USING TASK NETWORK MODELING TO PREDICT HUMAN ERROR

A Dissertation
Presented to
The Academic Faculty

by

Vlad Liviu Pop

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Psychology

Georgia Institute of Technology
December 2015

USING TASK NETWORK MODELING TO PREDICT HUMAN ERROR

Approved by:

Dr. Francis T. Durso, Advisor
School of Psychology
Georgia Institute of Technology

Dr. Karen Feigh
School of Aerospace Engineering
Georgia Institute of Technology

Dr. Rickey P. Thomas
School of Psychology
Georgia Institute of Technology

Dr. Rustin Meyer
Psychology
Georgia Institute of Technology

Dr. Bruce Walker
School of Psychology
Georgia Institute of Technology

Date Approved: August 31, 2015

ACKNOWLEDGEMENTS

I would like to thank my advisor, Frank Durso, for his guidance, teaching, and encouragement throughout my graduate career. I would also like to thank my committee members, Rick Thomas, Bruce Walker, Rustin Meyer, and Karen Feigh for their knowledge and advice through this dissertation. Lastly, I would like to thank my parents, my brother, my girlfriend, and my dog for supporting me and bearing with me through this dissertation.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

SUMMARY

Human error taxonomies have been implemented in numerous safety critical industries. These taxonomies have provided invaluable insight into understanding the underlying causes of human error; however, their utility for actually predicting future errors remains in question. A need has been identified for another approach to supplement what we can extrapolate from taxonomies and better predict human error. Task network modeling is a promising approach to human error prediction that had yet to be empirically evaluated. This study tested a task network modeling approach to predicting human error in the context of automotive assembly. The task network modeling architecture was expanded to include a set of predictors from the human error literature, and used to model part of an operational automotive assembly plant. This manuscript contains three studies. Study 1 tested separate task network models for two different target areas of an active automotive assembly line. Study 2 tested the validity of predictions made by the models from Study 1, both within and across samples. Study 3 tested predictions across both models on a larger sample of vehicles. The expanded architecture accounted for 21.9% to 36.5% of the variance in human error and identified 12 explanatory variables that significantly predicted the occurrence of human error. Model outputs were used to compute prediction equations that were tested using binary logistic regression and then cross-validated twice using both split-half and cross-sample validation. The predictors of Time Pressure, Visual Workload, Auditory Workload, Cognitive Workload, Psychomotor Workload, Task Frequency, Information Flow, Teamwork, and Equipment Feedback were significant predictors of human error in all three models that were tested. The variables of

Information Presentation and Task Dependency varied in significance across samples, but both were significant in two out of the three models. The variables of Shift and Hour into Shift were never significant in any of the three models. The variables that were greatly stable across studies were all related to the tasks being performed by each worker at each station. The variables related to the timing of errors, on the other hand, were never significant. The results indicate that an expanded task network architecture is a great tool for predicting the situations and circumstances in which human errors will occur, but not the timing of when they will occur. Nevertheless, task network modeling demonstrated to provide useful, valid, and accurate predictions of human error and should continue to be developed as an error prediction tool.

CHAPTER 1:

INTRODUCTION

Studies have reported that human error contributes between 70% and 90% of adverse events in the highly engineered world surrounding people (Department of Defense, 2005; Isaac, Shorrock, Kennedy, Kirwan, Andersen, & Bove, 2002; Shappell & Wiegmann, 1997; 2000; Stanton & Salmon, 2009). Human error is defined as something that has been done which was either: "*not intended by the actor; not desired by a set of rules or an external observer;* or *that led the task or system outside its acceptable limits"* (Senders & Moray, 1991; pg. 25). Numerous taxonomies have been developed to understand and manage the contributions of human error to safety in complex systems. Human error taxonomies have been implemented in a wide variety of safety critical industries including aviation, maritime, defense, mining, chemical process, nuclear power, electric power, and nuclear chemical reprocess.

The implementation of taxonomies has provided valuable insight into understanding the causal effects underlying human error (Leiden, Laughery, Keller, French, Warwick, & Wood, 2001). However, this insight has not demonstrated any reduction in human error. A recent study (Belland, Olsen, & Lawry, 2010) compared the number of aviation mishaps that occurred in the 10 years before the U.S. Navy implemented the HFACS (Shappell & Wiegmann, 2000) taxonomy, and the number that occurred in the 10 years after (1988-1998, and 1998-2008, respectively). The study found no significant decrease in aviation mishaps for the entire U.S. Navy fleet, even when accounting for differences in the number of flight hours. The same result was found in

several other cases. Studies have assessed the impact of the HFACS taxonomy by analyzing 1,020 commercial aviation accidents over a 13 year timeframe (Shappell, Detwiler, Holcomb, Hackworth, Boquet, & Wiegmann, 2007), 14,436 general aviation accidents over an 11 year timeframe (Wiegmann, Faaborg, Boquet, Detwiler, Halcomb, & Shappell, 2005), and 17,808 aviation accidents over a 13 year timeframe (Detwiler, Hackworth, Holcomb, Boquet, Pfleiderer, Wiegmann, & Shappell, 2006). All of these studies found that despite implementation of the HFACS taxonomy, both the percentage and rate of human errors that led to aviation accidents remained constant. These findings were confirmed by another study that concluded that the percentage of accidents associated with human error had not changed over the 15 years after HFACS was implemented (Shappell & Wiegmann, 2009). The lack of error reduction has been described as surprising and disconcerting, especially given the great deal of resources invested by the Federal Aviation Administration and the aviation industry to specifically target the causes of human errors (Wiegmann & Shappell, 2001).

A review for NASA's System-Wide Accident Prevention Program identified the need for another approach to supplement what we can extrapolate from taxonomies and better predict human error (Leiden et al., 2001). Human performance modeling has been used to explain and predict human behavior in a variety of domains.

1.1 Human Performance Modeling

Human performance modeling approaches can be classified into three general architecture types: cognitive, vision, and task network (Leiden et al., 2001). Cognitive architectures model human performance using representations of mechanisms that underlie human cognition. Vision architectures model human performance in visual

processing using computational algorithms. Task network architectures model human performance using a reductionist, top-down approach to human behavior. Human behavior within a complex system is successively decomposed into smaller elements until human-system interaction can be described as a closed-loop function. The individual elements of human behavior are then connected to tasks from a task analysis and organized according to task sequence (Leiden et al., 2001).

The cognitive, vision, and task network architectures of human performance modeling have been compared in the scientific literature. A literature review of human performance models found that the three approaches had different strengths and capabilities, but overall, the breakdown of complex human performance gave the task network approach a distinct advantage over cognitive and vision approaches to modeling human performance (Leiden et al., 2001). Thus, task network modeling is further discussed as a possible approach to human error prediction.

1.2 Task Network Modeling

Task network modeling (Laughery & Corker, 1997) has been used in a variety of complex systems and provided useful, efficient, and valid input for task scheduling and planning. Examples include, using task networking modeling to increase efficiency, reduce wait times, and reduce costs around patient flows in healthcare (Barnes & Quiason, 1997), determine crew size and shift schedules for controllers of U.S. Army Unmanned (sic) Aerial Vehicles (Walters, French, & Barnes, 2000), design tasks in Automated Teller Systems (ATMs; Laughery, 1998), develop flight attendant training for evacuation of transport category aircraft (Peacock, Savage, & Waldock, 2009), optimize

supply chains in vehicle manufacturing (Schunk & Plott, 2000), and allocate tasks in next generation U.S. Navy destroyers (Wetteland, Miller, French, O'Brien, & Spooner, 2000).

The task network modeling architecture is also particularly well suited to be used for prediction of human error. Task network models are used to break down complex human performance into smaller elements of behavior and organize them by human and system task sequences. This type of breakdown of human system behavior matches the current systems view of human error. In other words, human error is not just a matter of the actions of the operator, but connected to features of the tools, tasks, and operating environment. From a human factors perspective, human error results from a mismatch between human capabilities and task demands. Task network modeling describes human interaction with the system in a step by step, closed loop manner that is useful for identifying when these types of mismatches occur and errors are likely.

Task network modeling seemed to be a promising approach to human error prediction that had not been empirically evaluated. This study tested a task network modeling approach to predicting human error in automotive assembly. The automotive assembly context was conducive to testing this type of approach for several reasons. First and foremost, task network models depend on accurate and complete task analyses. Automotive assembly processes are heavily documented and include the full range of tasks that could be performed, greatly supporting the task analysis process. Task network models also require accurate timing information for each task identified in the task analysis. Automotive assembly tasks already have detailed timing data available from time-motion studies with a granularity down to the time it takes to restock each individual screw from its corresponding parts bin. Lastly, modeling human error requires large

amounts of error data. Quality control checks in automotive assembly are widespread and highly documented, providing abundant data about human errors.

1.3 Expanding Modeling Architecture

The task network modeling architecture already includes parameters for predicting errors that occur from lack of time, simultaneous scheduling of tasks, lack of resources, or miss-coordination with other workers or automation. The current study expanded the task network modeling architecture with variables from the scientific literature to also be able to predict human error. Predicting human error requires understanding the systematic connections between errors and contextual factors (Dekker, 2002). Sharit (2006) developed a modeling framework for understanding human error that encompasses human factors, cognitive engineering, and sociotechinical perspectives in the literature. Sharit's modeling framework, illustrated in Figure 1, demonstrates how human error arises from an interplay between fundamental human limitations and contextual factors. The contextual component of Sharit's model contains situational variables that can effect the occurrence of errors. These variables were used to develop a set of predictors that were added to the task network modeling architecture in an effort to predict human error. The added predictors and hypothesized effects are discussed in the subsequent section.

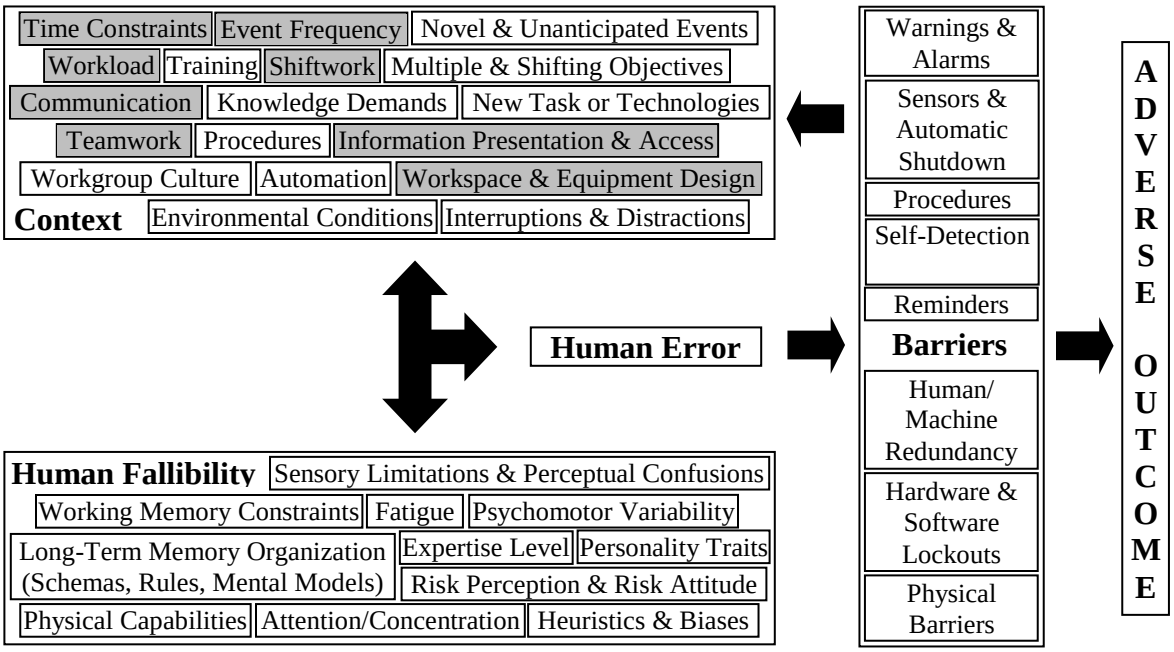

Figure 1. Framework for understanding human error. Situational variables that were used to develop the predictors in the current study are shaded in gray. Overall framework was recreated from Sharit (2006).

