluding Compound Causal Factors .....	79
5.1 Kappa Coefficient by Number of Causal Factors and Number of Response Categories Under Random Rating .....	117

## CHAPTER 1: INTRODUCTION

### 1.1 The Impact of Accidents

Industrial facilities and plants continually experience serious accidents and incidents, specifically during their construction and operations phases. In the U. S. in 2011, industrial accidents caused approximately 3,600 fatalities and 5.1 million disabling injuries (Bureau of Labor Statistics, 2011), representing on average a death rate of 1 every 2.5 hours and an injury rate of 1 every 6 seconds. These accidents and injuries have a significant impact.

The estimated total costs of industrial accidents in 2009 were approximately \$168.9 billion, including wage and productivity loss (\$82.4 billion), medical (\$38.3 billion), administrative (\$33.1 billion), motor vehicle damage (\$2 billion), employers' uninsured costs (\$10.3 billion), and fire loss costs (\$2.8 billion) (National Safety Council, 2011). In addition, a central economic cost of industrial accidents is the insurance premiums. According to Liberty Mutual (2011), the estimated direct U.S. workers compensation costs for disabling workplace injuries and illnesses in 2009 totaled \$50.1 billion. According to Mossink and De Greef (2002), no matter the preliminary costs associated with the accident, the indirect costs go beyond the visible ones, 2-20 times larger (Occupational Safety and Health Administration, 2006). These expenses include, but are not limited to, lost production time, employment time lost by an injured employee, and Occupational Safety and Health Administration (OSHA) penalties. These data reveal that although knowledge and technology have reduced the number of



accidents and improved safety, industrial accidents are still a serious concern, one needing further research.

## 1.2 Causes of Accidents

Based on current thinking, the causes of accidents involve the interaction of technical, environmental, organizational, and human factors (Reason, 2008; Sharit, 2006; Shorrock, 2011). A well-known factor that has received significant research attention is the technical failure of a component in a system. These factors involve equipment malfunction and failure resulting from design flaws such that the system no longer meets its designed specifications.

Environmental factors involve such physical surroundings of the operators or equipment that could adversely affect performance as weather conditions, noise, and illumination. For instance, the analysis of General Aviation (GA) databases from 2003 through 2007 shows that of 8,657 aviation accidents, 1,740 were weather-related either as the primary cause or as a contributing factor (Federal Aviation Administration, 2010).

Recent disasters like Chernobyl and the Challenger crash have brought considerable awareness to the organizational factors contributing to an accident (von Thaden, Wiegmann, & Shappell, 2006). While the Chernobyl disaster was caused by a poor safety culture, specifically infringements of safety rules (Salge & Milling, 2006), a major contributing factor in the Challenger accident was NASA's poor communication system (Heimann, 1993). Organizational factors also include inadequate procedures and

training; insufficient standards, requirements, and processes; and company/management-induced pressure.

Human causal factors are associated with human error, defined by Reason (1990) as “encompassing all those occasions in which a planned sequence of mental or physical activities fails to achieve its intended outcome, and when these failures cannot be attributed to the intervention of some chance agency.” Similarly, Senders and Moray (1991) concluded that human error results when the actions that were intended by the operator generate a production status beyond the acceptable limits or are not required by certain standards. Examples of human error include, but are not limited to, inattention, memory lapses, complacency, and mistakes. While there have been significant reductions in accidents resulting from technological failures, industrial incidents/accidents due to human error have significantly increased, representing a contributing cause of up to 80% (Aas, 2008; Peters & Peters, 2006). Table 1.1 lists a wide range of industries including their percentage of human error contributing to accidents.

Table 1.1: Human Error Contribution to Accidents across Industries

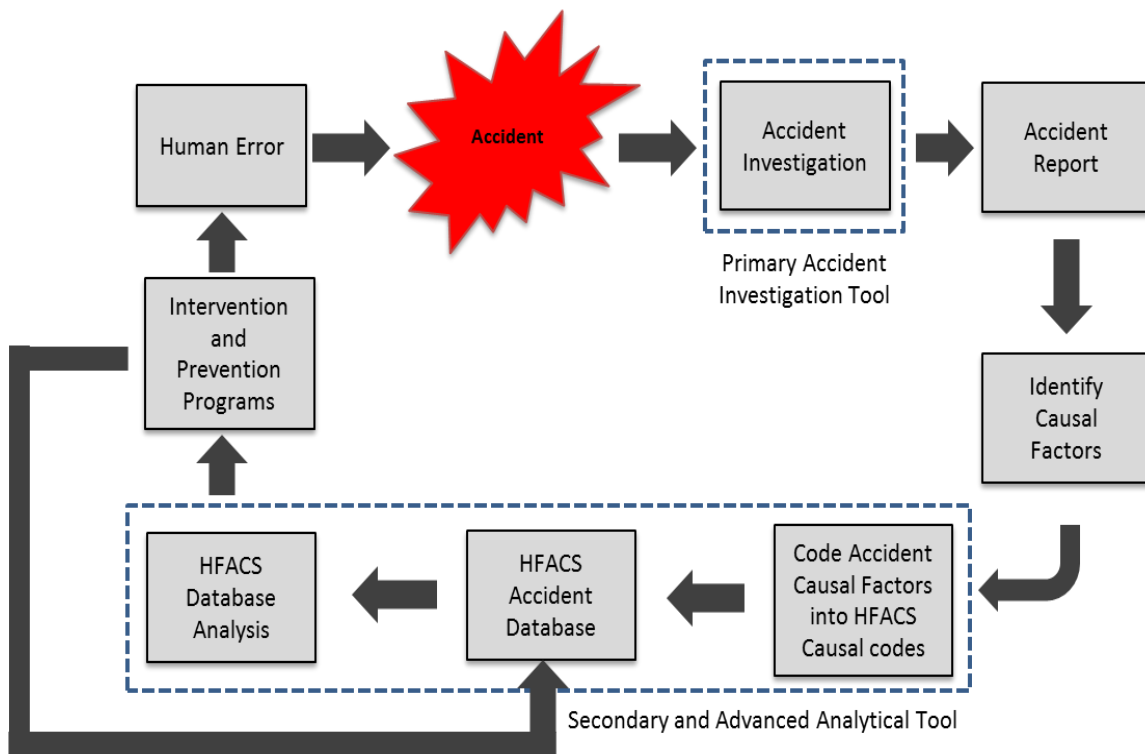
<b>Industry</b>	<b>Percentage</b>
<b>Aviation (GA)</b>	70 - 80 %
<b>Petrochemical</b>	41 %
<b>Marine</b>	74 %
<b>US Coast Guards</b>	80 - 90 %
<b>Healthcare (Anesthesia)</b>	82 %

Sources: (Shappell et al., 2007);  
 (Butikofer, 1986);  
 (Trucco, Cagno, Ruggeri, & Grande, 2008);  
 (Aas, 2008);  
 (Kohn, Corrigan, & Donaldson, 2000)

### 1.3 Managing Human Error

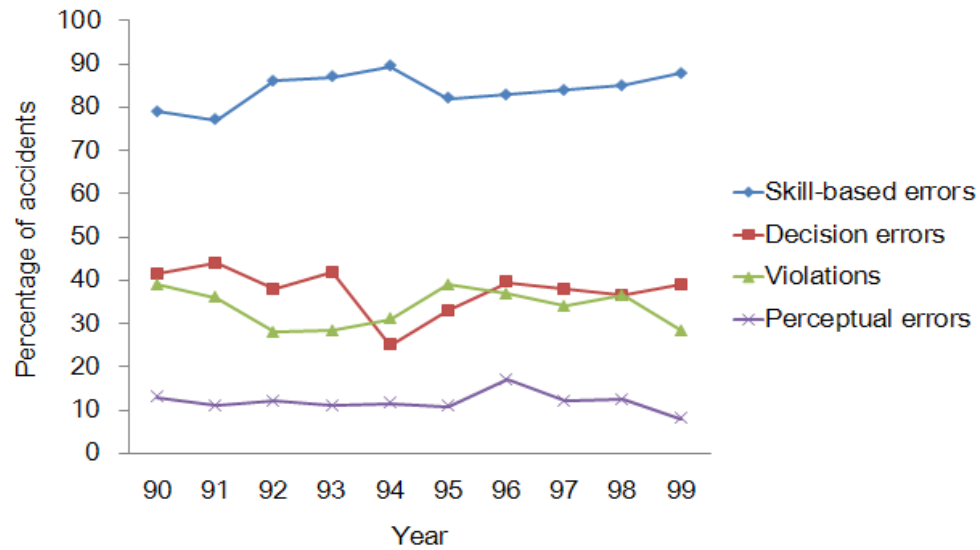
As humans are limited in their capabilities and are inherently fallible, the primary objective of any safety professional is to identify resulting errors, reduce their chances of occurrence, and minimize their impact. These can only be achieved by gaining information on the safety status of the organization or company, information usually collected in an incident/accident report. However, because textual data are difficult to analyze, the use of safety taxonomies is of vital importance. The advantage of such an approach enables safety professionals to develop a safety database, allowing them to efficiently analyze the information, searching for patterns, similarities, and trends among incidents/accidents. The resulting analysis can assist not only in the development of data-driven safety interventions and mitigation strategies, but also in evaluating their effectiveness.

One of the most important safety taxonomies is the Human Factors Analysis and Classification System (HFACS) (Harris & Li, 2011), a comprehensive accident investigation and analysis tool which focuses not only on the act of the individual preceding the accident but on other contributing factors in the system (environmental, supervisory, and organizational). Figure 1.1 highlights the areas where HFACS can be applied in the error process loop. As a primary accident investigation tool, it assists accident investigators in their search for causal factors, active and latent, within each level of HFACS, thereby serving as a checklist for determining possible contributing factors.



**Figure 1.1: HFACS Application Areas in the Human Error Process Loop**

As a secondary analysis tool, HFACS evaluates a collection of accidents, looking for trends which point to weaknesses in certain areas. For example, Wiegmann & Shappell (2003) analyzed 14,571 GA accidents from 1990 to 1999 using HFACS, finding that skill-based errors were the dominant type of aircrew errors as seen in Figure 1.2; therefore, safety strategies need to be directed towards reducing such errors. In addition, despite slight fluctuations, the data indicate that the error trends have not changed significantly over time, suggesting that the safety intervention efforts directed towards any of these errors have had no significant effect on them.



**Figure 1.2: Percentage of Nonfatal US GA Accidents Associated with Unsafe Acts (Wiegmann & Shappell, 2003)**

As an advanced analytical tool, HFACS can identify recurrent error pathways among its categories, providing the system safety professional with additional information to guide resources towards a more focused intervention. For example, Company X has identified a significant error pathway which includes resource management, failure to correct known problem, technical environment, and skill-based errors. As a result, it can allocate intervention resources towards resource management to prevent errors from propagating down the described pathway.

#### 1.4 Research Problem and Contributions

One of the fundamental issues concerning the utility of HFACS involves the classification of the incident/accident causal factors to the HFACS causal categories. The accuracy of this process, referred to as coding, which is accomplished through multiple raters, reflects the reliability of HFACS. Differences may occur between coders at a specific time (inter-rater reliability) or within coders (intra-rater reliability) over time.

If the same incident/accident is coded differently by more than one person or the coding results vary for the same person over a certain time frame, the detection of unique events becomes unachievable, implying that the frequency counts of events derived from the coding process are meaningless leading to ineffective mitigation/prevention plans, and in the end reducing the margin of safety of the system. The aim of coding via HFACS is to obtain frequencies that reflect the safety status of the system irrespective of who arrives at the classification (Wallace, 2008).

The reliability of HFACS has been called into question because of the limited research studies investigating its inter-rater and intra-rater reliability, a concern as more industries and organizations are adopting HFACS as an accident investigation analysis tool. Although its developers (Shappell and Wiegmann, 2003) achieved high reliability, recent reliability assessments have been less reassuring (Olsen, 2011). Methodologies assessing the reliability of HFACS vary across past studies in terms of the number and experience of the coders, the statistical methods, and the industrial sector accident

database used, making it difficult to synthesize their results and draw practical conclusions.

To address this issue, the primary goal in this research is to assess the reliability of HFACS, both intra-rater and inter-rater, as a general accident investigation analysis tool using accident data from a variety of industries. In addition, this study aims to use a large number of coders, more than 120, all having a similar level of experience. Furthermore, to thoroughly assess and investigate the reliability of HFACS, four statistical procedures – the percent agreement, Krippendorff's Alpha, Cohen's Kappa, and Fleiss' Kappa – were used. In addition, this research aims to evaluate the reliability of all HFACS tiers and categories, thereby identifying specifically which of these need further attention and improvement.

The second chapter of this dissertation analyzes the previous research in the fields of human error and human error taxonomies and frameworks, concentrating on the HFACS taxonomy and its validation criteria. The third chapter presents the methodology used in this study, including the subjects, instruments, data collection and the statistical methods and packages used. While the fourth chapter provides the results of this research, the fifth presents a detailed description and discussion of the significance of the results. The final chapter of this dissertation includes concluding statements of the research findings and prospects for future work.

## CHAPTER 2: LITERATURE REVIEW

### 2.1 Human Error

In general, human error is viewed as an inappropriate or unacceptable human decision or action that degrades, or has the potential of degrading, efficiency, safety, or system performance (Sanders and McCormick 1993). Currently, human error is frequently cited in safety sources as the major contributing cause of several significant industrial disasters such as Bhopal, and Chernobyl (Helmreich, 2000; Sarter & Alexander, 2000; Shappell & Wiegmann, 2001), one that will continue as humans are inherently fallible. Much of this research has focused on the theoretical and empirical study of human error. While some were cognitively oriented, others have taken a more holistic approach.

Significant research on human error conducted by Rasmussen (1982) defined three levels of human behavior based on the level of cognition involved, knowledge-based, rule-based, and skill-based behaviors. Knowledge-based activities are those involved in creating a plan to solve a new situation or problem. While rule-based behaviors are activities that are conducted using a set of stored instructions or procedures, skill-based behaviors are routine activities conducted spontaneously. With experience and practice, performance shifts from knowledge-based to skill-based; further, the level of conscious demand increases as we transfer from skill-based to rule-based reaching the highest conscious control levels for knowledge-based activities.



Reason (1990; 1995) supplemented Rasmussen's work by defining the error associated with the human behavior as "unsafe acts" committed by an operator at the front line preceding an adverse event. Unsafe acts take many forms including slips, lapses, mistakes, and violations, the first two being execution failures that usually occur when the plan of action is adequate but the actions performed are not carried out as intended. These two are related to failures of attention, recognition, memory, or selection. On the other hand, mistakes occur when a plan is completed as anticipated, but it proves to be inadequate to achieve its intended outcome.

The last form of unsafe acts, violations, which are classified as either routine or exceptional, include deviations from the established rules and regulations that increase the probability of committing an error resulting in a negative outcome (Reason et al., 1998). While routine violations represent less serious departures from rules and regulations tolerated by authority personnel, thus habitual in nature, exceptional violations are severe departures from rules and protocols that are not condoned by such personnel.

More recently, Sarter and Alexander (2000) categorized human error based on operator task performance as either errors of omission or commission. Whereas errors of omission occur when an operator fails to execute a necessary task at the intended time, errors of commission occur when the operator carries out an action in the inappropriate way or at the imprecise time, such classification affects the likelihood to detect errors. While human errors have been categorized in various ways to identify actions that threaten the safety of both the employees and plant, the lack of common definitions and

criteria for coding them has limited the ability to compare data across companies and industries, perhaps contributing to the continuing frequency of accidents due to human error (O'hare, 2000).

## 2.2 Human Error Models

As a concept, human error has traditionally been viewed in two ways: the earlier persons approach and the more recent systems approach described by Reason (2000). In the mid-twentieth century, the persons approach was dominant, with systems being considered error-free and needing to be protected from the unreliable humans committing errors and violations at the sharp end, operational level, causing failures (Woods, Dekker, Cook, Johannesen, & Sarter, 2010). In this approach, errors occur due to such psychological factors in an individual as forgetfulness, poor motivation, inattention, carelessness and complacency. Since this responsibility lies solely on the individual, the recommendations for addressing such errors included automation, training, employee selection, development of in-depth procedures, and the firing of the operator whose actions led to the accident; however, these steps were ineffectual as human error continued to be a major cause of accidents (Shappell & Wiegmann, 2001).

To address this situation, many accident models have been proposed to understand accidents and the role of human error within them. These single element models included the physiological perspective (Suchman, 1961), the behavioral perspective (Peterson, 1971), the organizational perspective (Bird, 1974), the cognitive perspective (Rasmussen, 1982; Wickens & Flach, 1988), and the psychosocial

perspective (Helmreich & Foushee, 1993). Table 2.1 presents a description of these perspectives:

Table 2.1: Perspectives of Human Error Models (Wiegmann & Shappell, 2003)

Perspective	Focus	Advantage	Limitation
<b>Physiological</b>	Focuses on the physical and/or physiological conditions of the operator that influence performance	<ul style="list-style-type: none"> <li>Highlighted the role of the physical status of the operator in safe performance</li> <li>Shaped military and industry view of fatigue</li> <li>Aided the development of scheduling, shift-rotations, and crew-rest policies</li> </ul>	<ul style="list-style-type: none"> <li>Lack of consensus concerning the role of physiological conditions in accidents</li> <li>Difficulty in identifying the presence of physiological factors and whether these factors caused the error</li> </ul>
<b>Behavioral</b>	Is based on the identification of incentives that reward safe behavior and/or punishes unsafe acts. Considers that errors are often due to unsafe acts that result from misplaced motivation.	<ul style="list-style-type: none"> <li>Emphasizes the role that motivation plays in influencing safe behavior</li> <li>Suggests accidents occur when individuals lack the motivation to perform safely</li> </ul>	<ul style="list-style-type: none"> <li>Motivation to be safe is self-directed because the result of unsafe acts are frequently fatal</li> <li>No distinction between unsafe acts that are motivation-driven such as violations and those that are cognitive driven such as errors.</li> </ul>
<b>Organizational</b>	Considers that errors are often due the rules, regulations, and procedures that are set by the organization; focuses on accident failures within the organization	<ul style="list-style-type: none"> <li>Broadens the field of inquiry in studying and preventing human error</li> <li>Suggests the ability to manage human error within the context of risk.</li> </ul>	<ul style="list-style-type: none"> <li>Lacking information about the types of organizational variables that cause specific errors.</li> <li>Focuses on a single type of causal factor.</li> </ul>

<b>Cognitive</b>	Treats the mind as an information processing system, the goal being to detect issues of this information processing	<ul style="list-style-type: none"> <li>• Addresses the underlying causes beyond simple classification</li> <li>• Identifies specific error trends to develop intervention strategies</li> </ul>	<ul style="list-style-type: none"> <li>• Not easily applicable</li> <li>• Focuses only on the human, disregarding other factors such as task-related factors and organizational factors that impact performance</li> </ul>
<b>Psychosocial</b>	Considers that most activities involve interaction and communication among individuals and/or teams and errors are identified within this context	<ul style="list-style-type: none"> <li>• Emphasizes the role of interpersonal aspects of human performance</li> <li>• Led to the development of Crew Resource Management (CRM) in aviation, which aids in improving coordination and communication between crew members in the cockpit.</li> </ul>	<ul style="list-style-type: none"> <li>• Limited information for formulating and testing psychosocial models of human error.</li> <li>• CRM is comprised of the psychosocial model among others</li> </ul>

However, during the last two decades, the view of human error has shifted from the persons to the systems approach. In this approach, human error is viewed as a symptom of deeper failures in the system rather than the failure of the human who is essential in creating safe systems (Woods, Johannesen, Cook, & Sarter, 1994). As a result, safety professionals focus on examining the system to reveal the latent factors, the organizational and technical elements, that created the conditions causing the operator to commit an error. Examples of latent factors include poor design, maintenance failure, ineffectual automation, inadequate supervision, manufacturing defects, inadequate training, inappropriate or poorly defined procedures, and inadequate equipment (Reason, 1997). Therefore, human error is no longer considered the major

cause of incidents/accidents; instead, it is viewed as an outcome of the latent conditions in the system. Comparisons of these two approaches can be seen in Table 2.1 below:

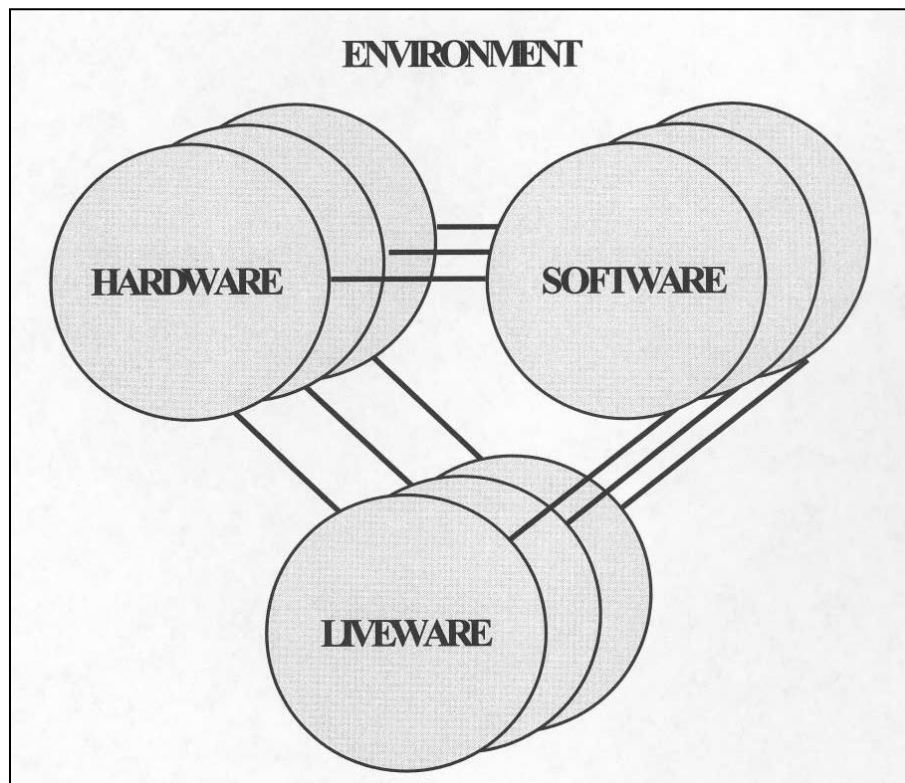
Table 2.2: Comparison between Persons and Systems Approach to Human Error

<b>Element</b>	<b>Persons</b>	<b>Systems</b>
<b>System</b>	Safe	Unsafe
<b>People</b>	Unreliable	Reliable and central to creating safety Operator errors are
<b>Cause of accident</b>	Operator at the front end (Human error)	indications of deeper failures in the system farther up-stream

As in the persons approach, many models have also been proposed in the systems approach to understand the role of human error. While some models aid in the investigation process of accidents, others provide a systematic way of understanding them (Toft, Dell, Klockner, & Hutton, 2012). The systems approach models, which include a combination effect of many factors including human error contribution to accidents, include the SHEL Model (Edwards, 1988; in Wiegmann & Shappell, 2003), the Swiss-cheese model (SCM) (Reason, 1990; 2008), the wheel of misfortune (O'hare, 2000), the incident cause analysis method (ICAM) (Gibb, Hayward, & Lowe, 2001), and the human factors analysis and classification system (HFACS) (Wiegmann & Shappell, 1997; 2001a).

### 2.2.1 SHEL Model

One of the most familiar system models is Edwards's (1973) Software, Hardware, Environmental Conditions, and Liveware (SHEL) model seen in Figure 2.1. The software is concerned with the rules and regulations that manage and run the systems operations, while the hardware involves the tools, equipment, material, and physical supplies. Third, the environmental conditions include the physical conditions such as ambient temperature and illumination, and finally, the liveware refers to the people working in the system. When any of these components or their connections fail, system failure results. Although the model includes the primary components of the system, a major drawback is its lack of specificity.



**Figure 2.1 SHEL Model (Edwards, 1988 in Wiegmann and Shappell, 2003)**

### 2.2.2 Swiss Cheese Model (SCM)

One of the most largely regarded system models of accident causation is Reason's (1990) Swiss Cheese Model (SCM) (Aas, 2008; Perneger, 2005). In this model, Reason proposes a systems approach to human error, which takes into consideration that humans are prone to error; thus, barriers and safeguards are developed to prevent system breakdown. Reason (1990) explains that accidents can be tracked to four levels of failure: unsafe acts, preconditions for unsafe acts, unsafe supervision, and organizational influences. The ideal system resembles a stack of slices of Swiss cheese, as seen in Figure 2.2, the cheese representing the barriers and safeguards against failure, while the holes represent the errors still remaining. The system is prone to an accident when the holes, errors, in each level in the system line up.

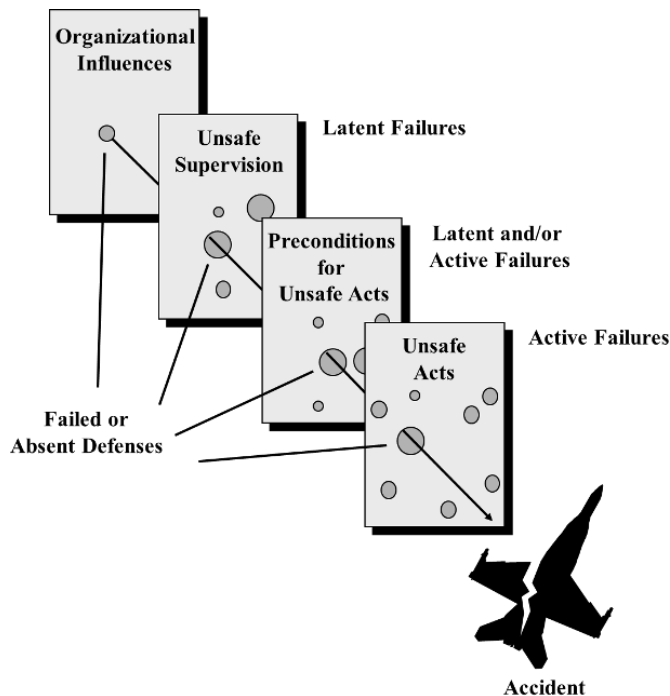


Figure 2.2: Reason's SCM for Human Error Causation (Wiegmann & Shappell, 2003)

The holes or errors in the defensive system are the active and latent failures that cause nearly all accidents (Reason, 2000). An active failure, which is the act of the operator resulting in an immediate accident/incident, is usually apparent, meaning it can quickly be attributed as the cause of an accident. On the other hand, latent errors, which are hard to detect, usually occur at higher organizational levels and may reside in the system for an extended period of time.

The second version of the SCM, Mark II, shown in Figure 2.3, reduced the number of levels, and hence, the defenses to three: organization, task/environment, and individual. Further, it included a latent failure path leading from the organization directly to the defenses, a path that takes into account accidents not involving active failures, for example the Challenger accident (Reason et al., 2006).

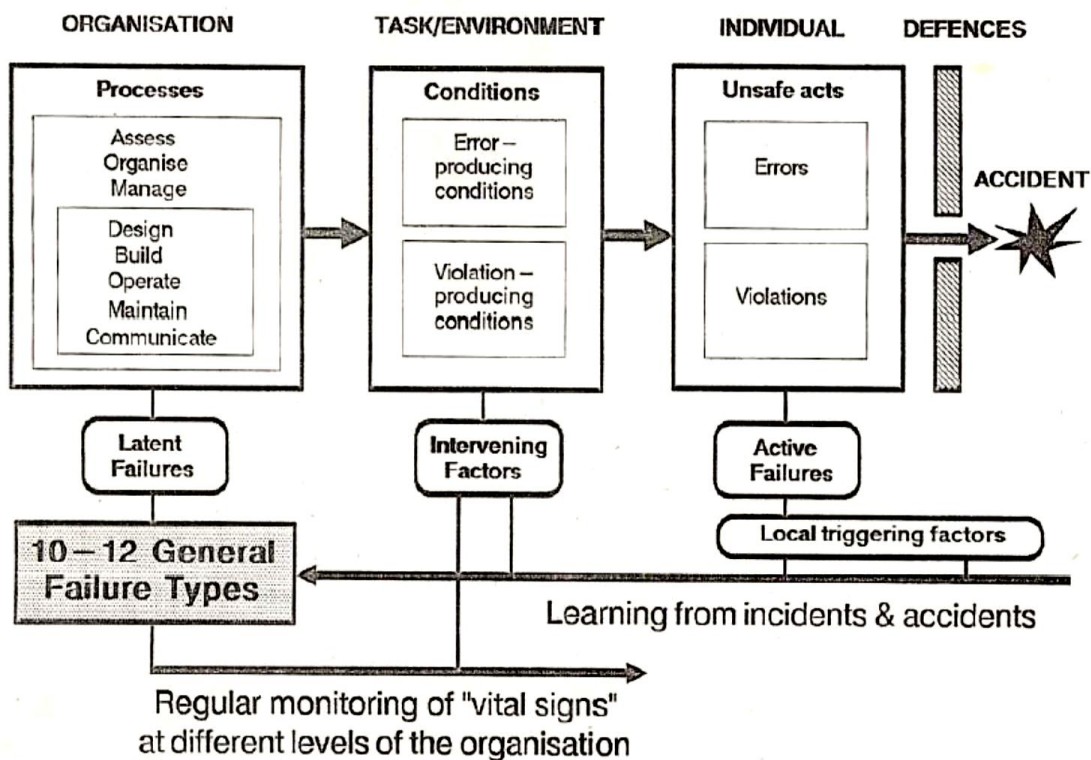


Figure 2.3: Mark II Version of the SCM (Reason, 2006)



In 1997, Reason developed a third version of the SCM, Mark III, shown in Figure 2.4. In this version, the top rectangle represents the components of an incident/accident with undefined defenses, whereas the lower triangle illustrates the system producing the event: unsafe acts of operator, local workplace conditions, and organizational factors. The arrows differentiate the directions in which an accident occurs and in which it is investigated. The main concept in the three versions is that incidents/accidents are a result of latent and active failures within the system composed of the organization, environment, and individuals that interact negatively with one another, thus breaching the defenses of the system and producing loss.

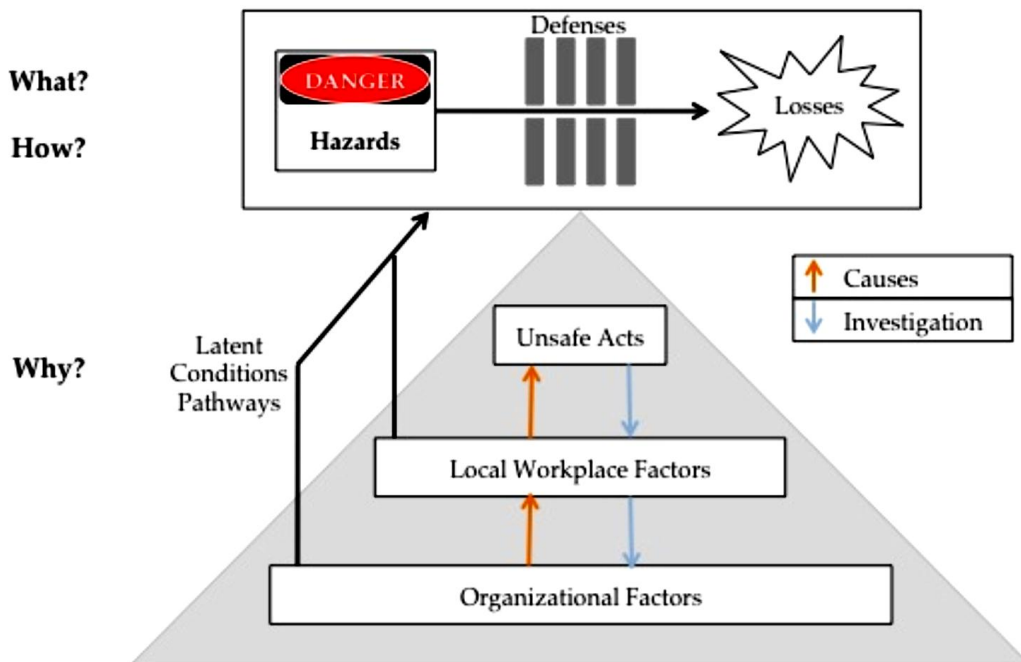


Figure 2.4: Mark 3 Version of the SCM (Reason, 1997)

The wide acceptance of the Swiss Cheese Model results from the fact that it integrates a majority of the human error perspectives previously described. However, it lacks practicability: the lack of identification of the failures, the nature of the holes, at all levels cause the model to be purely theoretical, benefitting academicians rather than practitioners. Accordingly, practitioners such as analysts and safety investigators encountered problems when applying this model to real incidents/accidents (Wiegmann & Shappell, 2003).

### 2.2.3 Wheel of Misfortune

The Wheel of Misfortune taxonomy is a general classification framework proposed by O'Hare (2000) to be used as a guideline to reveal the causes of an accident during an accident investigation. This model is primarily based on Helmreich's (1990 as cited in O'Hare, 2000) and Reason's (1990) work and is composed of three concentric spheres as illustrated in Figure 2.5. The innermost sphere represents the actions of the individual operator, based on Rasmussen's (1982) skill-rule-knowledge behavior classification. The middle sphere represents the local conditions that affect operator performance, including such external conditions as weather conditions and the internal state of the operator including excessive fatigue, distraction, and alcohol consumption, to mention a few. The outermost sphere represents the overall conditions created by the organization in which the task activity takes place, for example organizational policies. The innermost sphere, local actions, explains the results, i.e. the accident, based on the causes suggested in the middle and outermost spheres.



Figure 2.5: The Wheel of Misfortune (O'Hare, 2000)

#### 2.2.4 Incident Cause Analysis Method

The Incident Cause Analysis Method (ICAM), also based on Reason's Swiss Cheese Model (1990; 1997), is a reactive investigation tool developed by BHP Billiton (Gibb, Hayward, & Lowe, 2001), that identifies the local and latent factors contributing to an incident within the system and organization. In addition, it develops recommendations and solutions to system deficiencies and vulnerable organizational processes to prevent future incidents/accidents. The ICAM model classifies causal factors into four elements: absent/failed defenses, individual/team actions, task/environmental conditions, and organizational factors (De Landre, Gibb, & Walters, 2006), as shown in Figure 2.6. First, the absent/failed defenses identify the factors that failed to detect and protect the system against technical and human failures or the control

measures that did not prevent the incident or limit its consequences. Second, the individual/team actions include the errors or violations of the operator that led directly to the incident. Third, task/environmental conditions involve those circumstances that directly impact human and equipment performance prior to or at the time of the incident/accident. Finally, organizational factors are the underlying means generated by the organization that influence the performance of employees in the workplace. Through the analysis of the four elements, ICAM enhances the ability to identify the causes of the incident/accident and to develop improvement strategies aimed at building error-tolerant defenses to prevent future incidents.