1.4 Proposed Predictors

The proposed predictors were based on situational variables from Sharit's (2006) model. The specific variables were chosen using findings from the literature that indicated that these variables may be useful predictors of human error. The variables included in the models were selected based on their applicability to the automotive assembly context. For example, time constraints related to the amount of time each vehicle is in a workstation may be a great indicator of human error, but training may not be since all associates receive the same amount and type of training before being allowed to work on the assembly line. The proposed predictors, supporting literature, and related hypotheses are discussed in separate subsections.

1.4.1 Time Constraints

Time constraints are a major influence on human performance in complex systems. In the automotive assembly context, time constraints can determine whether performance is deemed successful or not. The pace of the assembly line is automated and each vehicle enters a workstation for a set amount of time. Workers must perform all required tasks within that time, otherwise the vehicle leaves the workstation and all uncompleted tasks are considered errors. The amount of time each vehicle spends in a workstation is constant; however, the tasks that need to be performed and the time they require varies depending on vehicle model and options. For example, all vehicles enter a workstation for the same amount of time; but some require the installation of two sets of seatbelts and some require three sets for an extra third row seating option.

The impact of time constraints in automotive assembly can be assessed by dividing the amount of time required to complete the necessary tasks by the amount of time available; in other words, computing the proportion of time that is utilized. The proportion of time utilized may serve as an indicator of the level of time stress different combinations of options impose on workers. The effect of stress on human performance has been described in the literature (Swain & Guttman, 1983). Low levels of stress can lead to reduced attention, decreased job involvement, and poor performance. High levels of stress can lead to overload, freezing, and errors. The relationship between stress and performance resembles the ∩-shaped Yerkes-Dodson curve (Yerkes & Dodson, 1908) and has been evidenced by measures of the stress hormone glucorticoid (Lupien, Maheu, Tu, Fiocco, & Schramek, 2007). The pattern between stress and performance in the

literature suggests that time pressure may be able to predict human error in the current database. The following effect of time pressure on human error was hypothesized:

Hypothesis 1: There is an optimum percentage of time utilization, such that the probability of human error is higher when the percentage of time utilized is either lower or higher.

1.4.2 Workload

Workload is an important contextual component that must be understood in order to predict human error. The impact of workload on human performance is typically depicted as a ∩-shaped function (Wickens, 1981), similar to stress in the previous section. Low levels of workload lead to decreases in arousal, situation awareness, and performance. High levels of workload lead to overload, missed signals, and task shedding. The mental workload levels of automotive assembly tasks have not been evaluated; however, anecdotal accounts from workers suggest that workload varies from station to station. The effect of workload was hypothesized as follows:

Hypothesis 2: There is an optimum level of workload, such that the probability of human error is higher when workload is either low or high.

1.4.3 Task Frequency

Task frequency can influence human performance in two ways. Tasks that are performed frequently and repetitively become boring and mindless (Rasmussen, 1982). On the other hand, tasks that are rare and infrequent become 'black swans' and may fail to be noticed, be misdiagnosed, and inappropriately executed (Wickens, Hooey, Gore, Sebok, Koenecke, & Salud, 2009). The frequency of tasks in automotive assembly widely varies. Some parts are installed on every vehicle on the assembly line. Some rare options

are only installed once every six months on average. The following effect of task frequency was hypothesized:

Hypothesis 3: There is an optimum frequency of tasks, such that the probability of human error is higher when tasks are performed either repeatedly or infrequently.

1.4.4 Shiftwork

Automotive assembly runs on a 24 hour schedule. Numerous studies have investigated the relationship between work hours and human performance. These studies have consistently found that human performance is majorly impaired during night shifts (see Folkard & Monk, 1980). The temporal distribution of human error accidents, catastrophes, and disasters has also been investigated. A major peak has been found in the number of errors that occur between 2 a.m. and 4 a.m., and a secondary more minor peak between 2 p.m. and 4 p.m. (Mitler, Carskadon, Czeisler, Dement, Dinges, & Graeber, 1988). Based on these findings, the following effects were hypothesized:

Hypothesis 4: Overall, the probability of human error is higher during night shift than day shift.

Hypothesis 5: There is a major peak of time in each shift during which the probability of human error is highest.

1.4.5 Information Flow

Human performance can be influenced by the flow of information needed to perform a task. The flow of information indicating the correct course of action can be explicit or implicit. Explicit information flows directly specify the correct course of action. Implicit information flows imply the correct course of action. For example, the number of seatbelts that need to be installed on a vehicle can be explicitly displayed on a

screen (e.g., 4 or 6), or implicitly displayed by the presence of a third-row seating option on a vehicle build sheet. Implicit information requires mental processing to interpret data and determine the correct course of action. The processing demands of implicit information result in increased response times and higher likelihoods of errors (Seminara, Gonzalez, & Parsons, 1976; Welford, 1976). Thus, the following was hypothesized:

Hypothesis 6: The probability of human error is higher when the flow of information is implicit rather than explicit.

1.4.6 Information Presentation

The way information is presented can affect human performance. The modality of information presentation influences the perceptual demands of tasks. For example, requiring workers to read vehicle information for tasks that require inspecting, checking, locating, or aligning can result in high visual demands that may degrade performance. The modality of information presentation also influences the likelihood of errors. For example, Human Reliability Assessments have found that if information is given auditorily, it is very unlikely that operators will fail to perform at least the first action that is required (Kryter, 1972). The following effect of information presentation was hypothesized:

Hypothesis 7: The probability of human error is higher when information is only presented visually rather than visually and auditorily.

1.4.7 Task Dependency

Tasks that require coordination with other workers have special implications for human performance. Some assembly tasks require workers to communicate with other workers to perform certain actions concurrently. Performance on these types of tasks is

highly interdependent, such that correct performance by one of the workers increases the probably of correct performance by the other worker, and incorrect performance by one of the workers increases the probability of incorrect performance by the other worker (Nuclear Regulatory Commission, 1975). This interdependence can be especially problematic if one of the workers fails to read the information required and correct performance depends on a single worker. Based on these implications, task dependency was hypothesized to have the following effect of human error:

Hypothesis 8: The probability of human error is higher when tasks are interdependent rather than independent.

1.4.8 Teamwork

Teamwork is an important contextual factor to consider in predicting human error. Teamwork puts workers in a position to observe each others' work and detect errors. It also provides an opportunity for workers to help each other recover from errors and catch up when they get behind. In this applied automotive context, workstations contain between 1 and 5 workers depending on the number of tasks that need to be performed. The effect of workers on human error was hypothesized as follows:

Hypothesis 9: The probability of human error decreases as the number of workers per station increases.

Hypothesis 9 and Hypothesis 8 are related. Both hypotheses propose that human error probabilities are not independent and can vary depending on performance of other system tasks. Hypothesis 8 and Hypothesis 9 both assess the relationship between different workers and human error; however, they do so at different levels of task dependence. Task dependence can be broken down into two types: direct and indirect

(Bell & Swain, 1983). Hypothesis 8 applies to tasks where workers directly depend on each other to perform the task and directly influence the performance of other workers. Hypothesis 9 applies to team tasks where workers perform separate tasks at the same station and may indirectly influence the performance of other workers. Both hypotheses are important for understanding the relationship between task dependence and human error and increasing model accuracy using joint probabilities.

1.4.9 Equipment Design

Most of the equipment on the assembly line was designed to reduce human motor requirements; however, this equipment also influences other aspects of human performance by providing key feedback information. For example, automated fastening tools provide workers with immediate feedback about whether a task was performed correctly, or where and how it failed. Immediate feedback, such as this, fosters learning that leads to reduction of human error (Reason, 1990; Senders & Moray, 1991). Thus, the following effect of feedback on human error was hypothesized:

> *Hypothesis 10: The probability of human error is lower at workstations with equipment that provides immediate performance feedback, than at stations without such equipment.*

1.5 Human Error Prediction

The current studies tested an expanded task network modeling approach to human error prediction in automotive assembly. The overall process is summarized in Figure 2. As Figure 2 shows, the process began with a task analysis of automotive assembly. After all tasks were identified, timing information and workload information were added for each task. This information was collectively used to construct a task network model for a

section of the automotive assembly line. Data for vehicles with human error were then entered into the task network model. The output of the model was the value of each proposed predictor and the error outcome, for every vehicle, at every workstation in that section of the assembly line. These output data were then entered into a logistic regression prediction equation and used to make predictions about human error. The following was hypothesized:

Hypothesis 11: Prediction equations constructed from the outputs of expanded task network models account for a significant percent of variance in human error.

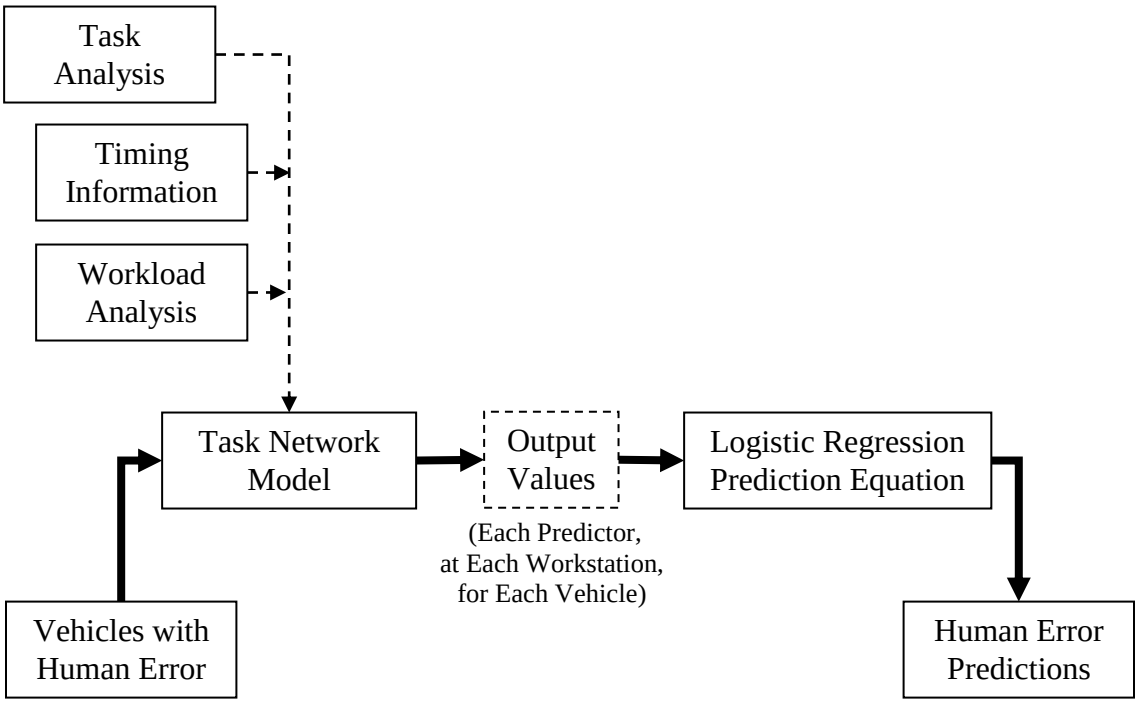

Figure 2. The overall process used for human error prediction in the current studies.

This manuscript contains three studies. Study 1 tested separate task network models for two different target areas of an active automotive assembly plant. Study 2

tested the validity of predictions made by the models from Study 1, both within and across samples. Study 3 tested predictions across both models on a larger sample of vehicles. Although the three studies were conducted chronologically, the methods, results, and discussions are presented concurrently for ease of comparison. The methods for all three studies are discussed in Chapter 2. The results for each prediction equation and model fit are presented and discussed in Chapter 3. The results and influence of individual predictor variables are presented and discussed in Chapter 4. Concluding thoughts across all equations, predictors, and studies are discussed in Chapter 5.

CHAPTER 2:

METHODS

2.1 Study 1: Models A & B

2.1.1 Participants

Participants were vehicles with human error from two target areas of an operational automotive assembly plant. Target area A contained 36 workstations where the vehicle wiring harness and key safety features were installed, including airbags and seatbelts. Target area A was selected because preventing human error has important implications for the safety of the vehicle driver and passengers. Target area B contained 27 workstations where the interior cabin components and electronics occupants interact with were installed. Target area B was selected because preventing human error has important implications for the satisfaction of vehicle users.

Observations for each vehicle were made at each workstation within the target area. Although each vehicle went through all stations, the observations made at each station were independent. In other words, the tasks performed at each workstation were different, did not repeat, and the outcome did not depend on other workstations. Nevertheless, both models stills controlled for any variance explained by the different stations and their order on the assembly line.

There was no direct interaction with workers. Experimental data were obtained by retrospectively reviewing quality control records. A power analysis indicated that 992 cases were necessary to be able to detect small effect sizes (ηp^2 = .010; Cohen, 1988)

with a power of 0.80 and alpha of 0.05. Thus, simple random sampling was used to select 1,000 vehicle quality control records from each of the two target areas. These records contained the outcome information for each workstation as each vehicle passed down the assembly line. Specifically, the records included: which workstation(s) made an error, while installing what part, for which task, on what vehicle build, and at what time.

The records originated from quality control gates on the assembly line. The original purpose of the records was to track errors so they can be repaired before vehicles leave the plant. As a result, there are no records for vehicles without errors. This is fairly typical in the human error analysis domain. For example: the National Transportation Safety Board aviation database, Federal Aviation Administration database, NASA Aviation Safety Reporting System, National Highway Traffic Administration database, Federal Railroad Administration database, and the European Maritime Safety Agency database all contains reports from accidents, incidents, or casualties. Such abundant and detailed data are rarely available for events that go without error or incident.

Although all records used in the current study were contingent on at least one workstation with an error, they also provided the opportunity to investigate the other workstations that did not make an error, in the same level of detail. For example target area A had 36 different workstations. As each of the 1,000 vehicles went down the assembly line from station to station, when one worker at one workstation made an error it also provided data for all the workers at the other 35 workstations that did not make an error. Looking ahead, this resulted in the model for area A containing 1,050[1] cases where an error occurred and 34,950 cases where no error occurred. The model for area B

[1] In Model A, 50 of the 1,000 vehicles contained errors at two stations.

contained 1,017[2] cases where an error occurred and 25,983 cases where no error occurred. The goal of the modeling was to predict which variables were present in situations in which errors occurred based on the available error data. Because of the inherent sampling bias in the data available, the frequencies reported above should not be considered representative of the total vehicle population.

2.1.2 Apparatus

The task network models were constructed using Micro Saint Sharp by Micro Analysis and Design, Inc. Micro Saint Sharp is a discrete event simulation tools that runs on the Windows platform. Task network models were built and organized using a point, click, and drag graphical interface. An example of a task network model in Micro Saint Sharp is illustrated in Figure 3. The simulation software has built-in features for modeling human performance constraints such as time, simultaneous tasks, and coordination with other workers. The simulation software is also very flexible, allowing custom objects, variables, functions, and algorithms to be designed and added to the model using C#. This flexibility made it possible to expand the task network modeling architecture to include the predictor variables discussed in the introduction. The exact method for how the architecture was expanded is described in section 2.1.5.4 in the Procedure section.

[2] In Model B, 15 of the 1,000 vehicles contained errors at two workstations stations and 1 contained errors at three workstations.

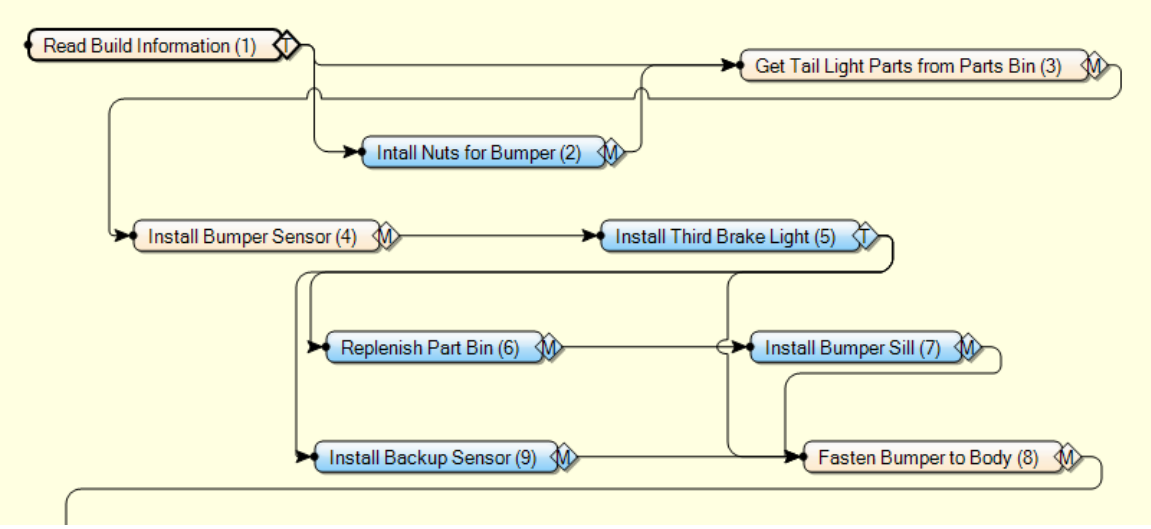

Figure 3. Example of a task network model in Micro Saint Sharp.

2.1.3 Measures

2.1.3.1 Time Pressure

The assembly line was constantly moving and each vehicle was in a workstation for a precise period of time. The amount of time a vehicle was in a workstation was the *available time* workers had to complete all necessary tasks. As previously mentioned, vehicles were highly customizable and the number of necessary tasks at a station could greatly depend on the options ordered on the vehicle. Detailed timing information was available for each task from time and motion studies. These timing data not only included the time required for workers to perform each task, but also the time required by the tools and machines used. For example, fastening the battery cable required 0.102 minutes of worker time and 0.100 minutes for the torque tool to complete the number of revolutions needed to achieve the specified torque. The *required time* was computed by adding the worker and machine times of all tasks that needed to be completed at a station given the set of options on the vehicle in the station. Time pressure was measured by the

percentage of time utilized. *Time Utilized* was computed by dividing the amount of time required by the amount of time available and then multiplying by 100.

- $Time\ Utilized = \frac{time\ required\ for\ task}{time\ available\ for\ task} \times 100$

For example, a worker had 3 minutes available to work on each vehicle that entered the station. A base model vehicle required 1.8 minutes of work, and a vehicle with several extra options required 2.9 minutes of work, resulting in time utilized percentages of 60% and 97%, respectively. Higher percentages of time utilized indicated higher time pressure.

2.1.3.2 Workload Demands

The demands that tasks placed on workers were assessed using a mental workload scale based on multiple resource theory: the Visual Auditory Cognitive Psychomotor (VACP) scale (McCracken & Aldrich, 1984). The VACP scale was developed by the US Army Research Laboratory to assess mental workload demands of different soldier and equipment configurations as military systems became increasingly more complex (Mitchell, 2000). The VACP scale has been used in a variety of systems and demonstrated to be useful for identifying peak levels of workload indicating tasks that should be reallocated, redesigned, or automated (Cain, 2007). For example, the VACP scale has been used to determine the allocation of functions on U.S. Navy destroyers (Archer & Lockett, 1997), the number of operators needed in Special Operations command stations (Malkin, Allender, Kelley, O'Brien, & Graybill, 1997), the crew size for the U.S. Army's automated artillery system (Beideman, Munro, & Allender, 2001),

and performance effects of the U.S. Army's Land Warrior integrated fighting system (Adkins, Murphy, Hemenway, Archer, & Bayless, 1996). Non military examples include, use by the Federal Aviation Administration to determine alternate function and crew allocations for air traffic controllers (Archer & Locke, 1997), and more recently, use by the U.S. Depart of Energy to increase prediction accuracy in human reliability analyses of nuclear power plants (Boring, 2006; Hugo & Gertman, 2012). Validation studies have found that VACP scale has good predictive validity (Mitchell, 2000), providing predictions that accounted for 77% of the variance between experimental conditions (Sarno & Wickens, 1995), has shown correlations with subjective workload ratings from participants, and has predicted the same performance differences observed in experimental data (Mitchell, 2000).

The VACP scale contains separate scales for each mental resource. Each resource was evaluated on a different 7 point interval scale with verbal anchors. The numerical values for each item originate from the original McCracken and Aldrich (1984) study. The full VACP scale is included in Table 1. Tasks were given separate visual, auditory, cognitive, and psychomotor ratings depending on the demands they placed on each component. For instance, if a task required operators to identify whether a system is ready by detecting a light, it was given a visual workload rating of 1.0, but if a task required operators to identify whether a system is ready by reading, it was given a visual workload rating of 5.9.

Table 1

The Visual Auditory Cognitive Psychomotor (VACP) Workload Scale.