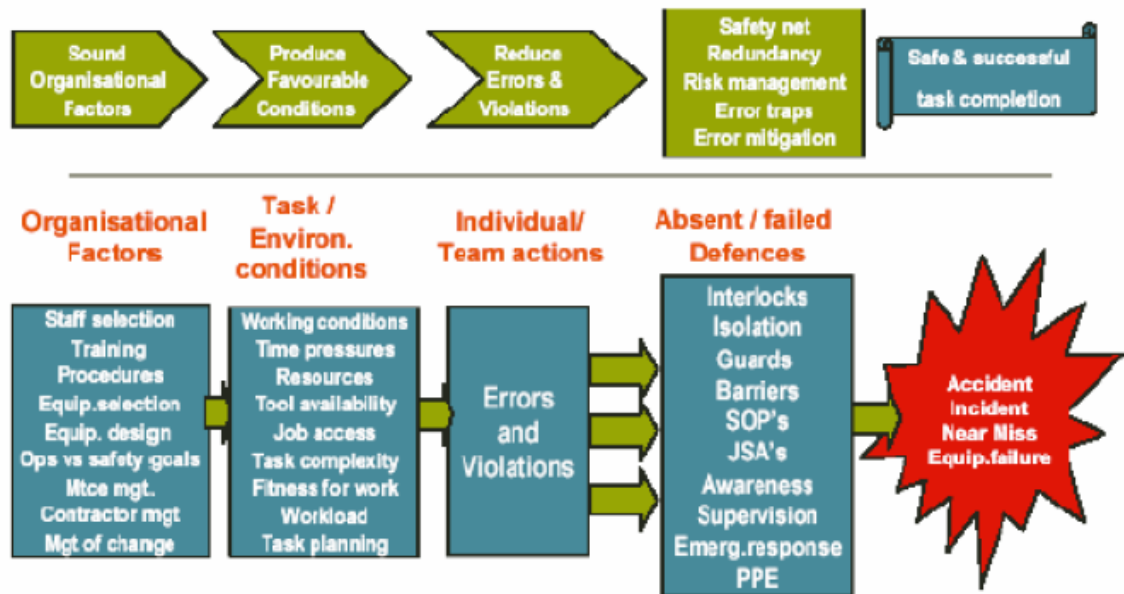
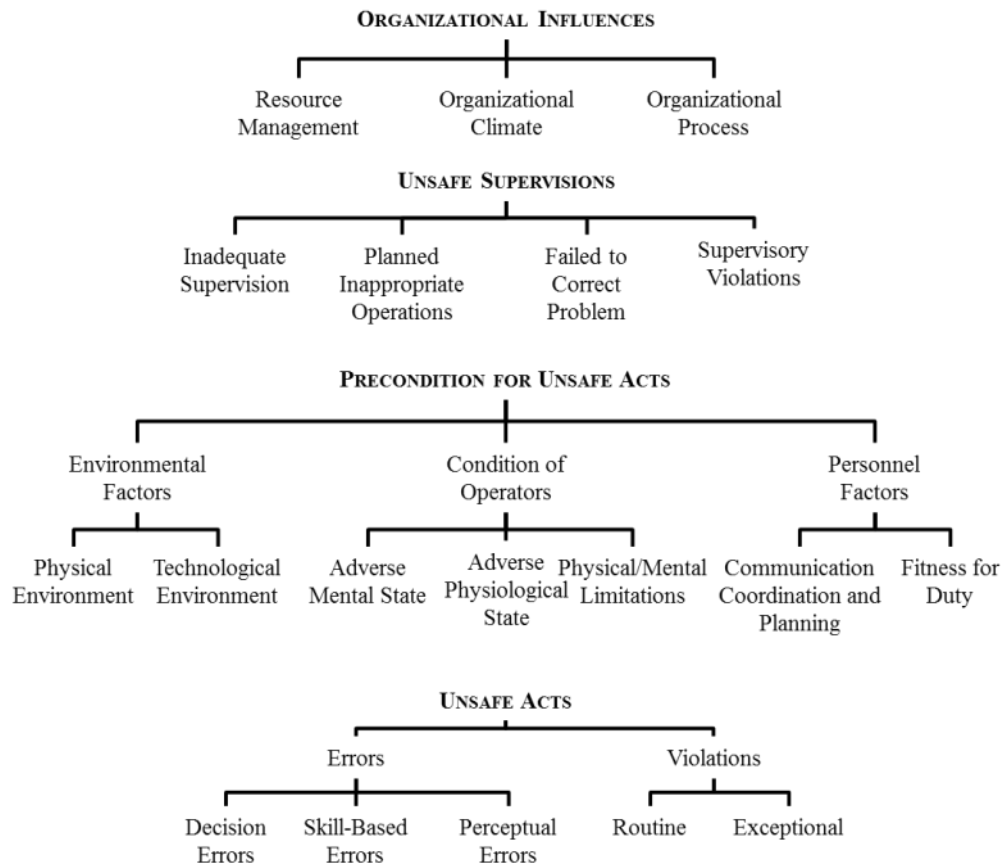


Figure 2.6: The ICAM Model of Incident Causation (De Landre, Gibb, & Walters, 2006)

### 2.2.5 Human Factors Analysis and Classification System

The Human Factors Analysis and Classification System (HFACS) is a human error taxonomy developed to provide a comprehensive framework to identify and classify causal factors of incidents/accidents; then based on these data, safety interventions can be developed and subsequently, their effectiveness evaluated (Shappel & Wiegmann, 2000). Currently, HFACS is the most extensively used human factors accident analysis framework (Harris & Li, 2011). Using Reason's SCM (1990) as a basis, Wiegmann and Shappell (1997; 2000) developed this system. HFACS identifies the holes, the failures, in the SCM, thus, providing a means of methodologically categorizing the causes of incidents/accidents. Therefore, it serves as a practical tool for accident investigators, analysts, and safety professionals in real-world settings (Wiegmann & Shappell, 2003). Similar to the barriers identified in Reason's SCM, the structure of HFACS is hierarchical, categorizing nineteen causal factors into four levels of failure. The four levels include active and latent failures; while the first level represents the active failures, unsafe acts, the other three levels include the latent failures, preconditions for unsafe acts, unsafe supervision, and organizational influences, each dependent on the previous one as illustrated in Figure 2.7.



**Figure 2.7: HFACS Framework (Shappell, 2009)**

### 2.2.5.1 HFACS Framework

#### **Unsafe Acts**

The first level of HFACS represents the unsafe acts of an operator leading to an incident/accident. Similar to the persons approach, this level focuses on the individual, putting the responsibility for the accident on the operator. These unsafe acts are classified into two categories: errors and violations. Errors or mistakes are actions of the operator that fail to carry out the desired outcomes and are extended to include three basic error types: skill-based, decision, and perceptual. Violations, which are the intentional neglect

of the established rules and regulations by the operator, are divided into two, routine and exceptional.

The most common types of errors are skill-based errors (Patterson & Shappell, 2010; Shappell et al., 2007). These physical errors occur with little or no conscious thought during highly automated tasks. The more familiar the task becomes to an individual, the more automated it becomes. For instance, a pilot examines a mechanical repair which has been performed recently during the walk-around. However, during this check, he becomes involved in the routine activities of the walk-around, totally forgetting to check the structural repair, hence, committing a skill-based error (Airbus, 2005). In general, skill-based errors are primarily due to failures of memory and/or attention, and often appear as forgetting or missing steps in a checklist, or misplacing a step in a sequence of steps.

The second type, decision errors, describes intentional actions of an individual that proceed as planned but the results indicate that they are inadequate or inappropriate for the situation. Decision errors involve three types: knowledge-based, rule-based, and problem-solving (Wiegmann et al., 2005). Knowledge-based errors occur when an operator selects an action plan that proves to be the incorrect procedure for the situation; factors such as inexperience, time, and stress enhance such errors. Rule-based errors, often referred to as procedural decision errors, occur when a situation is either not recognized or is misdiagnosed, and the wrong procedure is applied (Rasmussen, 1982). In many situations, an individual is confronted with a problem that is not well understood or

for which no formal procedure exists yet requiring a novel solution. In such situations, the time needed to arrive at a good solution is rarely available.

Perception errors, the third type of error, occur when the sensory input, whether visual, auditory, or olfactory, is degraded. These are caused by the misinterpretation of the input itself, not by the input being used. Therefore, there is a disparity between a person's perception of the situation and its reality.

Violations are actions of the operator disregarding the established rules and regulations, are therefore, considered intentional. Violations can be routine or exceptional based on their etiology (Wiegmann et al., 2005; Wiegmann & Shappell, 2003). Routine violations represent less serious departures from rules and regulations tolerated by authority personnel, thus becoming habitual in nature; on the other hand, exceptional violations are severe departures from rules and protocols that are not condoned by such personnel. While a pilot neglecting to use Air Traffic Control (ATC) radar advisories is an example of a routine violation (Wiegmann & Shappell, 2003), flying a commercial airplane without the mandatory co-pilot is an example of an exceptional violation.

### **Precondition for Unsafe Acts**

The second level, and the first latent tier, is the precondition for unsafe acts, including environmental factors, conditions of the operator, and personnel factors. Environmental factors are categorized into two causal factors: the physical environment and the technological environment. The physical environment describes both the operational (tools, machinery, etc.) and ambient (temperature, weather, etc.) conditions. Examples of physical environment causal factors include weather, housekeeping, and



lighting. The technological environment takes into consideration the design of equipment and controls, the interaction between operators and equipment and the display/interface characteristics, a critical issue in human error.

The second classification of the preconditions for unsafe acts, the conditions of operators, is categorized into three causal factors: adverse mental state, adverse physiological state, and physical/mental limitations. The adverse mental state of the operator deals with such mental conditions as the mental fatigue, distraction, inattention, and complacency that can adversely affect the performance of an operator. The adverse physiological state of the operator covers such medical and physical conditions as medical illness, physiological incapacitation, and physical fatigue. The physical/mental limitations category refers to situations where the operators' long-term capabilities are exceeded by the demands of the job such as incompatible intelligence/aptitude and incompatible physical capability for safely executing an occupation.

The last classification of the preconditions for unsafe acts tier, the personnel factors component, is categorized into two causal factors: communication coordination and planning and Fitness for Duty. Communication coordination and planning between personnel, management, crews and teams include such instances as the failure of an individual to use all available resources. The Fitness for Duty category involves off-duty activities that affect operator readiness to perform as proposed, including self-medication, alcohol, and violation of crew rest requirements.

### **Unsafe Supervision**

The third level, unsafe supervision, deals with performances and decisions of supervisors and managers that can affect the performance of operators in the frontline. It is categorized into four categories: inadequate supervision, planned inappropriate operations, failure to correct a known problem, and supervisory violations category.

Inadequate supervision includes those times when supervision either fails to or provides inappropriate or improper guidance, oversight, and/or training. The planned inappropriate operations category involves those situations when supervisors fail to evaluate the risk associated with a task, thereby placing employees at an unacceptable level of risk; these include improper staffing, mission not in accordance with rules/regulations, and inadequate opportunity for crew rest. The failure to correct a known problem refers to those instances where unacceptable conditions of equipment, training or behaviors are identified, yet actions or conditions remain uncorrected, meaning supervisors fail to initiate corrective actions or report such unsafe situations. The supervisory violations category is the willful disregard of the established rules and regulations by those in positions of leadership.

### **Organizational Influences**

The fourth level, and final latent tier, involves the organizational influences where deficiencies and failures can be traced to the highest levels of the organization. This tier is categorized into three causal factors: resource/acquisition management, organizational climate, and organizational process. Resource/acquisition management includes top management decisions related to the allocation of such resources as equipment, facilities,

money, and humans. The organizational climate category refers to those variables, such as the organizational structure, culture, and policies, which affect worker performance. The organizational process category refers to the decision-making that governs the day-to-day operations of an organization, such as operations, procedures, and oversight. Often latent conditions within the organizational level are overlooked during accident investigations; however, HFACS provides a mean of considering such factors in the investigation and analysis process.

#### 2.2.5.2 Validation of HFACS

Although research evaluating human error classification systems is limited, HFACS is an exception, its effectiveness having been investigated during its development; yet Beaubien and Baker (2002) criticized these studies because only the founders of HFACS conducted them. However, multiple researchers in addition to the developers have now researched its utility. The effectiveness of any human error framework is based on its validity, which refers to the extent to which a framework is well-grounded and corresponds accurately to the real world (Fleishman, Quaintance, & Broedling, 1984).

Mainly, two types of validity are important in scientific research, external validity and internal validity. While external validity refers to the extent to which an instrument can be generalized to other contexts, internal validity represents the extent to which an instrument is valid within a specific setting (Fleishman, Quaintance, & Broedling, 1984). The most important types of internal validity that are relevant to human error frameworks are content validity, face validity, and construct validity (Weigmann & Shappell, 2003).

While content validity represents whether a given framework covers all the major issues within the topic, face validity answers the question: does the framework have a reasonable approach and common sense in the eyes of those who would use it (Weigmann & Shappell, 2003). Construct validity refers to the extent to which a particular instrument (e.g., HFACS) performs in accordance with theoretical expectations (Carmines & Zeller, 1979). Construct validity is most likely the most difficult type of validity to verify.

From a practical view point, face validity and content validity are related to the evaluation criterion usefulness and comprehensiveness, respectively as proposed by Kirwan (1998). In addition to usefulness and comprehensiveness, Kirwan (1998) suggested a broad set of evaluation criteria for human error identification techniques, listed in Table 2.3. Besides Kirwan (1998), few researchers have proposed objective criteria for establishing the validity of human error frameworks in practical settings (O'Conner & Hardiman, 1996; Hollnagel, 1998). Furthermore, the founders of HFACS suggested that at least four criteria need to be considered when evaluating a human error Framework, comprehensiveness, usability, diagnosticity, and reliability (Wiegmann & Shappell, 2001b). The following sections discuss in detail the four criteria, comprehensiveness, usability, diagnosticity, and reliability used to validate the HFACS framework.

Table 2. 3: Validation Criteria for Human Error Identification Techniques (Kirwan, 1998)

<b>Criteria</b>	<b>Explanation</b>
<b>Comprehensiveness</b>	The ability to distinguish and classify a broad form of errors.
<b>Reliability</b>	The extent of how the framework is structured, which leads to consistent results between different users at a specific time (inter-rater reliability) and within coders (intra-rater reliability) over time.
<b>Theoretical Validity</b>	Whether the framework is built on a human performance model with a theoretically acceptable internal structure.
<b>Contextual Validity</b>	The extent to which the framework efficiently identifies the circumstances of an event occurs.
<b>Flexibility</b>	The ability of the framework to include different levels of analysis respect to project requirements and information and user experience.
<b>Usefulness</b>	Whether the framework recommends, or can promote, effective error reduction or mitigation strategies. This incorporates the criterion of <b>Diagnosticity</b> which refers to the ability of the framework to arrive at the causes of the error, permitting diagnostic resolution of error reduction strategies.
<b>Training Requirement</b>	The time spent to develop expertise on the framework.
<b>Resource Usage</b>	The total time involved to collect primary and auxiliary information and perform the analysis.
<b>Usability</b>	Refers to how easy it is to use the framework.
<b>Auditability</b>	The extent to which the framework supports auditable documentation.

#### 2. 2.5.2.1 Comprehensiveness

Comprehensiveness is the framework's ability to define and/or identify all significant information relating to an incident/accident. Since no statistical methods exist to quantify this criterion, it is investigated by mapping the human error framework onto an existing accident database of an organization (Wiegmann & Shappell, 2003), the framework being considered comprehensive if all causes of an incident/accident are incorporated in it. Initially, the comprehensiveness of HFACS was validated based on its application to USA civil and military aviation databases (Wiegmann & Shappell, 2003). Subsequently, it was applied to other applications including but are not limited to mining, construction, railroads, oil and gas, marine, and security. These studies also found that the causal factors associated to accidents could be classified using the HFACS distinct causal categories.

The analysis by such classification provides insights into possible tactics for preventing accidents. For example, the analysis of the majority of industrial accidents using HFACS, whether national or international, revealed that at the level of unsafe acts of operators, the most prevalent category was skill-based errors, including memory lapse, distraction, and poor technique (e.g., aviation: Boquet, Detwiler, Hackworth, Holomb, & Pfleiderer, 2007; Li, Harris, & Yu, 2008; S. Shappell et al., 2007; S. A. Shappell & Wiegman, 2003; Wiegmann & Shappell, 1997; 2001c, railroad: Baysari, McIntosh, & Wilson, 2008; Reinach & Viale, 2006 healthcare: ElBardissi, Wiegmann, Dearani, Daly, & Sundt, 2007, shipping: Celik & Cebi, 2009, mining: Lenne, Salmon, Liu, & Trotter, 2012; Patterson & Shappell, 2010; Patterson, 2009). The second highest percentage in

this level was decision errors (e.g., aviation: Boquet, Detwiler, Hackworth, Holomb, & Pfleiderer, 2007; Li, Harris, & Yu, 2008; Shappell et al., 2007; Shappell & Wiegman, 2003; Wiegmann & Shappell, 1997; 2001c, railroad: Baysari, McIntosh, & Wilson, 2008; Reinach & Viale, 2006, mining: Lenne, Salmon, Liu, & Trotter, 2012; Patterson & Shappell, 2010; Patterson, 2009) and violations (e.g., healthcare: ElBardissi, Wiegmann, Dearani, Daly, & Sundt, 2007, and mining: Lenne, Salmon, Liu, & Trotter, 2012; Patterson, 2009). In contrast to the majority of other industries, the most frequent category was the violation category in the construction industry seen in the improper use of the personal protective equipment (Hale, Walker, Walters, & Bolt, 2012).

At the second level, precondition for unsafe acts, adverse mental states have routinely been found to be the leading type of failures for many industries, specifically mental fatigue and stress (e.g., aviation: Boquet, Detwiler, Hackworth, Holomb, & Pfleiderer, 2007; Li, Harris, & Yu, 2008; Shappell et al., 2007; Shappell & Wiegman, 2003; Wiegmann & Shappell, 1997; 2001c, and healthcare: Portaluri et al., 2010). However, for the mining (Lenne, Salmon, Liu, & Trotter, 2012; Patterson & Shappell, 2010; Patterson, 2009) and maritime industries (Celik & Cebi, 2009) the highest percentage of accidents within this level is the physical and technical environment category, respectively, perhaps because of the harsh and continually changing environment of miners, and mariners. The most frequent causal factor identified in the cardiovascular surgery operating room was the communication and coordination causal code (ElBardissi, Wiegmann, Dearani, Daly, & Sundt, 2007).

The next two levels of HFACS are often infrequently investigated and, therefore, are underrepresented in most accident reports or narrative summary reports, meaning causal factors at the unsafe supervision and organizational tier are associated with fewer incident/accident cases than those at other tiers of HFACS (e.g., Wiegmann & Shappell, 2001a; Shappell et al., 2007). For the research available, the leading causal factors for each of these levels are inadequate supervision (e.g., Baysari, McIntosh, & Wilson, 2008; Gaur, 2005; Li, Harris, & Yu, 2008; Portaluri et al., 2010) and organizational processes (e.g., Baysari, McIntosh, & Wilson, 2008; Lenne, Salmon, Liu, & Trotter, 2012; Li, Harris, & Yu, 2008; Patterson & Shappell, 2010; Patterson, 2009; Portaluri et al., 2010). Aas (2008) in a study on oil and gas of the Norwegian offshore accidents found that 74% of the accidents examined had at least one contributing factor at the organizational level, 15% at the supervisory level, 7% at the preconditions for unsafe acts, and 4% at the unsafe acts level, a distinctive finding compared to other HFACS studies in which causal factors at the top two levels were relatively rare (e.g., Shappell et al., 2007; Wiegmann & Shappell, 2001a).

In addition, the comprehensiveness of HFACS as an investigative and analysis tool for accident causation has led to the development of the Aviation System Risk Model (ASRM), an analytical framework incorporating both data and expert judgments for projecting system risk, which evaluates the impact of technology interventions. This risk model, developed by Luxhoj (2003), involves three analytical approaches, HFACS, Bayesian Belief Networks (BBNs), and case studies and expert opinions. The analysis begins with discussion of accident cases with subject matter experts. Then the causal



factors are identified using the HFACS taxonomy. Influence diagrams are used to model interactions among the HFACS causal factors identified that are reviewed by subject matter experts. Next, conditional probability tables are created based on the opinions of these subject matter experts and integrated into a Bayesian Belief Network representing the industry of aviation maintenance. Subsequently, the efficiency of targeted interventions on HFACS causal factors is obtained from experts through sensitivity analysis. Finally, a user interface displays the expected risk on the relative risk intensity graph, a sample of which can be seen in Figure 2.8.

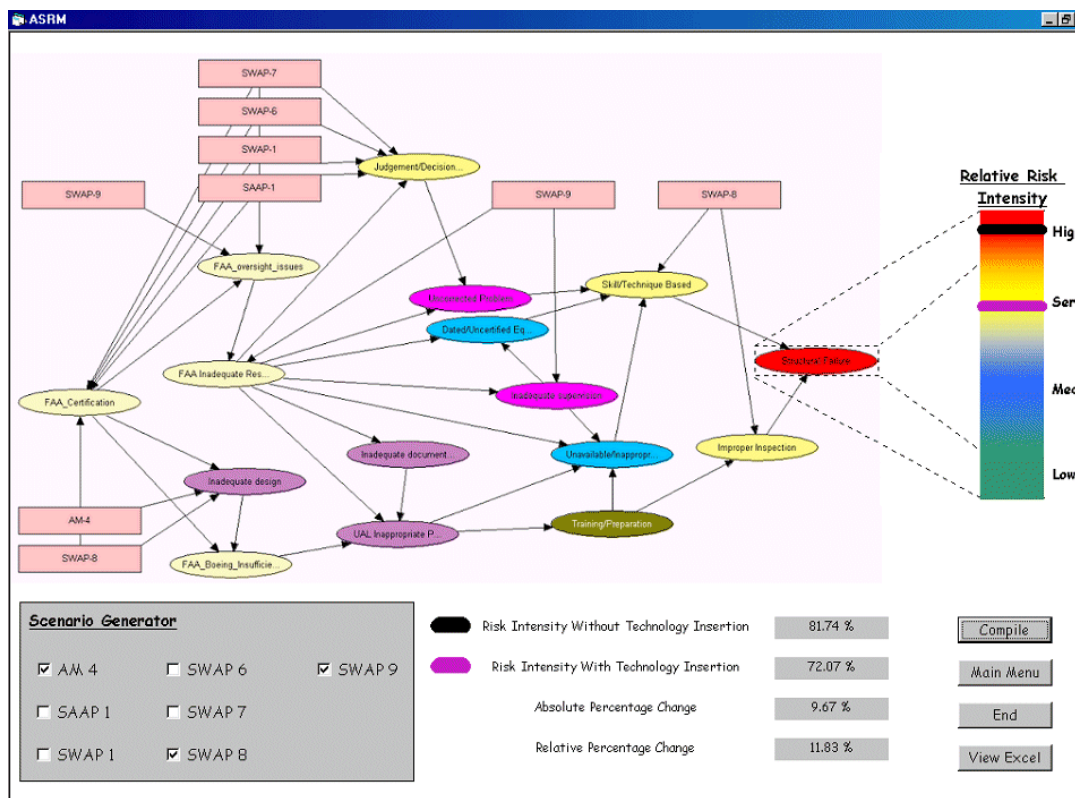
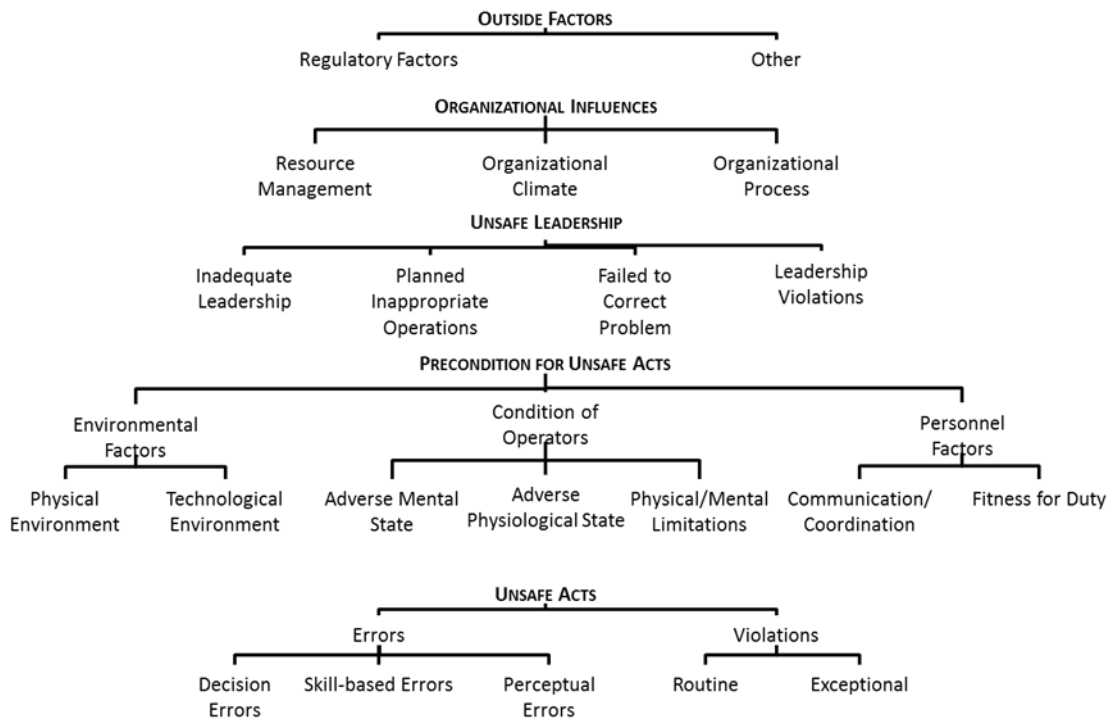
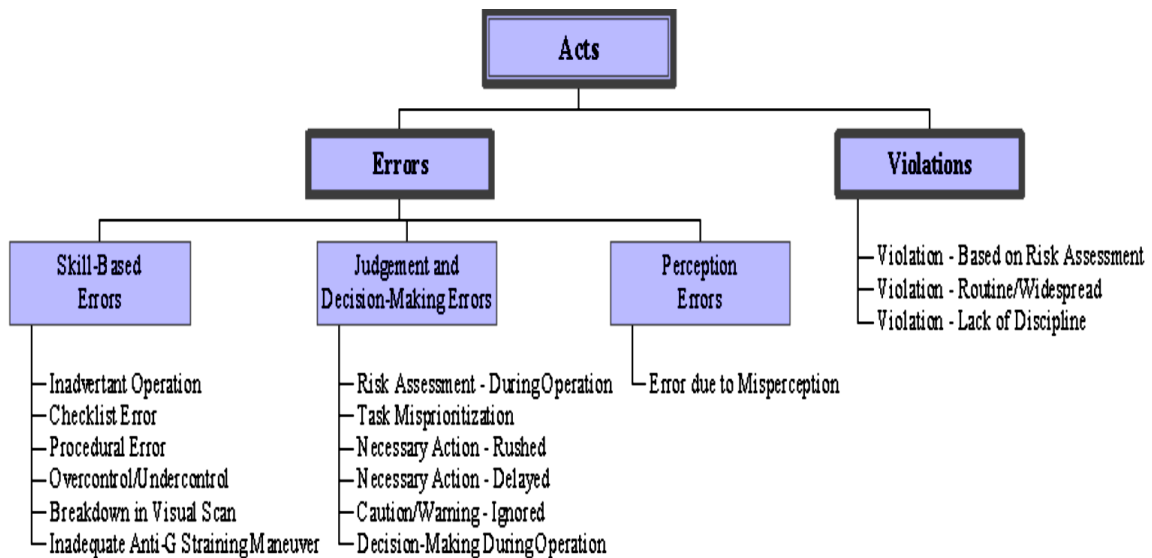


Figure 2.8: ASRM User Interface, Evaluation Form (Luxhøj & Kauffeld, 2003)

In an attempt to improve the comprehensiveness of HFACS, several HFACS derivatives have been developed, including HFACS maintenance (Krulak, 2004), HFACS railway operations (Baysari, Caponecchia, McIntosh, & Wilson, 2009; Reinach & Viale, 2006), Department of Defense HFACS (O'Connor, 2008; O'Connor, Walliser, & Philips, 2010), HFACS air traffic control (Scarborough & Pounds, 2001), HFACS mining (Patterson & Shappell, 2010), and HAFCS Australian Defense Force (Olsen & Shorrock, 2010). The structure of such derivatives is identical to the basic HFACS framework with slight variations appropriate for a specific industry. For instance, the HFACS Mining (HFACS-MI) structure is identical to the structure of HFACS except that a fifth level, Outside Factors, was added to account for external factors such as pressure from environmental groups and legal governmental influences as seen in Figure 2.9. Similarly, the Department of Defense HFACS (DoD- HFACS) adds a level of fine grain classification called nano codes, as shown below in Figure 2.10; the nano codes for violation category are violations based on risk assessment, violations that are routine/widespread, and violations due to lack of discipline. Although such derivatives have enhanced the comprehensiveness of HFACS for a particular industry, its comprehensiveness as a general accident investigation and analysis tool is maintained, since the majority of its categories are included in these derivative frameworks.



**Figure 2.9: Human Factors Analysis and Classification System-Mining Industry (HFACS-MI) Framework (Patterson & Shappell, 2010)**



**Figure 2.10: Department of Defense - Human Factors Analysis and Classification System (DoD-HFACS) Framework (Dept. of Defense, 2005)**

#### 2. 2.5.2.2 Usability

Usability is the framework's ability to be applied for practical use in industry. Similar to comprehensiveness, the usability of HFACS was suggested by its adoption by organizations like the U.S. Navy/Marine and the U.S. Army as an investigative and analysis tool for accident causation (Shappell & Wiegmann, 2001). Subsequently, it has seen successful applications in diverse industries including air traffic control (Broach & Dollar, 2002), civil aviation (Inglis & McRandle, 2007; Lenne, Ashby, & Fitzharris, 2008; Li, Harris, & Yu, 2008; Shappell et al., 2007; Ting & Dai, 2011; Wiegmann et al., 2005; Wiegmann & Shappell, 2001a), aviation maintenance (Krulak, 2004; Rashid, Place, & Braithwaite, 2010), mining (Lenne, Salmon, Liu, & Trotter, 2012; Patterson & Shappell, 2010), construction (Garrett & Teizer, 2009), railroads (Baysari, McIntosh, & Wilson, 2008; Baysari, Caponecchia, McIntosh, & Wilson, 2009; Reinach & Viale, 2006), healthcare (ElBardissi, Wiegmann, Dearani, Daly, & Sundt, 2007), oil and gas (Aas, 2008; Wang, Faghieh Roohi, Hu, & Xie, 2011), marine (Celik & Cebi, 2009; Schröder-Hinrichs, Baldauf, & Ghirxi, 2011), and security (Wertheim, 2010).

#### 2. 2.5.2.3 Diagnosticity

Diagnosticity is the framework's ability to show the relationships among errors and their trends and causes (Shappell and Wiegmann, 2001). Although the diagnosticity of the HFACS framework was originally verified case-by-case using aviation datasets (Shappell & Wiegmann, 2001), Dekker (2001) questioned the extent of the connection in the HFACS taxonomy between human error and the operational environment as it does not explain why an operator committed an error, only shifting it from the front end, at the

operator level, to higher up the organizational chain. However, recent research has investigated the statistical associations between the levels and the causal categories within HFACS (Berry, Stringfellow, & Shappell, 2010; Li & Harris, 2006; Li, Harris, & Yu, 2008; Tvaryanas & Thompson, 2008). These analyses have begun to describe statistically how actions and decisions at higher managerial levels propagate throughout the organization, resulting in active errors and, thus, accidents occur.

For instance, Li and Harris (2006) conducted an empirical study analyzing 523 accidents in the Republic of China (ROC) Air Force between 1978 and 2002 through the application of the HFACS framework. This study uses Goodman and Kruskal's lambda ( $\lambda$ ) to find the relationships, the links, between the lower categories and the immediately higher level in the framework. Based on those results, the study found various error pathways linking all four levels of the HFACS taxonomy. For instance, poor decisions at the organizational level significantly affects supervisory performance, thereby affecting preconditions for the unsafe acts level and, hence, indirectly affecting the performance of pilots at the operational level, Figure 2.9.

In a subsequent effort, Li, Harris and Yu (2008) analyzed 41 civil aviation accidents in the Republic of China (ROC) between 1999 and 2006 using the HFACS framework. This study identified paths relating errors at the operational level to the three levels above it, preconditions for unsafe acts, unsafe supervision, and organizational influences as seen in Figure 2.10. Specifically, at the HFACS highest level, the organizational process category is associated with the inadequate supervision category at level 3, and the latter is associated with crew resource management category, which

among many other categories at the second level, is associated with the immediate causes of many operational errors preceding accidents. The results support Reason's (1990) model that suggests active failures result from latent conditions in the organization.