Scale Value	Scale Descriptor
Visual	
0.0	No Visual Activity
1.0	Visually Register/Detect (detect occurrence of image)
3.7	Visually Discriminate (detect visual differences)
4.0	Visually Inspect/Check (discrete inspection/static condition)
5.0	Visually Locate/Align (selective orientation)
5.4	Visually Track/Follow (maintain orientation)
5.9	Visually Read (symbol)
7.0	Visually Scan/Search/Monitor (continuous/serial inspection, multiple conditions)
Auditory	
0.0	No Auditory Activity
1.0	Detect/Register Sound (detect occurrence of sound)
2.0	Orient to Sound (general orientation/attention)
4.2	Orient to Sound (selective orientation/attention)
4.3	Verify Auditory Feedback (detect occurrence of anticipated sound)
4.9	Interpret Semantic Content (speech)
6.6	Discriminate Sound Characteristics (detect auditory differences)
7.0	Interpret Sound Patterns (pulse rates, etc.)
Cognitive	
0.0	No Cognitive Activity
1.0	Automatic (simple association)
1.2	Alternative Selection
3.7	Sign/Signal Recognition
4.6	Evaluation/Judgment (consider single aspect)
5.3	Encoding/Decoding, Recall
6.8	Evaluation/Judgment (consider several aspects)
7.0	Estimation, Calculation, Conversion
Psychomotor	
0.0	No Psychomotor Activity
1.0	Speech
2.2	Discrete Actuation (button, toggle, trigger)
2.6	Continuous Adjustive (flight control, sensor control)
4.6	Manipulative
5.8	Discrete Adjustive (rotary, vertical thumbwheel, lever position)
6.5	Symbolic Production (writing)
7.0	Serial Discrete Manipulation (keyboard entries)

2.1.3.3 Task Frequency

The relative frequency of assembly tasks was computed using actual production data from the plant. Task frequency was computed as a ratio of the number of times a task was performed to the total number of vehicles produced, multiplied by 100. These task frequency computations resulted in empirical probability percentages for each task. For example, take the tasks of installing a sun roof and a clutch pedal. Production data from 10,000 vehicles indicated that 6,729 had a sunroof option and 2,738 had a manual transmission option, resulting in empirical probability percentage estimates of 67% and 27%, respectively.

- $Task\ Frequency = \frac{number\ of\ times\ task\ was\ performed}{total\ number\ of\ vehicles\ built} \times 100$

2.1.3.4 Shift

The shift during which errors occurred was included in the model as a predictor. The shift measure was binary, indicating whether the error occurred during the day shift or night shift.

2.1.3.5 Hours into Shift

The time at which errors occurred was measured as a function of the number of hours into the worker's shift. The assembly plant runs on day shifts and night shifts that start 12:00 hours apart. Interestingly, the error peaks identified in the literature occur after the same amount of time on task in each shift. As discussed in the introduction of Hypothesis 5, human errors peak between the hours of 2 a.m. to 4 a.m. and 2 p.m. to 4 p.m.. Both of these peaks fell exactly between 7.5 to 9.5 hours into each shift at the

assembly plant. Thus, the timing of errors was entered as a function of hours into shift to evaluate whether there would be corresponding peaks that could be used to predict the occurrence of human errors.

2.1.3.6 Information Flow

The information flow for each task was a binary measure of whether it was explicit or implicit. Workstations with visual or auditory displays that directly indicated the part that needed to be installed or the action that needed to be taken were marked explicit. For example, "install roof rails" or "fasten 6 fasteners." Workstations with visual or auditory displays that indicated the options on a vehicle, requiring workers to interpret the information and determine which part needed to be installed or action needed to be taken, were marked as implicit. For example, "option 429" or "option 3AF."

2.1.3.7 Information Presentation

At some workstations information is provided visually with screens or vehicle build sheets. At some workstations information is provided auditorily with directional speakers. Information presentation was a binary measure of whether necessary task information was presented visually or auditorily.

2.1.3.8 Task Dependency

Task dependence was a binary measure indicating whether a task required workers to directly work with other workers to complete the task. Tasks that required coordination were identified using the task analysis discussed in the subsequent procedure section. The task analysis specified which tasks required coordination. For example, "work with the right side associate and place the front seat in the car" or "work

with the left side associate to place the auxiliary harness cable insulation through the trunk wall."

2.1.3.9 Teamwork

Teamwork was a measure of the number of workers located at each station. Some stations contained a single worker whereas other stations contained up to five workers. The number of workers at a station was a function of manufacturing processes and remained constant across the study time.

2.1.3.10 Equipment Feedback

Equipment feedback was a binary measurement of whether the equipment at a workstation provided workers with performance feedback. Stations that contained automated tools, hand scanners, electrical tests, or automated checks that indicated whether a task was performed correctly or incorrectly were marked as feedback stations. Stations that did not contain any such equipment were marked as no feedback stations.

2.1.4 Design

The experiment was a between subjects polynomial regression design (Hill & Lewicki, 2007). Shift (day, night), Information Flow (implicit, explicit), Information Presentation (visual, auditory), Task Dependency (independent, interdependent), Teamwork (1-5 people), and Equipment Feedback (not provided, provided) were included as first-order effects. Time Pressure (0-100% percent of time utilized), Visual Workload (0.0-7.0 points), Auditory Workload (0.0-7.0 points), Cognitive Workload (0.0-7.0 points), Psychomotor Workload (0.0-7.0 points), and Task Frequency (0-100%

percent of vehicles) were included as second-order effects. Hours into Shift (0.0-12.0 hours into shift) was included as a third-order effect.

2.1.5 Procedure

2.1.5.1 Task Analysis

Task network modeling begins with understanding the system and tasks that will be simulated. A task analysis was performed on the two target areas of the assembly line. The task analysis was initially constructed using process sheets from the assembly plant. These processes sheets included detailed step-by-step information about the process and sequence of installing every single part at every single station on the assembly line. They also included detailed timing data for each task. A total of 2,103 process sheets were examined and broken down into 4,979 tasks; 2,950 tasks in target area A and 2,027 tasks in target area B. The task analysis was constructed in Microsoft Excel. Each task description, along with timing data, was recorded on a separate row. The tasks were organized according to the order of stations on the assembly line, and then the sequence of tasks within each station. The task analysis was then verified on both target areas of the assembly line to ensure that the tasks and sequences from the process sheets matched the work actually being performed by the workers at each station.

2.1.5.2 Workload Analysis

After the task analysis was completed and all tasks were identified, the next step was determining the demands that each task placed on workers. The tasks were evaluated using the VACP mental workload scale described in the Measurements section. Given the large number of 4,979 tasks, the scoring was split up among three independent raters. The

three raters were employees of the Georgia Tech Research Institute. The raters were not directly paid for participating but were compensated for their time by charging the hours they worked to a project account. All raters had a background in human factors. One rater had a doctorate in Engineering Psychology and the other two raters both had master's in Human Factors. Each rater was given an Excel version of the VACP scale and a practice workstation with 180 tasks to rate. Each rater scored the tasks independently. The three raters then met to discuss disagreements and attempt to reach consensus. Given the level measurement of the VACP scale, interscorer reliability for overlapped ratings was calculated using Pearson's *r*. The interscorer reliability for the practice workstation was $r = 0.67$. The three raters then independently rated the practice station again. Once again, the three raters met to discuss disagreements and attempt to reach consensus. Interscorer reliability for the second time rating the practice session was $r = 0.87$. The three raters were then given another practice station with 157 tasks. The three raters performed the ratings independently and then met to discuss disagreements and reach consensus. The interscorer reliability of the second practice station was $r = 0.74$. This marked the end of the practice. The three raters were each given 1,660 tasks from the task analysis to independently score. The raters were not given a time limit but were told to finish as soon as they could. The task ratings spanned a 2 month pay. The three raters overlapped on 1,052 ratings. Given the ratio level of measurement of the VACP scale, interscorer reliability for the overlapping tasks was calculated using Pearson's *r*. The correlation between scores from rater 1 and rater 2 was $r = 0.863$, rater 1 and rater 3 was $r = 0.808$, and rater 2 and rater 3 was $r = 0.806$. Overall, interscorer reliability was found to be acceptable because all correlations were greater than $r = .80$.

The ratings for each task were recorded in Microsoft Excel. The workload ratings from the three raters were compiled into the Excel worksheet that originated from the task analysis. This resulted in each task description being on a separate row, sequenced by station, and then the order performed at each station, with separate columns for the time duration, visual workload score, auditory workload score, cognitive workload score, and psychomotor workload score of every single task in the task analysis.

2.1.5.3 Other Data for Simulation

The existing Excel worksheet was then expanded to record additional data required for this type of modeling approach. Additional columns were added for task frequency, information flow, information presentation, task dependency, teamwork, and equipment feedback. Task frequency was computed using one month of production data from the assembly plant. The number of times each task was performed was identified and divided by the total number of vehicles produced to record frequency. The other columns were populated by visiting the assembly line itself. Each station was observed to determine and record if the flow of information was implicit or explicit, if the information was presented visually or auditorally, if the task depended on other workers or was independent, the number of works present at the station, and whether the equipment provided performance feedback.

2.1.5.4 Expanding the Task Network Modeling Architecture

After the data required to build the simulation were compiled, the next step was expanding the task network modeling architecture to be able to model the data. As previously mentioned, the task network modeling architecture in Micro Saint Sharp already includes parameters for identifying errors that occur from lack of time,

simultaneous scheduling of tasks, lack of resources, or miss-coordination with other workers or automation. In Micro Saint Sharp, this capability is achieved using five system variables that keep track of the state of the system. The five variables that are automatically created for each model and their function are as follows:

- *Clock* - records the time elapsed within the model
- *Distributions* - contains the distributions used to generate random values for the model (i.e., logistic)
- *Entity* - all the items that travel through a task network model
- *Model* - the functions of the model used to perform procedures, such as starting and stopping tasks
- *Animator* - used to keep track of image components to animate the execution of the task network model

To be able to predict human error, 14 additional variables were added to the architecture. Parameters for variable type and value range were then entered for each variable. The names and parameters of the variables that were added to the architecture are reported in Table 2. The detailed procedure for how these variables were entered at a keystroke level is also included in Appendix A.

Table 2

Variables added to the task network modeling architecture.

Variable Name	Type	For Values
Time_Utilization	Floating Point	0.00 - Continuous
Visual_Workload	Floating Point	0.00 - Continuous
Auditory_Workload	Floating Point	0.00 - Continuous
Cognitive_Workload	Floating Point	0.00 - Continuous
Psychomotor_Workload	Floating Point	0.00 - Continuous
Task_Frequency	Floating Point	0.00 - 1.00
Shift	Integer	Day, Night
Hours_into_Shift	Floating Point	0.00 - 12.00 hours
Information_Flow	Integer	Implicit, Explicit
Information_Presentation	Integer	Visual, Auditory
Task_Dependency	Integer	Yes, No
Teamwork	Integer	1 - 5 workers per station
Equipment_Feedback	Integer	Yes, No
Human_Error	Integer	Yes, No

2.1.5.5 Constructing the Task Network Models

The expanded Micro Saint Sharp architecture was used to construct separate task network models for each target area based on the data compiled in the Excel worksheet. The model for target area A was constructed first. The point and click graphical interface was used to turn each of the 2,950 tasks identified in the task analysis into separate tasks within the network. The tasks were organized by station and then connected according to the sequence of tasks in the assembly process. Parameters were entered and then used to represent how each task influenced the 14 variables being tracked in an effort to predict

human error. The specific code for each parameter is discussed at a keystroke level in Appendix A.

The task parameters outlined in this section were used to increment the 14 variables that were proposed as predictors of human error. These 14 variables were tracked throughout the simulation of each vehicle and their values were recorded for each worker at each station on the assembly line. These data were recorded by defining snapshots within the execution of the model to collect the value of each variable at the end of the last task being performed by each worker at each station. The end result was a tab delimited '.res' file for each snapshot that included the value of each of the 14 variables, for each worker, at each station, for each vehicle simulated.

After the model was compiled with all 2,950 tasks from target area A, it was checked for errors using the error checker built into Micro Saint Sharp. After all syntax errors were fixed, the model was run several times to check for logic errors. Logic errors were fixed using the line debugger tool in Micro Saint Sharp. After the model was error free, the values of the 14 variables exported by the model for each worker at each station were compared to manual calculations for the same vehicle. Computational errors were fixed and model accuracy was checked again. After the model was ready for execution, the entire process described in this section was repeated to construct a separate task network model for the 2,027 tasks from target area B.

2.1.5.6 Task Network Modeling Human Error Data

After the task network model for each target area was constructed, human error data were entered into the models. Each model only received error data from its corresponding target area to prevent contamination. Quality control records were used to

identify and randomly select 1,000 vehicles with human errors from each target area. The quality control data were highly detailed and included the build specifications for each vehicle with all options ordered, the exact error that occurred, the time the error occurred, the location on the assembly line, and the shift during which the error occurred. The error data were organized in an Excel worksheet. Each error was entered on a separate row with separate columns indicating the station in which the error occurred, all the options ordered on the vehicle, the time of day when the vehicle was built, and the shift during which the error occurred.

To enter the error data into the models, custom entity attributes were added. The parameters for each entity attribute were then entered. The names and parameters of the variables that were added to the architecture are reported in Table 3. The detailed procedure for how these variables were entered at a keystroke level is also included in Appendix A.

Table 3

Entity attributes added to the task network modeling architecture.

Entity Attribute Name	Type	For Values
Entity.Build_Information	String	All Build and Option Codes
Entity.Error _Location	Integer	Station 1 - Last Station
Entity.Build_Time	Floating Point	HH:MM Military Time
Entity.Error_Shift	Integer	1 Day, 2 Night

The error data were entered into each corresponding model by importing the Excel worksheet of errors and their Entity Attributes into Micro Saint Sharp in batches of 100 vehicles. Each batch of 100 vehicles was simulated through the task network and the snapshots built into each model exported the value of each of the 14 variables, for each

worker, at each station, for each vehicle simulated. After all 10 batches of 100 vehicles were simulated in the corresponding model, the tab delimited .res file from each snapshot was imported into Excel, organized, and compiled into a final output data file for each target area model. The end result was an Excel worksheet for each target area that contained the value of each of the 13 predictors proposed in the measurements section and whether or not a human error occurred, for each worker, at each station, for each of the 1,000 vehicles simulated. These Excel worksheets were then imported into SPSS 23.0 and used to compute logistic regression equations that are described in Chapter 3.

2.2 Study 2: Validation

The validity of predictive models A & B from Study 1 was assessed using double cross validation. Double cross-validation is a validation technique that involves splitting the original data sample in two, conducting separate analyses, and comparing the results to determine replicability (Reinhardt, 1992). Double cross-validation is more advantageous than single cross-validation because it is a more rigorous approach to validation and does not waste data (Reinhardt, 1992). The double cross-validation technique was observed throughout all phases of Study 1. The original data sample obtained from quality control records were split into two target areas from the very beginning. Separate task-network models were built for each target area and all variable measures were performed separately. Each model was only populated with data from its corresponding target area and all data were analyzed separately, including all predictors variables.

The double cross technique was used to assess two measures of cross-validation; split-half cross-validation and out of sample cross-validation. The double cross-validation

procedure used in this study is illustrated in Figure 4. As Figure 4 shows, the prediction models from each target area were first validated by randomly sub-sampling half of the respective dataset as training data and half as validation data. Each model was fit to the training data and then used to make predictions about the validation data. This procedure was repeated a total of 20 times for each model. The prediction models were then validated by using the full dataset from each target area to make predictions about the other target area. The procedure for both split-half and out of sample cross-validation is discussed in the subsequent method section. The results from both procedures are then discussed in section 3.2 in Chapter 3 according to the stability of model predictors and the accuracy and sensitivity of model predictions.

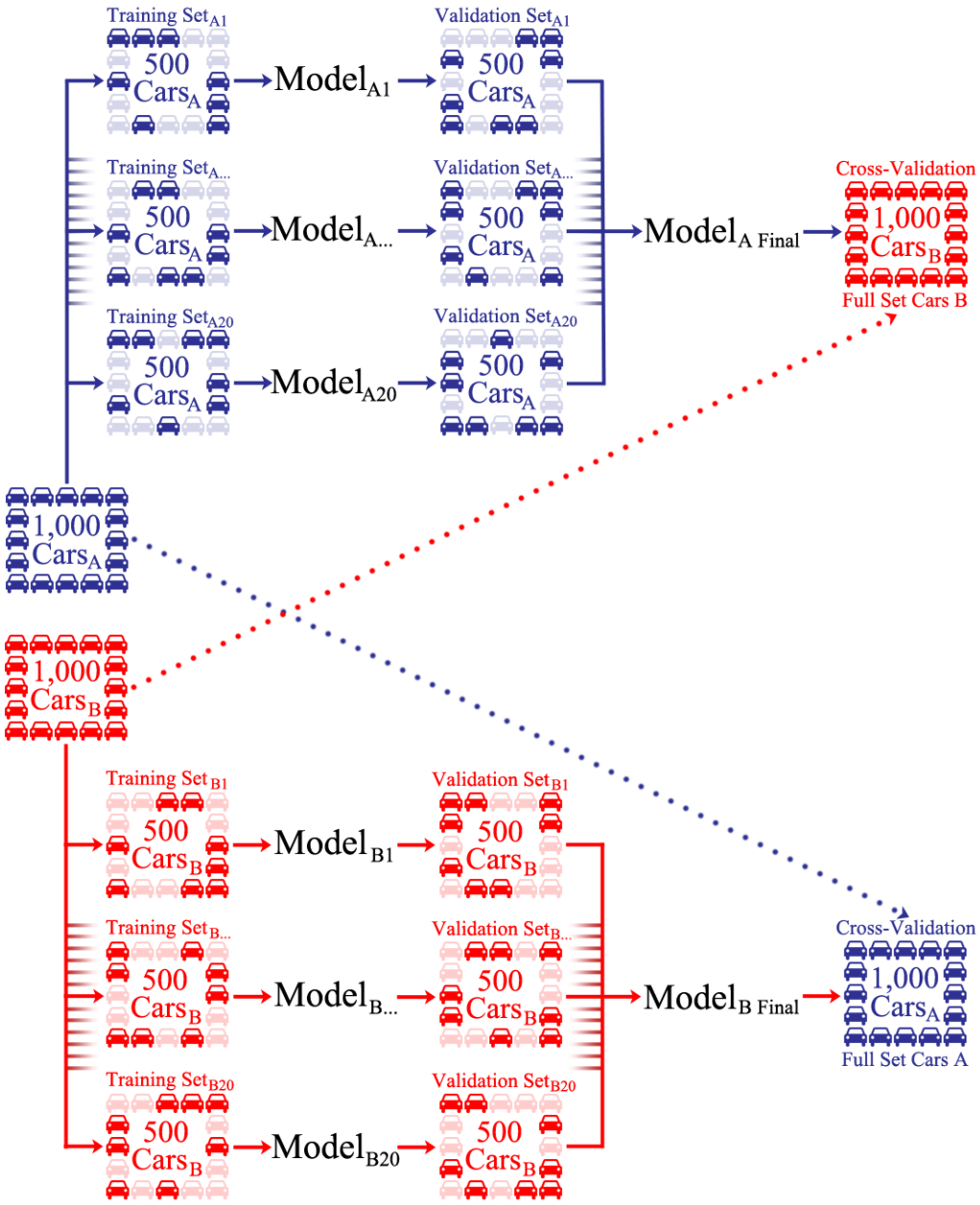

Figure 4. An illustration of the double cross-validation procedure used in this study.

2.2.1 Split-Half Validation

The prediction equations from each target area were first validated within sample. The sample of 1,000 vehicles with human errors from target area A was divided into two

halves, a training set and a testing set. This was done by randomly selecting 500 vehicles out of the dataset of 1,000 vehicles from target area A to designate as Training Set A_1. The remaining 500 cars in the dataset from target area A were designated as Validation Set A_1. Training Set A_1 was then used to compute regression coefficients for Model A_1 using maximum likelihood estimations. Model A_1 was then used to make human error predictions for the 500 vehicles in Validation Set A_1. These predictions were then compared to the actual human errors that occurred on the 500 vehicles in Validation Set A_1. All vehicles were then returned to the single sample set of 1,000 vehicles from target area A. This procedure was repeated a total of 20 times for Model A.

The same procedure was then used for within sample validation of Model B. The sample of 1,000 vehicles with human errors from target area B was divided in half for a training set and a testing set. This was done by randomly selecting 500 vehicles out of the dataset of 1,000 vehicles from target area B to designate as Training Set B_1. The remaining 500 cars in the dataset from target area B were designated as Validation Set B_1. Training Set B_1 was then used to compute regression coefficients for Model B_1 using maximum likelihood estimations. Model B_1 was then used to make human error predictions for the 500 vehicles in Validation Set B_1. These prediction were then compared to the actual human errors that occurred on the 500 vehicles in Validation Set B_1. Then once again, all vehicles were returned to the single sample set of 1,000 vehicles from target area B, and the procedure was repeated a total of 20 times.

2.2.2 Cross Sample Validation

The prediction equations from both models were also validated across samples. The original data set was split up by target area. The final prediction equation for target

area A was identified from the analysis summarized in section 3.1.1 and all regression coefficients were computed using maximum likelihood estimations. The final prediction equation from target area A was then used to predict the probability of human error in each of the 27 workstations in target area B using the values of predictors from each of the 1,000 vehicles from target area B. For example, for the first vehicle in target area B, the value of each predictor (i.e., the proportion of time utilized, visual workload, information presentation, information flow, and shift) was entered into the prediction equation from target area A and the probability of human error was computed for the first workstation. The same computation was then done using the value of each predictor in workstation 2, 3, ... 27. The process was then repeated for the 2nd, 3rd, ... 1,000th vehicle from target area B. In the end, the prediction equation from target area A was used to make 27,000 predictions about human error in target area B. These predictions were then compared to the actual human errors that occurred on the 1,000 vehicles from target area B.