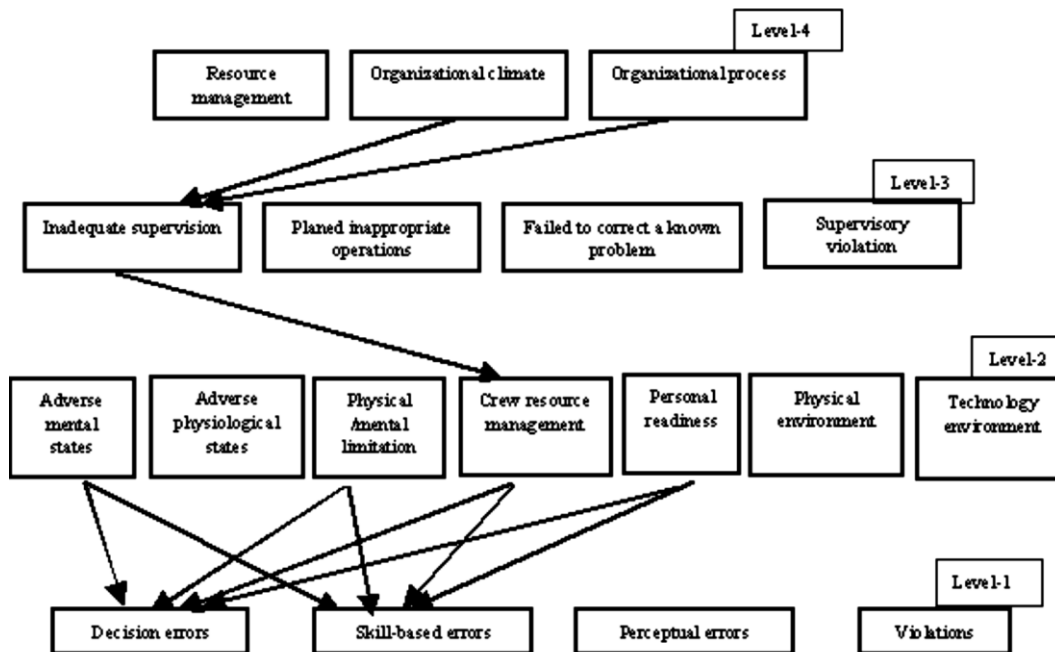
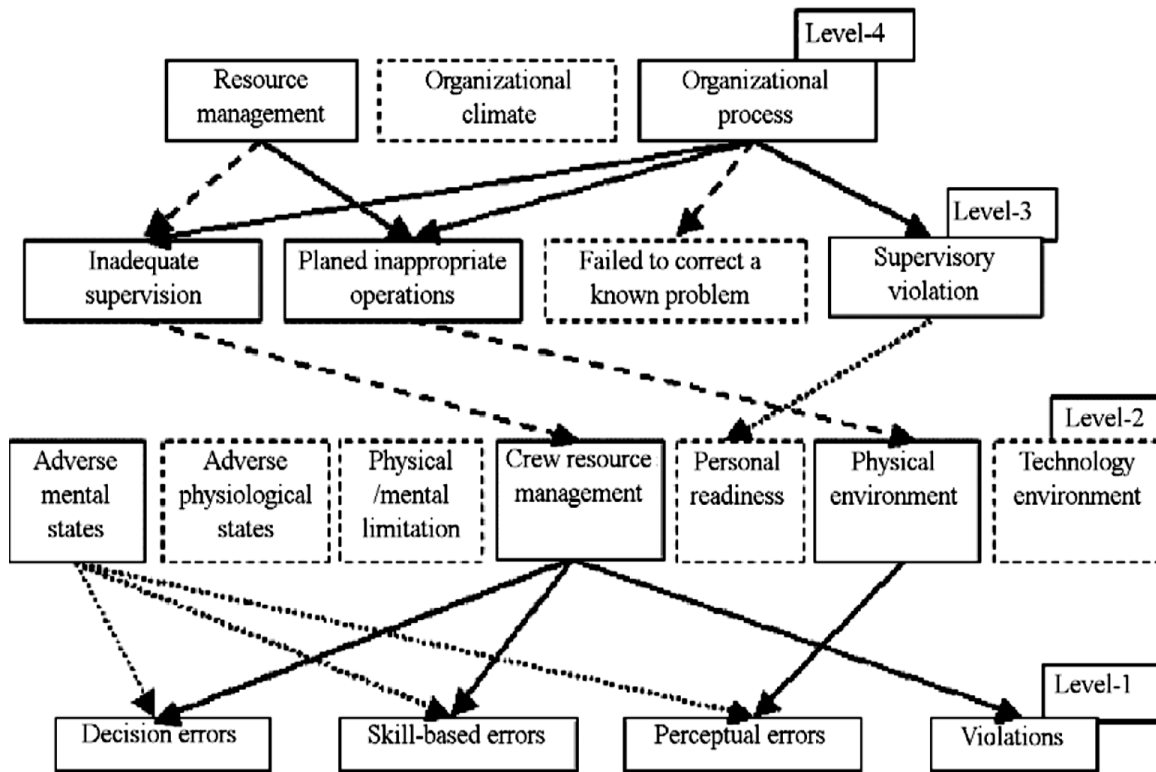


Figure 2.11: Significant Paths between Categories at the Four levels in the HFACS Framework (Li & Harris, 2006)



**Figure 2.12: Failure Paths between HFACS Categories (Li & Harris, 2008)**

Note: Solid lines indicate lambda in excess of 50%. Dashed thick lines indicate lambda in excess of zero. Dashed fine lines indicate Chi-Square is significant but lambda zero. Dashed rectangle indicates the category has no significant association with any lower level categories.

In a similar study, Tvaryanas and Thompson (2008) identified recurrent error pathways using the HFACS framework, analyzing 95 remote piloted aircraft (RPA) mishaps and safety incidents reported to the Air Force Safety Center 1997 – 2005. An interesting aspect of this study is the utilization of a tree diagram that quantitatively assesses the associations between active and latent failures including the identification of error pathways. Four recurrent error pathways associated with four types of HFACS active failures were identified. Two of these were related to situation awareness errors

associated with perception of the environment, 57% of which involved crew member mishaps.

While most of the studies that investigated the associations between active errors and latent conditions based on HFACS framework were aviation related, Berry, Stringfellow and Shappell (2010) conducted a study beyond this industry. They focused on identifying relationships between active errors and latent conditions in seven industries ranging from maintenance to mining to entertainment, looking for common human error pathways. Using Pearson's Chi-square test, odds ratio and the relative risk, significant causal factor pairings emerged from the analysis of adjacent and non-adjacent tiers as seen in Figure 2.11. Fifteen causal category pairs were found to be significant, twelve in the adjacent tier analysis and three in the non-adjacent. For the former, four associations were found between the unsafe supervision and the preconditions for unsafe acts tiers and eight associations were found between the preconditions for unsafe acts and the unsafe acts tiers; this high percentage of associations between these two tiers is due to their ease of investigation and classification. For the latter three, two associations were found between the unsafe supervision and the unsafe acts tiers and one between the organizational influences and the unsafe acts tiers.



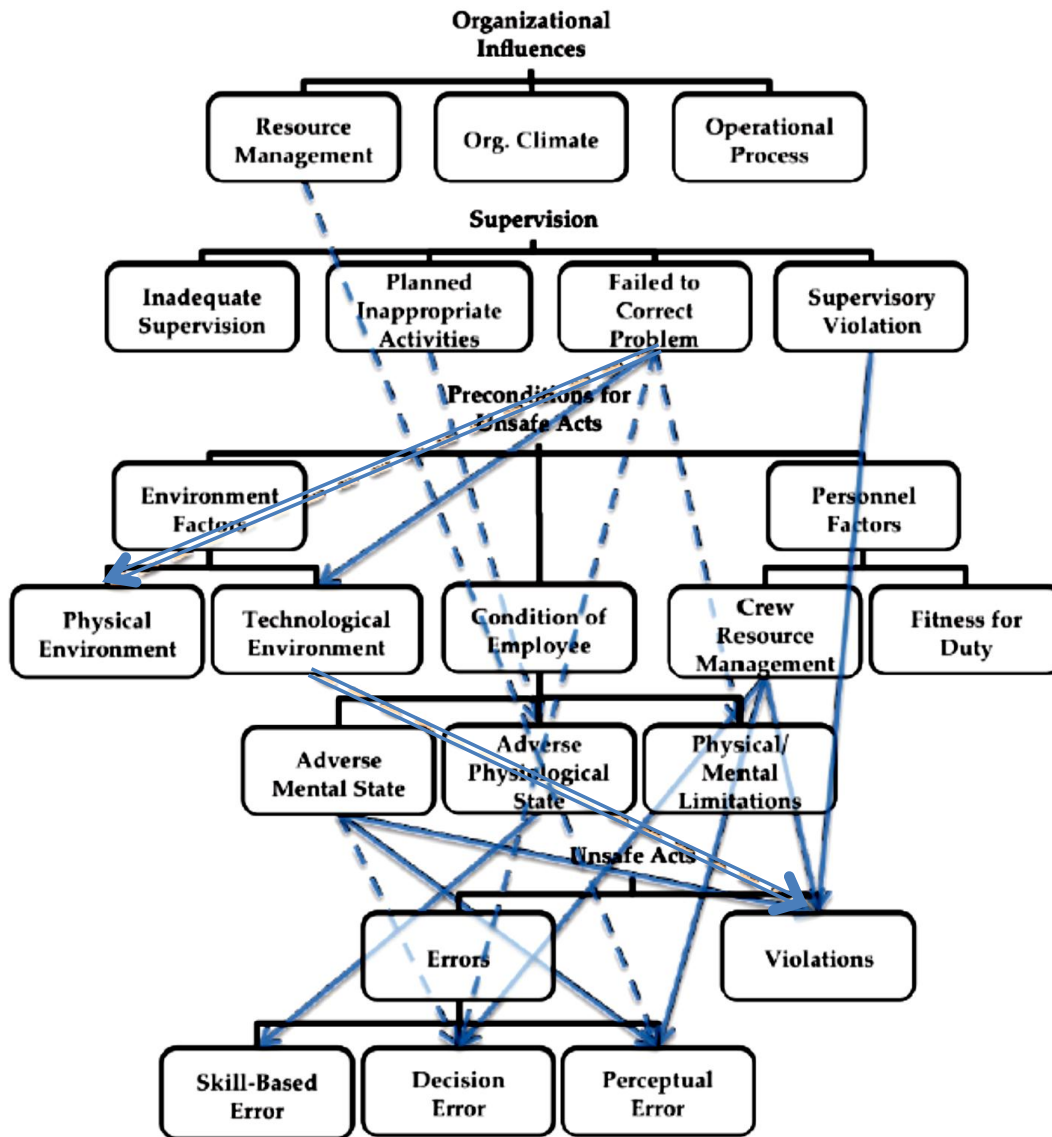


Figure 2.13: Pictorial Associations between HFACS Categories (Berry, 2010)

Note: Solid lines indicate associations with significant Chi-Square, Odds ratio, and lower relative risk results. Dashed lines indicate associations with significant Chi-Square and Odds ratio results. Double lines indicate associations with significant Chi-Square results.

#### 2. 2.5.2.4 Reliability

In addition to the previous criteria, comprehensiveness, usability, and diagnosticity, it is also very important that the HFACS satisfies a specific reliability standard. In general, reliability refers to the extent to which a framework, experiment, test or measuring instrument yields the same result over repeated trials (Carmines and Zeller, 1979). Evaluating reliability is a primary concern for many fields including the behavioral, psychological, medical, and social sciences, particularly as new methods, tests, devices, and instruments are developed.

Basically, there are two types of reliability important in scientific research: inter-rater and intra-rater reliability. While inter-rater reliability refers to the framework's ability to obtain the same results irrespective to who conducts the analysis, intra-rater reliability refers to the consistency of each rater. It is anticipated that minor variations may occur in both cases; however, in general, the more stable the results, the more confident that the results are reproducible and trustworthy. Of the two, inter-rater reliability is considered the most crucial, explaining why it has received the most research attention.

The inter-rater reliability assessment of the HFACS, which can be traced back to its development phase, was initially investigated using the military aviation accident database, specifically the Marine Corps Controlled Flight into Terrain accidents (Rabbe, 1996; Walker, 1996 as cited in Shappell & Wiegmann, 2001), Navy Tactical Aircraft accidents and Rotary Wing accidents (Plourde, 1997; Ranger, 1997 as cited in Shappell & Wiegmann, 2001), and 77 A-10 accidents (Johnson, 1997; Plourde, 1997 as cited

in Shappell & Wiegmann, 2001). In these studies, three raters classified several accident causal factors, with the inter-rater reliability being determined for every pair of raters using Cohen's Kappa . In these initial studies, Cohen's Kappa ranged between 0.65 to 0.70 in the early studies increasing to 0.93 to 0.95 in the later ones; this improvement was probably due to the continuous enhancement in defining the HFACS causal categories, indicating that a reliable framework has been developed to be used in this field.

Subsequently, Weigmann (2000) extended the inter-rater reliability of HFACS to two sets of commercial aviation accident data, the first involving 44 air carrier accidents and the second 79 commercial aviation accidents; again, the measure was Cohen's Kappa ( $K = 0.65$  and  $K = 0.75$ , respectively). Expanding the sample size of this study, Weigmann and Shappell (2001c) conducted the largest inter-rater reliability assessment of HFACS to date using a dataset involving 2,500 general aviation accidents associated with more than 6,000 causal factors classified by five raters. This study showed an average Cohen's Kappa value of 0.72, implying a substantially reliable framework (Landis & Koch, 1977).

Similar to the developers of HFACS, Li and Harris (2005) investigated the inter-rater reliability of 523 ROC Air Force aviation accidents associated with more than 1,762 HFACS causal factors. An instructor pilot and an aviation psychologist classified these factors independently and reliability was assessed using Cohen's Kappa and percent agreement. While Cohen's Kappa ranged from 0.44 to 0.83, revealing moderate to satisfactory agreement, percent agreement fluctuated between 72% and 96.4%, demonstrating acceptable inter-rater reliability without considering agreement by chance.

More recently, Olsen (2011) evaluated the inter-rater reliability of HFACS in the Australian military air traffic control (ATC) environment. Two groups of coders, three human factors ATC specialists and four air traffic controllers self-trained through a self-paced workbook independently classified causal factors of 14 incident reports pre-identified by the researcher using HFACS. The results revealed low inter-rater reliability for both groups; percentage agreement for both the category level and the tier level for each group was 36.1% for the ATCO group and 34.5% for the HF specialist group, and 64.8% for the ATCO group and 56.4% for the HF specialist group, respectively.

In addition, two studies have investigated the reliability of several HFACS derivatives, Department of Defense HFACS (DoD-HFACS) and HAFCS Australian Defense Force (HFACS-ADF) (O'Connor, 2008; O'Connor, Walliser, & Philips, 2010; Olsen & Shorrock, 2010). O'Connor (2008) investigated the inter-rater reliability of the DoD-HFACS framework, by determining the within-group inter-rater reliability coefficient ( $r_{wg}$ ) and percent agreement of 123 coders. These coders, students at the Navy/Marine Corps School of Aviation Safety, identified and classified the human factors causes of two aviation mishap scenarios. While at the categorical level percent agreement ranged from 53% to 99%, fluctuating between fair to excellent inter-rater reliability, at the nano level percent agreement ranged from 24% to 43%, indicating acceptable levels of inter-rater reliability was not achieved.

Using the same derivative, O'Connor (2010) used multi-rater Kappa free ( $K_{free}$ ) to evaluate the inter-rater reliability of the DoD-HFACS; in this study 22 military officers classified causes of an aviation incident by interviewing a U.S. Navy officer involved in

the incident. The results showed an average Fleiss' Kappa of 0.76 at the categorical level, a moderate level of inter-rater reliability.

In the second study, Olsen and Shorrock (2010) investigated the inter-rater and Intra-rater reliability of the HFACS Australian Defense Force (HFACS-ADF) framework by calculating percent agreement. First, to investigate the inter-rater reliability, 11 air traffic control officers (ATCOs) from the Royal Australian Air Force with different levels of training and experience with HFACS-ADF classified two randomly chosen ATC incident reports. Percentage agreement at the category level and the nano level were 39.9% and 19.8%, respectively, both considered unsuitable levels of inter-rater reliability. Second, to investigate the Intra-rater reliability, four members of the ATC classified five incident reports within a 4-to-20 month time period. The results showed that percent agreement at the category level ranged from 36.2% to 46.2% and at the nano level from 26.7% to 43.8%. Both the inter-rater and Intra-rater reliability were very low, suggesting that the HFACS-ADF is unreliable.

In addition to the limited number of studies investigating the reliability of HFACS, four limitations in their approaches make the comprehensive comparison across these studies difficult. The first limitation is the number of coders, which fluctuates from two to twenty-two, meaning that most of these studies used only a few raters, perhaps indicating sample bias. The second limitation is the level of experience of the coders, ranging from students to human factors specialists, a factor that might also affect the generalizability of such studies. Another limitation is that the majority of these studies used only aviation datasets, and the number of causal factors that were classified differed.

Moreover, only one to two types of statistical measures were used to assess the reliability of HFACS. Finally, the comparison of intra-rater reliability of HFACS in these studies is limited since only one part of one study of the seven considered Intra-rater reliability.

As more and more industries are adopting HFACS framework as an investigation and analysis tool for incidents/accidents, safety professionals must be confident that the data are valid and reliable. The study proposed here addresses these limitations by evaluating both the inter-rater and intra-rater reliability of HFACS. Initially, the targeted number of raters is more than 80; additionally, this study attempts to ensure that the coders have had standardized training and similar experiences in the real-world use of HFACS prior to participating in this study. Moreover, to ensure comprehensiveness of the data used, accident causal factors will be populated from various datasets ranging from lodging to mining to construction. Finally, because some statistical measures are more appropriate for nominal data, the percent agreement, Krippendorff's Alpha ( $\alpha$ ), Cohen's Kappa (K), and Fleiss' Kappa ( $K_F$ ) will be used to investigate the reliability of HFACS; in addition, using multiple measures is more likely to ensure a comprehensive evaluation of the reliability of HFACS.

## CHAPTER 3: METHODOLOGY

To investigate and evaluate the inter-rater and intra-rater reliability of the HFACS, this study proposed to use a within-subject design. This experimental design was selected because of its statistical benefit, as the number of subjects increase, statistical power increases. A description of the participants, the instrument, and the procedure, all of which have been IRB-approved through Clemson University, is provided in this chapter. In addition, it covers the data collection techniques and the statistical procedures and packages used in the study.

### 3.1 Participants

One hundred and twenty five safety professionals from various industries considering implementing the HFACS as an alternative for the current accident investigation and analysis system in their workplaces participated in this study. Two days of instruction on HFACS was provided to all participants through HFACS Inc. This entry-level training, designed for safety specialists engaged in the investigation and/or analysis of accidents, included a comprehensive description of the HFACS structure, the nineteen causal codes, to enable the participants to classify mishap/accident causal factors accurately in relation to the relevant human error level and to the appropriate HFACS code. The participants were recruited through a face-to-face presentation during the training. The study was explained to the participants, and were asked if they wished to participate; participation in the study was voluntary.

### 3.2 Instrument

The self-developed survey using a Google form found in Appendix A was used as the primary data-gathering instrument in this study. This survey was divided into two sections: the user identifier and the survey. While the user identifier was the participant's email address, which provides accurate differentiation between participants, the survey statements were structured using a multiple choice format. In the Google form, each of the 95 causal factors was formatted as a statement, followed by the 19 HFACS causal codes. The participants were required to read and identify each causal factor, and then attribute it to the causal code that best described it by checking the appropriate box.

The self-developed survey used here to measure the reliability of HFACS provides a focus different from the majority of reliability studies (Olsen & Shorrock, 2010; Wallace, Ross, Davies, Wright, & White, 2002). Frequently in these studies, the participants were provided with two or more incident reports needing to be coded. This approach not only measures the ability of a coder to code a causal factor to the right code but also incorporates the ability of the coder to identify the presence of a certain causal factor, extract it, and then classify it. As a result, reliability is tested on both selection and coding of events, while in this study reliability was tested on coding of events only.

The causal factors represented by the statements on the survey were extracted from actual accident reports from the National Transportation Safety Board (NTSB), Occupational Safety and Health Administration (OSHA), and other HFACS accident databases such as mining and lodging. For example, the causal factor statement in the survey, "The captain chose to continue fishing despite the severe weather predictions and



the exposed location of the ship Katmai,” was extracted from the NTSB accident report number DCA-09-CM-001:

National Weather Service data indicated that at the time of the accident, winds were from the east at 60–70 knots; the air temperature was 38 F; the water temperature was 43 F; the wave height was 20–30 feet; and prevailing conditions were rain with no icing. Despite severe weather predictions and the exposed location of the Katmai, the master chose to continue fishing. The master told the marine board that, at about 0200 on October 21, the Katmai had completed fishing operations and crew members had begun to store their gear in preparation for the return to Dutch Harbor (NTSB, 2011).

This approach was adopted to ensure content validity of the survey statements. Moreover, to ensure coders will remain focused and alert, each causal code had a total of five causal factors randomly ordered in the Google survey, meaning 95 causal factors were coded by all participants. Additionally, to prevent training bias, the HFACS instructors were not involved in any stage of the development of the survey and did not have access to the Google form.

The multiple choice format was selected as it enables the respondents to answer the survey easily with minimum error as opposed to a drop box or fill in the blank. In addition, the Google form allowed the researcher to perform the computations efficiently as the results of the survey were compiled automatically in a spreadsheet. To ensure the survey statements used for the study were clear, the researcher tested it twice using ten

total respondents each participating twice. The respondents' responses were used only for testing purposes and did not form part of the study.

For this pilot study, ten graduate students in the Industrial Engineering Department at Clemson University were offered 2-hour refresher course on HFACS. Subsequently, the respondents were given the self-developed survey in which each HFACS causal code involved a total of six causal factors, meaning the respondents coded 114 causal factors. Further, during this first round, the researcher asked the respondents to record any statement number that caused confusion and at what point, if any, they experienced fatigue. The researcher revised any survey statement having a difficulty index above 60% and if 40% or more found it ambiguous. A difficulty index is the percentage of students who submitted an incorrect answer. The researcher modified the vague and compound causal factor statements into simpler ones to guarantee comprehension.

For the second round of the pilot study, the researcher reduced the number of survey statements to 95 from the original 114 since the majority of the respondents indicated fatigue at approximately statement number 100. The same ten respondents and testing criteria were used as in round one. In addition, to ensure the suitability of the survey statements for the study, item analysis was performed on the survey statements that had high difficulty index which allowed for further refinement to these statements.

### 3.3 Procedures

The majority of the 125 participants participated in two experimental sessions, the first immediately at the end the HFACS training and the second two weeks from the first session. The first session did not exceed an hour and a half and was conducted in the location where the training took place; all participants were required to bring their personal computers. For the second session, conducted 2-weeks later, the participants completed the reliability trial at their convenience, submitting the survey within a 72-hour window.

For the first session, the initial five minutes were devoted to clarifying the instructions and distributing the consent form (Appendix B). Oral and written instructions emphasized that the participants are to work individually. While the participants started their personal computers and signed in to their email, the researcher emailed a link of the Google form to each of them. Once the Google form has been accessed, the participants (i.e., the coders) read and classified each causal factor into the causal code which was the best description. Upon completion of the coding of all the causal factors, the participants submitted the Google form. The participants' responses along with their unique user IDs were combined into a spreadsheet having the same name as the Google form. When the session was over, the researcher deleted the Google form to prevent further participant accessibility, and the data was converted to an Excel spreadsheet.

### 3.4 Data Management and Statistical Analysis

The data obtained from the classification of each survey statement into the HFACS causal code by all participants was not in appropriate form for the analyses of evaluating the intra-rater and inter-rater reliability of HFACS. The raw data obtained from the two sessions were converted into numerical notations using an Microsoft Excel macro. For example, a skill-based error was converted to a 1 and a decision error to a 2; Appendix C provides the lists of the 19 HFACS categories and their numerical notations for each HFACS tier and category level, respectively.

Reliability of HFACS is established by demonstrating agreement among coders. Many agreement measures have been proposed; for example, for nominal data Popping (1988) listed 43 measures. In addition, the lack of consensus among statisticians and researchers on which measures are appropriate further increases the complexity of the decision on which to use. This study used four measures to analyze the data for this study: percent agreement (PA), Krippendorff's Alpha ( $\alpha$ ), Cohen's Kappa (K), and Fleiss' Kappa ( $K_F$ ). The percent agreement and Krippendorff's Alpha ( $\alpha$ ) were used to assess both the intra-rater and inter-rater reliability of HFACS. Cohen's Kappa was used to evaluate the intra-rater reliability of HFACS and Fleiss' Kappa ( $K_F$ ) was used to assess the inter-rater reliability. Table 3.1 summarizes the agreement measures used in this study to determine and evaluate the inter-rater and intra-rater reliability of HFACS.

Table 3.1: Agreement Coefficients with Respect to Different Measures

Criteria	Percent Agreement (PA)	Cohen's Kappa (K)	Fleiss' Kappa (K <sub>F</sub> )	Krippendorff's Alpha (α)
<b>Tests for</b>	Inter-rater and Intra-rater	Intra-rater	Inter-rater	Inter-rater and Intra-rater
<b>Number of coders</b>	2	2	>2	>=2
<b>Type of data</b>	Nominal	Nominal	Nominal	Nominal, ordinal, interval, ratio
<b>Corrects for chance agreement</b>	No	Yes	Yes	Yes
<b>Independent coders</b>	Yes	Yes	Yes	Yes
<b>Accounts for Missing observations</b>	No	No	No	Yes
<b>Value</b>	0 - 100 %	-1.0 – 1.0	0.0 -1.0	0.0 -1.0
<b>Independent of the number of observers employed</b>	Yes	No	Yes	Yes
<b>Allows observers to be freely permutable or interchangeable</b>	Yes	No	Yes	Yes
<b>Confounded by the number of categories</b>	Yes	Yes	Yes	No

To verify the reliability of HFACS, the following are defined:

- The set of items (i) that are coded,  $i = 1, 2, 3, \dots, I$ ;
- The set of categories (C) into which the items are coded,  $c = 1, 2, 3, \dots, C$ ; and

- The set of coders ( $R$ ) who designate for each item a distinctive category,  $r = 1, 2, 3, \dots, R$ .

For this study,  $I$  represents the survey statements ( $I=95$ ),  $C$  the HFACS codes ( $C=19$  at the category level,  $C=4$  at the tier level), and  $R$  the number of participants or coders ( $R=125$  for inter-rater,  $R=59$  for intra-rater). Additional notations used in this study are listed in Table 3.2.

Table 3.2: List of Notations

<b>Notation</b>	<b>Definition</b>
$A_o$	Observed agreement
$D_o$	Observed disagreement
$PA$	Percent agreement
$p$	Pair of coders
$D_e$	Expected disagreement
$P_o$	Proportion of observed agreement
$P_e$	Proportion of expected agreement
$N$	Total number of items to be classified
$n_{ic}$	Number of coders who assigned item $i$ to category $c$
$N_{rc}$	Number of items assigned by coder $r$ to category $c$
$n_c$	Total number of items assigned by all coders to category $c$

The first measure used to assess the inter-rater and intra-rater reliability of HFACS was percent agreement (PA). Scott (1955) defines percent agreement as the percentage of instances on which two independent coders agree when coding the same data. For example, for a pair of coders who code all items,  $I$ , percent agreement is computed as follows

$$PA = \frac{\sum_{i=1}^I A_{oi}}{I} * 100, \quad (1)$$

where

$$A_{oi} = \begin{cases} 1 & \text{if the two coders assign item } i \text{ to the same category, } c \\ 0 & \text{if the two coders assign item } i \text{ to different categories} \end{cases}$$

For inter-rater reliability, the total number of pairs, P, depends on the number of coders who participated in each session, where R=125 for the first session and R = 59 for the second, which was calculated using the formula  $P = R*(R - 1)/2$ . Percent agreement was determined for each pair of coders, meaning that the response of each coder was paired with another coder's response and PA was determined using Equation 1; then an overall average percent agreement was determined for all pair of coders using a Microsoft Excel macro. The overall average percent agreement for inter-rater was calculated using the following formula:

$$\text{overall average } PA_{(\text{inter-rater})} = \frac{\sum_{p=1}^P PA_p}{(R*(R - 1)/2)} * 100. \quad (2)$$

For intra-rater reliability, p represents a specific coder's response from the first session compared to his response for the second; PA was again determined using Equation 1. Thus, the total number of pairs equals the number of coders who participated in both reliability sessions, (P=R=59). The overall average intra-rater percent agreement was calculated using the following formula:

$$\text{overall average } PA_{(\text{intra-rater})} = \frac{\sum_{p=1}^P PA_p}{R} * 100. \quad (3)$$

In addition, the intra-rater reliability for each tier and category using percent agreement was determined. The observed agreement of a coder's response from the first session to the second with respect to each category (or tier),  $c$ , is

$$A_{oi(c)} = \begin{cases} 1 & \text{if the coder in both sessions assigns item } i \text{ to the same category } c \\ 0 & \text{if the coder in both sessions assigns item } i \text{ to different categories} \end{cases}$$

Also, the number of items assigned by coder,  $r$ , to a specific category (or tier),  $c$ , in the first session,  $N_{rcFS}$ , was determined. The percent agreement for each coder,  $r$ , for each category (or tier),  $c$ , was calculated using

$$PA_{r(c)} = \frac{\sum_{i=1}^i A_{oi(c)}}{N_{rcFS}} * 100. \quad (4)$$

Subsequently, an overall average percent agreement for each category (or tier),  $c$ , was determined using

$$\text{overall average } PA_{(c)} = \frac{\sum_{r=1}^R PA_{r(c)}}{R}. \quad (5)$$

For this research, percent agreement was also used to detect rogue coders that may have an impact on the inter-rater reliability results. Given the large sample size of the percent agreement values of all coders who participated in each session ( $P=7750$  for



first session, P=1711 for second session), rogue coders were identified using the empirical rule, which states that in a symmetric distribution, approximately 95% of observations lie within 2 standard deviations of the mean:  $\bar{x}_{\text{percent agreement}} \pm 2 * s_{\text{percent agreement}}$  (Wilcox, 2010). For example, in the first session every coder was paired with the other 124, and percent agreement for the 7750 pairs, the PA sample mean, and the PA sample standard deviation were determined along with the 95% cutoff values based on the empirical rule. For each coder, all PA values were compared with the lower limit of the 95% cutoff. The fraction of times the PA value was less than the lower limit of the 95% cutoff out of the 125 was determined. A coder with at least 22% of PA values below the cutoff was considered rogue; this procedure was implemented using a Microsoft Excel macro. The analysis of the data obtained for this study was conducted with and without rogue coders for each session.

Because percent agreement is easily calculated, this method has seen wide-spread use; however, most researchers do not rely on it solely as it does not take into consideration that a proportion of coder agreement may be due to chance. For example, rater X may use one set of guidelines to distinguish between the presence or absence of the physical environment as a cause and a second rater Y, using a different set of guidelines, may arrive at the same conclusion. In addition, coders might simply agree just by guessing. Such observed agreements may be explained by chance; for this reason Krippendorff's Alpha ( $\alpha$ ), Cohen's Kappa (K), and Fleiss' Kappa ( $K_F$ ) were also used in this study.

Krippendorff's Alpha, widely considered to be a robust and versatile reliability coefficient, was also used in assessing the inter-rater and intra-rater reliability of HFACS. Krippendorff's Alpha can be applied to large and small sample sizes, any number of coders, and any number of categories, while adjusting to various types of measurement (e.g., nominal, ordinal, or ratio). More importantly, this measure can compensate for missing data and yet, its results are considered viable and accurate (Krippendorff, 2012).

The computation of Krippendorff's Alpha depends on the observed coincidence matrix in which the number of values that participate in pair comparisons are tabulated. Consider, the following observed coincidence matrix,

$$\begin{array}{cccccc}
 & & 1 & \cdot & c & \cdot & C & & \\
 1 & & \boxed{\begin{array}{cccc} o_{11} & \cdot & o_{1c} & \cdot & o_{1C} \end{array}} & & n_{1.} & & \\
 \cdot & & \cdot & \cdot & \cdot & \cdot & \cdot & & \\
 q & & \boxed{\begin{array}{cccc} o_{q1} & \cdot & o_{qc} & \cdot & o_{qC} \end{array}} & & n_{q.} = \sum_{c=1}^C o_{qc} & & \\
 \cdot & & \cdot & \cdot & \cdot & \cdot & \cdot & & \\
 Q=C & & \boxed{\begin{array}{cccc} \cdot & \cdot & \cdot & \cdot & \cdot \end{array}} & & \cdot & & \\
 & & n_{\cdot 1} & \cdot & n_{\cdot c} & \cdot & \cdot & & n_{\cdot.} = \sum_{q=1}^Q n_{q.}
 \end{array}$$

the columns are the set of categories (C) into which the items are coded,  $c = 1, 2, 3, \dots, C$  and the rows, which are identified as (Q), are also the set of categories ( $Q=C$ );  $o_{qc}$  represents the number of observed coincidences for the two values  $q$  and  $c$  when, for example  $x_{cq}$  is the number of times a particular coder uses  $q$  while the second uses  $c$ , the number of coincidences  $o_{qc} = x_{qc} + x_{cq}$ . Thus, the coincidence matrix is symmetrical around the diagonal,  $o_{qc} = o_{cq}$ , and the margins of the coincidence matrices enumerate the values used by all coders (Krippendorff, 2006). In contrast, agreement tables tabulate the

numbers of units being coded, and thus are not symmetrical around the diagonal and the summation of the margins of the agreement table are equal to the number of items being coded.

Krippendorff's Alpha is computed as

$$\alpha = 1 - \frac{D_o}{D_e}, \quad (6)$$

where the observed disagreement ( $D_o$ ) and the expected disagreement ( $D_e$ ) are represented as

$$D_o = \frac{1}{n_{..}} \sum_{q=1}^Q \sum_{c=1}^C o_{qc} \text{ nominal} \delta_{qc}^2 \quad (7)$$

and

$$D_e = \frac{1}{n_{..}(n_{..}-1)} \sum_{q=1}^Q \sum_{c=1}^C n_{q.} n_{.c} \text{ nominal} \delta_{qc}^2, \quad (8)$$

where,  $\text{nominal} \delta_{qc}^2 = \begin{cases} 0 & \text{iff } q = c \\ 1 & \text{iff } q \neq c \end{cases}$  (Krippendorff, 2007).

Two reliability scale values are identified with Krippendorff's Alpha, with  $\alpha = 1$  representing perfect reliability and  $\alpha = 0$  representing the absence of reliability, thereby denoting that the categories are statistically unrelated to the items they describe. While a Krippendorff's alpha value of 0.8 and above is considered reliable, a value between 0.667 and 0.8, although not deemed reliable, can be used to draw tentative conclusions (Krippendorff, 2006). Krippendorff's Alpha ( $\alpha$ ) was computed to determine the inter-rater reliability in this research using the KALPHA macro in SAS v. 9.2. An advantage of using this macro is that it computes the distribution of Krippendorff's Alpha through bootstrapping, thus providing two additional measures: a confidence interval for Alpha at

a defined level of statistical significance and a probability that Alpha could be less than a chosen minimum required for data to be deemed sufficiently reliable (Hayes & Krippendorff, 2007).

In addition, to determine the inter-rater reliability for a single category or tier, the overall coincidence matrix for all items from the KALPHA macro output was also used in this study. The Alpha agreement for a single category (or tier),  $c$ , was determined by

$$\alpha_{(c)} = 1 - (n_{..} - 1) \frac{\sum_q^c o_{qc \text{ nominal}} \delta_{qc}^2}{n_{.c} \sum_q^c n_{q \cdot \text{ nominal}} \delta_{qc}^2} \quad (\text{Krippendorff, 2013}). \quad (9)$$

Similarly, the intra-rater reliability for each tier and category using Krippendorff's Alpha was determined for each coder who participated in the two sessions (R=59). For each coder, an overall Krippendorff's Alpha and coincidence matrix including the data from the first session and the second session were determined for all items, I, generating R = 59 Krippendorff's Alpha values and coincidence matrices. Additionally, using Equation 9, a Krippendorff's Alpha for each category and tier,  $c$ , was determined for each coder which included the data from the first session and the second session, generating R = 59 Krippendorff's Alpha values for each C=4 tiers and C=19 categories. Then, an overall average Krippendorff's Alpha for each tier and category was determined by

$$\text{Overall Average Krippendorff's Alpha}_{(c)} = \frac{\sum_{r=1}^R \alpha_r(c)}{R} . \quad (10)$$

HFACS intra-rater reliability was also assessed by Cohen's Kappa (K). This measure, which is regarded as the most widely used reliability coefficient (Kolbe & Burnett, 1991; Vach, 2005; Zwick, 1988), estimates the degree of agreement between two coders across different categories after adjusting for the agreement that could be attributed to chance alone. The computation of Cohen's Kappa is based upon the values of the marginal distributions (MD) of the coders (i.e., a distribution for the categorical variables indicating the total frequency of each outcome) thus, it is known to be marginal or prevalence dependent (Nelson & Pepe, 2000).