The same process was then repeated for the prediction equation from Model B. The final prediction equation for target area B was identified from the analysis in section 3.1.2, and all regression coefficients were computed using maximum likelihood estimations. The final prediction equation from target area B was then used to predict the probability of human error in each of the 36 workstations in target area A using the values of predictors from each of the 1,000 vehicles from target area A. This resulted in the prediction equation from target area B being used to make 36,000 predictions about human error in target area A. These prediction were then compared to the actual human errors that occurred on the 1,000 vehicles from target area A.

Cross sample validation was conducted for two reasons. The first reason was to prevent statistical overfitting within the binary logistic regression models. The second reason was to observe which variables were more generalizable across target areas. In other words, to observe which variables were more universal indicators of the situations in which human errors occur, rather than just within the respective dataset. The ultimate goal was to identify which variables were stable across datasets and models to make human error intervention recommendations for the assembly plant.

The results from both the split-half and cross sample cross-validation procedures performed on both models are discussed in separate subsections in terms of the stability of model predictors in section 3.2.1 and the accuracy and sensitivity of model predictions in section 3.2.2.

2.3 Study 3: Application

After Models A and B were completed and analyzed, the task network architecture was tested on a larger sample of human errors. This larger application study was conducted for several reasons. First and foremost, the data from Models A and B had some variable data ranges that were never observed and thus could not be modeled. These missing ranges included both cases where the values observed did not span the low or high end of the score distribution as well as some cases where an entire range of values in the middle of the distribution were not observed in the data set. The situations in which missing variable ranges were observed in Models A and B are further discussed in Chapter 4.

The second reason for conducting the larger application study was to yet again observe the stability of predictor variables to be able to make recommendations for

human error reduction with more confidence. As will be discussed in section 3.2.1, the stability of predictors both Study 1 and Study 2 varied between the datasets from Model A and Model B. A larger dataset was needed to observe stability of variables across both target areas. This larger dataset also made it possible to assess how much each variable contributed to the actual occurrence of human error within the target areas in the assembly plant.

The final reason for conducting Study 3, and the main reason for calling it an application study, is because the results from this study were used to make recommendations for reducing human error in the two target areas of the automotive assembly plant from which the error data originated. The individual variable results from the application study are included in Chapter 4, paired with recommendations for whether changing the values of each variable within the stations in the plant would reduce human errors, and if so, how to change each variable. Although outside the scope of this manuscript, these recommendations were then implemented into the active assembly plant and as of this time the actual reduction in human error is being measured over a two month period.

The methods that were used for Study 3 are discussed in this chapter. The results of Study 3 are reported and discussed together with the results from Models A and B in Chapters 3 and 4, for ease of comparison.

2.2.1 Participants

Once again, participants were vehicles with human error from target areas A and B described in section 2.1.1. Experimental data were obtained by retrospectively reviewing quality control records from an active automotive assembly plant. For the

application study, a larger sample of 4,188 vehicles with human error was analyzed. This larger sample included the 1,000 vehicles from Model A and 1,000 vehicles from model B. The vehicles in this sample contained an average of $M = 1.149$ human errors with a standard deviation of $SD = 0.398$. The majority of vehicles, $n = 3{,}625$, contained only one error. Of the rest of the vehicles, $n = 509$ contained two errors, $n = 48$ contained three errors, and $n = 6$ contained four errors. A total of 4,811 human errors were included in the application study.

2.2.2 Apparatus

The task network models for the application study were also constructed using Micro Saint Sharp by Micro Analysis and Design, Inc. The expanded task network architecture described in sections 2.1.5.4 and 2.1.5.5 served as the basis of these models.

2.2.3 Measures

The measures included in the application study were the same measures included in the initial prediction study. The following variables were included and measured exactly as described in section 2.1.3 Measures: 2.1.3.1 Time Pressure, 2.1.3.2 Workload Demands, 2.1.3.3 Task Frequency, 2.1.3.4 Shift, and 2.1.3.5 Hours into Shift, 2.1.3.6 Information Flow, 2.1.3.7 Information Presentation, 2.1.3.8 Task Dependency, 2.1.3.9 Teamwork, 2.1.3.10 Equipment Feedback.

2.2.4 Design

The experiment was a between subjects polynomial regression design (Hill & Lewicki, 2007). Shift (day, night), Information Flow (implicit, explicit), Information Presentation (visual, auditory), Task Dependency (independent, interdependent),

Teamwork (1-5 people), and Equipment Feedback (not provided, provided) were included as first-order effects. Time Pressure (0-100% percent of time utilized), Visual Workload (0.0-7.0 points), Auditory Workload (0.0-7.0 points), Cognitive Workload (0.0-7.0 points), Psychomotor Workload (0.0-7.0 points), and Task Frequency (0-100% percent of vehicles) were included as second-order effects. Hours into Shift (0.0-12.0 hours into shift) was included as a third-order effect.

The design of the application study was the same as the design of Study 1 summarized in section 2.1.4, with one exception; the design of Study 3 also controlled for the different target areas and stations to ensure that the results would not be simply caused by these factors. This control was done by including the between subject control variables of Target Area (A, B), Station Number (1-36), and the interaction of Target Area × Station Number.

2.2.5 Procedure

The overall outline of the application procedure followed the initial prediction procedure in section 2.1.5. The portions for 2.1.5.1 Task Analysis, 2.1.5.2 Workload Analysis, 2.1.5.3 Other Data for Simulation, 2.1.5.4 Expanding the Task Network Modeling Architecture, and 2.1.5.5 Constructing the Task Network Models were the same in Study 3.

The difference was in section 2.1.5.6 Modeling Human Error Data; how human error data were entered into each model. In Study 3, each model received the full set of data of 4,188 vehicles with human error. The data were entered using the same entity attributes described in section 2.1.5.6 Modeling Human Error Data. Once again, the data were then imported from Excel into Micro Saint Sharp in batches. The snapshots build

into the models then exported the value of each variable, for each worker, at each station, for each vehicle into a tab delimited .res file. These files were compiled together in Excel and imported into SPSS 23.0 to compute the application prediction equation described in the subsequent section.

CHAPTER 3:

PREDICTION RESULTS

3.1 Study 1: Models A & B

Prediction equations were computed to predict the probability of human error using the variables described in section 2.1.3 Measures. Each equation controlled for different stations by entering Station Number into Block 1. The first first-order effects for all 13 predictors were then entered into Block 2: Time Pressure, Visual Workload, Auditory Workload, Cognitive Workload, Psychomotor Workload, Task Frequency, Shift, Hours into Shift, Information Flow, Information Presentation, Task Dependency, Teamwork, and Equipment Feedback. The second-order polynomials for Time Pressure, Visual Workload, Auditory Workload, Cognitive Workload, Psychomotor Workload, Task Frequency, and Hours into Shift were entered in Block 3 to test for quadratic effects. The third-order polynomial for Hours into Shift was entered into Block 4 to test for cubic effects. The complete equation, with all first, second, and third order effects included, contained 22 terms: the 21 predictors and 1 control variable. Human error was a dichotomous outcome variable; errors either occurred or they did not. Thus, the prediction equations were tested using binary logistic regression. The full logistic regression equation used was as follows:

$$\pi_i = \Pr(Y_i = 1 | X_i = x_i) = \frac{\begin{array}{c}\exp(\beta_0 + \beta_1(Time\ Pressure) + \beta_2(Time\ Pressure)^2 + \\ \beta_3(Visual\ Workload) + \beta_4(Visual\ Workload)^2 + \\ \beta_5(Auditory\ Workload) + \beta_6(Auditory\ Workload)^2 + \\ \beta_7(Cognitive\ Workload) + \beta_8(Cognitive\ Workload)^2 + \\ \beta_9(Psychomotor\ Workload) + \beta_{10}(Psychomotor\ Workload)^2 + \\ \beta_{11}(Task\ Frequency) + \beta_{12}(Task\ Frequency)^2 + \\ \beta_{13}(Shift) + \beta_{14}(Time) + \beta_{15}(Time)^2 + \beta_{16}(Time)^3 + \\ \beta_{17}(Information\ Flow) + \beta_{18}(Information\ Presentation) + \\ \beta_{19}(Task\ Dependency) + \beta_{20}(Teamwork) + \\ \beta_{21}(Equipment\ Feedback) + \beta_{22}(Station))\end{array}}{\begin{array}{c}1 + \exp(\beta_0 + \beta_1(Time\ Pressure) + \beta_2(Time\ Pressure)^2 + \\ \beta_3(Visual\ Workload) + \beta_4(Visual\ Workload)^2 + \\ \beta_5(Auditory\ Workload) + \beta_6(Auditory\ Workload)^2 + \\ \beta_7(Cognitive\ Workload) + \beta_8(Cognitive\ Workload)^2 + \\ \beta_9(Psychomotor\ Workload) + \beta_{10}(Psychomotor\ Workload)^2 + \\ \beta_{11}(Task\ Frequency) + \beta_{12}(Task\ Frequency)^2 + \\ \beta_{13}(Shift) + \beta_{14}(Time) + \beta_{15}(Time)^2 + \beta_{16}(Time)^3 + \\ \beta_{17}(Information\ Flow) + \beta_{18}(Information\ Presentation) + \\ \beta_{19}(Task\ Dependency) + \beta_{20}(Teamwork) + \\ \beta_{21}(Equipment\ Feedback) + \beta_{22}(Station))\end{array}}$$

The prediction equations were analyzed separately for Model A and for Model B. Both analyses were performed in SPSS 23.0 using the following statistical approach from Cohen, Cohen, West, & Aiken's (2003) textbook of Applied Multiple Regression. Regression coefficients were computed using maximum likelihood estimation. The goodness of fit of the full model was assessed using Nagelkerke (R^2) and tested for significance using an Omnibus Chi-Square test (χ^2). The contribution of individual predictor variables was assessed using Odds Ratios and tested for statistical significance using Wald Chi-Square (χ^2). Variables that did not account for significant variance were deleted from the model and the change in model fit was tested for statistical significance using the likelihood-ratio test (*LRT*). The results of each analysis are discussed in separate subsections below.

3.1.1 Model A

The prediction equation for Model A was computed by applying the full 22-term model to the data sample of 1,000 vehicles from target area A. The results of the analysis did not indicate any evidence of over dispersion and all Pearson residuals were ≤ 1.00. The goodness of fit for the full model was significant, with an Omnibus Chi-Square of $\chi^2(22) = 3{,}178.476$, $p < .001$. Hypothesis 11 proposed that prediction equations constructed from the outputs of expanded task network models would account for a significant percent of variance in human error. The findings from Model A revealed a Nagelkerke R^2 was .365, indicating that the model accounted for 36.5% of the variance in human error. Hypothesis 11 was supported.

The individual contribution and significance of each variable is summarized in Table 4. As Table 4 shows, the Wald test results indicated that 15 out of the 21 predictor variables included in the model accounted for a significant portion of the variance in human error. The -2 Log Likelihood, indicating the relative deviance of the model, was 6,313.527. In an effort to reduce deviance and improve model fit, the six variables that did not account for a significant portion of variance were excluded and the model was reanalyzed.