Cohen's Kappa employs square cross-classifications of the judgments of two coders' known as agreement tables. For example, Table 3.3 illustrates an agreement table for two coders A and B who code a specific set of items,  $i = 1, 2, 3, \dots, I$ , to  $c = 1, 2, 3, \dots, C$  categories. The frequencies ( $m$ ) in the agreement table in Table 3.3 give the number of instances in which both coders A and B identified a particular category. For instance,  $m_{1c}$  is the number of instances coder A used category 1 and coder B used Category c. The cells, along the diagonal, display the number of incidences in which the two coders used the same category; these are called agreement cells.

Table 3.3: Agreement Table of Two Coders

		Coder B				MD	P <sub>a</sub>	
		1	·	c	·			C
Coder A	1	$m_{11}$	·	$m_{1c}$	·	$m_{1C}$	$n_{1.}$	$P_{a1} = n_{1./N}$
	·	·	·	·	·	·	·	·
	c	$m_{c1}$	·	$m_{cc}$	·	$m_{cC}$	$n_{c.} = \sum_1^C m_{c.}$	$P_{ac} = n_{c./N}$
	·	·	·	·	·	·	·	·
	C	$m_{C1}$	·	$m_{Cc}$	·	$m_{cC}$	·	·
MD		$n_{.1}$	·	$n_{.c}$	·	·	N=I	
P <sub>b</sub>		$P_{b1} = n_{.1}/N$	·	$P_{bc} = n_{.c}/N$	·	·		

Based on this agreement table, Cohen's Kappa (K) is computed by

$$K = \frac{P_o - P_e}{1 - P_e}, \quad (11)$$

where the proportion of expected agreement ( $P_e$ ) and the proportion of observed agreement ( $P_o$ ), are represented as

$$P_e = \sum_{c=1}^C P_{ac} * P_{bc} \quad (12)$$

and

$$P_o = \frac{1}{N} \sum_{c=1}^C m_{cc}, \quad (13)$$

where  $P_{ac}$  and  $P_{bc}$  are the marginal probabilities – the probability of a specific category, c, regardless of the values of the other categories – for coders A and B, respectively. While  $P_o - P_e$  is the actual amount of agreement beyond chance,  $1 - P_e$  is the largest possible discrepancy between  $P_o$  and  $P_e$ . Cohen's Kappa (K) was computed in this research using the FREQ procedure in SAS v. 9.2.

Kappa can range from -1 to 1, with  $K = 0$  representing agreement at the chance level,  $K = 1$  representing perfect agreement beyond chance, and a negative value, agreement less than chance. Landis and Koch (1977) recommended guidelines for interpreting values of both Cohen's and Fleiss' Kappa (Table 3.4). Although these guidelines have been recommended by a number of sources (Agresti, 2007; Stokes, Davis, & Koch, 2000), there are limited studies supporting their accuracy. Fleiss (1981) suggested a similar interpretation of Kappa: a Kappa value less than 0.40 indicating poor agreement, a Kappa value between 0.40 and 0.75 good agreement, and a Kappa value above 0.75 excellent agreement.

Table 3.4: Kappa Interpretations (Landis & Koch, 1977)

<b>K</b>	<b>Interpretation</b>
<b>&lt; 0</b>	Poor agreement
<b>(0.00 – 0.20]</b>	Slight agreement
<b>(0.20 – 0.40]</b>	Fair agreement
<b>(0.40 – 0.60]</b>	Moderate agreement
<b>(0.60 – 0.80]</b>	Substantial agreement
<b>(0.80 – 1.00]</b>	Almost perfect agreement

While Cohen's Kappa is widely used, its results are criticized for yielding what is known as the Kappa paradox (Nelson & Pepe, 2000; A. von Eye & von Eye, 2008; Warrens, 2010). Two such paradoxes have been identified in the literature: the first, high levels of observed agreement may yield low Kappa values, a result dependent on the characteristics of the sample being coded. The second, more probable paradox is that for a fixed observed agreement, Kappa can have different values depending on the symmetry

of observations in the disagreement categories (Warrens, 2010). However, Vach (2005) emphasizes that the second paradox is not a serious disadvantage for this measure, provided that the results are carefully analyzed and interpreted.

HFACS inter-rater reliability was also assessed using Fleiss' Kappa ( $K_F$ ), which measures the degree of agreement for more than 2 coders beyond that which would be expected by chance using pairwise agreement. As a result, the agreement for a specific item is defined as the proportion of coded pairs agreeing of the total number of coded pairs for that item (Fleiss, 1981).

Although, Equation 11 used to compute Cohen's Kappa is also used to determine Fleiss' Kappa, the corresponding agreement table (Table 3.3) is not suitable for displaying the data for Fleiss' Kappa ( $K_F$ ). In general, the classification of multiple coders for Fleiss' Kappa ( $K_F$ ) is displayed in a table similar to Table 3.5, which shows the classification of 5 coders classifying  $I$  items into 4 categories. However, to determine this agreement coefficient, this table is reordered to create Table 3.6 emphasizing the items,  $I$  and categories  $C$  rather than the coders.

Table 3.5: Classification of Multiple Coders

Item	Coders R=5				
	Coder1	Coder2	Coder3	Coder4	Coder5
<b>1</b>	2	2	3	2	4
<b>2</b>	1	1	1	1	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
<b>I-1</b>	2	3	3	3	2
<b>I</b>	c	c	c	c	c



Table 3.6: Distribution of Coders by Item Number and Classification

Item	Classification				Total Coders
	c=1	2	3	C=4	
<b>1</b>	0	3	1	1	5
<b>2</b>	5	0	0	0	5
<b>i</b>	.	$n_{ic}$	$n_{33}$	.	5
.	.	.	.	.	5
.	.	.	.	.	5
<b>I-1</b>	0	2	3	0	5
<b>I</b>	c	c	c	c	5
<b>Total</b>	$n_c$	$n_2$	$n_3$	$n_C$	

In addition, the proportion of observed agreement ( $P_o$ ) and the proportion of expected agreement ( $P_e$ ) from Equation 11, which are determined differently from Cohen's Kappa, are defined for Fleiss' Kappa as follows:

$$P_o = \frac{1}{NR(R-1)} \sum_{i=1}^I \sum_{c=1}^C n_{ic}(n_{ic} - 1) \quad (14)$$

and

$$P_e = \frac{1}{(NR)^2} \sum_{c=1}^C n_c^2, \quad (15)$$

where,  $n_{ic}$  represents the number of coders who assigned item  $i$  to category  $c$  and  $n_c$  the total number of items assigned by all coders to category  $c$  both of which come from Table 3.6.

Fleiss' Kappa can range from 0 to 1, with  $K_F = 0$  representing agreement not better than chance, and  $K_F = 1$  representing perfect agreement beyond chance. The guidelines used for interpreting values of Cohen's Kappa (Table 3.3) are also recommended for interpreting values of Fleiss' Kappa ( $K_F$ ). Fleiss' Kappa ( $K_F$ ) for this research was computed using the MAGREE macro in SAS v. 9.2.

Since to date there is no universally accepted method for measuring the reliability of safety taxonomies, the four methods -- percent agreement, Krippendorff's Alpha ( $\alpha$ ), Cohen's Kappa (K), and Fleiss' Kappa ( $K_F$ ) -- used in this study provide a thorough analysis of the reliability of HFACS. Moreover, the approach for measuring the reliability of HFACS with the corresponding training described here provides an efficient way of measuring its reliability without involving the variable of experience in extracting causal factors from incident/accident reports.

## CHAPTER 4: RESULTS

Data from 125 participants who coded 95 causal factors into HFACS causal categories were collected in the first reliability session, while 59 participated in both reliability sessions. Although the Google survey was tested twice to ensure the exclusion of compound causal factors, 3 such causal factors among the 95 were identified:

- 24 – The two monorail trains were identical, which caused confusion to the operator.
- 92 – The electrical operator got distracted by an external noise and forgot to take readings on the main transformer.
- 93 – The warehouse forklift driver was suffering from a severe head cold, took OTC drugs, became groggy and dropped a load of boxes.

The analyses reported here were conducted both with and without these 3 compound causal factors.

The four agreement measures -- percent agreement (PA), Krippendorff's Alpha ( $\alpha$ ), Cohen's Kappa (K), and Fleiss' Kappa ( $K_F$ ) -- were computed to analyze the data for this study. First, percent agreement, Cohen's Kappa and Krippendorff's Alpha were used to evaluate the individual intra-rater reliability of HFACS, with percent agreement, and Krippendorff's Alpha being used to evaluate the overall intra-rater reliability of each HFACS tier and causal category level. Second, the overall inter-rater reliability of HFACS was assessed using percent agreement, Fleiss' Kappa ( $K_F$ ), and Krippendorff's Alpha ( $\alpha$ ), and the overall inter-rater reliability of each HFACS tier and causal category

were assessed using Fleiss' Kappa ( $K_F$ ), and Krippendorff's Alpha ( $\alpha$ ). The guidelines for these four agreement measures are included in Table 4.1.

Table 4.1: Table Key

Font	Percent Agreement		Cohen's / Fleiss' Kappa		Krippendorff's Alpha	
	Value	Conclusion	Value	Conclusion	Value	Conclusion
Regular	[70% - 100%]	Reliable	>0.8	Almost Perfect Reliability	[0.8 - 1]	Reliable
Bold and Single Underlined	[60% - 70%]	Almost Reliable	(0.60 - 0.8]	Substantial Reliability	[0.667 - 0.8)	Tentative Reliability
Shaded	(0, 60%)	Unreliable	(0.40- 0.60]	Moderate Reliability	(0, 0.667)	Unreliable

#### 4.1 Intra-rater Reliability Analysis

Percent agreement, Cohen's Kappa and Krippendorff's Alpha were determined individually for each of the 59 participants who participated in both sessions using SAS v. 9.3 statistical software and Microsoft Excel macros. The intra-rater agreement results of these measures are tabulated in Table 4.2 and 4.3 for each participant at the HFACS tier level with and without the 3 compound causal factors, respectively, and in Table 4.4 and 4.5 for each participant at the HFACS causal category level with and without the 3 compound causal factors, respectively. Comparing the results of the agreement measures with and without the 3 compound causal factors ( $I = 92$ ) – Table 4.2 with Table 4.3 and Table 4.4 with Table 4.5 – indicated no substantial differences between the 2 data sets; thus, this discussion focuses only on the results including all 95 causal factors.

Table 4.2: Intra-rater; Tier Level; PA, K, and  $\alpha$ ; Individual Coders and Overall; Whole Data Set

Coder	PA	K	95% CI for K	$\alpha$	95% CI for $\alpha$
1	88.42	0.83	(0.74, 0.93)	0.83	(0.73, 0.92)
2	94.74	0.93	(0.87, 0.99)	0.93	(0.86, 0.99)
3	88.42	0.84	(0.75, 0.93)	0.84	(0.76, 0.92)
4	85.26	<b><u>0.79</u></b>	(0.7, 0.89)	0.80	(0.69, 0.88)
5	83.16	<b><u>0.77</u></b>	(0.67, 0.67)	<b><u>0.77</u></b>	(0.66, 0.87)
6	86.32	0.81	(0.72, 0.91)	0.82	(0.72, 0.91)
7	95.79	0.94	(0.89, 1.0)	0.94	(0.89, 0.99)
8	76.84	<b><u>0.68</u></b>	(0.57, 0.80)	<b><u>0.68</u></b>	(0.55, 0.80)
9	92.63	0.90	(0.83, 0.97)	0.90	(0.81, 0.97)
10	94.74	0.93	(0.87, 0.99)	0.93	(0.86, 0.99)
11	89.47	0.85	(0.77, 0.94)	0.85	(0.75, 0.93)
12	86.32	0.81	(0.72, 0.91)	0.81	(0.71, 0.90)
13	86.32	0.81	(0.72, 0.91)	0.81	(0.71, 0.90)
14	86.32	0.81	(0.72, 0.91)	0.81	(0.71, 0.91)
15	90.53	0.87	(0.79, 0.95)	0.87	(0.79, 0.94)
16	100	1	(1, 1)	1	(1, 1)
17	87.37	0.83	(0.74, 0.92)	0.83	(0.74, 0.91)
18	86.32	0.81	(0.71, 0.9)	0.81	(0.72, 0.90)
19	98.95	0.99	(0.96, 1)	0.99	(0.96, 1)
20	90.53	0.87	(0.79, 0.95)	0.87	(0.78, 0.94)
21	91.58	0.89	(0.81, 0.96)	0.89	(0.80, 0.96)
22	85.26	<b><u>0.8</u></b>	(0.70, 0.89)	0.80	(0.70, 0.90)
23	82.11	<b><u>0.76</u></b>	(0.66, 0.86)	<b><u>0.76</u></b>	(0.65, 0.86)
24	90.53	0.87	(0.79, 0.95)	0.87	(0.78, 0.94)
25	88.42	0.84	(0.75, 0.93)	0.84	(0.75, 0.93)
26	94.74	0.93	(0.86, 0.99)	0.93	(0.87, 0.99)
27	80.0	<b><u>0.72</u></b>	(0.61, 0.83)	<b><u>0.72</u></b>	(0.61, 0.84)
28	97.89	0.97	(0.93, 1)	0.97	(0.93, 1)
29	95.79	0.94	(0.89, 1)	0.94	(0.88, 0.99)
30	91.58	0.88	(0.81, 0.96)	0.88	(0.80, 0.96)
31	88.42	0.84	(0.75, 0.93)	0.84	(0.75, 0.93)
32	97.89	0.97	(0.93, 1)	0.97	(0.93, 1)
33	87.37	0.82	(0.73, 0.92)	0.83	(0.74, 0.91)
34	94.74	0.93	(0.86, 0.99)	0.93	(0.85, 0.99)
35	91.58	0.88	(0.80, 0.96)	0.88	(0.80, 0.96)

36	97.89	0.97	(0.93, 1)	0.97	(0.93, 1)
37	94.74	0.93	(0.87, 0.99)	0.93	(0.86, 0.99)
38	94.74	0.93	(0.87, 0.99)	0.93	(0.86, 0.99)
39	86.32	0.81	(0.72, 0.90)	0.81	(0.71, 0.90)
40	86.32	0.81	(0.72, 0.90)	0.81	(0.70, 0.90)
41	98.95	0.99	(0.96, 1)	0.99	(0.96, 1)
42	93.68	0.91	(0.85, 0.98)	0.91	(0.84, 0.97)
43	89.47	0.85	(0.77, 0.94)	0.85	(0.77, 0.93)
44	85.26	<b>0.8</b>	(0.71, 0.9)	0.8	(0.70, 0.90)
45	83.16	<b>0.77</b>	(0.67, 0.87)	<b>0.77</b>	(0.66, 0.87)
46	92.63	0.90	(0.83, 0.97)	0.90	(0.81, 0.96)
47	92.63	0.90	(0.83, 0.97)	0.90	(0.83, 0.97)
48	95.79	0.94	(0.89, 1)	0.94	(0.88, 0.99)
49	87.37	0.83	(0.73, 0.92)	0.83	(0.73, 0.91)
50	93.68	0.91	(0.85, 0.98)	0.91	(0.84, 0.97)
51	92.63	0.90	(0.82, 0.97)	0.90	(0.83, 0.96)
52	84.21	<b>0.78</b>	(0.68, 0.88)	<b>0.78</b>	(0.68, 0.88)
53	82.11	<b>0.76</b>	(0.66, 0.86)	<b>0.76</b>	(0.65, 0.86)
54	88.42	0.84	(0.75, 0.93)	0.84	(0.75, 0.93)
55	91.58	0.88	(0.81, 0.96)	0.89	(0.8, 0.96)
56	89.47	0.86	(0.77, 0.94)	0.86	(0.77, 0.94)
57	98.95	0.99	(0.96, 1)	0.99	(0.96, 1)
58	84.21	<b>0.78</b>	(0.68, 0.88)	<b>0.78</b>	(0.68, 0.87)
59	92.63	0.90	(0.83, 0.97)	0.90	(0.83, 0.97)
Average	90.22	0.87		0.87	
95% CI	(88.89, 91.56)	(0.85, 0.88)		(0.85, 0.88)	

Table 4.3: Intra-rater; Tier Level; PA, K, and  $\alpha$ ; Individual Coders and Overall; Excluding Compound Causal Factors

Coder	PA	K	95% CI for K	$\alpha$	95% CI for $\alpha$
1	85.26	0.83	(0.74, 0.92)	0.83	(0.74, 0.92)
2	91.58	0.93	(0.86, 0.99)	0.93	(0.86, 0.99)
3	85.26	0.84	(0.75, 0.93)	0.84	(0.76, 0.93)
4	83.16	<b>0.80</b>	(0.71, 0.90)	0.80	(0.70, 0.89)
5	81.05	<b>0.78</b>	(0.67, 0.87)	<b>0.78</b>	(0.66, 0.87)
6	84.21	0.82	(0.73, 0.92)	0.82	(0.74, 0.91)
7	92.63	0.94	(0.88, 1)	0.94	(0.88, 0.99)

8	75.79	<b><u>0.70</u></b>	(0.59, 0.81)	<b><u>0.70</u></b>	(0.59, 0.82)
9	89.47	0.90	(0.82, 0.97)	0.90	(0.82, 0.97)
10	91.58	0.93	(0.86, 0.99)	0.93	(0.85, 0.99)
11	87.37	0.86	(0.78, 0.95)	0.87	(0.76, 0.94)
12	84.21	0.82	(0.73, 0.92)	0.82	(0.72, 0.91)
13	85.26	0.84	(0.74, 0.93)	0.84	(0.75, 0.93)
14	84.21	0.82	(0.73, 0.92)	0.82	(0.74, 0.91)
15	87.37	0.87	(0.78, 0.95)	0.87	(0.78, 0.94)
16	96.84	1	(1, 1)	1	(1, 1)
17	84.21	0.83	(0.73, 0.92)	0.82	(0.72, 0.91)
18	84.21	0.82	(0.72, 0.91)	0.81	(0.71, 0.91)
19	95.79	0.99	(0.96, 1)	0.99	(0.96, 1)
20	87.37	0.87	(0.78, 0.95)	0.87	(0.78, 0.94)
21	89.47	0.90	(0.82, 0.97)	0.90	(0.81, 0.97)
22	83.16	0.81	(0.71, 0.90)	0.81	(0.70, 0.90)
23	78.95	<b><u>0.75</u></b>	(0.65, 0.86)	<b><u>0.76</u></b>	(0.65, 0.86)
24	87.37	0.86	(0.78, 0.95)	0.87	(0.78, 0.94)
25	87.37	0.87	(0.78, 0.95)	0.87	(0.78, 0.94)
26	92.63	0.94	(0.88, 1)	0.94	(0.88, 0.99)
27	76.84	<b><u>0.71</u></b>	(0.60, 0.83)	<b><u>0.72</u></b>	(0.60, 0.82)
28	94.74	0.97	(0.93, 1)	0.97	(0.93, 1)
29	92.63	0.94	(0.88, 1)	0.94	(0.88, 0.99)
30	89.47	0.90	(0.82, 0.97)	0.90	(0.81, 0.97)
31	85.26	0.83	(0.74, 0.93)	0.83	(0.73, 0.92)
32	94.74	0.97	(0.93, 1)	0.97	(0.93, 1)
33	84.21	0.82	(0.73, 0.92)	0.82	(0.72, 0.91)
34	92.63	0.94	(0.88, 1)	0.94	(0.86, 0.98)
35	89.47	0.90	(0.82, 0.97)	0.90	(0.82, 0.97)
36	94.74	0.97	(0.93, 1)	0.97	(0.93, 1)
37	91.58	0.93	(0.86, 0.99)	0.93	(0.85, 0.99)
38	92.63	0.94	(0.88, 1)	0.94	(0.88, 0.99)
39	83.16	0.81	(0.71, 0.90)	0.81	(0.70, 0.90)
40	83.16	0.81	(0.71, 0.90)	0.81	(0.71, 0.90)
41	95.79	0.99	(0.96, 1)	0.99	(0.96, 1)
42	91.58	0.93	(0.86, 0.99)	0.93	(0.85, 0.97)
43	86.32	0.85	(0.76, 0.94)	0.85	(0.74, 0.94)
44	83.16	0.81	(0.72, 0.91)	0.81	(0.71, 0.90)
45	80	<b><u>0.76</u></b>	(0.66, 0.87)	<b><u>0.76</u></b>	(0.66, 0.87)
46	90.53	0.91	(0.84, 0.98)	0.91	(0.83, 0.97)

47	90.53	0.91	(0.84, 0.98)	0.91	(0.84, 0.97)
48	92.63	0.94	(0.88, 1)	0.94	(0.88, 0.99)
49	84.21	0.82	(0.73, 0.91)	0.82	(0.71, 0.91)
50	90.53	0.91	(0.84, 0.98)	0.91	(0.84, 0.97)
51	89.47	0.90	(0.82, 0.97)	0.90	(0.82, 0.97)
52	82.11	<b><u>0.79</u></b>	(0.69, 0.89)	<b><u>0.79</u></b>	(0.69, 0.88)
53	80	<b><u>0.77</u></b>	(0.66, 0.87)	<b><u>0.77</u></b>	(0.66, 0.87)
54	86.32	0.85	(0.76, 0.94)	0.85	(0.76, 0.93)
55	89.47	0.90	(0.82, 0.97)	0.90	(0.82, 0.97)
56	87.37	0.87	(0.78, 0.95)	0.87	(0.78, 0.94)
57	95.79	0.99	(0.96, 1)	0.99	(0.96, 1)
58	82.11	<b><u>0.79</u></b>	(0.69, 0.89)	<b><u>0.79</u></b>	(0.68, 0.89)
59	89.47	0.90	(0.82, 0.97)	0.90	(0.82, 0.96)
Average	87.6	0.87		0.87	
95% CI	(86.32, 88.88)	(0.85, 0.89)		(0.85, 0.89)	

Table 4.4: Intra-rater; Category Level; PA, K, and  $\alpha$ ; Individual Coders and Overall; Whole Data Set

Coder	PA	K	95% CI for K	$\alpha$	95% CI for $\alpha$
1	<b><u>68.42</u></b>	<b><u>0.66</u></b>	(0.56, 0.76)	0.66	(0.56, 0.75)
2	91.58	0.91	(0.87, 0.99)	0.91	(0.84, 0.96)
3	<b><u>68.42</u></b>	<b><u>0.66</u></b>	(0.57, 0.76)	<b><u>0.67</u></b>	(0.56, 0.76)
4	<b><u>67.37</u></b>	<b><u>0.65</u></b>	(0.55, 0.75)	0.65	(0.55, 0.75)
5	<b><u>64.21</u></b>	0.57	(0.47, 0.68)	0.62	(0.52, 0.72)
6	71.58	<b><u>0.70</u></b>	(0.60, 0.79)	<b><u>0.67</u></b>	(0.60, 0.80)
7	89.47	0.89	(0.82, 0.95)	0.89	(0.82, 0.96)
8	56.84	0.54	(0.44, 0.65)	0.54	(0.43, 0.66)
9	84.21	0.83	(0.76, 0.91)	0.83	(0.76, 0.91)
10	83.16	0.82	(0.74, 0.90)	0.82	(0.75, 0.90)
11	84.21	0.83	(0.76, 0.91)	0.83	(0.75, 0.91)
12	75.79	<b><u>0.74</u></b>	(0.65, 0.83)	<b><u>0.74</u></b>	(0.64, 0.83)
13	74.74	<b><u>0.73</u></b>	(0.64, 0.83)	<b><u>0.73</u></b>	(0.63, 0.82)
14	<b><u>68.42</u></b>	<b><u>0.67</u></b>	(0.57, 0.76)	<b><u>0.67</u></b>	(0.56, 0.76)
15	70.53	<b><u>0.69</u></b>	(0.59, 0.78)	<b><u>0.69</u></b>	(0.58, 0.79)
16	98.95	0.99	(0.97, 1)	0.99	(0.97, 1)
17	71.58	<b><u>0.70</u></b>	(0.60, 0.79)	<b><u>0.70</u></b>	(0.61, 0.80)
18	70.53	<b><u>0.69</u></b>	(0.59, 0.78)	<b><u>0.69</u></b>	(0.59, 0.78)



19	89.47	0.89	(0.82, 0.95)	0.89	(0.82, 0.94)
20	75.79	<b><u>0.74</u></b>	(0.65, 0.83)	<b><u>0.74</u></b>	(0.66, 0.83)
21	83.16	0.82	(0.74, 0.90)	0.82	(0.73, 0.90)
22	72.63	<b><u>0.71</u></b>	(0.62, 0.80)	<b><u>0.71</u></b>	(0.60, 0.81)
23	<b><u>64.21</u></b>	<b><u>0.62</u></b>	(0.52, 0.72)	0.62	(0.52, 0.72)
24	80	<b><u>0.79</u></b>	(0.70, 0.87)	<b><u>0.79</u></b>	(0.70, 0.88)
25	78.95	<b><u>0.78</u></b>	(0.69, 0.86)	<b><u>0.78</u></b>	(0.68, 0.86)
26	86.32	0.86	(0.78, 0.93)	0.86	(0.78, 0.92)
27	70.53	<b><u>0.69</u></b>	(0.59, 0.78)	<b><u>0.69</u></b>	(0.6, 0.79)
28	94.74	0.94	(0.90, 0.99)	0.94	(0.89, 0.99)
29	84.21	0.833	(0.76, 0.91)	0.83	(0.75, 0.90)
30	76.84	<b><u>0.75</u></b>	(0.66, 0.84)	<b><u>0.76</u></b>	(0.67, 0.84)
31	78.95	<b><u>0.78</u></b>	(0.69, 0.86)	<b><u>0.79</u></b>	(0.69, 0.87)
32	90.53	0.90	(0.84, 0.96)	0.90	(0.83, 0.96)
33	76.84	<b><u>0.75</u></b>	(0.66, 0.84)	<b><u>0.76</u></b>	(0.67, 0.83)
34	83.16	0.82	(0.74, 0.90)	0.82	(0.74, 0.90)
35	76.84	<b><u>0.75</u></b>	(0.66, 0.84)	<b><u>0.76</u></b>	(0.67, 0.84)
36	93.68	0.93	(0.88, 0.98)	0.93	(0.88, 0.98)
37	84.21	0.83	(0.76, 0.91)	0.83	(0.75, 0.90)
38	85.26	0.84	(0.77, 0.92)	0.85	(0.77, 0.92)
39	75.79	<b><u>0.74</u></b>	(0.65, 0.83)	<b><u>0.75</u></b>	(0.66, 0.83)
40	75.79	<b><u>0.74</u></b>	(0.65, 0.83)	<b><u>0.74</u></b>	(0.64, 0.84)
41	92.63	0.92	(0.87, 0.98)	0.92	(0.87, 0.97)
42	<b><u>64.21</u></b>	<b><u>0.62</u></b>	(0.52, 0.72)	0.62	(0.52, 0.73)
43	77.89	<b><u>0.75</u></b>	(0.66, 0.84)	<b><u>0.76</u></b>	(0.67, 0.83)
44	<b><u>68.42</u></b>	<b><u>0.67</u></b>	(0.66, 0.77)	<b><u>0.67</u></b>	(0.57, 0.77)
45	<b><u>68.42</u></b>	<b><u>0.66</u></b>	(0.57, 0.76)	<b><u>0.67</u></b>	(0.57, 0.77)
46	83.16	0.82	(0.74, 0.90)	0.82	(0.75, 0.90)
47	87.37	0.87	(0.80, 0.94)	0.87	(0.79, 0.93)
48	86.32	0.86	(0.78, 0.93)	0.86	(0.78, 0.92)
49	70.53	<b><u>0.69</u></b>	(0.59, 0.78)	<b><u>0.69</u></b>	(0.59, 0.78)
50	80	<b><u>0.79</u></b>	(0.70, 0.87)	<b><u>0.79</u></b>	(0.70, 0.88)
51	72.63	<b><u>0.71</u></b>	(0.62, 0.81)	<b><u>0.71</u></b>	(0.61, 0.80)
52	<b><u>68.42</u></b>	<b><u>0.67</u></b>	(0.57, 0.76)	<b><u>0.67</u></b>	(0.57, 0.77)
53	70.53	<b><u>0.69</u></b>	(0.59, 0.78)	<b><u>0.69</u></b>	(0.59, 0.78)
54	82.11	0.81	(0.73, 0.89)	0.81	(0.72, 0.89)
55	77.89	<b><u>0.77</u></b>	(0.68, 0.85)	<b><u>0.77</u></b>	(0.68, 0.86)
56	73.68	<b><u>0.72</u></b>	(0.63, 0.81)	<b><u>0.72</u></b>	(0.63, 0.81)
57	97.89	0.98	(0.95, 1)	0.98	(0.94, 1)

58	71.58	<u><b>0.7</b></u>	(0.60, 0.79)	<u><b>0.70</b></u>	(0.59, 0.79)
59	77.89	<u><b>0.77</b></u>	(0.68, 0.85)	<u><b>0.77</b></u>	(0.68, 0.86)
Average	78.45	<u><b>0.77</b></u>		<u><b>0.77</b></u>	
95% CI	(76.09, 80.80)	(0.74, 0.79)		(0.75, 0.80)	

Table 4.5: Intra-rater; Tier Category; PA, K, and  $\alpha$ ; Individual Coders and Overall; Excluding Compound Causal Factors

Coder	PA	K	95% CI for K	$\alpha$	95% CI for $\alpha$
1	<u><b>68.48</b></u>	<u><b>0.66</b></u>	(0.56, 0.76)	<u><b>0.67</b></u>	(0.56, 0.75)
2	91.3	0.91	(0.85, 0.97)	0.91	(0.84, 0.97)
3	<u><b>67.39</b></u>	<u><b>0.65</b></u>	(0.55, 0.76)	0.66	(0.54, 0.75)
4	<u><b>67.39</b></u>	<u><b>0.65</b></u>	(0.55, 0.75)	0.65	(0.55, 0.76)
5	<u><b>65.22</b></u>	0.58	(0.48, 0.69)	0.63	(0.53, 0.72)
6	72.83	<u><b>0.71</b></u>	(0.62, 0.81)	<u><b>0.71</b></u>	(0.61, 0.80)
7	91.30	0.91	(0.85, 0.97)	0.91	(0.84, 0.97)
8	58.7	0.56	(0.46, 0.67)	0.56	(0.47, 0.67)
9	83.7	0.83	(0.75, 0.91)	0.83	(0.75, 0.91)
10	82.61	0.82	(0.73, 0.90)	0.82	(0.74, 0.90)
11	86.96	0.86	(0.79, 0.93)	0.86	(0.78, 0.93)
12	76.09	<u><b>0.75</b></u>	(0.65, 0.84)	<u><b>0.75</b></u>	(0.66, 0.84)
13	76.09	<u><b>0.75</b></u>	(0.65, 0.84)	<u><b>0.75</b></u>	(0.65, 0.83)
14	<u><b>69.57</b></u>	<u><b>0.68</b></u>	(0.58, 0.78)	<u><b>0.68</b></u>	(0.57, 0.78)
15	<u><b>69.57</b></u>	<u><b>0.68</b></u>	(0.60, 0.78)	<u><b>0.68</b></u>	(0.57, 0.77)
16	98.91	0.99	(0.97, 1)	0.99	(0.95, 1)
17	71.74	<u><b>0.70</b></u>	(0.60, 0.80)	<u><b>0.70</b></u>	(0.60, 0.81)
18	70.65	<u><b>0.69</b></u>	(0.60, 0.79)	<u><b>0.69</b></u>	(0.60, 0.78)
19	90.22	0.90	(0.83, 0.96)	0.90	(0.83, 0.95)
20	75	<u><b>0.73</b></u>	(0.64, 0.83)	<u><b>0.74</b></u>	(0.63, 0.82)
21	83.7	0.83	(0.75, 0.91)	0.83	(0.75, 0.90)
22	73.91	<u><b>0.72</b></u>	(0.63, 0.82)	<u><b>0.73</b></u>	(0.62, 0.83)
23	<u><b>64.13</b></u>	<u><b>0.62</b></u>	(0.51, 0.72)	0.62	(0.51, 0.72)
24	79.35	<u><b>0.78</b></u>	(0.69, 0.87)	<u><b>0.78</b></u>	(0.69, 0.86)
25	80.43	<u><b>0.79</b></u>	(0.71, 0.89)	<u><b>0.79</b></u>	(0.70, 0.87)
26	86.96	0.86	(0.79, 0.93)	0.86	(0.78, 0.93)
27	<u><b>69.57</b></u>	<u><b>0.68</b></u>	(0.58, 0.78)	<u><b>0.68</b></u>	(0.57, 0.78)
28	94.57	0.94	(0.89, 0.99)	0.94	(0.89, 0.99)
29	85.87	0.85	(0.78, 0.93)	0.85	(0.78, 0.92)

30	77.17	<b><u>0.76</u></b>	(0.67, 0.85)	<b><u>0.76</u></b>	(0.67, 0.85)
31	79.35	<b><u>0.78</u></b>	(0.69, 0.86)	<b><u>0.78</u></b>	(0.69, 0.86)
32	90.22	0.9	(0.83, 0.96)	0.9	(0.83, 0.95)
33	76.09	<b><u>0.75</u></b>	(0.66, 0.83)	<b><u>0.75</u></b>	(0.65, 0.84)
34	84.78	0.84	(0.76, 0.92)	0.84	(0.76, 0.91)
35	79.35	<b><u>0.78</u></b>	(0.69, 0.87)	<b><u>0.78</u></b>	(0.69, 0.86)
36	93.48	0.93	(0.88, 0.98)	0.93	(0.87, 0.98)
37	85.87	0.85	(0.78, 0.93)	0.85	(0.77, 0.92)
38	86.96	0.86	(0.79, 0.93)	0.86	(0.79, 0.93)
39	75	<b><u>0.74</u></b>	(0.64, 0.83)	<b><u>0.74</u></b>	(0.63, 0.83)
40	75	<b><u>0.74</u></b>	(0.64, 0.83)	<b><u>0.74</u></b>	(0.64, 0.83)
41	94.57	0.94	(0.89, 0.99)	0.94	(0.89, 0.99)
42	<b><u>66.3</u></b>	<b><u>0.64</u></b>	(0.54, 0.74)	0.64	(0.54, 0.75)
43	77.17	<b><u>0.76</u></b>	(0.67, 0.85)	<b><u>0.76</u></b>	(0.66, 0.85)
44	<b><u>69.57</u></b>	<b><u>0.68</u></b>	(0.58, 0.78)	<b><u>0.68</u></b>	(0.57, 0.77)
45	70.65	<b><u>0.69</u></b>	(0.59, 0.79)	<b><u>0.69</u></b>	(0.57, 0.78)
46	83.7	0.83	(0.75, 0.91)	0.83	(0.75, 0.91)
47	88.04	0.87	(0.8, 0.94)	0.87	(0.79, 0.94)
48	86.96	0.86	(0.79, 0.93)	0.86	(0.79, 0.93)
49	70.65	<b><u>0.69</u></b>	(0.59, 0.79)	<b><u>0.69</u></b>	(0.59, 0.78)
50	81.52	<b><u>0.8</u></b>	(0.72, 0.89)	0.81	(0.73, 0.87)
51	72.83	<b><u>0.71</u></b>	(0.62, 0.81)	<b><u>0.71</u></b>	(0.62, 0.81)
52	<b><u>69.57</u></b>	<b><u>0.68</u></b>	(0.58, 0.78)	<b><u>0.68</u></b>	(0.58, 0.77)
53	71.74	<b><u>0.70</u></b>	(0.60, 0.80)	<b><u>0.70</u></b>	(0.61, 0.79)
54	82.61	0.82	(0.73, 0.90)	0.82	(0.74, 0.90)
55	79.35	<b><u>0.78</u></b>	(0.70, 0.87)	<b><u>0.78</u></b>	(0.69, 0.87)
56	73.91	<b><u>0.72</u></b>	(0.63, 0.82)	<b><u>0.73</u></b>	(0.63, 0.81)
57	97.83	0.98	(0.95, 1)	0.98	(0.94, 1)
58	72.83	<b><u>0.71</u></b>	(0.62, 0.81)	<b><u>0.71</u></b>	(0.63, 0.80)
59	78.26	<b><u>0.77</u></b>	(0.68, 0.86)	<b><u>0.77</u></b>	(0.68, 0.85)
Average	78.7	<b><u>0.78</u></b>		<b><u>0.78</u></b>	
95 % CI	(76.35, 81.06)	(0.75, 0.80)		(0.75, 0.80)	

As Tables 4.2 and 4.4 show, the percent agreement ranged from 76.84% to 100% at the HFACS tier level, while at the HFACS causal category level, the range decreased, ranging from 56.84% to 98.95%. The overall average percent agreement at both the tier

and the causal category level were 90.22% and 78.45%, respectively. According to Wallace and Ross (2006), a 70% agreement between coders is considered a reasonable minimum for data to be deemed reliable, suggesting that at the tier level all coders were within the reliable level while at the causal category level, 11 coders were below the acceptable level with Coder 8 being well below this level.




The examination of the Cohen's Kappa confidence intervals reveals that the values at both levels, tier and category, were all positive with no zero values, meaning that agreement exceeded chance at the 95% confidence level. Cohen's Kappa ranged from 0.68 to 1.00 at the tier level, while at the category level it ranged from 0.54 to 0.99. Based on Landis and Koch (1977), the estimated Kappa values ranged from "substantial" to "perfect" agreement at the tier level, and "moderate" to "perfect" at the causal category level.

Similar to Cohen's Kappa values, Krippendorff's Alpha ranged from 0.68 to 1.00 at the tier level and 0.54 to 0.99 at the category level. According to Krippendorff (2006) while a Krippendorff's Alpha value above 0.79 is considered reliable, a value between 0.667 and 0.800 can be used to draw tentative conclusions. In addition, the results of the analysis of all three reliability coefficients -- percent agreement, Cohen's Kappa, and Krippendorff's Alpha -- agree that Coder 16 exhibits the highest agreement and Coder 8 the lowest.

The data included in Tables 4.2 to 4.5, are graphically presented in Figures 4.1 to 4.4. These figures compare the frequency and distribution of the agreement coefficient values at both the HFACS tier level and HFACS causal category level. The figures show

similar trends and distributions for Cohen's Kappa and Krippendorff's Alpha. Table 4.6 includes the key for these figures.

Table 4.6: Key Table for Figures 4.1 to 4.4

Pattern	Conclusion		
	Percent Agreement	Cohen's / Fliess' Kappa	Krippendorff's Alpha
	Reliable	Almost Perfect Reliability	Reliable
	Almost reliable	Substantial Reliability	Tentative Reliability
	Unreliable	Moderate Reliability	Unreliable

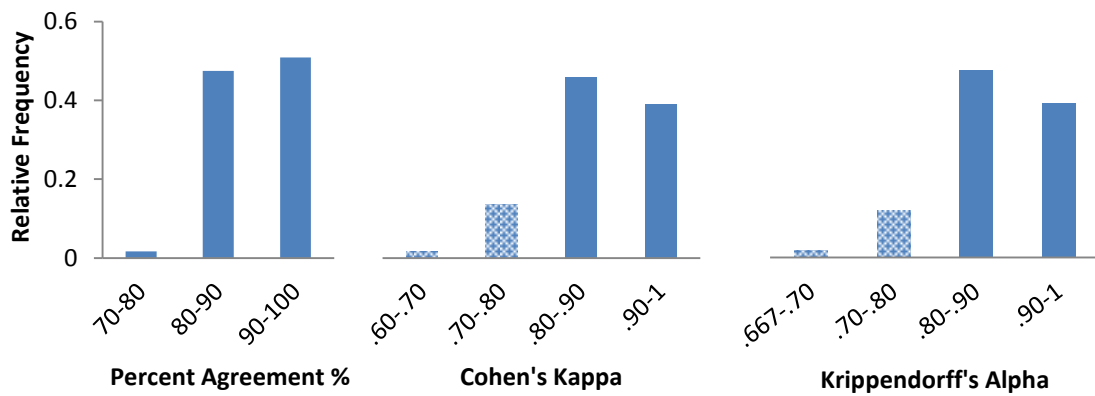


Figure 4.1: Frequency and Distribution of PA, K, and α; Intra-rater; Tier Level; Individual Coders; Whole Data Set

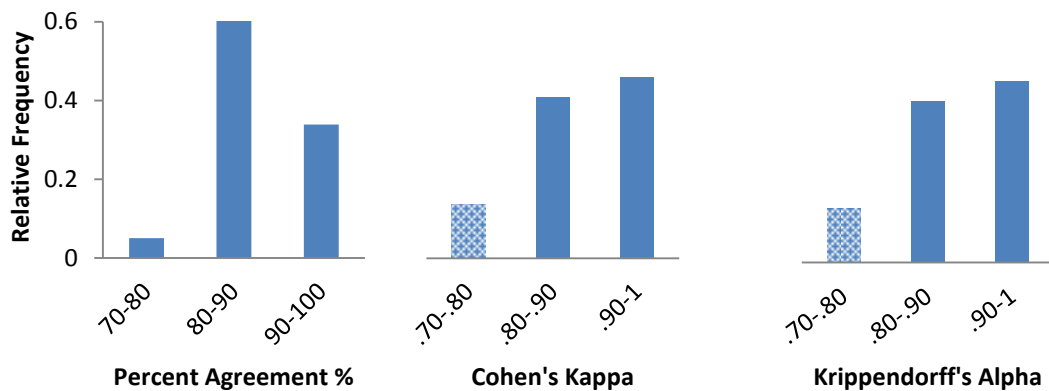
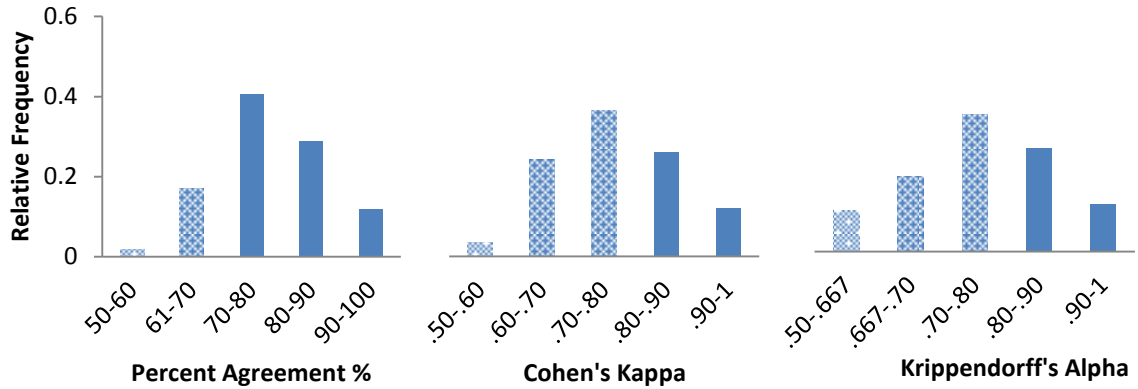
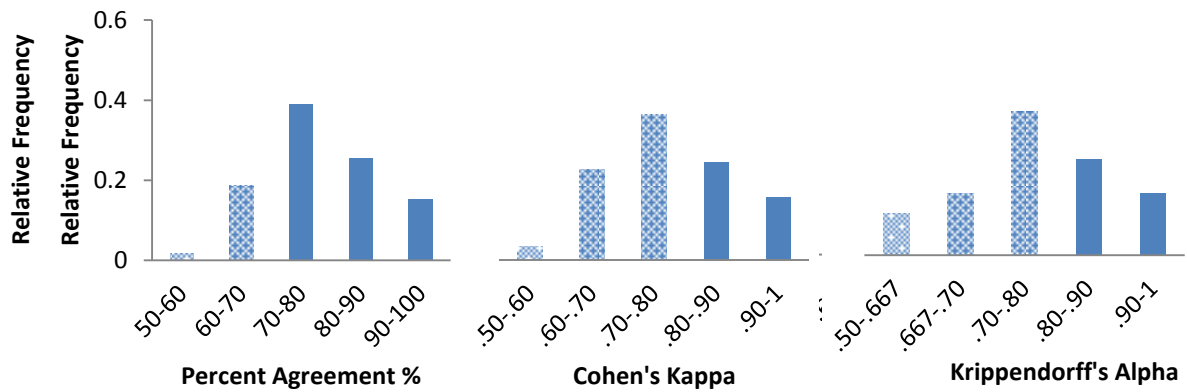


Figure 4.2: Frequency and Distribution of PA, K, and α; Intra-rater; Tier Level; Individual Coders; Excluding Compound Causal Factors



**Figure 4.3: Frequency and Distribution of PA, K, and  $\alpha$ ; Intra-rater; Category Level; Individual Coders; Whole Data Set**



**Figure 4.4: Frequency and Distribution of PA, K, and  $\alpha$ ; Intra-rater; Category Level; Individual Coders; Excluding Compound Causal Factors**

Second the overall intra-rater reliability of each HFACS tier and each HFACS causal category was assessed using percent agreement and Krippendorff's Alpha. Although it was initially proposed to also use Cohen's Kappa, this analysis produced misleading results, negative Kappa values with very high agreement, a situation known as

sample prevalence, a widely cited limitation (paradox) of Cohen’s Kappa; therefore, it was not used in this study. These results are tabulated in Tables 4.7 to 4.10, with Tables 4.7 and 4.8 showing that although every tier met the reliability criteria for both reliability coefficients, the unsafe supervision tier exhibited the least agreement value among the four, indicating that it is the most problematic.

Table 4.7: Intra-rater; Individual HFACS Tiers; Overall Average PA, and  $\alpha$ ; Whole Data Set

<b>HFACS Tier</b>	<b>Average PA</b>	<b>95 % CI Average PA</b>	<b>Average <math>\alpha</math></b>	<b>95 % CI Average <math>\alpha</math></b>
Unsafe Acts Tier	91.92	(90.38, 93.46)	0.88	(0.86, 0.90)
Preconditions of Unsafe Acts Tier	91.85	(90.01, 93.70)	0.87	(0.85, 0.90)
Unsafe Supervision Tier	87.74	(85.42, 90.06)	0.83	(0.79, 0.86)
Organizational Influences Tier	88.72	(85.19, 92.24)	0.87	(0.84, 0.90)

Table 4.8: Intra-rater; Individual HFACS Tiers; Overall Average PA, and  $\alpha$ ; Excluding Compound Causal Factors

<b>HFACS Tier</b>	<b>Average PA</b>	<b>95 % CI Average PA</b>	<b>Average <math>\alpha</math></b>	<b>95 % CI Average <math>\alpha</math></b>
Unsafe Acts Tier	92.71	(91.19, 94.22)	0.89	(0.87, 0.90)
Preconditions of Unsafe Acts Tier	91.92	(90.07, 93.77)	0.88	(0.86, 0.90)
Unsafe Supervision Tier	87.43	(85.15, 89.71)	0.83	(0.8, 0.85)
Organizational Influences Tier	88.75	(85.07, 92.43)	0.87	(0.85, 0.90)

Table 4.9: Intra-rater; Individual HFACS Categories; Overall Average PA, and  $\alpha$ ; Whole Data Set

	HFACS Category	Average PA	95 % CI Average PA	Average $\alpha$	95 % CI Average $\alpha$
Unsafe Acts	Skill Based Error	71.21	(65.15, 77.28)	0.66	(0.61, 0.72)
	Decision Error	<b><u>62.90</u></b>	(54.61, 71.18)	0.57	(0.49, 0.65)
	Perceptual Error	82.96	(78.07, 87.84)	0.82	(0.79, 0.86)
	Routine Violation	84.03	(80.12, 87.94)	0.82	(0.79, 0.86)
	Exceptional Violation	80.57	(75.62, 85.53)	<b><u>0.75</u></b>	(0.71, 0.80)
Preconditions of Unsafe Acts	Physical Environment	88.38	(85.08, 91.68)	0.87	(0.84, 0.90)
	Technological Environment	82.51	(77.34, 87.69)	<b><u>0.75</u></b>	(0.71, 0.80)
	Adverse Mental State	81.67	(76.86, 86.48)	<b><u>0.78</u></b>	(0.74, 0.82)
	Adverse Physiological State	72	(65.5, 78.49)	<b><u>0.68</u></b>	(0.62, 0.74)
	Physical / Mental Limitations	83.72	(78.39, 89.06)	0.82	(0.77, 0.86)
	Communication Coordination and Planning	84.67	(78.91, 90.43)	0.83	(0.79, 0.88)
	Fitness for Duty	76.76	(70.43, 83.09)	<b><u>0.73</u></b>	(0.68, 0.79)
Unsafe Supervision	Inadequate Supervision	73.27	(67, 79.55)	0.66	(0.60, 0.73)
	Planned Inappropriate Operations	72.73	(65.09, 80.36)	0.64	(0.58, 0.71)
	Failed to Correct a Known Problem	87.74	(83.85, 91.62)	0.85	(0.82, 0.89)
	Supervisory Violation	<b><u>64.98</u></b>	(57.42, 72.55)	0.62	(0.55, 0.70)
Organizational Influences	Resource / Acquisition Management	79.29	(74.68, 83.90)	<b><u>0.75</u></b>	(0.71, 0.79)
	Organizational Climate	91.25	(87.70, 94.79)	0.89	(0.86, 0.92)
	Organizational Process	80.74	(74.07, 87.41)	0.80	(0.74, 0.85)



Table 4.10: Intra-rater; Individual HFACS Categories; Overall Average PA, and  $\alpha$ ; Excluding Compound Causal Factors

	<b>HFACS Category</b>	<b>Average PA</b>	<b>95 % CI Average PA</b>	<b>Average <math>\alpha</math></b>	<b>95 % CI Average <math>\alpha</math></b>
Unsafe Acts	Skill Based Error	72.22	(66.20, 78.25)	<b><u>0.67</u></b>	(0.62, 0.72)
	Decision Error	<b><u>62.90</u></b>	(54.61, 71.18)	0.57	(0.49, 0.65)
	Perceptual Error	84.71	(79.32, 90.09)	0.83	(0.79, 0.87)
	Routine Violation	84.03	(80.12, 87.94)	0.82	(0.79, 0.86)
	Exceptional Violation	81.53	(76.71, 86.36)	<b><u>0.75</u></b>	(0.71, 0.80)
Preconditions of Unsafe Acts	Physical Environment	90.05	(87.01, 93.08)	0.88	(0.85, 0.91)
	Technological Environment	82.49	(77.06, 87.92)	<b><u>0.76</u></b>	(0.72, 0.80)
	Adverse Mental State	83.11	(78.25, 87.97)	<b><u>0.79</u></b>	(0.75, 0.83)
	Adverse Physiological State	73.08	(66.34, 79.83)	<b><u>0.69</u></b>	(0.63, 0.75)
	Physical / Mental Limitations	83.88	(78.57, 89.18)	0.82	(0.77, 0.86)
	Communication Coordination and Planning	84.67	(78.91, 90.43)	0.83	(0.79, 0.88)
	Fitness for Duty	78.63	(72.47, 84.8)	<b><u>0.75</u></b>	(0.69, 0.8)
Unsafe Supervision	Inadequate Supervision	72.46	(66.27, 78.65)	0.66	(0.60, 0.73)
	Planned Inappropriate Operations	72.75	(65.21, 80.30)	0.64	(0.58, 0.71)
	Failed to Correct a Known Problem	87.74	(83.85, 91.62)	0.85	(0.82, 0.89)
	Supervisory Violation	<b><u>64.98</u></b>	(57.42, 72.55)	0.62	(0.55, 0.70)
Organizational Influences	Resource / Acquisition Management	76.72	(71.85, 81.59)	<b><u>0.75</u></b>	(0.71, 0.78)
	Organizational Climate	90.68	(87.15, 94.22)	0.89	(0.86, 0.92)
	Organizational Process	82.25	(76.5, 88)	0.8	(0.74, 0.85)

The overall intra-rater agreement values for each HFACS causal category is seen in Tables 4.8 and 4.9 show a decline in percent agreement and Krippendorff's Alpha values in comparison with the HFACS tiers. In addition, the variability increased, as emphasized by an increase in the 95% confidence interval. Specifically, 8 categories -- Perceptual Error, Routine Violation, Physical Environment, Physical/Mental Limitations, Communication Coordination and Planning, Failed To Correct a Known Problem, Organizational Climate, and Organizational Process -- exhibited percent agreement and Krippendorff's Alpha values within the required reliability criteria. While two categories -- Decision Error, and Supervisory Violation -- exhibited percent agreement values below the required reliability criteria, the categories -- Exceptional Violation, Technological Environment, Adverse Mental State, Adverse Physiological State, Fitness for Duty, and Resource/Acquisition Management -- exhibited tentative values of Krippendorff's Alpha and the categories -- Decision Error, Skill Based Error, Inadequate Supervision, Planned Inappropriate Operations, and Supervisory Violation -- exhibited Krippendorff's Alpha values well below the criterion, signifying that these five are the most problematic categories.

#### 4.2 Inter-rater Reliability Analysis

Inter-rater reliability assessment receives the most research attention because intra-rater reliability alone is considered insufficient (Krippendorff, 2006). For this study, inter-rater reliability was determined separately for each session, R=125 participants from the first session and R= 59 from the second. In addition to analyzing the results with and

without the compound causal factors, the inter-rater reliability analysis also involved identifying 5 rogue coders, the analysis of which is shown in Appendix D, and conducting the analysis both with and without these coders. First, the overall inter-rater reliability for both the HFACS tier and the causal category levels was determined using percent agreement, Fleiss' Kappa, and Krippendorff's Alpha. Second, the overall inter-rater reliability of each HFACS tier and causal category was determined using Fleiss' Kappa, and Krippendorff's Alpha. In addition, diagnostic analyses were conducted to assist in identifying problematic areas in the HFACS structure. These three analyses were conducted separately for the data obtained from each session.

### **First Session**

Assessing the overall inter-rater reliability of HFACS for the tier and category levels involved determining the overall percent agreement, Fleiss' Kappa, and Krippendorff's Alpha for each level, which included the analysis of 3 data sets: I = 95 and R = 125, I = 95 and R = 120, and I=92 and R =125. The results are tabulated in Tables 4.11 to 4.13 for the tier level and Tables 4.14 to 4.16 for the category level. Comparing the results of the agreement measures for the 3 analyses -- Table 4.11 to Table 4.13 and Table 4.14 to Table 4.15 -- indicated no significant differences between the 3 data sets; thus, this discussion focuses only on the results of the analysis of the first data set, Tables 4.11 and 4.14 (I = 95 and R = 125).

Table 4.11: Inter-rater; Tier Level; Overall PA,  $K_F$ , and  $\alpha$ ; Whole Data Set; Including Rogue Coders, First Session

<b>Agreement Measure</b>	<b>Overall</b>	<b>95% CI for Overall</b>
Average Percent Agreement	84.77%	(84.67, 84.87)
Fleiss' Kappa	<u><b>0.79</b></u>	(0.79, 0.79)
Krippendorff's Alpha	<u><b>0.79</b></u>	(0.74, 0.83)

Table 4.12: Inter-rater; Tier Level; Overall PA,  $K_F$ , and  $\alpha$ ; Whole Data Set; Excluding Rogue Coders; First Session

<b>Agreement Measure</b>	<b>Overall</b>	<b>95% CI for Overall</b>
Average Percent Agreement	85.33%	(85.23, 85.46)
Fleiss' Kappa	<u><b>0.80</b></u>	(0.80, 0.80)
Krippendorff's Alpha	0.80	(0.75, 0.84)

Table 4.13: Inter-rater; Tier Level; Overall PA,  $K_F$ , and  $\alpha$ ; Excluding Compound Causal Factors; Including Rogue Coders; First Session

<b>Agreement Measure</b>	<b>Overall</b>	<b>95% CI for Overall</b>
Average Percent Agreement	85.37%	(85.27, 85.48)
Fleiss' Kappa	<u><b>0.80</b></u>	(0.80, 0.80)
Krippendorff's Alpha	0.80	(0.75, 0.85)

Table 4.14: Inter-rater; Category Level; Overall PA,  $K_F$ , and  $\alpha$ ; Whole Data Set; Including Rogue Coders, First Session

<b>Agreement Measure</b>	<b>Overall</b>	<b>95% CI for Overall</b>
Average Percent Agreement	<u><b>68.69%</b></u>	(68.52, 68.86)
Fleiss' Kappa	<u><b>0.67</b></u>	(0.67, 0.67)
Krippendorff's Alpha	<u><b>0.67</b></u>	(0.66, 0.68)

Table 4.15: Inter-rater; Category Level; Overall PA,  $K_F$ , and  $\alpha$ ; Whole Data Set; Excluding Rogue Coders; First Session

<b>Agreement Measure</b>	<b>Overall</b>	<b>95% CI for Overall</b>
Average Percent Agreement	<u><b>69.9%</b></u>	(69.74, 70.06)
Fleiss' Kappa	<u><b>0.68</b></u>	(0.68, 0.68)
Krippendorff's Alpha	<u><b>0.68</b></u>	(0.67, 0.69)

Table 4.16: Inter-rater; Category Level; Overall PA,  $K_F$ , and  $\alpha$ ; Excluding Compound Causal Factors; Including Rogue Coders; First Session

<b>Agreement Measure</b>	<b>Overall</b>	<b>95% CI for Overall</b>
Average Percent Agreement	<b><u>69.65%</u></b>	(69.47, 69.82)
Fleiss' Kappa	<b><u>0.68</u></b>	(0.68, 0.68)
Krippendorff's Alpha	<b><u>0.68</u></b>	(0.67, 0.69)

The overall average percent agreement at the tier level was 84.77%, suggesting an acceptable inter-rater reliability based on the 70% criterion. The results of both Fleiss' Kappa and Krippendorff's Alpha achieved a value of 0.79 although their computation differs for these two agreement coefficients.

With respect to the category level, the values of the 3 agreement coefficients were lower than for the tier level. While the overall average percent agreement for the causal category level was 68.69%, approaching the required level to be considered reliable, both Fleiss' Kappa and Krippendorff's Alpha values were  $K_F = \alpha = 0.67$ . Based on these results, the overall inter-rater reliability of HAFCS for the tier level is considered acceptable, while for the category level the overall inter-rater reliability is considered approximately reliable.

Second, assessing the inter-rater reliability for each HFACS tier and causal category involved determining Fleiss' Kappa, and Krippendorff's Alpha for each level, which included the analysis of 3 data sets: I = 95 and R = 125, I = 95 and R = 120, and I=92 and R =125. The results are presented in Tables 4.17 to 4.19 for each tier level and Tables 4.20 to 4.22 for each category level. Similar to the previous results, the results of the agreement measures for the 3 analyses -- Tables 4.17 to 4.19 and Tables 4.20 to 4.22 -- indicated no significant differences between the 3 data sets; thus, this discussion

focuses only on the results of the analysis of the first data set, Tables 4.17 and 4.22 ( $I = 95$  and  $R = 125$ ).

Table 4.17: Inter-rater; Individual HFACS Tiers; Overall  $K_F$  and  $\alpha$ ; Whole Data Set; Including Rogue Coders, First Session

<b>HFACS Tier</b>	<b><math>K_F</math></b>	<b>95 % CI <math>K_F</math></b>	<b><math>\alpha</math></b>
Unsafe Acts Tier	0.82	(0.81, 0.82)	0.82
Preconditions of Unsafe Acts Tier	<u><b>0.80</b></u>	(0.80, 0.80)	0.80
Unsafe Supervision Tier	<u><b>0.73</b></u>	(0.73, 0.73)	<u><b>0.73</b></u>
Organizational Influences Tier	<u><b>0.80</b></u>	(0.08, 0.81)	0.80

Table 4.18: Inter-rater; Individual HFACS Tiers; Overall  $K_F$  and  $\alpha$ ; Whole Data Set; Excluding Rogue Coders; First Session

<b>HFACS Tier</b>	<b><math>K_F</math></b>	<b>95 % CI <math>K_F</math></b>	<b><math>\alpha</math></b>
Unsafe Acts Tier	0.83	(0.82, 0.83)	0.83
Preconditions of Unsafe Acts Tier	0.81	(0.80, 0.81)	0.81
Unsafe Supervision Tier	<u><b>0.74</b></u>	(0.74, 0.74)	<u><b>0.74</b></u>
Organizational Influences Tier	0.81	(0.81, 0.82)	0.81

Table 4.19: Inter-rater; Individual HFACS Tiers; Overall  $K_F$  and  $\alpha$ ; Excluding Compound Causal Factors; Including Rogue Coders; First Session

<b>HFACS Tier</b>	<b><math>K_F</math></b>	<b>95 % CI <math>K_F</math></b>	<b><math>\alpha</math></b>
Unsafe Acts Tier	0.84	(0.83, 0.84)	0.84
Preconditions of Unsafe Acts Tier	0.82	(0.81, 0.82)	0.82
Unsafe Supervision Tier	<u><b>0.73</b></u>	(0.73, 0.73)	<u><b>0.73</b></u>
Organizational Influences Tier	<u><b>0.80</b></u>	(0.80, 0.81)	0.80

Table 4.20: Inter-rater; Individual HFACS Categories; Overall  $K_F$  and  $\alpha$ ; Whole Data Set; Including Rogue Coders, First Session

	<b>HFACS Category</b>	<b><math>K_F</math></b>	<b>95 % CI <math>K_F</math></b>	<b><math>\alpha</math></b>
Unsafe Acts	Skill Based Error (SBE)	0.56	(0.55, 0.56)	0.56
	Decision Error (DE)	0.46	(0.46, 0.47)	0.46
	Perceptual Error (PE)	<b><u>0.72</u></b>	(0.72, 0.72)	<b><u>0.72</u></b>
	Routine Violation (RV)	<b><u>0.76</u></b>	(0.76, 0.76)	<b><u>0.76</u></b>
	Exceptional Violation (EV)	<b><u>0.63</u></b>	(0.63, 0.63)	0.63
Preconditions of Unsafe Acts	Physical Environment (PhE)	0.82	(0.82, 0.82)	0.82
	Technological Environment (TE)	<b><u>0.65</u></b>	(0.65, 0.65)	0.65
	Adverse Mental State (AMS)	<b><u>0.68</u></b>	(0.67, 0.68)	<b><u>0.68</u></b>
	Adverse Physiological State (APS)	<b><u>0.63</u></b>	(0.63, 0.63)	0.63
	Physical / Mental Limitations (PML)	<b><u>0.73</u></b>	(0.73, 0.73)	<b><u>0.73</u></b>
	Communication Coordination & Planning (CC)	<b><u>0.78</u></b>	(0.78, 0.78)	<b><u>0.78</u></b>
	Fitness for Duty (FfD)	<b><u>0.73</u></b>	(0.72, 0.73)	<b><u>0.73</u></b>
Unsafe Supervision	Inadequate Supervision (IS)	0.51	(0.51, 0.51)	0.51
	Planned Inappropriate Operations (PIO)	0.49	(0.49, 0.49)	0.49
	Failed To Correct a Known Problem (FTCNP)	0.82	(0.81, 0.82)	0.82
	Supervisory Violation (SV)	0.53	(0.53, 0.54)	0.53
Organizational Influences	Resource / Acquisition Management (RAM)	<b><u>0.62</u></b>	(0.62, 0.62)	0.62
	Organizational Climate (OC)	<b><u>0.80</u></b>	(0.79, 0.80)	0.80
	Organizational Process (OP)	<b><u>0.69</u></b>	(0.69, 0.69)	<b><u>0.69</u></b>

Table 4.21: Inter-rater; Individual HFACS Categories; Overall  $K_F$  and  $\alpha$ ; Whole Data Set; Excluding Rogue Coders; First Session

	<b>HFACS Category</b>	<b><math>K_F</math></b>	<b>95 % CI <math>K_F</math></b>	<b><math>\alpha</math></b>
Unsafe Acts	Skill Based Error (SBE)	0.57	(0.57, 0.58)	0.57
	Decision Error (DE)	0.48	(0.48, 0.49)	0.48
	Perceptual Error (PE)	<b><u>0.73</u></b>	(0.73, 0.73)	<b><u>0.73</u></b>
	Routine Violation (RV)	<b><u>0.77</u></b>	(0.77, 0.77)	<b><u>0.77</u></b>
	Exceptional Violation (EV)	<b><u>0.65</u></b>	(0.65, 0.65)	0.65
Preconditions of Unsafe Acts	Physical Environment (PhE)	0.83	(0.83, 0.83)	0.83
	Technological Environment (TE)	<b><u>0.66</u></b>	(0.66, 0.66)	<b><u>0.66</u></b>
	Adverse Mental State (AMS)	<b><u>0.68</u></b>	(0.68, 0.68)	<b><u>0.68</u></b>
	Adverse Physiological State (APS)	<b><u>0.64</u></b>	(0.64, 0.65)	0.64
	Physical / Mental Limitations (PML)	<b><u>0.74</u></b>	(0.73, 0.74)	<b><u>0.74</u></b>
	Communication Coordination & Planning (CC)	<b><u>0.78</u></b>	(0.78, 0.79)	<b><u>0.78</u></b>
	Fitness for Duty (FfD)	<b><u>0.74</u></b>	(0.74, 0.75)	<b><u>0.74</u></b>
Unsafe Supervision	Inadequate Supervision (IS)	0.52	(0.52, 0.53)	0.52
	Planned Inappropriate Operations (PIO)	0.50	(0.5, 0.50)	0.50
	Failed To Correct a Known Problem (FTCNP)	0.83	(0.83, 0.83)	0.83
	Supervisory Violation (SV)	0.55	(0.55, 0.56)	0.55
Organizational Influences	Resource / Acquisition Management (RAM)	<b><u>0.63</u></b>	(0.63, 0.63)	0.63
	Organizational Climate (OC)	0.81	(0.81, 0.81)	0.81
	Organizational Process (OP)	<b><u>0.71</u></b>	(0.71, 0.71)	<b><u>0.71</u></b>



Table 4.22: Inter-rater; Individual HFACS Categories; Overall  $K_F$  and  $\alpha$ ; Excluding Compound Causal Factors; Including Rogue Coders; First Session

	<b>HFACS Category</b>	<b><math>K_F</math></b>	<b>95 % CI <math>K_F</math></b>	<b><math>\alpha</math></b>
Unsafe Acts	Skill Based Error (SBE)	0.57	(0.57, 0.58)	0.57
	Decision Error (DE)	0.46	(0.46, 0.47)	0.46
	Perceptual Error (PE)	<b><u>0.74</u></b>	(0.74, 0.74)	<b><u>0.74</u></b>
	Routine Violation (RV)	<b><u>0.76</u></b>	(0.76, 0.76)	<b><u>0.76</u></b>
	Exceptional Violation (EV)	<b><u>0.64</u></b>	(0.63, 0.64)	0.64
Preconditions of Unsafe Acts	Physical Environment (PhE)	0.86	(0.85, 0.86)	0.86
	Technological Environment (TE)	<b><u>0.69</u></b>	(0.68, 0.69)	<b><u>0.69</u></b>
	Adverse Mental State (AMS)	<b><u>0.72</u></b>	(0.71, 0.72)	<b><u>0.72</u></b>
	Adverse Physiological State (APS)	<b><u>0.64</u></b>	(0.64, 0.65)	0.64
	Physical / Mental Limitations (PML)	<b><u>0.73</u></b>	(0.73, 0.74)	<b><u>0.73</u></b>
	Communication Coordination & Planning (CC)	<b><u>0.78</u></b>	(0.78, 0.78)	<b><u>0.78</u></b>
	Fitness for Duty (FfD)	<b><u>0.76</u></b>	(0.76, 0.76)	<b><u>0.76</u></b>
Unsafe Supervision	Inadequate Supervision (IS)	0.51	(0.51, 0.51)	0.51
	Planned Inappropriate Operations (PIO)	0.49	(0.49, 0.49)	0.49
	Failed To Correct a Known Problem (FTCNP)	0.82	(0.81, 0.82)	0.82
	Supervisory Violation (SV)	0.53	(0.53, 0.54)	0.53
Organizational Influences	Resource /Acquisition Management (RAM)	<b><u>0.62</u></b>	(0.62, 0.62)	0.62
	Organizational Climate (OC)	<b><u>0.80</u></b>	(0.79, 0.8)	0.80
	Organizational Process (OP)	<b><u>0.69</u></b>	(0.69, 0.69)	<b><u>0.69</u></b>

Table 4.17 indicates that the estimated Fleiss' Kappa values for each tier ranged from 0.73 to 0.82, suggesting "substantial" to "near perfect" reliability for the individual tier levels according to Landis and Koch (1977); specifically, the Unsafe Supervision tier exhibited the lowest estimated Fleiss' Kappa value. Similar to the overall inter-rater analysis, the estimated Fleiss' Kappa for each causal category decreased. Tables 4.20 to 4.22 show a decline in Fleiss' Kappa and Krippendorff's Alpha values in comparison with the HFACS tiers. Fleiss' Kappa ranged from 0.46 to 0.82 for the individual categories. According to Landis and Koch (1977), these results suggest "moderate" to "near perfect" reliability for the individual causal categories. While two causal categories -- Physical Environment and Failed To Correct a Known Problem -- exhibited "near perfect" reliability levels, 5 causal categories -- Skill Based Error, Decision Error, Inadequate Supervision, Planned Inappropriate Operations, and Supervisory Violation -- exhibited "moderate" reliability levels, suggesting that these are the problematic categories; the remaining causal categories exhibited "substantial" reliability levels.