Table 4

Results for the individual variables in the full prediction equation for Model A.

Variable	B	S.E.	Wald	df	Sig.	Exp(B)
Auditory Workload	1.882	.673	7.816	1	.005*	6.569
Auditory Workload^2	-.218	.122	3.209	1	.073	.804
Cognitive Workload	.055	.019	8.597	1	.003*	1.057
Cognitive Workload^2	.000	.000	9.224	1	.002*	1.000
Equipment Feedback	5.896	.466	159.904	1	<.001*	363.750
Hours into Shift	.023	.118	.039	1	.843	1.024
Hours into Shift^2	-.003	.022	.025	1	.875	.997
Hours into Shift^3	.000	.001	.018	1	.894	1.000
Information Flow	-.971	.201	23.446	1	<.001*	.379
Information Presentation	-3.275	.644	25.877	1	<.001*	.038
Psychomotor Workload	.085	.018	22.218	1	<.001*	1.089
Psychomotor Workload^2	.000	.000	6.794	1	.009*	1.000
Shift	-.072	.073	.976	1	.323	.930
Task Dependency	-.584	.386	2.293	1	.130	.558
Task Frequency	1.916	.754	6.456	1	.011*	6.796
Task Frequency^2	-4.742	1.042	20.701	1	<.001*	.009
Teamwork	-.578	.153	14.330	1	<.001*	.561
Time Pressure	-3.145	1.288	5.964	1	.015*	.043
Time Pressure^2	1.099	.529	4.320	1	.038*	3.002
Visual Workload	-.043	.014	9.979	1	.002*	.958
Visual Workload^2	.000	.000	8.940	1	.003*	1.000

Note. *$p < .05$.

The prediction equation for Model A was reduced by removing the six predictors that did not account for a significant portion of variance. Specifically, the variables of Shift, Hours into Shift, Hours into Shift ^{2}, Hours into Shift ^{3}, Task Dependency, and Auditory Workload^2 were removed from the model. The remaining equation with 15 predictor variables was retested on the sample of 1,000 vehicles from target area A. The results indicated no evidence of over dispersion and all Pearson residuals were ≤ 1.00. The goodness of fit for the reduced model was significant, with an Omnibus Chi-square of $\chi^2(16) = 3{,}008.371$, $p < .001$. The Nagelkerke R^2 was .346, once again indicating that model accounted for 34.6% of the variance in human error. The results for the

contribution of each predictor in the reduced model are included in Table 5. The relative deviance of the model was -2 Log Likelihood of 6,483.633. The goodness of fit between the two models was compared using a likelihood ratio test. The likelihood ratio test was significant, $LRT = 170.106$, $p < .001$, indicating that the reduced 15-predictor model did not provide as good of a fit as the full 21-predictor model. The full model was retained.

Table 5

Results for the individual variables in the reduced prediction equation for Model A.

Variable	B	S.E.	Wald	df	Sig.	Exp(B)
Auditory Workload	-.052	.024	4.617	1	.032*	.950
Cognitive Workload	.110	.017	40.456	1	<.001*	1.117
Cognitive Workload^2	.000	.000	53.737	1	<.001*	1.000
Equipment Feedback	5.530	.416	176.851	1	<.001*	252.206
Information Flow	-.419	.214	3.821	1	.051	.658
Information Presentation	.932	.193	23.320	1	<.001*	2.539
Psychomotor Workload	.202	.015	176.490	1	<.001*	1.224
Psychomotor Workload^2	-.001	.000	89.344	1	<.001*	.999
Task Frequency	2.728	.759	12.930	1	<.001*	15.307
Task Frequency^2	-6.131	1.046	34.372	1	<.001*	.002
Teamwork	.531	.126	17.805	1	<.001*	1.700
Time Pressure	-4.031	1.235	10.655	1	.001*	.018
Time Pressure^2	1.439	.515	7.805	1	.005*	4.215
Visual Workload	-.110	.012	83.998	1	<.001*	.896
Visual Workload^2	.000	.000	75.640	1	<.001*	1.000

Note. *$p < .05$.

3.1.2 Model B

The prediction equation for Model B was computed by applying the full 22-term model to the data sample of 1,000 vehicles from target area B. The results indicated no evidence of over dispersion and all Pearson residuals were ≤ 1.00. The goodness of fit for the full model was significant, with an Omnibus Chi-Square of $\chi^2(21) = 4,484.156$, $p < .001$. The Nagelkerke R^2 was .219, indicating that the model accounted for 21.9% of the variance in human error. Hypothesis 11 was again supported by Model B. The fit of

Model B was poorer than the fit of Model A; a -2 Log Likelihood of 18,611.264 indicated that deviance was higher. Results for the individual contribution of each predictor are included in Table 6. As Table 6 shows, the results of the Wald test indicated that only 13 out of the 21 predictor variables accounted for a significant portion of the variance in human error. The variables of Cognitive Workload^2, Hours into Shift, Hours into Shift^2, Hours into Shift^3, Information Presentation, Psychomotor Workload^2, Shift, and Visual Workload^2 did not account for significant portions of the variance in human error in Model B. These eight variables were removed from the equation and model fit was reanalyzed. The results are discussed in the subsequent paragraph.

Table 6

Results for the individual variables in the full prediction equation for Model B.

Variable	B	S.E.	Wald	df	Sig.	Exp(B)
Auditory Workload	-.534	.243	4.832	1	.028*	.586
Auditory Workload^2	.312	.057	30.223	1	<.001*	1.366
Cognitive Workload	.091	.017	29.893	1	<.001*	1.095
Cognitive Workload^2	.000	.000	.000	1	.997	1.000
Equipment Feedback	1.611	.189	72.710	1	<.001*	5.008
Hours into Shift	-.095	.121	.613	1	.434	.910
Hours into Shift^2	.023	.022	1.009	1	.315	1.023
Hours into Shift^3	-.001	.001	1.115	1	.291	.999
Information Flow	-.963	.180	28.749	1	<.001*	.382
Information Presentation	-17.147	824.345	.000	1	.983	.000
Psychomotor Workload	.038	.015	6.158	1	.013*	1.038
Psychomotor Workload^2	.000	.000	1.988	1	.159	1.000
Shift	-.025	.075	.114	1	.736	.975
Task Dependency	2.799	.267	109.646	1	<.001*	16.431
Task Frequency	2.733	.561	23.755	1	<.001*	15.383
Task Frequency^2	-4.504	.673	44.732	1	<.001*	.011
Teamwork	.645	.111	33.598	1	<.001*	1.907
Time Pressure	-12.115	1.146	111.748	1	<.001*	.000
Time Pressure^2	5.788	.578	100.368	1	<.001*	326.244
Visual Workload	-.024	.010	5.865	1	.015*	.976
Visual Workload^2	.000	.000	1.601	1	.206	1.000

Note. *$p < .05$.

The prediction equation for Model B was reduced to the 13 predictor variables that accounted for significant portions of the variance. The reduced model was then reapplied to the sample of 1,000 vehicles from target area B. Again, the results indicated no evidence of overdispersion and all Pearson residuals were ≤ 1.00. The goodness of fit for the reduced model was significant, with an Omnibus Chi-square of $\chi^2(13) = 1{,}566.792$, $p < .001$. The Nagelkerke R^2 was .205, indicating that the reduced model accounted for 20.5% of the variance in human error. The full results of the contribution of each predictor in the reduced model are included in Table 7. A -2 Log Likelihood of 7,097.856 indicated that deviance was still high and model fit was still poor. The goodness of fit between the two models was compared using a likelihood ratio test. The likelihood ratio test was significant, $LRT = 614.223$, $p < .001$, indicating that the reduced 13-predictor model did not provide as good of a fit as the full 21-predictor model. The full model was retained.

Table 7

Results for individual variables in the reduced prediction equation for Model B.

Variable	B	S.E.	Wald	df	Sig.	Exp(B)
Auditory Workload	-.578	.212	7.460	1	.006*	.561
Auditory Workload^2	.308	.050	37.862	1	<.001*	1.360
Cognitive Workload	.087	.006	232.989	1	<.001*	1.090
Equipment Feedback	2.182	.162	182.365	1	<.001*	8.861
Information Flow	-.819	.132	38.571	1	<.001*	.441
Psychomotor Workload	.016	.002	42.307	1	<.001*	1.016
Task Dependency	2.155	.205	110.466	1	<.001*	8.624
Task Frequency	2.382	.547	19.000	1	<.001*	10.828
Task Frequency^2	-4.037	.650	38.563	1	<.001*	.018
Teamwork	1.072	.096	125.065	1	<.001*	2.921
Time Pressure	-13.843	1.064	169.290	1	<.001*	.000
Time Pressure^2	6.600	.548	144.987	1	<.001*	734.798
Visual Workload	-.030	.003	131.488	1	<.001*	.970

Note. *$p < .05$.

3.2 Study 2: Validation

The results from both the split-half and cross sample cross-validation procedures performed on both models are discussed in separate subsections in terms of the stability of model predictors and the accuracy and sensitivity of model predictions.

3.2.1 Stability of Predictors

As part of the validation procedure, the stability of the 21 predictors was tracked across the 20 split-half validation trials for each target area. The results from these analyses are summarized in Table 8. The full results from each individual validation trial are reported in Tables 17 – 22 in Appendix B. The results from the significance testing of the full models discussed in sections 3.1.1 and 3.1.2 are also compiled in Table 8. As Table 8 shows, the significance and thus the stability of predictors varied greatly across validation trials. Some variables, such as Cognitive Workload, Task Frequency^2, Information Flow, and Equipment Feedback, were very stable and highly significant across most, if not all trials. The stability of other variables greatly depended on the model. For example, Time Pressure, Visual Workload, Task Dependency, and Teamwork had high probabilities of retention in one of the models, usually 1.00, and somewhat low probabilities of retention in the other model, sometimes as low .05 or even .00. Other variables were never significant at all. For example the variable of Hours into Shift was never significant, either as a first, second, or third-order effect, and also had a .00 probability of retention for each. The same was true for the variable of Shift, again with a probability of retention of .00 in all validation trials.

Table 8

The stability of predictors across validation trials.