While Krippendorff's Alpha values were numerically identical to the Fleiss' Kappa values for both the tier and causal category levels, the reliability criteria differ for these two agreement coefficients. The individual Krippendorff's Alpha value for each HFACS tier, as shown in Table 4.17, ranged from 0.73 to 0.82. According to Krippendorff (2006), the overall Alpha values for all individual tiers are considered reliable except for the Unsafe Acts tier, which is considered "tentatively" reliable. Similar to Fleiss' Kappa values, Krippendorff's Alpha values for the causal category level were lower than the tier level, these values ranging from 0.46 to 0.82, suggesting a

heterogeneous outcome. Only 3 categories -- Physical Environment, Failed To Correct a Known Problem and Organizational Climate -- are considered reliable, while 7 categories -- Perceptual Error, Routine Violation, Adverse Mental State, Physical/Mental Limitations, Communication Coordination and Planning, Fitness for Duty, and Organizational Process -- exhibited Krippendorff's Alpha values between 0.67 and 0.79, also considered "tentatively" reliable. The remaining 9 categories -- Skill Based Error, Decision Error, Exceptional Violation, Technological Environment, Adverse Physiological State, Inadequate Supervision, Planned Inappropriate Operations, Supervisory Violation, and Resource/Acquisition Management -- are considered unreliable.

The reliability determined by both of Fleiss' Kappa and Krippendorff's Alpha, which take into account chance in their calculations, was in agreement for 3 HFACS tiers (Table 4.17) -- Unsafe Acts, Unsafe Supervision, and Organizational Influences -- and 14 HFACS causal categories (Table 4.20) -- Skill Based Error, Decision Error, Perceptual Error, Routine Violation, Physical Environment, Adverse Mental State, Physical/Mental Limitations, Communication Coordination and Planning, Fitness for Duty, Inadequate Supervision, Planned Inappropriate Operations, Failed To Correct a Known Problem, Supervisory Violation, and Organizational Process. However, these two reliability criteria did not agree on the remaining 5 causal categories.

In general, the overall and the individual inter-rater reliability for the tier level exhibited acceptable levels; however, the overall and the individual inter-rater reliability

at the causal category level was less consistently acceptable, suggesting the need for further analyses for the individual causal categories.

Diagnostic analysis using item analysis was conducted on all of the causal factors (I = 95) included in the survey including all coders (R=125) to determine the most and less frequently chosen causal categories for each causal factor. Such knowledge has the potential to indicate common misconceptions and misunderstandings of particular categories among coders, providing insight for appropriate remediation. Item analysis shows, for each causal factor, represented as a row, the distribution of coders responses with respect to HFACS categories. The results of such analysis are presented by percentage in Table 4.23. For each causal category, beginning with Skill Based Error, the causal factor item in the survey with the highest percentage referring to a particular category is arranged in descending order; these percentages are shaded in Table 4.23.

Table 4.23: Percentage of Coders Responses to Each Statement for First Session

Statement Number	SBE	DE	PE	RV	EV	PhE	TE	AMS	APS	PML	CC	FfD	IS	PIO	FTCNP	SV	RAM	OC	OP
49	86.4	6.4	2.4	1.6	1.6	0	0	1.6	0	0	0	0	0	0	0	0	0	0	0
65	82.4	6.4	0	0	0	0	0	0	0	4	0.8	0	5.6	0	0	0	0.8	0	0
6	80.8	17.6	0	0	0.8	0	0	0	0	0.8	0	0	0	0	0	0	0	0	0
51	77.6	10.4	4	0	1.6	0	0	6.4	0	0	0	0	0	0	0	0	0	0	0
* 92	44	0	0.8	0	3.2	30.4	0	21.6	0	0	0	0	0	0	0	0	0	0	0
79	17.6	76.8	0.8	0	0.8	0	0	0	0	0	0	0	0	0	4	0	0	0	0
12	4	72	4.8	2.4	7.2	0	0	8	0	0	0	0	0	0.8	0	0.8	0	0	0
72	31.2	66.4	1.6	0	0	0	0	0.8	0	0	0	0	0	0	0	0	0	0	0
83	24	60.8	0	0	2.4	0	0	0	0	8	0	0.8	2.4	1.6	0	0	0	0	0
64	2.4	40	0	0.8	14.4	6.4	0	0.8	0	0	0.8	0	0	8	4.8	21.6	0	0	0
81	1.6	0	97.6	0	0	0	0	0	0	0.8	0	0	0	0	0	0	0	0	0
62	4	1.6	93.6	0	0	0	0	0	0	0.8	0	0	0	0	0	0	0	0	0
84	7.2	20	72.8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	24	3.2	72	0	0	0	0.8	0	0	0	0	0	0	0	0	0	0	0	0
91	32.8	2.4	64.8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
* 24	2.4	0.8	58.4	0	0	1.6	26.4	4.8	0	4	0	0	0	0	0	0	0.8	0	0.8
67	0	1.6	0	95.2	0.8	0	0	0	0	0	0	0	0.8	0.8	0	0.8	0	0	0
78	0	0.8	0	94.4	0	0	0	0	0	0	4	0	0	0.8	0	0	0	0	0
18	0	0	0	93.6	0.8	0	0	0	0	0	0	0	1.6	0	0.8	1.6	0	0.8	0.8
29	1.6	2.4	0	92	2.4	0	0	0	0	0	0	0	0	0	0	0	0	0	1.6
61	0	0	0	83.2	0.8	0	0	0	0.8	0	0	0	0.8	0.8	7.2	1.6	0	4.8	0
86	0	0.8	0	7.2	91.2	0	0	0	0	0	0	0	0.8	0	0	0	0	0	0
68	3.2	6.4	0	1.6	88	0	0	0.8	0	0	0	0	0	0	0	0	0	0	0
54	0	3.2	0	17.6	79.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Category Number	SBE	DE	PE	RV	EV	PhE	TE	AMS	APS	PML	CC	FfD	IS	PIO	FTCNP	SV	RAM	OC	OP
52	1.6	4.8	0	15.2	77.6	0	0	0	0	0	0	0	0	0	0	0.8	0	0	0
46	0	0.8	0	8	49.6	0	0	14.4	16.8	0	0	7.2	3.2	0	0	0	0	0	0
53	0	0	0	0	0	98.4	0.8	0	0	0	0	0	0	0	0	0	0	0	0.8
34	0	0	0	0	0	94.4	3.2	0	0.8	0	0	0	0	0	0	0	0	0	1.6
13	0	0	0	0	0.8	93.6	4.8	0	0	0	0	0	0	0	0	0	0.8	0	0
77	0	0	0	0.8	0	91.2	5.6	0	0	0.8	0	0	0	0	0	0	0.8	0	0.8
39	0	0	0	0.8	0	90.4	1.6	0	0	2.4	0	0	0	0	2.4	0.8	0	0	1.6
44	0	0	1.6	0	0	0	98.4	0	0	0	0	0	0	0	0	0	0	0	0
95	0.8	0	7.2	0	0	4	80.8	0	0	0.8	1.6	0	0	0	0.8	0	0	0	4
14	0	0	0	0.8	0	0.8	74.4	0	0	0	0	0	0	0	4	0	18.4	0	1.6
2	0	0.8	0	5.6	0	28.8	35.2	0	0	0	0.8	0	0.8	8	6.4	0.8	5.6	0.8	6.4
47	0	0	0	0	0	0	0	97.6	2.4	0	0	0	0	0	0	0	0	0	0
88	0	0	0	0	0	0	0	94.4	2.4	0	0	0	0	0	0	0	1.6	1.6	0
69	0.8	0	0	0	0.8	0	0	91.2	4	0.8	0.8	1.6	0	0	0	0	0	0	0
7	1.6	0	2.4	0	0	0.8	0.8	76.8	1.6	9.6	0	0	0.8	4	0	0	1.6	0	0
17	0	0	0	0	0	0	0	75.2	11.2	3.2	0	2.4	0	6.4	0	0	1.6	0	0
89	0	0	0	0	0	0	0	0	95.2	4.8	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	4.8	84	3.2	0	8	0	0	0	0	0	0	0
27	0.8	8	2.4	0	0.8	4	0	0	72	1.6	0	9.6	0	0	0	0	0.8	0	0
66	0	0	0	0	0	0	0	10.4	67.2	13.6	0	0.8	1.6	6.4	0	0	0	0	0
* 93	0.8	0	0	0	0	0	0	9.6	57.6	0	0	32	0	0	0	0	0	0	0
90	0	0	0	0	0	0.8	0.8	0	0.8	96	0	0	0	0	0	0	1.6	0	0
82	0	0	0	0	0	1.6	2.4	0	0	95.2	0	0	0	0.8	0	0	0	0	0
5	0	0	2.4	0	0	0	0	0	1.6	94.4	0	0.8	0.8	0	0	0	0	0	0
71	0	0	3.2	0	0	0	0	0	4.8	92	0	0	0	0	0	0	0	0	0
75	8.8	2.4	0	0	0.8	0	0	0	0	50.4	0	2.4	15.2	2.4	0	0	16	0	1.6

Category Number	SBE	DE	PE	RV	EV	PhE	TE	AMS	APS	PML	CC	Ffd	IS	PIO	FTCNP	SV	RAM	OC	OP
43	0.8	0	0	0	0	0	0	0	0	0	99.2	0	0	0	0	0	0	0	0
11	0.8	0	0	0	0	0	0.8	0.8	0	0.8	96.8	0	0	0	0	0	0	0	0
73	0.8	0	0	0	0	0	0	0.8	0	0	96.8	0	0.8	0.8	0	0	0	0	0
32	4	0.8	0	0	0	0	0	0	0	0	90.4	0	3.2	0	0	0	0	0	1.6
22	0.8	8.8	1.6	0	0	0	0	0	0	0.8	60	0	28	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0.8	3.2	3.2	0	92.8	0	0	0	0	0	0	0
58	0	1.6	0	0	2.4	0	0	4	1.6	0	0	88	0.8	1.6	0	0	0	0	0
50	0	0	0	0	0	0	0	8	4.8	0.8	0	86.4	0	0	0	0	0	0	0
80	0	0	0	0	2.4	0	0	8.8	7.2	0	0	81.6	0	0	0	0	0	0	0
70	0.8	4	0.8	0	0	0	0	3.2	28.8	0	0.8	61.6	0	0	0	0	0	0	0
16	0	0	0	0	0.8	0	0	0	0	0	0.8	0	87.2	5.6	0	5.6	0	0	0
63	0	0	0	0	0	0	0.8	0	0	4	0.8	0	81.6	0	0	1.6	8.8	0.8	1.6
9	0	0	0	0	0	0	0	0	0	0	0.8	0	76	12.8	0	8	0.8	0	1.6
41	0	0	0	0	0	0	0	7.2	0	0	4	0	58.4	0.8	1.6	1.6	0	25.6	0.8
8	0	0.8	0	0	0	0	0	0.8	0	0	35.2	0	54.4	1.6	2.4	4	0	0	0.8
19	0	0	0	0	0	0	0	0	0	0	3.2	0	4.8	87.2	3.2	0	1.6	0	0
31	0	1.6	0	0	0	0	0	0	0	4.8	0	0	12.8	64.8	0.8	12	3.2	0	0
38	0	0	0	0	0	0	0	0	0	0	0.8	0	17.6	60	5.6	12	0	0	4
60	0	1.6	0	0	0	0	0	0	0	0	0	0	7.2	60	0.8	4.8	24.8	0	0.8
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	96.8	3.2	0	0	0
55	0	0	0	0	0	0	0	0	0	0	0	0	1.6	0.8	95.2	2.4	0	0	0
94	0	0	0	0	0	2.4	0.8	0	0	0	0	0	0	0	95.2	1.6	0	0	0
45	0	0.8	0	0	0.8	0	0	0	0	0	0	0	0.8	0.8	94.4	2.4	0	0	0
87	0	0	0	0	0	0	0.8	0	0	0	0	0	0	0	93.6	2.4	3.2	0	0
48	0	0.8	0	0	0	0	0	0	0	4	14.4	0	33.6	9.6	36	0	0.8	0.8	0
20	0	0	0	0.8	8.8	0	0	0	0	0	0	0	0.8	4	1.6	84	0	0	0

Category Number	SBE	DE	PE	RV	EV	PhE	TE	AMS	APS	PML	CC	FfD	IS	PIO	FTCNP	SV	RAM	OC	OP
25	0	0.8	0	9.6	0	0	0	0	0	0	0	0	1.6	7.2	1.6	76.8	0.8	1.6	0
23	0	0.8	0	11.2	8.8	0	0	0	0	0	0	0	0	5.6	0.8	72.8	0	0	0
74	2.4	5.6	0	8.8	19.2	0	0.8	0.8	0	0	0	0	0.8	0	0	60	0	0.8	0.8
33	2.4	0	0	3.2	21.6	0	0	0	0	13.6	0	3.2	8	1.6	0	22.4	16.8	0.8	6.4
40	0	0	0	0	0	0	0.8	0	0	0	0	0	0	0.8	0	0	96	0.8	1.6
37	0	0	0	0	0.8	0	0	0	0	0	0	0	0	1.6	0	0.8	92	0.8	4
1	0	0.8	0	0	0	0	2.4	0	0	0	0	0	0	0	0.8	4	89.6	0.8	1.6
57	0	0	0	0.8	0	0	0	0	0	1.6	0	0	10.4	1.6	0	4	61.6	4.8	15.2
85	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	58.4	2.4	39.2
76	0	0	0	0	0	0	0	0	0	0	0	0	0	0.8	0	1.6	0	96	1.6
4	0	0	0.8	0	0	0	0	4.8	0	0	0	0	0	0	0	0	0	94.4	0
30	0	0	0.8	0	0	0	0	2.4	0	0	1.6	0	0	0	0	0	0.8	93.6	0.8
21	0	0	0	0	0	0	0	0	0	0	0	0	0.8	1.6	0	0	2.4	87.2	8
35	0	0	0	0	0	0	0	6.4	0	0	12	0	2.4	5.6	0	0	1.6	72	0
36	0	0	0	0	0	0	1.6	0	0	0	0	0	0.8	0	0	1.6	0.8	1.6	93.6
59	0	0	0	0	0	0	0	0	0	0.8	0.8	0	0.8	0	0	0	7.2	1.6	88.8
3	0	0	0	0	0	0	0	0	0	0	1.6	0	4.8	1.6	0.8	0	2.4	3.2	85.6
56	0	0	0	0	0	0	14.4	0	0	0	0.8	0	0	0	1.6	0	3.2	0.8	79.2
42	0	0	0	0	0	0	0	0	0	0	0	0	7.2	2.4	0	0	12	1.6	76.8

\* Indicates compound causal factor



## Second Session

The 3 analyses conducted on the data obtained from the first session were also conducted for the data obtained from the second, R=59 participants in which 4 coders were identified as rogue. Additionally, the analyses were also conducted with and without the 3 compound causal factors ((I = 95 and I = 92). First, the overall inter-rater reliability of HFACS for the tier and category levels was determined using the overall average percent agreement, Fleiss' Kappa, and Krippendorff's Alpha for each level, which included the analysis of 3 data sets: I = 95 and R = 59, I = 95 and R =55, and I=92 and R =59. The results are tabulated in Tables 4.24 to 4.26 for the tier level and Tables 4.27 to 4.29 for the category level. As for the first session, the comparison of the results of the agreement measures for the data sets indicated no significant differences between the 3 data sets; thus, this discussion focuses only on the results of the analysis of the first data set, Tables 4.24 and 4.27 (I = 95 and R = 59).

Table 4.24: Inter-rater; Tier Level; Overall PA,  $K_F$ , and  $\alpha$ ; Whole Data Set; Including Rogue Coders, Second Session

<b>Agreement Measure</b>	<b>Overall</b>	<b>95% CI for Overall</b>
Average Percent Agreement	85.25%	(84.99, 85.55)
Fleiss' Kappa	<b><u>0.80</u></b>	(0.79, 0.80)
Krippendorff's Alpha	0.80	(0.75, 0.84)

Table 4.25: Inter-rater; Tier Level; Overall PA,  $K_F$ , and  $\alpha$ ; Whole Data Set; Excluding Rogue Coders; Second Session

<b>Agreement Measure</b>	<b>Overall</b>	<b>95% CI for Overall</b>
Average Percent Agreement	86.20%	(85.95, 85.46)
Fleiss' Kappa	0.81	(0.81, 0.81)
Krippendorff's Alpha	0.81	(0.76, 0.86)

Table 4.26: Inter-rater; Tier Level; Overall PA,  $K_F$ , and  $\alpha$ ; Excluding Compound Causal Factors; Including Rogue Coders; Second Session

Agreement Measure	Overall	95% CI for Overall
Average Percent Agreement	85.86%	(85.59, 86.13)
Fleiss' Kappa	0.81	(0.81, 0.81)
Krippendorff's Alpha	0.81	(0.76, 0.85)

Table 4.27: Inter-rater; Category Level; Overall PA,  $K_F$ , and  $\alpha$ ; Whole Data Set; Including Rogue Coders, Second Session

Agreement Measure	Overall	95% CI for Overall
Average Percent Agreement	<b><u>68.10%</u></b>	(67.97, 68.52)
Fleiss' Kappa	<b><u>0.66</u></b>	(0.66, 0.66)
Krippendorff's Alpha	<b><u>0.66</u></b>	(0.65, 0.67)

Table 4.28: Inter-rater; Category Level; Overall PA,  $K_F$ , and  $\alpha$ ; Whole Data Set; Excluding Rogue Coders; Second Session

Agreement Measure	Overall	95% CI for Overall
Average Percent Agreement	<b><u>69.52%</u></b>	(69.10, 69.94)
Fleiss' Kappa	<b><u>0.68</u></b>	(0.68, 0.68)
Krippendorff's Alpha	<b><u>0.68</u></b>	(0.67, 0.69)

Table 4.29: Inter-rater; Category Level; Overall PA,  $K_F$ , and  $\alpha$ ; Excluding Compound Causal Factors; Including Rogue Coders; Second Session

Agreement Measure	Overall	95% CI for Overall
Average Percent Agreement	<b><u>69.05%</u></b>	(68.91, 69.48)
Fleiss' Kappa	<b><u>0.67</u></b>	(0.67, 0.67)
Krippendorff's Alpha	<b><u>0.67</u></b>	(0.66, 0.68)

The overall average percent agreement at the tier level was 85.25%, suggesting an acceptable inter-rater reliability based on the 70% criterion. In addition, the results of both Fleiss' Kappa and Krippendorff's Alpha values, which take agreement by chance into consideration, were 0.80. However, for the category level, the values of the 3 agreement coefficients were lower than for the tier level. While the overall average

percent agreement for the causal category level was 68.10%, approaching the required level to be considered reliable, both Fleiss' Kappa and Krippendorff's Alpha values were  $K_F = \alpha = 0.66$ . Based on these results, the overall inter-rater reliability of HFACS for the tier level is considered acceptable, while for the category level the overall inter-rater reliability is considered approximately reliable.

Second, Fleiss' Kappa, and Krippendorff's Alpha for each level were determined to assess the inter-rater reliability for each HFACS tier and causal category, which included analyzing 3 data sets: I = 95 and R = 59, I = 95 and R = 55, and I=92 and R =59. The results are presented in Tables 4.30 to 4.32 for each tier level and Tables 4.33 to 4.35 for each category level. Similar to the previous results, the results of the agreement measures for the 3 analyses – Tables 4.30 to 4.32 and Tables 4.33 to 4.35 – indicated no substantial differences between the 3 data sets; thus, this discussion focuses only on the results of the analysis that included whole dataset; including Rogue Coders, Tables 4.30 and 4.33 (I = 95 and R = 59).

Table 4.30: Inter-rater; Individual HFACS Tiers; Overall  $K_F$  and  $\alpha$ ; Whole Data Set; Including Rogue Coders, Second Session

<b>HFACS Tier</b>	<b><math>K_F</math></b>	<b>95 % CI <math>K_F</math></b>	<b><math>\alpha</math></b>
Unsafe Acts Tier	0.82	(0.82, 0.83)	0.82
Preconditions of Unsafe Acts Tier	0.81	(0.81, 0.82)	0.80
Unsafe Supervision Tier	<b><u>0.74</u></b>	(0.74, 0.75)	<b><u>0.73</u></b>
Organizational Influences Tier	<b><u>0.80</u></b>	(0.79, 0.80)	0.80

Table 4.31: Inter-rater; Individual HFACS Tiers; Overall  $K_F$  and  $\alpha$ ; Whole Data Set; Excluding Rogue Coders; Second Session

<b>HFACS Tier</b>	<b><math>K_F</math></b>	<b>95 % CI <math>K_F</math></b>	<b><math>\alpha</math></b>
Unsafe Acts Tier	0.84	(0.83, 0.84)	0.84
Preconditions of Unsafe Acts Tier	0.82	(0.82, 0.83)	0.82
Unsafe Supervision Tier	<b><u>0.76</u></b>	(0.76, 0.77)	<b><u>0.76</u></b>
Organizational Influences Tier	0.81	(0.80, 0.81)	0.81

Table 4.32: Inter-rater; Individual HFACS Tiers; Overall  $K_F$  and  $\alpha$ ; Excluding Compound Causal Factors; Including Rogue Coders; Second Session

HFACS Tier	$K_F$	95 % CI $K_F$	$\alpha$
Unsafe Acts Tier	0.84	(0.84, 0.85)	0.84
Preconditions of Unsafe Acts Tier	0.83	(0.82, 0.83)	0.83
Unsafe Supervision Tier	<b><u>0.74</u></b>	(0.74, 0.75)	<b><u>0.74</u></b>
Organizational Influences Tier	<b><u>0.80</u></b>	(0.79, 0.80)	0.80

Table 4.33: Inter-rater; Individual HFACS Categories; Overall  $K_F$  and  $\alpha$ ; Whole Data Set; Including Rogue Coders, Second Session

	HFACS Category	$K_F$	95 % CI $K_F$	$\alpha$
Unsafe Acts	Skill Based Error (SBE)	0.54	(0.53, 0.54)	0.54
	Decision Error (DE)	0.46	(0.45, 0.46)	0.45
	Perceptual Error (PE)	<b><u>0.72</u></b>	(0.71, 0.72)	<b><u>0.71</u></b>
	Routine Violation (RV)	<b><u>0.76</u></b>	(0.75, 0.76)	<b><u>0.76</u></b>
	Exceptional Violation (EV)	<b><u>0.66</u></b>	(0.66, 0.67)	<b><u>0.66</u></b>
Preconditions of Unsafe Acts	Physical Environment (PhE)	0.83	(0.82, 0.83)	0.83
	Technological Environment (TE)	<b><u>0.62</u></b>	(0.61, 0.62)	0.61
	Adverse Mental State (AMS)	<b><u>0.69</u></b>	(0.69, 0.70)	<b><u>0.69</u></b>
	Adverse Physiological State (APS)	0.58	(0.57, 0.58)	0.58
	Physical / Mental Limitations (PML)	<b><u>0.76</u></b>	(0.75, 0.76)	<b><u>0.75</u></b>
	Communication Coordination & Planning (CC)	<b><u>0.78</u></b>	(0.78, 0.79)	<b><u>0.78</u></b>
	Fitness for Duty (FfD)	<b><u>0.61</u></b>	(0.60, 0.61)	0.61
Unsafe Supervision	Inadequate Supervision (IS)	0.53	(0.52, 0.53)	0.52
	Planned Inappropriate Operations (PIO)	0.52	(0.52, 0.53)	0.52
	Failed To Correct a Known Problem (FTCNP)	0.81	(0.80, 0.81)	0.81
	Supervisory Violation (SV)	0.50	(0.49, 0.50)	0.50
Organizational Influences	Resource / Acquisition Management (RAM)	<b><u>0.66</u></b>	(0.66, 0.67)	<b><u>0.66</u></b>
	Organizational Climate (OC)	0.82	(0.81, 0.82)	0.82
	Organizational Process (OP)	<b><u>0.67</u></b>	(0.67, 0.68)	<b><u>0.67</u></b>

Table 4.34: Inter-rater; Individual HFACS Categories; Overall  $K_F$  and  $\alpha$ ; Whole Data Set; Excluding Rogue Coders; Second Session

	<b>HFACS Category</b>	<b><math>K_F</math></b>	<b>95 % CI <math>K_F</math></b>	<b><math>\alpha</math></b>
Unsafe Acts	Skill Based Error (SBE)	0.55	(0.54, 0.55)	0.55
	Decision Error (DE)	0.50	(0.50, 0.51)	0.50
	Perceptual Error (PE)	<b><u>0.74</u></b>	(0.74, 0.75)	<b><u>0.74</u></b>
	Routine Violation (RV)	<b><u>0.77</u></b>	(0.76, 0.77)	<b><u>0.77</u></b>
	Exceptional Violation (EV)	<b><u>0.67</u></b>	(0.67, 0.68)	<b><u>0.67</u></b>
Preconditions of Unsafe Acts	Physical Environment (PhE)	0.84	(0.83, 0.84)	0.84
	Technological Environment (TE)	<b><u>0.64</u></b>	(0.63, 0.64)	0.64
	Adverse Mental State (AMS)	<b><u>0.70</u></b>	(0.70, 0.71)	<b><u>0.70</u></b>
	Adverse Physiological State (APS)	0.59	(0.58, 0.59)	0.59
	Physical / Mental Limitations (PML)	<b><u>0.76</u></b>	(0.76, 0.77)	<b><u>0.76</u></b>
	Communication Coordination & Planning (CC)	<b><u>0.79</u></b>	(0.79, 0.80)	<b><u>0.79</u></b>
	Fitness for Duty (FfD)	<b><u>0.63</u></b>	(0.62, 0.63)	0.63
Unsafe Supervision	Inadequate Supervision (IS)	0.54	(0.53, 0.54)	0.54
	Planned Inappropriate Operations (PIO)	0.54	(0.53, 0.55)	0.54
	Failed To Correct a Known Problem (FTCNP)	0.82	(0.81, 0.82)	0.82
	Supervisory Violation (SV)	0.53	(0.52, 0.53)	0.53
Organizational Influences	Resource /Acquisition Management (RAM)	<b><u>0.66</u></b>	(0.66, 0.67)	<b><u>0.66</u></b>
	Organizational Climate (OC)	0.82	(0.82, 0.83)	0.82
	Organizational Process (OP)	<b><u>0.70</u></b>	(0.70, 0.71)	<b><u>0.70</u></b>

Table 4.35: Inter-rater; Individual HFACS Categories; Overall  $K_F$  and  $\alpha$ ; Excluding Compound Causal Factors; Including Rogue Coders; Second Session

	<b>HFACS Category</b>	<b><math>K_F</math></b>	<b>95 % CI <math>K_F</math></b>	<b><math>\alpha</math></b>
Unsafe Acts	Skill Based Error (SBE)	0.56	(0.55, 0.56)	0.56
	Decision Error (DE)	0.47	(0.46, 0.47)	0.47
	Perceptual Error (PE)	<b><u>0.75</u></b>	(0.74, 0.75)	<b><u>0.75</u></b>
	Routine Violation (RV)	<b><u>0.76</u></b>	(0.75, 0.76)	<b><u>0.76</u></b>
	Exceptional Violation (EV)	<b><u>0.66</u></b>	(0.66, 0.67)	<b><u>0.66</u></b>
Preconditions of Unsafe Acts	Physical Environment (PhE)	0.86	(0.85, 0.86)	0.86
	Technological Environment (TE)	<b><u>0.65</u></b>	(0.64, 0.65)	0.65
	Adverse Mental State (AMS)	<b><u>0.73</u></b>	(0.73, 0.74)	<b><u>0.73</u></b>
	Adverse Physiological State (APS)	0.58	(0.57, 0.58)	0.58
	Physical / Mental Limitations (PML)	<b><u>0.76</u></b>	(0.75, 0.76)	<b><u>0.76</u></b>
	Communication Coordination & Planning (CC)	<b><u>0.78</u></b>	(0.78, 0.79)	<b><u>0.78</u></b>
	Fitness for Duty (FfD)	<b><u>0.63</u></b>	(0.62, 0.63)	0.63
Unsafe Supervision	Inadequate Supervision (IS)	0.52	(0.52, 0.53)	0.52
	Planned Inappropriate Operations (PIO)	0.52	(0.52, 0.53)	0.52
	Failed To Correct a Known Problem (FTCNP)	0.81	(0.80, 0.81)	0.81
	Supervisory Violation (SV)	0.49	(0.49, 0.50)	0.49
Organizational Influences	Resource /Acquisition Management (RAM)	<b><u>0.66</u></b>	(0.66, 0.67)	<b><u>0.66</u></b>
	Organizational Climate (OC)	0.82	(0.81, 0.82)	0.82
	Organizational Process (OP)	<b><u>0.67</u></b>	(0.67, 0.68)	<b><u>0.67</u></b>

Table 4.30 indicates that the estimated Fleiss' Kappa values for each tier ranged from 0.74 to 0.82, suggesting “substantial” to “near perfect” reliability for the individual

tier levels according to Landis and Koch (1977). As in the first session, the Unsafe Supervision tier exhibited the lowest estimated Fleiss' Kappa value. In addition, Tables 4.33 to 4.35 show a decline in Fleiss' Kappa and Krippendorff's Alpha values in comparison with the HFACS tiers. Fleiss' Kappa ranged from 0.46 to 0.83 for the individual categories. According to Landis and Koch (1977), these results suggest "moderate" to "near perfect" reliability for the individual causal categories. While three causal categories -- Physical Environment, Failed To Correct a Known Problem, and Organizational Process -- exhibited "near perfect" reliability levels, 6 causal categories -- Skill Based Error, Decision Error, Adverse Physiological State, Inadequate Supervision, Planned Inappropriate Operations, and Supervisory Violation -- exhibited "moderate" reliability levels, suggesting that these are the problematic categories; the remaining causal categories exhibited "substantial" reliability levels.

Similarly, the individual Krippendorff's Alpha values for each HFACS tier, as shown in Table 4.33, ranged from 0.73 to 0.82. According to Krippendorff (2006), the overall Alpha values for all individual tiers are considered reliable except for the Unsafe Acts tier, which is considered "tentatively" reliable. Similar to Fleiss' Kappa values, Krippendorff's Alpha values for the causal category level were lower than the tier level, these values ranging from 0.45 to 0.83, suggesting a heterogeneous outcome. Only 3 categories -- Physical Environment, Failed To Correct a Known Problem and Organizational Climate -- are considered reliable, while 8 categories -- Perceptual Error, Routine Violation, Exceptional Violation, Adverse Mental State, Physical/Mental Limitations, Communication Coordination and Planning, Resource/Acquisition Management, and Organizational Process -- exhibited Krippendorff's Alpha values

between 0.67 and 0.79, also considered “tentatively” reliable. The remaining 8 categories -- Skill Based Error, Decision Error, Technological Environment, Adverse Physiological State, Fitness for Duty, Inadequate Supervision, Planned Inappropriate Operations, and Supervisory Violation -- are considered unreliable.

As in the first session, diagnostic analysis using item analysis was conducted on all of the causal factors (I = 95) included in the survey including all coders (R=59) to determine the most and least frequently chosen causal categories for each causal factor. The results of this analysis are presented by percentage in Table 4.36. For each causal category, beginning with Skill Based Error, the causal factor item in the survey with the highest percentage referring to a particular category is arranged in descending order; these percentages are shaded in Table 4.36.



Table 4.36: Percentage of Coders Responses to each Statement for Second Session

Statement Number	SBE	DE	PE	RV	EV	PhE	TE	AMS	APS	PML	CC	FfD	IS	PIO	FTCNP	SV	RAM	OC	OP
65	89.8	6.8	0	0	0	0	0	0	0	1.7	0	0	1.7	0	0	0	0	0	0
49	84.7	5.1	5.1	1.7	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	78	19	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
51	71.2	19	5.1	0	2	0	0	3.4	0	0	0	0	0	0	0	0	0	0	0
* 92	40.7	5.1	3.4	0	0	20	0	30.5	0	0	0	0	0	0	0	0	0	0	0
79	11.9	81	0	0	0	0	0	0	0	0	0	0	1.7	0	3.4	1.7	0	0	0
12	3.4	68	5.1	3.4	5	0	0	10.2	0	0	0	0	0	0	0	5.1	0	0	0
72	32.2	68	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
83	30.5	56	1.7	0	2	0	0	0	0	5.1	0	0	5.1	0	0	0	0	0	0
64	5.1	54	0	1.7	10	1.7	0	0	0	0	0	0	1.7	10.2	0	15	0	0	0
81	3.4	1.7	95	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
62	3.4	3.4	93	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
84	5.1	8.5	85	0	0	0	0	0	0	1.7	0	0	0	0	0	0	0	0	0
28	28.8	3.4	66	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0
91	32.2	0	64	0	2	0	2	0	0	0	0	0	0	0	0	0	0	0	0
* 24	0	3.4	58	0	0	3.4	27	5.1	0	1.7	0	0	0	0	0	0	1.7	0	0
67	0	0	0	98	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
78	0	0	0	97	0	0	0	0	0	0	1.7	0	0	0	0	0	0	0	1.7
18	0	0	0	90	0	0	0	0	0	0	0	0	0	1.7	3.4	0	0	5.1	0
29	0	5.1	0	88	2	0	0	0	0	0	0	0	0	1.7	0	3.4	0	0	0
61	0	0	0	81	2	0	0	0	0	0	0	0	0	0	6.8	5.1	1.7	3.4	0
68	3.4	5.1	0	0	92	0	0	0	0	0	0	0	0	0	0	0	0	0	0
86	0	0	0	10	90	0	0	0	0	0	0	0	0	0	0	0	0	0	0
54	0	1.7	0	17	81	0	0	0	0	0	0	0	0	0	0	0	0	0	0
52	3.4	1.7	0	15	80	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Category Number	SBE	DE	PE	RV	EV	PhE	TE	AMS	APS	PML	CC	FfD	IS	PIO	FTCNP	SV	RAM	OC	OP
46	0	1.7	0	14	59	0	0	3.4	8.5	0	0	10.2	1.7	0	0	1.7	0	0	0
33	0	0	0	1.7	24	0	0	0	0	11.9	0	8.5	5.1	0	0	19	18.6	0	12
39	0	0	0	0	0	98	0	0	0	0	0	0	0	0	0	0	1.7	0	0
34	0	0	0	0	0	97	2	0	0	0	0	0	0	0	0	0	0	0	1.7
53	0	0	0	0	0	97	2	0	0	0	0	0	0	0	0	0	1.7	0	0
13	0	0	0	0	0	95	2	0	0	0	0	0	0	0	0	0	0	0	3.4
77	0	0	0	0	0	86	9	0	0	0	0	0	0	0	1.7	0	3.4	0	0
44	0	0	0	0	0	1.7	98	0	0	0	0	0	0	0	0	0	0	0	0
95	0	0	6.8	0	0	14	75	0	0	0	0	0	0	1.7	1.7	0	1.7	0	0
14	0	0	0	0	0	0	68	0	0	0	0	0	0	0	3.4	0	25.4	1.7	1.7
2	0	0	0	0	0	29	41	0	0	0	0	0	0	10.2	8.5	0	1.7	1.7	8.5
7	1.7	0	1.7	0	0	0	0	84.7	3.4	6.8	0	0	0	1.7	0	0	0	0	0
17	0	0	0	0	0	0	0	69.5	14	5.1	0	3.4	0	3.4	0	5.1	0	0	0
47	1.7	0	0	0	0	0	0	96.6	0	0	0	0	0	1.7	0	0	0	0	0
69	0	0	0	0	2	0	0	91.5	1.7	0	0	5.1	0	0	0	0	0	0	0
88	0	0	0	0	0	0	0	94.9	3.4	1.7	0	0	0	0	0	0	0	0	0
89	0	0	0	0	0	0	0	0	92	3.4	0	5.1	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	3.4	78	3.4	0	15.3	0	0	0	0	0	0	0
27	0	8.5	0	0	0	1.7	0	0	73	3.4	0	13.6	0	0	0	0	0	0	0
* 93	0	0	0	0	0	0	0	6.8	63	0	0	30.5	0	0	0	0	0	0	0
66	0	0	0	0	0	0	0	13.6	61	16.9	0	5.1	0	3.4	0	0	0	0	0
82	0	0	0	0	0	1.7	2	0	0	96.6	0	0	0	0	0	0	0	0	0
90	0	0	0	0	0	0	3	0	0	96.6	0	0	0	0	0	0	0	0	0
71	0	0	5.1	0	0	0	0	0	1.7	93.2	0	0	0	0	0	0	0	0	0
5	0	0	5.1	0	0	0	0	0	3.4	88.1	0	0	1.7	0	0	0	1.7	0	0
75	5.1	1.7	0	1.7	0	0	0	0	0	57.6	0	3.4	8.5	1.7	0	1.7	18.6	0	0

Category Number	SBE	DE	PE	RV	EV	PhE	TE	AMS	APS	PML	CC	FfD	IS	PIO	FTCNP	SV	RAM	OC	OP
43	3.4	0	0	0	0	0	0	0	0	0	97	0	0	0	0	0	0	0	0
73	1.7	0	0	0	0	0	0	0	0	0	97	0	1.7	0	0	0	0	0	0
11	3.4	1.7	0	0	0	0	0	1.7	0	0	92	0	1.7	0	0	0	0	0	0
32	10.2	0	0	0	0	0	0	0	0	0	88	0	1.7	0	0	0	0	0	0
22	0	10	0	0	0	0	0	0	0	0	58	1.7	27	3.4	0	0	0	0	0
10	0	0	0	0	0	0	0	0	8.5	3.4	0	88.1	0	0	0	0	0	0	0
50	0	0	0	0	0	0	0	5.1	12	0	0	83.1	0	0	0	0	0	0	0
58	0	1.7	0	0	3	0	0	5.1	12	0	0	76.3	0	1.7	0	0	0	0	0
80	0	0	0	0	2	0	0	15.3	14	0	0	69.5	0	0	0	0	0	0	0
70	1.7	3.4	0	0	0	0	0	1.7	37	0	0	55.9	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	1.7	88	3.4	0	6.8	0	0	0
63	0	0	0	0	0	0	0	0	0	3.4	0	0	83	0	0	0	8.5	0	5.1
9	0	0	0	0	0	0	0	0	0	0	0	0	76	18.6	0	1.7	3.4	0	0
41	0	0	0	0	0	0	0	5.1	0	0	3.4	0	63	1.7	0	0	0	27.1	0
8	3.4	5.1	1.7	0	0	0	0	0	0	0	24	0	61	1.7	0	3.4	0	0	0
48	0	0	0	0	0	0	0	0	0	3.4	10	0	36	13.6	32.2	5.1	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	5.1	83.1	10.2	1.7	0	0	0
31	0	1.7	0	0	0	0	0	0	0	0	0	0	10	78	0	8.5	1.7	0	0
60	0	0	0	0	0	0	0	1.7	0	0	0	0	15	69.5	0	0	13.6	0	0
38	0	0	0	0	0	0	0	0	0	0	0	0	14	61	3.4	15	0	0	6.8
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	96.6	3.4	0	0	0
55	0	1.7	0	0	0	0	0	0	0	0	0	0	0	0	96.6	1.7	0	0	0
94	0	0	0	0	0	5.1	0	0	0	0	0	0	0	0	94.9	0	0	0	0
45	0	1.7	0	0	0	0	0	0	0	0	0	0	0	3.4	93.2	1.7	0	0	0
87	0	1.7	0	0	0	0	0	0	0	0	0	0	0	0	88.1	3.4	5.1	1.7	0
20	0	1.7	0	1.7	5	0	0	0	1.7	0	0	0	1.7	6.8	1.7	80	0	0	0

Category Number	SBE	DE	PE	RV	EV	PhE	TE	AMS	APS	PML	CC	FfD	IS	PIO	FTCNP	SV	RAM	OC	OP
23	0	3.4	0	8.5	5	0	0	0	0	0	0	0	1.7	3.4	1.7	76	0	0	0
25	0	0	0	6.8	2	0	0	0	0	0	0	0	5.1	8.5	5.1	71	0	1.7	0
74	3.4	8.5	0	10	27	0	0	1.7	0	0	0	0	0	0	0	49	0	0	0
40	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	98.3	0	0
1	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	94.9	0	1.7
37	0	0	0	1.7	0	0	0	0	0	0	0	0	0	0	0	0	94.9	0	3.4
57	0	0	0	0	0	0	0	0	0	0	0	0	5.1	1.7	0	6.8	71.2	3.4	12
85	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.7	57.6	0	41
76	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	98.3	1.7
4	0	0	0	0	0	0	0	3.4	0	0	0	0	0	0	0	0	0	96.6	0
30	0	0	0	0	0	0	0	3.4	0	0	0	0	0	0	0	0	0	96.6	0
21	0	0	0	0	0	0	0	0	0	0	0	0	1.7	0	0	0	3.4	89.8	5.1
35	0	0	0	0	0	0	0	8.5	1.7	0	14	0	0	5.1	0	0	1.7	69.5	0
3	0	0	0	0	0	0	0	0	0	0	3.4	0	3.4	3.4	0	0	0	1.7	88
36	0	0	0	0	0	0	2	0	0	0	0	0	3.4	0	0	1.7	3.4	1.7	88
59	0	0	0	0	0	0	2	0	0	0	0	0	8.5	0	0	0	3.4	1.7	85
56	0	0	0	0	0	0	15	0	0	0	1.7	0	1.7	1.7	1.7	0	1.7	1.7	75
42	0	0	0	0	0	0	0	0	0	0	0	0	14	1.7	0	0	10.2	1.7	73

\* Indicates compound causal factor

In general, in both reliability sessions, the overall and the individual inter-rater reliability for the tier level exhibited acceptable levels; however, the overall and the individual inter-rater reliability at the causal category level was less consistently acceptable. Furthermore, the overall intra-rater reliability of HFACS were 5% higher than the overall inter-rater reliability for the tier level, while for the category level this percentage increased to approximately 14%. In addition, the intra-rater reliability levels, which ranged from 0.57-0.89 for each HFACS category, were higher than the inter-rater reliability levels, which ranged from 0.46-0.82. Based on the results of both the intra-rater and inter-rater agreement coefficient values, the HFACS categories can be grouped into four groups based on their level of reliability, those exhibiting acceptable intra-rater and inter-rater reliability levels, those exhibiting acceptable intra-rater reliability levels and “tentatively”/“substantially” inter-rater reliability levels, those exhibiting “tentatively”/“substantially” levels for both intra-rater and inter-rater reliability, and finally, those exhibiting low and very low intra-rater and inter-rater reliability levels.

## CHAPTER 5: DISCUSSION

This chapter discusses the main findings of the overall intra-rater and inter-rater reliability of HFACS, in terms of its analysis, relation to other research, and contributions. Furthermore, it also considers the intra-rater and inter-rater reliability of HFACS for each tier and category. The 4 HFACS tiers are referred here as the macro-scale, corresponding to the 4 basic levels of the Swiss cheese model (Reason, 1990), while the finer level of the HFACS taxonomy, representing the 19 categories is referred to as the micro-scale. Section one considers in detail the intra-rater reliability of HFACS including the overall, macro-scale, and micro-scale, while section two discusses the inter-rater reliability across all levels, covering the overall, macro-scale, and micro-scale.

### 5.1 Intra-rater Reliability Discussion

In this study, the overall intra-rater reliability of HFACS at the macro-scale achieved acceptable levels based on percent agreement, Cohen's Kappa, and Krippendorff's Alpha values, while at the micro-scale these values declined; although considered reliable based on percent agreement values, according to Krippendorff's Alpha and Cohen's Kappa values, it achieved "tentative" and "substantial" reliability levels, respectively. While the studies in the safety literature on the reliability of HFACS in general are limited, as can be seen in Table 5.1, specifically, test-retest reliability has received the least attention. Olsen and Shorrock (2010) investigated the intra-rater

reliability of the HFACS-ADF derivative using percent agreement achieving 41%, well below the 70% agreement criterion considered reliable, while in the study reported here a 78.45% was achieved.

This contrast in the results between the two studies is perhaps due to several factors. First, the instrument used differed. While this study focused on the HFACS, Olsen and Shorrock (2010) used the HFACS-ADF derivative, a framework, although similar to the basic HFACS structure, includes additional causal categories. Another important factor is that in Olsen and Shorrock's (2010) study, the coders were given accident reports containing several causal factors that needed to be identified before they could be coded, whereas in this study the causal factors were pre-identified. Ross, Wallace, & Davies (2004) emphasize that the results of reliability studies using actual reports are 10% lower than those using pre-identified causal factors. Moreover, the duration between the two sessions differed; while in Olsen and Shorrock's (2010) study, the duration varied from 4 to 11 months; in this study it was much shorter, only 2 weeks, perhaps indicating that over time as memory fades, the positive effects of training on test-retest reliability deteriorates.

Table 5. 1: Comparison of HFACS Reliability Studies Reported in the Literature with the Current Study

		Taxonomy of Unsafe Operations	HFACS					HFACS Derivatives			
Aspect	Study	Rabbe, 1996 Walker, 1996 Ranger, 1997 Plourde, 1997	Johnson, 1997	Weigmann et al., 2000	Weigmann and Shappell, 2001c	Li and Harris, 2005	Olsen, 2011	This Study	DoD-HFACS O'Connor, 2008	DoD-HFACS O'Connor (2010)	HFACS-ADF Olsen and Shorrock (2010)
	Reliability	Inter-rater	✓	✓	✓	✓	✓	✓	✓	✓	✓
Intra-rater							✓			✓	
Measure Used	Percent Agreement					✓	✓	✓		✓	
	Cohen's Kappa	✓	✓	✓	✓	✓	✓				
	Fliess' kappa						✓				
	Krippendorff's Alpha						✓				
	Multi-rater Kappa Free								✓		
Dataset	Transportation						✓				
	Mining						✓				
	Construction						✓				
	Aviation	✓	✓	✓	✓	✓	✓	✓		✓	
	Food						✓				
	Lodging						✓				

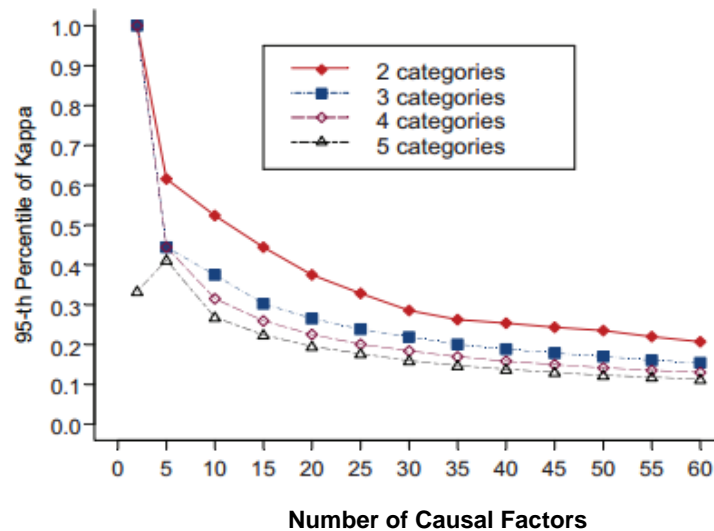


<b>Causal Factors</b>	Report										✓
	Pre-identified	✓	✓	✓	✓	✓	✓	✓	✓	✓	Interviewing U.S. Navy Officer
<b>Number of Coders</b>	Inter-rater	3	3	2	5	2	3 (HFS) 4 (ATC)	125	123	22	11
	Intra-rater							59			4
<b>Level of Experience</b>	Safety Specialists: Basic HFACS ASO students Human factors specialist (HFS) Air traffic controllers (ATC) Military Officers Pilots Aviation psychologist Instructor pilot	✓	✓	✓	✓		self-trained  self-trained	✓	✓	✓	✓
	Tier Level						Overall: PA: - 64.8% (ATCO) - 56.4% (HFS)	Overall: PA: 84.77% K <sub>F</sub> : 0.79 $\alpha$ : 0.79 Ranging K <sub>F</sub> : 0.73 - 0.82 $\alpha$ : 0.73 - 0.82			
<b>Inter-rater Results</b>											

	Category level	Overall 0.65 - 0.70 to 0.85 – 0.89	Overall 0.93 - 0.95	Overall 0.65 (air carrier) 0.75 (commercial)	Overall 0.72	Ranging PA: 72% - 96.4% K: 0.44 - 0.826	Overall: PA: - 36.1% (ATCO) - 34.5% (HFS) Ranging PA: - 0 - 41.2% (ATCO) - 0 – 33.3% (HFS)	Overall: PA: 68.69% K <sub>F</sub> : 0.67 α: 0.67 Ranging K <sub>F</sub> : 0.46- 0.82 α: 0.46- 0.82	Ranging 53% - 99%	Overall 0.76	Overall 39.9%
Intra-rater Results	Tier Level							Overall: PA: 90.22% K: 0.87 α: 0.87 Ranging Pa: 87.74- 91.92% α: 0.83-0.87			
	Category level							Overall: PA: 78.45% K: 0.77 α: 0.77 Ranging PA: 62.90- 91.25% α: 0.57-0.89			Overall PA:44.6%

In addition, the intra-rater reliability of HFACS at the macro scale was also considered acceptable, indicating that all tiers achieved acceptable levels of test-retest reliability. This result suggests that at the macro scale an internal consistency within the coders was achieved and the same level of understanding was maintained after a two-week period. However, the intra-rater reliability levels of HFACS at the micro-scale were generally lower than the macro-scale, the results ranging from reliable, tentatively reliable and unreliable. Eight categories – Perceptual Error, Routine Violation, Physical Environment, Physical/Mental Limitations, Communication Coordination and Planning, Failed To Correct a Known Problem, Organizational Climate, and Organizational Process – exhibited acceptable levels of intra-rater reliability, whereas 6 categories – Exceptional Violation, Technological Environment, Adverse Mental State, Adverse Physiological State, Fitness for Duty, and Resource/Acquisition Management – exhibited tentative levels of intra-rater reliability, indicating that for these 14 categories at the micro scale the coders maintained a consistent level of understanding after a two-week period. However, the remaining 5 categories – Skill Based Error, Decision Error, Inadequate Supervision, Planned Inappropriate Operations and Supervisory Violation – exhibited unacceptable intra-rater reliability levels, indicating that the coders were inconsistent in their responses of classifying the causal factors corresponding to these categories. As inter-rater reliability combines both stability and reproducibility, the factors that may have contributed to these results will be discussed in conjunction with inter-rater reliability.

The reduction of the intra-rater reliability levels from the macro scale to the micro scale is expected because as past research has emphasized, as the number of categories increases, reliability decreases. For example, Gwent (2011) demonstrated through a Monte-Carlo experiment, the results which are presented in Figure 5.1, that Kappa's critical value decreases as the number of categories increases, while the number of causal factors (I) are kept constant. In addition, Figure 5.1 shows that as the number of causal factors (I) increases in a test-retest reliability study, the Kappa values become more accurate and stable, when the number of categories are kept constant.



**Figure 5. 1: Kappa Coefficient by Number of Causal Factors and Number of Response Categories Under Random Rating (Gwent, 2011)**

An additional factor that might have contributed to the lower levels of the intra-rater reliability at the micro-scale may be related to the limitation of conducting the two reliability trials in two different settings. While the first session was conducted in an unconstrained timed classroom setting immediately at the end the HFACS training at the location of the training, the second was conducted in a work and/or home environment setting with a 72-hour timeframe restriction. This limitation was uncontrollable due to the remote distance of the participants, making it impractical to set up a reliability trial similar to the first.

Intra-rater reliability, also referred to as stability, is limited as it investigates only whether the coder was consistent in his responses in the first and second trial; thus, no judgment can be made as to whether inconsistency was due to improvement or regression in the coding. Thus, it is considered the weakest form of reliability; however, the first step in investigating the reliability HFACS was to evaluate its stability as internal inconsistencies may limit its inter-rater reliability. Evaluating the inter-rater reliability, also known as reproducibility, incorporates investigating both intra-coder inconsistencies and inter-coder differences (Krippendorff, 2006), thus it is a significantly stronger measure of reliability.

## 5.2 Inter-rater Reliability Discussion

The overall inter-rater reliability of HFACS at the macro-scale tier level was considered acceptable by percent agreement, substantial by Fleiss' Kappa, and tentatively reliable by Krippendorff's Alpha. Although similar results were achieved at the

micro-scale category level, the agreement coefficient value for the latter two was 0.67, in the lower ranges of substantially reliable for Fleiss' Kappa and tentatively reliable for Krippendorff's Alpha

While several studies used different agreement coefficients in evaluating the overall inter-rater reliability of HFACS, the ones comparable to this research using percent agreement are Olsen (2011) and Olsen and Shorrock (2010). In Olsen's (2011) study, the overall inter-rater reliability at the macro-scale tier level was 23.35% to 33.55%, lower than the results found here, while at the micro-scale category level this percentage difference increased to 47.4% to 50%. This variation in the results perhaps may have been due to differences in training method; while in Olsen's (2011) study the coders were self-trained, using a training workbook that included definitions for each category in the HFACS taxonomy, examples, and a solved example to practice coding, in this study the coders attended a 2-day face-to-face training workshop on HFACS. The face-to-face training is supported by Shohamy, Gordon and Kraemer (1992) who found that the overall inter-rater reliability levels were higher for classroom trained coders than they were for the untrained ones. Similarly, in Olsen and Shorrock (2010) the overall inter-rater reliability at the macro-scale was almost 40%, also lower than the results obtained in this study. This difference may be due to their use of the HFACS-ADF derivative and/or to their methodology of coders coding actual accident reports rather than pre-identified causal factors.

Subsequent to evaluating the overall inter-rater reliability, a finer assessment was conducted evaluating the inter-rater reliability of HFACS at the macro-scale for each tier. These reliability results for all tiers were considered acceptable except for the Unsafe Supervision tier which was found to be substantially and tentatively reliable according to Fliess' Kappa and Krippendorff's Alpha, respectively, perhaps suggesting that at the micro-scale one or more categories in this tier lack adequate inter-rater reliability levels.

Similar to the intra-rater reliability results of HFACS at the macro-scale, the inter-rater reliability at the micro-scale ranged from 0.46 - 0.82 based on both Fliess' Kappa and Krippendorff's Alpha. Although these agreement coefficients used in this study take into account agreement by chance, these results are almost 3 times higher than Olsen's (2011) using percent agreement. This difference is probably due to the training method and/or because percent agreement is confounded by the number of categories coded, meaning the denominator of percent agreement for some pairs differed in Olsen's (2011).

Based on the inter-rater agreement coefficient values for each causal category, the HFACS categories can be grouped into four levels: acceptable, substantial/tentative, mixed, and low/very low. The acceptable group consists of the 3 categories – Physical Environment, Failed To Correct a Known Problem and Organizational Climate – indicating both an internal consistency for each coder and that all coders had a consistently similar understanding of these 3. The substantial/tentative group includes 7 categories – Perceptual Error, Routine Violation, Adverse Mental State, Physical/Mental

Limitations, Communication and Coordination Planning, Fitness for Duty, and Organizational Process. The mixed group, exhibiting substantial and unreliable levels based on Fleiss' Kappa and Krippendorff's Alpha, respectively, includes 4 categories – Exceptional Violation, Technological Environment, Adverse Physiological State, and Resource/Acquisition Management. The fourth low/very low group consists of the remaining 5 categories – Skill Based Error, Decision Error, Inadequate Supervision, Planned Inappropriate Operations, and Supervisory Violation; these categories also suffered low intra-rater reliability levels.

This decline in reliability probably involves a combination: a decrease in the percentage of coders agreeing on classifying a certain number of causal factors into a particular category and an increase in the percentage of coders agreeing on classifying other causal factors into a particular category. These disagreements can be referred to as horizontal and vertical, based on Tables 4.23 and 4.36, which perhaps indicates coding difficulties. More specifically, these coding difficulties might be due to several factors, the primary ones being related to the lack of detail in the phrasing of the causal factors, the inattention of coders when coding, ambiguous factors, and/or the training.

Although this research addressed the already known issue of compound causal factors, its results suggest a lack of clarity in the phrasing. For example, one might argue that causal factor 84 – the forklift driver under-estimated the container's weight, which resulted in the forklift tipping over – can be classified into the Decision Error category, suggesting the truck driver created a plan that proved to be inappropriate, while another



coder may think that the forklift driver used his vision to estimate the weight, coding it as a Perceptual Error. As a result, this causal factor appears to lack enough detail to determine with certainty and confidence the appropriate category. This problem is anticipated in the real world because causal factors that form mishap/accident reports are usually prepared by personnel who may not be as thorough or specific as needed.

In addition, the inattention of some coders might have also contributed to low inter-rater and intra-rater reliabilities, because of the large number of causal factors, 95, used in this study. This situation impacting particular causal factors Routine Violation, Exceptional Violation and Supervisory Violation, an indication that although they have identified it as a violation, some inattentively missed the phrasing in the causal factor differentiating them. For instance, some coders, approximately 11%, identified causal factors 23 and 25 as Routine Violation category, the phrasing – to control insects, the supervisor uses unauthorized pesticides in the hotel’s garden areas – and – the night shift supervisor encourages maintenance crews to “bend the rules” in order to complete work orders on time – suggested that these were Routine Violations since they might be tolerated by top management; however, these coders missed that they are committed by supervisors rather than operational workers, indicating them as Supervisory Violations. Similarly, the phrasing of causal factor 17 – the shift supervisor was mentally tired after working two shifts – was classified by 11% of the coders as an Adverse Physiological State category, suggesting that several coders missed the word “mentally” in front of tired classifying it as an Adverse Physiological State category, rather than Adverse Mental State.

In addition, ambiguous factors that may contribute to the low reliability levels, especially for the inter-rater type, may be due to differences in coders themselves, including age, gender, experience, educational background, and personal character. In addition, to a less extent, this decline in reliability may also be due to the possibility that the HFACS categories are not exhaustive, mutually exclusive, a situation not in the scope of this research. Furthermore, research has found that for taxonomies including a hierarchy of exhaustive, mutually exclusive categories, like HFACS, coder agreements are expected to vary according to specific training requirements (Annett & Duncan, 1967), affecting both the inter-rater and intra-rater reliability levels. To address this issue, the HFACS training should be designed to decrease coder agreement variation to a level that does not influence the reliability of each HFACS category.

As past research has found, training is a significant factor affecting the intra-rater and inter-rater reliability, with levels being found to be higher for trained coders than for the untrained ones, especially for intra-rater reliability (Weigle, 1998); more importantly, it has been found that reliability levels can be improved if coders receive classroom training (Shohamy, Gordon & Kraemer, 1992). The lack of control of the training in this study was intended to use real world HFACS training programs, which can be considered a limitation as the researcher did not have control over the training, including the material, the examples covered, and the level of the instructor. While various causal factors indicate that some coders had issues with how to code, others could not differentiate between certain HFACS categories, both of which perhaps may indicate training weaknesses.

Teaching coders how to code is a fundamental component of any HFACS training; although this was addressed in the training aligned with this study, some coders still had issues with how to code. The primary such issue involved coders coding were based on the consequence of the causal factor rather than the literal statement and the facts included in it. For example, factor 46 – the construction worker was smoking marijuana during work without knowledge of top management – was coded by some as Adverse Physiological State or Adverse Mental State indicating that they based their decision on the physical and/or mental consequence of smoking marijuana, rather than focusing on the illegal act of the worker at the operational level, suggesting an Exceptional Violation category. Similarly, causal factor 75 – the technical worker lacks the type of skills and performance levels required for an acceptable level of job competency – was coded by a few as Inadequate Supervision or Resource/Acquisition Management, indicating that they based their coding on the previous event of who was responsible for hiring this worker as opposed to focusing on the fact that this worker lacked the mental abilities to do the job successfully, denoting a Physical/Mental Limitation.

In addition, some coders could not distinguish between certain categories belonging to the same tier, perhaps implying training weaknesses represented by lack of emphasis, explanation or examples. For instance, some coders couldn't differentiate between Skill Based Errors and Decision Errors; although the majority of the coders, 89%, identified causal factors 6, 49, 51, and 65 as being an Unsafe Act committed at the operational level, 12% classified them as Decision Errors and the remaining classified

them as Skill Based Errors. Similarly, while 80% of the coders identified causal factors 12, 64, 72, 79, and 83 as an Unsafe Act, 20% of these coders classified them as Skill Based Errors and the remaining classified them as Decision Errors. These results suggest that these coders, 12% and 20%, realized these causal factors were Unsafe Acts committed at the operational level; however, they couldn't differentiate the different level of conscious demand required by the two: Decision Errors requires medium to high conscious demand, while the Skill Based Errors requires no conscious demand, as it is spontaneous in nature. Similarly, to a less extent some coders confused Perceptual Error with either Skill Based Error or Decision Error.

Another issue relating to training may involve providing simple examples to trainees, impacting the trainees' ability to think beyond obvious examples. For instance, coders classified causal factor 2 – the inadequate layout design of the equipment in the plant forces the workers to take routes other than the designated ones – to either Physical Environment or Technological Environment. The discussion of this causal factor with the training instructor revealed that complex examples such as layouts of equipment and design considerations were not included in the training.

Coder training appears to be the solution to the majority of the problems mentioned here. Research in the education assessment domain has found that properly designed training can improve reliability (Graham et al, 2012). In this domain, Frame of Reference (FOR) training was developed to foster common and consistent understanding among raters of the rating system. A similar method, FOR-HFACS training, could be developed to address the primary causes of coder disagreement for the HFACS

taxonomy. This FOR-HFACS training may specifically involve an explanation and process overview of using the HFACS as an accident investigation and analysis tool, a thorough explanation of each HFACS tier and category focusing on differences that distinguish similar categories, a discussion of common errors through examples of how to avoid bias, and training on the proper way of coding, covering the mental process and key words/concepts for example.

While one of the conclusions from this study is designing and developing a training program for HFACS, the most significant findings of this study was the consistency of the inter-rater reliability results between the two sessions. For example, the difference in the values of all agreement coefficients between the first and second session for the overall inter-rater reliability for the macro-scale tier level was below 1%, shown in Table 5.2. In addition, this conclusion is also supported by the 1% difference in the values of all agreement coefficients between the first and second session for the overall inter-rater reliability for the micro-scale category level, shown in Table 5.3, despite the time difference of 2 weeks between the two sessions.

Table 5. 2: Comparison of Overall Inter-rater Reliability Results Between First and Second Session for Tier Level

Agreement Measure	First Session		Second Session	
	Overall	95% CI for Overall	Overall	95% CI for Overall
Average Percent Agreement	84.77%	(84.67, 84.87)	85.25%	(84.99, 85.55)
Fleiss' Kappa	<b><u>0.79</u></b>	(0.79, 0.79)	<b><u>0.80</u></b>	(0.79, 0.80)
Krippendorff's Alpha	<b><u>0.79</u></b>	(0.74, 0.83)	0.80	(0.75, 0.84)

Table 5. 3: Comparison of Overall Inter-rater Reliability Results Between First and Second Session for Category Level

Agreement Measure	First Session		Second Session	
	Overall	95% CI for Overall	Overall	95% CI for Overall
Average Percent Agreement	<b><u>68.69%</u></b>	(68.52, 68.86)	<b><u>68.10%</u></b>	(67.97, 68.52)
Fleiss' Kappa	<b><u>0.67</u></b>	(0.67, 0.67)	<b><u>0.66</u></b>	(0.66, 0.66)
Krippendorff's Alpha	<b><u>0.67</u></b>	(0.66, 0.68)	<b><u>0.66</u></b>	(0.65, 0.67)

Similarly, the variability of all agreement coefficients values between the first and second sessions for the inter-rater reliability of HFACS at the macro scale for all tiers was also within 1%, as shown in Table 5.4. However, this variability increased reaching to 4% for all agreement coefficients values between the first and second sessions for the inter-rater reliability of HFACS at the micro scale including all categories. This low increment in variability supports the consistency of the inter-rater reliability results regardless of the time difference of 2-weeks between the two sessions, perhaps indicating the inter-rater reliability levels of HFACS for practicing coders.

Table 5. 4: Comparison of Inter-rater Reliability Results for Each HFACS Tier Between First and Second Session

HFACS Tier	First Session			Second Session		
	$K_F$	95 % CI $K_F$	$\alpha$	$K_F$	95 % CI $K_F$	$\alpha$
Unsafe Acts Tier	0.82	(0.81, 0.82)	0.82	0.82	(0.82, 0.83)	0.82
Preconditions of Unsafe Acts Tier	<b><u>0.80</u></b>	(0.80, 0.80)	0.80	0.81	(0.81, 0.82)	0.80
Unsafe Supervision Tier	<b><u>0.73</u></b>	(0.73, 0.73)	<b><u>0.73</u></b>	<b><u>0.74</u></b>	(0.74, 0.75)	<b><u>0.73</u></b>
Organizational Influences Tier	<b><u>0.80</u></b>	(0.08, 0.81)	0.80	<b><u>0.80</u></b>	(0.79, 0.80)	0.80

Table 5. 5: Comparison of Inter-rater Reliability Results for Each HFACS Category Between First and Second Session

	HFACS Category	First Session		Second Session	
		$K_F$	$\alpha$	$K_F$	$\alpha$
Unsafe Acts	Skill Based Error (SBE)	0.56	0.56	0.54	0.54
	Decision Error (DE)	0.46	0.46	0.46	0.45
	Perceptual Error (PE)	<b><u>0.72</u></b>	<b><u>0.72</u></b>	<b><u>0.72</u></b>	<b><u>0.71</u></b>
	Routine Violation (RV)	<b><u>0.76</u></b>	<b><u>0.76</u></b>	<b><u>0.76</u></b>	<b><u>0.76</u></b>
	Exceptional Violation (EV)	<b><u>0.63</u></b>	0.63	<b><u>0.66</u></b>	<b><u>0.66</u></b>
Preconditions of Unsafe Acts	Physical Environment (PhE)	0.82	0.82	0.83	0.83
	Technological Environment (TE)	<b><u>0.65</u></b>	0.65	<b><u>0.62</u></b>	0.61
	Adverse Mental State (AMS)	<b><u>0.68</u></b>	<b><u>0.68</u></b>	<b><u>0.69</u></b>	<b><u>0.69</u></b>
	Adverse Physiological State (APS)	<b><u>0.63</u></b>	0.63	0.58	0.58
	Physical / Mental Limitations (PML)	<b><u>0.73</u></b>	<b><u>0.73</u></b>	<b><u>0.76</u></b>	<b><u>0.75</u></b>
	Communication Coordination & Planning (CC)	<b><u>0.78</u></b>	<b><u>0.78</u></b>	<b><u>0.78</u></b>	<b><u>0.78</u></b>
	Fitness for Duty (FfD)	<b><u>0.73</u></b>	<b><u>0.73</u></b>	<b><u>0.61</u></b>	0.61
Unsafe Supervision	Inadequate Supervision (IS)	0.51	0.51	0.53	0.52
	Planned Inappropriate Operations (PIO)	0.49	0.49	0.52	0.52
	Failed To Correct a Known Problem (FTCNP)	0.82	0.82	0.81	0.81
	Supervisory Violation (SV)	0.53	0.53	0.50	0.50
Organizational Influences	Resource / Acquisition Management (RAM)	<b><u>0.62</u></b>	0.62	<b><u>0.66</u></b>	<b><u>0.66</u></b>
	Organizational Climate (OC)	<b><u>0.80</u></b>	0.80	0.82	0.82
	Organizational Process (OP)	<b><u>0.69</u></b>	<b><u>0.69</u></b>	<b><u>0.67</u></b>	<b><u>0.67</u></b>

In addition, the consistency established here is further supported by the similarity of the inter-rater reliability levels for the two sessions, although the number of coders who participated in each session differed, 125 in the first session and 59 in the second.

The large sample size of coders used in this study, may have contributed to this consistency, probably reflecting intra-rater and inter-rater reliability levels of HFACS for practicing coders. In addition, the large sample size also played an important role in reducing variability and achieving similar inter-rater reliability results within each session across all levels, both with and without rogue coders and with and without compound causal categories; because the larger the sample size, the more it represents the population mean and reduces the variability within the sample.

Furthermore, this consistency is also strengthened by the similarity of the values achieved using different agreement coefficients, especially for those agreement coefficients that take agreement by chance into account. The results of these agreement coefficients converge to similar interpretations, despite the difference in their computations and properties. Percent agreement results were highest among the 3 agreement measures used, while the other 2 which take agreement by chance into account, were lower; specifically, Krippendorff's Alpha was the most conservative between the latter two.



## CHAPTER 6: CONCLUSIONS AND FUTURE WORK

This research furthers the research field of HFACS and its validity as its goal was to investigate its reliability focusing on both intra-rater and inter-rater, including individual tiers and categories. This study supplements past research by using a large sample size of 125 coders rather than 3 or 4 coders. Furthermore, these coders were safety professionals from several industries who received similar HFACS training, while in other studies they were either human factors specialists or pilots. Moreover, this study used actual incident/accident data from various industries represented by 95 causal factors, whereas past research has focused mainly on accident data from the aviation sector.

More importantly, this study used more than one statistical measure to evaluate its reliability, allowing for a more detailed interpretation of the degree of the reliability, compared to a single statistical measure that provides limited information. The results of the 3 statistical measures – percent agreement, Krippendorff's Alpha ( $\alpha$ ), and Cohen's Kappa (K) – converge to suggest that the overall intra-rater reliability of HFACS is acceptable. Although, its inter-rater reliability determined using percent agreement, Krippendorff's Alpha ( $\alpha$ ), and Fleiss' Kappa (KF) is also reasonable, the values of these measures were close to the minimum threshold. In addition, while the findings also suggest that the 4 tiers of HFACS are reliable, not all of the 19 categories are reliable. This finding is the cause for considerable concern and further research is required.

Future studies on the reliability of HFACS should be designed to include coders coding actual causal factors from incident/accident reports in addition to the other factors

that were adopted here. In this study here pre-identified causal factors were used instead of actual incident/accident reports because the researcher had no control over training. In such potential studies, the evaluation of reliability could individually focus on the identification of causal factors from incident/accident reports, the coding of causal factors into HFACS causal categories, and then on both together.

The findings also suggest that additional consideration needs to be given to the HFACS training. In addition to the idea of developing FOR-HFACS training, the effectiveness of this training program is assessed first by evaluating reliability then identifying issues within this program and what elements need to be improved in the training to increase its coder consistency. Moreover, the effect of coder training using FOR-HFACS could be evaluated by comparing its reliability to 2 groups, no training and common training programs using between-subjects research design, hypothesizing FOR-HFACS training would improve reliability.

Specifically for training purposes, future work may include designing and establishing a tool, for example a flowchart, with the intention of increasing the reliability of HFACS. This flowchart could assist coders in the process of correctly classifying the mishap/accident causal factor into the appropriate HFACS causal category. The flowchart would begin by asking the coder sequential questions until he/she accurately identifies the tier to which the causal factor belongs; then for each tier this process is repeated with an additional set of questions until the correct HFACS causal category is determined. In this digital and distance education age, this tool could be adopted as an online tool and/or

an application. One of the advantages of this online tool is it can function as a refresher training, enhancing the coding process when needed.

Reliable HFACS data is essential for empirical research on safety systems and on the effectiveness of any mitigation and/or accident prevention plans and strategies. Once a company has classified its accident and near miss cases using the HFACS taxonomy, it can analyze these data searching for trends which point to weaknesses in certain areas of the system. In addition, conducting association analysis among HFACS categories can help identify additional areas for improvement. Information of this nature not only provides the safety professional with supplementary knowledge to guide limited resources towards a more focused intervention, but also offers benefits to worker health through lowering frequency and severity of work accidents, all of which have a positive impact on cost.

HFACS reliability studies have an important role in advancing safety practices, techniques, and training. While this study furthers the research field of HFACS and its validity, the design study adopted here, including the results and how to test reliability, is applicable not only to safety taxonomies in particular but to all taxonomies used in various industries including healthcare, computer science and education. Because it is crucial that the data derived using taxonomy be defect free from bias and noise and have the same meaning for all users. Conducting reliability studies would, by time, enhance confidence in existing tools and provide trustworthy and reliable data that reflects properties of the taxonomy.

## APPENDICES

## HFACS Reliability Study

Please classify all the causal factors to the appropriate causal code. You are required to work independently.

What is your primary email address?

1. The Fire Dept. failed to provide fire proof clothing for the firemen for they were expensive.

- Skill Based Error
- Perceptual Error
- Decision Error
- Routine Violation
- Exceptional Violation
- Physical Environment
- Technological Environment
- Adverse Mental State
- Adverse Physiological State
- Physical / Mental Limitations
- Communication Coordination and Planning
- Fitness for Duty
- Inadequate Supervision
- Planned Inappropriate Operations
- Failed To Correct a Known Problem
- Supervisory Violation
- Resource / Acquisition Management
- Organizational Climate
- Organizational Process

2. The inadequate layout design of the equipment in the plant forces the workers to take routes other than the designated ones.
3. No procedure exists to ensure that only vegetable oils are ordered for all kitchen operations.
4. Workers working for company X are afraid to get things wrong or to admit to making mistakes because of the blame attitude.
5. The operator in the control room was colorblind and was not able to distinguish between different lights of the control panel.
6. Although, the worker has experience working with the saw and scrap materials, the worker forgot to adequately purge the tank and test for vapors before beginning to cut.
7. The nurse's mental capability degraded as the number of critical patients increased in the ER.
8. The chief engineer provided incorrect performance feedback to the worker.
9. The plant supervisor established work quotas that only the most skilled employees could complete safely and effectively.
10. While off-duty, a miner went to the gym and overexerted himself.
11. The day shift maintenance crew failed to tell the swing shift operators that the valve line-up was completed.
12. The doctor based his decision on intuition rather than requesting extra investigations from the patient, such as x-rays, blood tests ...etc.
13. The tsunami caused water intrusion into the emergency diesel generator rooms.
14. The fire suppression system is outdated and does not accurately reflect modern, upgraded equipment status.
15. A night shift driver with a severe cold and congestion fell asleep while transporting a load.
16. The utility manager did not provide guidance before allowing new employees to operate the machinery.
17. The shift supervisor was mentally tired after working two shifts.

18. Habitually, the construction workers in company X do not wear personal protective equipment in the working area.
19. The boss assigned two waitresses with a history of personal quarrel to the same work shift.
20. The supervisor authorized blasting activities knowing full well that the established blast safety zone was not according to the mines blasting rules and procedures.
21. Top management only values employees current results, disregarding employees level of commitment, previous performance ...etc.
22. The construction manager's instructions were vague to the worker and the worker did not seek clarification.
23. To control insects, the supervisor uses unauthorized pesticides in the hotel's garden areas.
24. The two monorail trains were identical, which caused confusion to the operator.
25. The night shift supervisor encourages maintenance crews to "bend the rules" in order to complete work orders on time.
26. The nuclear plant equipment operator told his supervisor about the leaking pipe, but the supervisor took no action.
27. The forklift driver was poisoned after eating contaminated food.
28. The operator misread the gauge meter and recorded a false reading.
29. The doctor customarily does not wash his hands in between patients.
30. The night shift custodial crew is afraid to voice concerns due to threatened retaliation by management.
31. The chief engineer assigned a job that was beyond the capability of the crew.
32. The company events manager failed to inform hotel staff that a congressman was spending the night.
33. The captain of the ship did not hold a merchant mariner license.
34. A large gust of wind caused an opened gate to close unexpectedly setting off a catastrophic chain of events.

35. The poor working relationships between company employees lead to a sense of isolation among employees.
36. The mine lacks sufficient standard operating procedures (SOPs) and policy standards.
37. The frequency of sampling and testing is decreased to every four hours instead of every hour at the wastewater treatment plant, due to budget cuts.
38. The supervisor created a work plan that did not address some very important safety precautions and risks.
39. The floors in the kitchen areas are slippery and wet.
40. The company is trying to save money and bought improper equipment that has been improperly guarded.
41. The workshop supervisor always criticizes the work of his employees.
42. The company did not develop and establish a training program for employees on the proper procedures of safety rules.
43. The second shift workers failed to inform the third shift workers of a hazard found during the course of their shift.
44. The instruments failed to indicate that the vessel was drifting.
45. Although, the chief electrical engineer was notified of the hazard in the electric room he did not initiate a plan to eliminate the hazard.
46. The construction worker was smoking marijuana during work, without knowledge of top management.
47. The engineer got stressed as the project deadline approached.
48. The head of the mechanical group was unable to solve/manage a conflict between two of his group members.
49. The security officer missed a check on his normally scheduled rounds.
50. The pool lifeguard was out all night partying and fell asleep on the job.
51. The worker inattentively isolated the incorrect equipment/machinery during scheduled maintenance which resulted to an electrical shock fatality.



52. The worker signed off a maintenance sheet without performing the maintenance or inspecting the work, which is against the rules and regulations.
53. A tremendous hail storm destroyed the atrium's glass enclosure.
54. The waitress, against published rules and top management policy, reheated the dinner guest's prime rib in the microwave.
55. The supervisor was informed of brake issues for one of the two haul trucks, but did not make it a priority to ensure that it was fixed before it was used again.
56. Work order systems and processes have become too cumbersome and need revision.
57. The mine contracts with outside personnel without doing background checks or checking that their qualifications are up-to-date.
58. The worker only slept three hours the previous night even though he was required to obtain 8 hours of crew rest.
59. The company lacks a program for frequent and regular inspections of the job site, materials, and equipment by a competent person.
60. The lab supervisor did not staff the lab adequately for timely response to emergent and urgent issues due to competing priorities.
61. The road maintenance worker, with the full knowledge of management, always smokes while working.
62. The operator misjudged the length of the boom and inaccurately determined the load radius.
63. The new chef did not receive adequate mentoring and coaching when he was hired.
64. The captain chose to continue fishing despite the severe weather predictions and the exposed location of the ship "Katmai".
65. The new chef burned the dish due to his poor braising technique.
66. The brick mason assistant was tired, after lifting 30 lbs. of brick continuously for two hours without a break.
67. Although against the rules, housekeepers routinely use the indoor executive pool on their breaks.

68. Not to his nature, the worker intentionally misused his personal protective equipment.
69. The receptionist just had an argument with her spouse and snapped at a rather demanding guest.
70. The bulldozer driver missed his allergy medication the previous night and instead took it in the morning before work.
71. The maintenance man had hearing deficiencies and was not able to hear the fire alarm.
72. The mechanic selected the wrong procedure to fix the hydraulic pump.
73. The day shift foreman failed to inform the swing shift of a large baking order required for completion by next morning.
74. The chief engineer inspected the construction site without using personal protective equipment.
75. The technical worker lacks the type of skills and performance levels required for an acceptable level of job competency.
76. The organization rewards successful risk takers and punishes those who slow down or halt a process for safety concerns.
77. The lighting in the laundry facility is inadequate.
78. Plant equipment operators routinely enter the switchyard without first contacting the control room.
79. The new operator recognized a malfunction in the equipment, but chose the wrong remedy to fix it.
80. The operator showed up for work while the effect of alcohol is still present.
81. The head crane operator misjudged the correct position to load the steel bars, which resulted to a risky imbalanced horizontal transportation.
82. The apprentice nuclear plant equipment operator was not tall enough to read the site glass gauges.
83. The maintenance engineer, having limited knowledge of the new air conditioning system took a chance and aligned the ventilation ducts improperly.

84. The forklift driver under-estimated the container's weight, which resulted in the forklift tipping over.
85. The restaurant lacks an adequate employment selection process resulting in the hiring of personnel who have infectious diseases that are inappropriate for the job.
86. Against rules and regulations and without supervisor knowledge, a miner randomly jumped onto the back of a haul truck and used it as an unauthorized form of transportation across the work site.
87. The temperature gauge in the freezer has been broken for weeks, but the local restaurant management refuses to fix it.
88. The accountant was stressed about losing his job because of the downsizing in the company.
89. The worker experienced moderate trembling due to blood sugar insufficiency.
90. The new haul truck driver is not tall enough to reach the pedal controls on the haul truck.
91. The nurse misread the reading of the blood pressure monitor.
92. The electrical operator got distracted by an external noise and forgot to take readings on the main transformer.
93. The warehouse forklift driver was suffering from a severe head cold, took OTC drugs, became groggy and dropped a load of boxes.
94. On rainy days the hotel entryway gets slippery, but the supervisors haven't taken steps to fix the problem.
95. The font size and coloring schemes on control room labels creates confusion when multiple events take place.

## Appendix B: Consent

Information Concerning Participation in a Research Study Clemson University

### **Assessment of the Human Factors Analysis and Classification System (HFACS): Inter-rater and Intra-rater Reliability**

#### **Description of the Research and Your Participation**

You are invited to participate in a research study conducted by Awatef Ergai under the direction of Dr. Scott Shappell and Dr. Anand K. Gramopadhye. The purpose of this research is to evaluate the reliability of the Human Factors Analysis and Classification System (HFACS).

This study will take place over two coding sessions subsequent to training. The first session will occur immediately at the end of the HFACS training, and the second two weeks later. In each coding session you will be given a survey which is composed of various causal factors along with a list of HFACS causal codes and your task is to read and classify each factor into the causal code that best describes it by checking the appropriate box. The amount of time required for your participation in each session will not exceed one hour.

#### **Risks and Discomforts**

There are no known risks associated with this research.

#### **Potential Benefits**

This research will help us to evaluate the reliability of the HFACS framework and to identify any weaknesses in the structure of the HFACS and thus refine the HFACS framework to a more trustworthy and dependable framework.

#### **Protection of Confidentiality**

We will do everything we can to protect your privacy. Your identity will not be revealed in any publication that might result from this study.

#### **Voluntary Participation**

Your participation in this research study is voluntary. You may choose not to participate and you may withdraw your consent at any time. You will not be penalized in any way should you decide not to participate or to withdraw from this study.

**Contact Information**

If you have any questions or concerns about this study or any problems arise, please contact Dr. Dr. Anand K. Gramopadhye at Clemson University at 864-656-5540. If you have any questions or concerns about your rights as a research participant, please contact the Clemson University Office of Research Compliance (ORC) at 864-656-6460 or [irb@clemson.edu](mailto:irb@clemson.edu). If you are outside the Upstate South Carolina area, please use the ORC's toll-free number, 866-297-3071.

A copy of this consent form will be given to you.

Appendix C: Numerical Notations of HFACS Categories

<b>HFACS Category</b>	<b>Numerical Code</b>	
	<b>Tier</b>	<b>Category</b>
<b>Skill Based Error</b>	1	1
<b>Decision Error</b>	1	2
<b>Perceptual Error</b>	1	3
<b>Routine Violation</b>	1	4
<b>Exceptional Violation</b>	1	5
<b>Physical Environment</b>	2	6
<b>Technological Environment</b>	2	7
<b>Adverse Mental State</b>	2	8
<b>Adverse Physiological State</b>	2	9
<b>Physical/Mental Limitations</b>	2	10
<b>Communication Coordination and Planning</b>	2	11
<b>Fitness for Duty</b>	2	12
<b>Inadequate Supervision</b>	3	13
<b>Planned Inappropriate Operations</b>	3	14
<b>Failed To Correct Problem Known Problem</b>	3	15
<b>Supervisory Violations</b>	3	16
<b>Resource Management</b>	4	17
<b>Organizational Climate</b>	4	18
<b>Organizational Process</b>	4	19

## Appendix D: Identification of Rogue Coders

Rogue coders were identified and excluded from the analyses using percent agreement values of the 125 coders who participated in the first session. Every coder was paired with the other 124, and percent agreement was determined, yielding a sample of 15,500 percent agreement values. Given the size of this dataset and the visual representations of the dataset -- histogram, box plot, and normal probability plot -- seen in Figure D.1 suggest that the data are approximately normally distributed with a sample mean of percent agreement of 68.49% and a sample standard deviation of percent agreement of 7.78%:

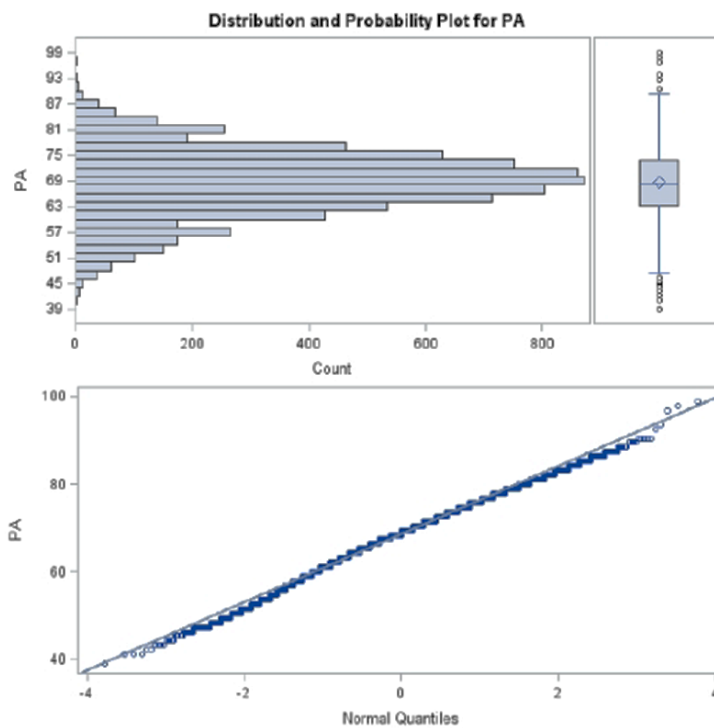


Figure D.1: Normal Plots of Percent Agreement Values

The empirical rule --  $\bar{x}_{\text{percent agreement}} - 2 * s_{\text{percent agreement}}$  -- was used to identify a rogue coder (Wilcox, 2010), which generated a lower cutoff value of 52.92% for this dataset. A coder exhibiting a percent agreement value less than this amount 22% or more of the time when paired with other coders was considered rogue. The five coders subsequently identified as rogue coders were not included in the analysis, reducing the sample size to 120 coders. These coders exhibited percent agreement values less than the criterion 72%, 56%, 22%, and 25% (for 2 of them) of the time when paired with other coders.



## REFERENCES

- Aas, A. (2008). The human factors assessment and classification system (HFACS) for the oil & gas industry. Paper presented at the *International Petroleum Technology Conference*,
- Agresti, A. (2007). *An introduction to categorical data analysis*. Wiley-Interscience.
- Airbus. (2005). *Flight operations briefing notes: Human performance, error management*. (No. FLT\_OPS – HUM\_PER – SEQ 07 – REV 01).
- Annett, J. & Duncan, K. (2005). *Task Analysis and Training Design*. (Report Resumes, Hull University).
- Baysari, M. T., McIntosh, A. S., & Wilson, J. R. (2008). Understanding the human factors contribution to railway accidents and incidents in Australia. *Accident Analysis and Prevention*, 40(5), 1750-1757.
- Baysari, M. T., Caponecchia, C., McIntosh, A. S., & Wilson, J. R. (2009). Classification of errors contributing to rail incidents and accidents: A comparison of two human error identification techniques. *Safety Science*, 47(7), 948-957.
- Beaubien, J. M., & Baker, D. P. (2002). A review of selected aviation human factors taxonomies, accident/incident reporting systems and data collection tools. *International Journal of Applied Aviation Studies*, 2(2), 11-36.
- Berry, K. A. (2010). *A meta-analysis of human factors analysis and classification system causal factors [electronic resource] : Establishing benchmarking standards and human error latent failure pathway associations in various domains*. (Clemson University Electronic Theses and Dissertations): (Doctor of Philosophy, Clemson University).
- Berry, K. A., Stringfellow, P. F., & Shappell, S. A. (2010). Examining error pathways: An analysis of contributing factors using HFACS in non-aviation industries. Paper presented at the *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 54(21) 1900-1904.
- Bird, F. E. (1974). *Management guide to loss control*. Institute Press Atlanta, Georgia.

- Boquet, A., Detwiler, C., Hackworth, C., Holomb, K., & Pfliegerer, E. (2007). Beneath the tip of the iceberg: A human factors analysis of general aviation accidents in Alaska versus the rest of the United States. Washington DC: Federal Aviation Administration.
- Broach, D. M., & Dollar, C. S.(2002). *Relationship of Employee Attitudes and Supervisor-Controller Ratio to En Route Operational Error Rates*, Oklahoma City, OK, Federal Aviation Administration, Civil Aeromedical Institute.
- Bureau of Labor Statistics .(2011). *Incident Rates*. Retrieved 10/2, 2012, from [www.census.gov/compendia/statab/2012/tables/12s0657.xls](http://www.census.gov/compendia/statab/2012/tables/12s0657.xls)
- Butikofer, R. E. (1986). *Safety digest of lessons learned*. ( No. 758). Washington DC: American Petroleum Institute.
- Carmines, E., & Zeller, R. (1979). *Reliability and Validity Assessment*. Thousands Oaks, CA: Sage Publications.
- Celik, M., & Cebi, S. (2009). Analytical HFACS for investigating human errors in shipping accidents. *Accident analysis & Prevention*, 41(1), 66-75.
- De Landre, J., Gibb, G., & Walters, N. (2006). Using incident investigation tools proactively for incident prevention. Paper presented at the *Meeting of the Australian and New Zealand Society of Air Safety Investigators*.
- Dekker, S. W. A. (2001). The disembodiment of data in the analysis of human factors accidents. *Human Factors and Aerospace Safety*, 1(1)
- Dept. of Defense. (2005). *A mishap investigation and data analysis tool*. (No. 16). Dept. of Defense.
- Edwards, E. (1973). Man and machine- systems for safety (man machine systems for flight safety, studying accidents, human factors in system design and implementation of personnel). *Outlook on Safety*, 21-36.
- ElBardissi, A. W., Wiegmann, D. A., Dearani, J. A., Daly, R. C., & Sundt, T. M. (2007). Application of the human factors analysis and classification system methodology to the cardiovascular surgery operating room. *Annals of Thoracic Surgery*, 83(4), 1412-1419.

- Federal Aviation Administration. (2010). *Weather-related aviation accident study 2003–2007*. Washington, DC: Federal Aviation Administration.
- Fleishman, E. A., Quaintance, M. K., & Broedling, L. A. (1984). *Taxonomies of human performance: The description of human tasks* Academic Press Orlando, FL.
- Fleiss, J. L. (1981). *Statistical methods for ratios and proportions* (2nd ed.). New York: John Wiley & Sons, Inc.
- Gaur, D. (2005). Human factors analysis and classification system applied to civil aircraft accidents in india. *Aviation, Space, and Environmental Medicine*, 76(5), 501-505.
- Gibb, G., Hayward, B., & Lowe, A. (2001). Applying reason to safety investigation: BHP Billiton's ICAM. *Sustaining Safety in the New Millennium*, 159.
- Graham, M., Milanowski, A., & Miller, J. (2012). *Measuring and Promoting Inter-Rater Agreement of Teacher and Principal Performance Ratings*, Center for Educator Compensation Reform
- Gwet, Kilem. (2010). *Handbook of Inter-rater Reliability*, (2) Advanced Analytics, LLC
- Hale, A., Walker, D., Walters, N., & Bolt, H. (2012). Developing the understanding of underlying causes of construction fatal accidents. *Safety Science*, 50(10), 2020-2027.
- 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.
- Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1), 77-89.
- Heimann, C. F. L. (1993). Understanding the challenger disaster: Organizational structure and the design of reliable systems. *American Political Science Review*, , 421-435.
- Helmreich, R. L. (2000). On error management: Lessons from aviation. *BMJ: British Medical Journal*, 320(7237), 781.
- Inglis, M. S. J., & McRandle, B. (2007). *Human Factors Analysis of Australian Aviation Accidents and Comparison with the United States*, Canberra ACT, Australia: Australian Transport Safety Bureau

- Johnson, W. (1997). *Classifying pilot human factor causes in A-10 class A mishaps*. Daytona Beach, Florida: Unpublished graduate research project, Embry Riddle Aeronautical University.
- Kirwan, B. (1998). Human error identification techniques for risk assessment of high risk Systems-Part 1: review and evaluation of techniques. *Applied Ergonomics*, 29(3), 157-177.
- Kohn, L. T., Corrigan, J., & Donaldson, M. S. (2000). *To err is human: Building a safer health system* Joseph Henry Press.
- Kolbe, R. H., & Burnett, M. S. (1991). Content-analysis research: An examination of applications with directives for improving research reliability and objectivity. *Journal of Consumer Research*, 18(2), 243-250. Retrieved from <http://www.jstor.org/stable/2489559>
- Krippendorff, K. (2013). *Personal contact*.
- Krippendorff, K. (2006). *Content analysis: An introduction to its methodology* Sage Publications, Incorporated.
- Krulak, D. C. (2004). Human factors in maintenance: Impact on aircraft mishap frequency and severity. *Aviation, Space, and Environmental Medicine*, 75(5), 429-432.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, , 159-174.
- Lenne, M. G., Ashby, K., & Fitzharris, M. (2008). Analysis of general aviation crashes in australia using the human factors analysis and classification system. *The International Journal of Aviation Psychology*, 18(4), 340-352.
- Lenne, M. G., Salmon, P. M., Liu, C. C., & Trotter, M. (2012). A systems approach to accident causation in mining: An application of the HFACS method. *Accident Analysis and Prevention*, 48, 111-117.
- Li, W. C., & Harris, D. (2005). HFACS analysis of ROC air force aviation accidents: Reliability analysis and cross-cultural comparison. *International Journal of Applied Aviation Studies*, 5(1), 65-81.

- Li, W. C., & Harris, D. (2006). Pilot error and its relationship with higher organizational levels: HFACS analysis of 523 accidents. *Aviation, Space, and Environmental Medicine*, 77(10), 1056-1061.
- Li, W. C., Harris, D., & Yu, C. S. (2008). Routes to failure: Analysis of 41 civil aviation accidents from the republic of china using the human factors analysis and classification system. *Accident Analysis and Prevention*, 40(2), 426-434.
- Liberty Mutual. (2011). *Workplace safety index*. (No.1). Hopkinton, MA: Liberty Mutual.
- Luxhøj, J. T. (2003). Probabilistic causal analysis for system safety risk assessments in commercial air transport. Workshop on investigating and reporting of incidents and accidents (IRIA), Williamsburg, VA
- Luxhøj, J. T., & Kauffeld, K. (2003). Evaluating the effect of technology insertion into the national airspace system. *The Rutgers Scholar*, 5
- Mossink, J., & de Greef, M. (2002). *Inventory of socioeconomic costs of work accidents* Office for Official Publications of the European Communities.
- National Safety Council. (2011). *Injury facts*. IL: Itasca.
- Nelson, J. C., & Pepe, M. S. (2000). Statistical description of interrater variability in ordinal ratings. *Statistical Methods in Medical Research*, 9(5), 475-496.
- NTSB. (2011). *Sinking of commercial fishing vessel katmai*. ( No. DCA-09-CM-001). Washington, DC: NTSB.
- Occupational Safety and Health Administration. (2006). *Safety and health management systems*. [http://www.osha.gov/SLTC/etools/safetyhealth/mod1\\_costs.html](http://www.osha.gov/SLTC/etools/safetyhealth/mod1_costs.html)
- O'Connor, P. (2008). HFACS with an additional layer of granularity: Validity and utility in accident analysis. *Aviation, Space, and Environmental Medicine*, 79(6), 599-606.
- O'Connor, P., Walliser, J., & Philips, E. (2010). Evaluation of a human factors analysis and classification system used by trained raters. *Aviation, Space, and Environmental Medicine*, 81(10), 957-960.

- O'hare, D. (2000). The 'Wheel of misfortune': A taxonomic approach to human factors in accident investigation and analysis in aviation and other complex systems. *Ergonomics*, 43(12), 2001-2019.
- Olsen, N. S. (2011). Coding ATC incident data using HFACS: Inter-coder consensus. *Safety Science*, 49(10), 1365-1370.
- Olsen, N. S., & Shorrock, S. T. (2010). Evaluation of the HFACS-ADF safety classification system: Inter-coder consensus and intra-coder consistency. *Accident Analysis & Prevention*, 42(2), 437-444.
- Patterson, J. M., & Shappell, S. A. (2010). Operator error and system deficiencies: Analysis of 508 mining incidents and accidents from Queensland, Australia using HFACS. *Accident Analysis & Prevention*, 42(4), 1379-1385.
- Patterson, J. M. (2009). *Human error in mining a multivariable analysis of mining accidents/incidents in queensland, australia and the united states of america using the human factors analysis and classification system framework* Retrieved from <http://etd.lib.clemson.edu/documents/1263397320/>
- Perneger, T. V. (2005). The swiss cheese model of safety incidents: Are there holes in the metaphor? *BMC Health Services Research*, 5.
- Peters, G. A., & Peters, B. J. (2006). *Human error : Causes and control*. Boca Raton, FL: CRC/Taylor & Francis.
- Plourde, G. (1997). *Human factor causes in fighter-bomber mishaps: A validation of the taxonomy of unsafe operations*. Daytona Beach, Florida: Unpublished graduate research project, Embry Riddle Aeronautical University.
- Popping, R. (1988). On agreement indices for nominal data. In W. E. Saris & I. N. Gallhofer (Eds.), *Sociometric research: Data collection and scaling* (pp. 90–105). Cambridge, MA: MIT Press
- Portaluri, M., Fucilli, F. I. M., Gianicolo, E. A. L., Tramacere, F., Francavilla, M. C., De Tommaso, C., . . . Pili, G. (2010). Collection and evaluation of incidents in a radiotherapy department A reactive risk analysis. *Strahlentherapie Und Onkologie*, 186(12), 693-699.

- Rabbe, L. (1996). *Categorizing air-force F-16 mishaps using the taxonomy of unsafe operations*. Daytona Beach, Florida: Unpublished graduate research project, Embry Riddle Aeronautical University.
- Ranger, K. (1997). *Inter-rater reliability of the taxonomy of unsafe operations*. Daytona Beach, Florida: Unpublished graduate research project, Embry Riddle Aeronautical University.
- Rashid, H., Place, C., & Braithwaite, G. (2010). Helicopter maintenance error analysis: Beyond the third order of the HFACS-ME. *International Journal of Industrial Ergonomics*, 40(6), 636-647.
- Rasmussen, J. (1982). Human errors: Taxonomy for describing human malfunction in industrial installations. *Journal of Occupational Accidents*, 4(2), 311-333.
- Reason, James. (1995). Understanding adverse events: Human factors. *Quality in Health Care*, 4(2), 80-89.
- Reason, James. (2000). Human error: Models and management. *Bmj*, 320(7237), 768-770.
- Reason, James. (1997). *Managing the risks of organizational accidents* Ashgate Aldershot.
- Reason, James. (1990). *Human error* New York: Cambridge University Press.
- Reason, James. (2008). *The human contribution: Unsafe acts, accidents and heroic recoveries* Ashgate Publishing.
- Reinach, S., & Viale, A. (2006). Application of a human error framework to conduct train accident/incident investigations. *Accident Analysis and Prevention* 38(2), 396-406.
- Ross, A., Wallace, B. & Davies J. (2004). Technical note: measurement issues in taxonomic reliability. *Safety Science* 42, 771-778.
- Salge, M., & Milling, P. M. (2006). Who is to blame, the operator or the designer? two stages of human failure in the chernobyl accident. *System Dynamics Review*, 22(2), 89-112.

- Sanders, M. and E. McCormick (1993). *Human Factors in Engineering and Design*. New York, McGraw-Hill.
- Sarter, N. B., & Alexander, H. M. (2000). Error types and related error detection mechanisms in the aviation domain: An analysis of aviation safety reporting system incident reports. *The International Journal of Aviation Psychology*, 10(2), 189-206.
- Scarborough, A., & Pounds, J. (2001). Retrospective human factors analysis of ATC operational errors. *Focusing Attention on Aviation Safety*,
- Schröder-Hinrichs, J. U., Baldauf, M., & Ghirxi, K. T. (2011). Accident investigation reporting deficiencies related to organizational factors in machinery space fires and explosions. *Accident Analysis & Prevention*, 43(3), 1187-1196.
- Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*,
- Senders, J. W., & Moray, N. P. (1991). Human error: Cause, prediction, and reduction. Paper presented at the *The Chapters in this Volume are Drawn Chiefly from Papers Presented at the Second Conference on the Nature and Source of Human Error, 1983, Held in Bellagio, Italy*.
- Shappell, S. A., & Wiegmann, D. A. (2000). *The human factors analysis and classification system--HFACS*. Springfield, Va: US Federal Aviation Administration, Office of Aviation Medicine.
- Shappell, S., Detwiler, C., Holcomb, K., Hackworth, C., Boquet, A., & Wiegmann, D. A. (2007). Human error and commercial aviation accidents: An analysis using the human factors analysis and classification system. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 49(2), 227-242.
- Shappell, S. A., & Wiegman, D. A. (2003). *A human error analysis of general aviation controlled flight into terrain accidents occurring between 1990-1998*. Washington, DC: Federal Aviation Administration.
- Shappell, S. A., & Wiegmann, D. A. (2001). Applying reason: The human factors analysis and classification system (HFACS). *Human Factors and Aerospace Safety*, 1(1), 59-86.



- Sharit, J. (2006). Human error. *Handbook of Human Factors and Ergonomics, Third Edition*, 708-760.
- Shohamy, E., Gordon, C., & Kraemer, R. (1992). The Effect of Rater's Background and Training on the Reliability of Direct Writing Tests. *The Modern Language Journal*, 76(1), 27-33.
- Shorrock, S. (2011). Aviation psychology and human factors. *Ergonomics*, 54(10), 983-984.
- Stokes, M. E., Davis, C. S., & Koch, G. G. (2000). *Categorical data analysis using the SAS system* Sas Inst.
- Ting, L. Y., & Dai, D. M. (2011). The identification of human errors leading to accidents for improving aviation safety. Paper presented at the *Intelligent Transportation Systems (ITSC), 2011 14th International IEEE Conference On*, 38-43.
- Toft, Y., Dell, G., Klockner, K., & Hutton, A. (2012). Models of causation: Safety. *Foundation science. in HaSPA (health and safety professionals alliance)*. Tullamarine, VIC: Safety Institute of Australia.
- Trucco, P., Cagno, E., Ruggeri, F., & Grande, O. (2008). A bayesian belief network modelling of organisational factors in risk analysis: A case study in maritime transportation. *Reliability Engineering & System Safety*, 93(6), 845-856.
- 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.
- Vach, W. (2005). The dependence of cohen's Kappa on the prevalence does not matter. *Journal of Clinical Epidemiology*, 58(7), 655-661.
- von Eye, A., & von Eye, M. (2008). On the marginal dependency of cohen's  $\kappa$ . *European Psychologist*, 13(4), 305.
- von Thaden, T. L., Wiegmann, D. A., & Shappell, S. A. (2006). Organizational factors in commercial aviation accidents. *The International Journal of Aviation Psychology*, 16(3), 239-261.

- Walker, S. (1996). *A human factors examination of U.S. naval controlled flight into terrain CFIT accidents*. Daytona Beach, Florida: Unpublished graduate research project, Embry Riddle Aeronautical University.
- Wallace, B., Ross, A., Davies, J., Wright, L., & White, M. (2002). The creation of a new minor event coding system. *Cognition, Technology & Work*, 4(1), 1-8.
- Wang, Y. F., Faghih Roohi, S., Hu, X. M., & Xie, M. (2011). Investigations of human and organizational factors in hazardous vapor accidents. *Journal of Hazardous Materials*, 191(1), 69-82.
- Warrens, M. (2010). A formal proof of a paradox associated with Cohen's Kappa . *Journal of Classification*, 27(3), 322-332.
- Wertheim, K. E. (2010). Human factors in large-scale biometric systems: A study of the human factors related to errors in semiautomatic fingerprint biometrics.4(2), 138-146.
- Wiegmann, D., Faaborg, T., Boquet, A., Detwiler, C., Holcomb, K., & Shappell, S. (2005). *Human error and general aviation accidents: A comprehensive, fine-grained analysis using HFACS*. Washington DC: Federal Aviation Administration.
- Wiegmann, D., Shappell, S., Cristina, F., & Pape, A. (2000). A human factors analysis of aviation accident data: An empirical evaluation of the HFACS framework. *Aviation, Space and Environmental Medicine*, 71, 328.
- Wiegmann, D. A., & Shappell, S. A. (1997). Human factors analysis of postaccident data: Applying theoretical taxonomies of human error. *International Journal of Aviation Psychology*, 7(1), 67-81.
- Wiegmann, D. A., & Shappell, S. A. (2001a). *A human error analysis of commercial aviation accidents using the human factors analysis and classification system (HFACS)*. (). Washington DC: Federal Aviation Administration.
- Wiegmann, D. A., & Shappell, S. A. (2001b). Human error analysis of commercial aviation accidents: Application of the human factors analysis and classification system (HFACS). *Aviation, Space, and Environmental Medicine*, 72(11), 1006-1016.

- Wiegmann, D. A., & Shappell, S. A. (2001c). Human error perspectives in aviation. *International Journal of Aviation Psychology*, 11(4), 341-357.
- Wiegmann, D. A., & Shappell, S. A. (2003). *A human error approach to aviation accident analysis: The human factors analysis and classification system* Ashgate Pub Limited.
- Woods, D. D., Dekker, S., Cook, R., Johannesen, L., & Sarter, N. (2010). *Behind human error* Ashgate Publishing Company.
- Woods, D. D., Johannesen, L. J., Cook, R. I., & Sarter, N. B. (1994). *Behind Human Error: Cognitive Systems, Computers and Hindsight*,
- Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin*, 103(3), 374.