Predictors	Model A		Model B	
	Derivation Set A (n = 1000)	Validation Set A (n = 500)	Derivation Set B (n = 1000)	Validation Set B (n = 500)
	p	$P_{Retention}$	p	$P_{Retention}$
Auditory Workload	.005*	.60	.028*	.95
Auditory Workload^2	.073	.00	<.001*	1.00
Cognitive Workload	.003*	1.00	<.001*	1.00
Cognitive Workload^2	.002*	1.00	.997	.00
Equipment Feedback	<.001*	1.00	<.001*	1.00
Hours into Shift	.843	.00	.434	.00
Hours into Shift ^{3}	.875	.00	.315	.00
Hours into Shift^2	.894	.00	.291	.00
Information Flow	<.001*	.80	<.001*	1.00
Information Presentation	<.001*	1.00	.983	.00
Psychomotor Workload	<.001*	1.00	.013*	.05
Psychomotor Workload^2	.009*	.85	.159	.00
Shift	.323	.00	.736	.00
Task Dependency	.130	.05	<.001*	1.00
Task Frequency	.011*	.50	<.001*	.95
Task Frequency^2	<.001*	1.00	<.001*	1.00
Teamwork	<.001*	.20	<.001*	1.00
Time Pressure	.015*	.05	<.001*	1.00
Time Pressure^2	.038*	.05	<.001*	1.00
Visual Workload	.002*	1.00	.015*	.00
Visual Workload^2	.003*	1.00	.206	.65

Taken together, the results summarized in Table 8 indicate that utility of each predictor variable greatly depends on the area and data the model is applied to. Some of the variables were very stable in both models. Some variables were only stable within one of the models, even though both models were in the same domain. Some variables were not even stable within the same model and greatly depended on the sample of vehicles. This may be why the reduced prediction equations in both models did not provide a significantly better fit. Nevertheless, the results of the validation trials suggest

that prediction equations should be domain and area specific. Each predictor should be evaluated in each application of the prediction equation. The stability of variables is also further discussed in Chapter 5, as a function of all three studies and all three models included in this manuscript.

3.2.2 Sensitivity and Specificity of Predictions

As part of the validation procedure, the accuracy and stability of predictions were also tracked across both the 20 split-half validation trials for each target area and the cross sample validation trials. As previously mentioned, human error is a dichotomous outcome variable, and determining the suitability of binary logistic prediction models requires binary classification of predictions (Myers & Forgy, 1963). Each prediction outcome computed, again in both the split-half and cross-sample trials, was set equal to 0 if predicted probabilities was less than actual probability of human error, and set equal to 1 if predicted probabilities are greater than the actual probability of human error. The actual probability of human error in this automotive assembly context was 0.05, thus this was chosen as the cutoff for predictions. This cutoff was used to arrange the predictions from the validation trials in separate 2 x 2 contingency tables, with four possible classifications; hit, miss, false alarm, or correct rejection. The 2 x 2 contingency tables were then used to compute the Sensitivity and Specificity of model predictions using the following equations:

- $$Sensitivity = \frac{number\ of\ true\ positives}{number\ of\ true\ positives + \ false\ negatives}$$

- $$Specificity = \frac{number\ of\ true\ negatives}{number\ of\ true\ negatives + \ false\ positives}$$

The sensitivity and specificity of model predictions are reported in Table 9. The sensitivity and specificity values were judged as follows:

90-100% *excellent*,

80-90% *good*,

70-80% *fair*,

60-70% *poor*, and

< 60% *failure*.

As Table 9 shows, within area sensitivity ranged from excellent to fair; 92.85% for Model A and 73.11% for model B. Within area specificity ranged from good, 82.03% for Model A, to fair, 75.64% for Model B. These results indicate that the expanded task network models can make accurate and valid predictions about the occurrence of human errors. However, Table 9 also shows that the accuracy and validity of predictions is highly model specific.

When Model B was used to predict human error in the data from area A, specificity was only 60.10%, poor, a borderline failure. The detection criterion for model B resulted in too many false alarms. When model A was used to predict human error in the data from area B, the detection criterion also resulted in too many false alarms. Specificity was 64.04%, again poor. In the cross-sample validation trials, both models were poor at correctly identifying the situations in which errors *did not* occur. Thus it appears that the utility of predictions of each expanded task network model is limited in generalization to the area or domain the model was based on. In other words, the models cannot predict the situations in which errors do not occur for tasks that were not specifically included in the task network.

Table 9

Results for the Sensitivity and Specificity of model predictions.

		Data from Area A	Data from Area B
Model A			
	Sensitivity	92.85%	77.39%
	Specificity	82.03%	64.04%
Model B			
	Sensitivity	72.86%	73.11%
	Specificity	60.10%	75.64%

3.3 Study 3: Application

Given the inconsistencies based on vehicle sample in Study 1 and lack of specificity across models in Study 2, a third study was conducted. Study 3 tested a prediction equation across both target areas on a larger sample of vehicles in order to develop recommendations for reducing human errors.

The prediction equation for Study 3 was also computed using the variables discussed in Measures section 2.1.3. This analysis controlled for the different target areas and stations to ensure that the results were not simply caused by these factors in the following way. The control variables of Target Area and Station Number were entered into Block 1. The interaction term of Target Area × Station Number was entered into Block 2. The first first-order effects for all 13 predictors were then entered into Block 3: Time Pressure, Visual Workload, Auditory Workload, Cognitive Workload, Psychomotor Workload, Task Frequency, Shift, Hours into Shift, Information Flow, Information Presentation, Task Dependency, Teamwork, and Equipment Feedback. The second-order polynomials for Time Pressure, Visual Workload, Auditory Workload, Cognitive Workload, Psychomotor Workload, Task Frequency, and Hours into Shift were entered in

Block 4 to test for quadratic effects. The third-order polynomial for Hours into Shift was entered into Block 5 to test for cubic effects. The complete equation used was as follows:

$$\pi_i = \Pr(Y_i = 1 | X_i = x_i) = \frac{\begin{array}{c}\exp(\beta_0 + \beta_1(Time\ Pressure) + \beta_2(Time\ Pressure)^2 + \\ \beta_3(Visual\ Workload) + \beta_4(Visual\ Workload)^2 + \\ \beta_5(Auditory\ Workload) + \beta_6(Auditory\ Workload)^2 + \\ \beta_7(Cognitive\ Workload) + \beta_8(Cognitive\ Workload)^2 + \\ \beta_9(Psychomotor\ Workload) + \beta_{10}(Psychomotor\ Workload)^2 + \\ \beta_{11}(Task\ Frequency) + \beta_{12}(Task\ Frequency)^2 + \\ \beta_{13}(Shift) + \beta_{14}(Time) + \beta_{15}(Time)^2 + \beta_{16}(Time)^3 + \\ \beta_{17}(Information\ Flow) + \beta_{18}(Information\ Presentation) + \\ \beta_{19}(Task\ Dependency) + \beta_{20}(Teamwork) + \\ \beta_{21}(Equipment\ Feedback) + \beta_{22}(Target\ Area) \\ \beta_{23}(Station) + \beta_{22}(Target\ Area \times Station))\end{array}}{\begin{array}{c}1 + \exp(\beta_0 + \beta_1(Time\ Pressure) + \beta_2(Time\ Pressure)^2 + \\ \beta_3(Visual\ Workload) + \beta_4(Visual\ Workload)^2 + \\ \beta_5(Auditory\ Workload) + \beta_6(Auditory\ Workload)^2 + \\ \beta_7(Cognitive\ Workload) + \beta_8(Cognitive\ Workload)^2 + \\ \beta_9(Psychomotor\ Workload) + \beta_{10}(Psychomotor\ Workload)^2 + \\ \beta_{11}(Task\ Frequency) + \beta_{12}(Task\ Frequency)^2 + \\ \beta_{13}(Shift) + \beta_{14}(Time) + \beta_{15}(Time)^2 + \beta_{16}(Time)^3 + \\ \beta_{17}(Information\ Flow) + \beta_{18}(Information\ Presentation) + \\ \beta_{19}(Task\ Dependency) + \beta_{20}(Teamwork) + \\ \beta_{21}(Equipment\ Feedback) + \beta_{22}(Target\ Area) \\ \beta_{23}(Station) + \beta_{22}(Target\ Area \times Station))\end{array}}$$

The prediction equation was analyzed in SPSS 23.0 using the statistical approach from Cohen, Cohen, West, & Aiken's (2003) textbook of Applied Multiple Regression. Raw scores for each variable were standardized by converting to z-scores. The goodness of fit of model was assessed using Nagelkerke (R^2) and tested for significance using an Omnibus Chi-Square test (χ^2). The contribution of individual predictor variables was assessed using Odds Ratios and tested for statistical significance using Wald Chi-Square (χ^2). This made it possible to not only identify which variables were significant risk factors for human error, but also compare the magnitude on a standard scale and rank the variables (Szumilas, 2010).

There was no evidence of over dispersion and all Pearson residuals were ≤ 1.00. The goodness of fit for the model was significant; Omnibus Chi-Square of $\chi^2(24)$ =

8,205.141, $p < .001$. The deviance of the model was a -2 Log Likelihood of 33,739.602. The Nagelkerke R^2 was .220, indicating that the model accounted for 22% of the variance in human error. Once again, Hypothesis 11 was supported. The individual contribution and significance of each variable is summarized in Table 10. As Table 10 shows, the Wald test results indicated that 17 out of the 21 predictor terms entered into the application equation accounted for a significant portion of the variance in human error.

Table 10

Results for the individual variables in the full prediction equation for Study 3.

Variable	B	S.E.	Wald	df	Sig.	Exp(B)
Auditory Workload	1.257	.066	359.026	1	<.001*	3.514
Auditory Workload^2	-.238	.019	153.227	1	<.001*	.788
Cognitive Workload	1.220	.077	248.008	1	<.001*	3.387
Cognitive Workload^2	-.402	.026	245.554	1	<.001*	.669
Equipment Feedback	2.395	.081	865.527	1	<.001*	10.971
Hours into Shift	-.011	.056	.039	1	.843	.989
Hours into Shift^2	.005	.010	.199	1	.655	1.005
Hours into Shift^3	.000	.001	.256	1	.613	1.000
Information Flow	-.918	.050	337.858	1	<.001*	.399
Information Presentation	-1.554	.108	205.261	1	<.001*	.211
Psychomotor Workload	1.651	.066	633.011	1	<.001*	5.211
Psychomotor Workload^2	-.238	.017	201.166	1	<.001*	.788
Shift	-.038	.034	1.257	1	.262	.963
Task Dependency	1.031	.086	144.294	1	<.001*	2.803
Task Frequency	.851	.236	13.026	1	<.001*	2.342
Task Frequency^2	-2.887	.287	100.996	1	<.001*	.056
Teamwork	.460	.035	173.889	1	<.001*	1.583
Time Pressure	-7.013	.418	281.058	1	<.001*	.001
Time Pressure^2	3.074	.188	267.527	1	<.001*	21.636
Visual Workload	-1.716	.091	352.758	1	<.001*	.180
Visual Workload^2	.415	.026	259.327	1	<.001*	1.515

Note. *$p < .05$.

In an effort to reduce deviance and improve model fit, the four variables that did not account for a significant portion of variance were excluded and the model was reanalyzed. Specifically, the variables of Shift, Hours into Shift, Hours into Shift^2, and Hours into Shift^3 were removed from the equation. Once again there was no evidence of over dispersion and all Pearson residuals were ≤ 1.00. The goodness of fit for the reduced model was also significant, with an Omnibus Chi-Square of $\chi^2(20) = 8{,}200.738$, $p < .001$. The deviance of the model was a -2 Log Likelihood of 33,744.005. The Nagelkerke R^2 was .220, indicating that the reduced model also accounted for 22% of the variance in human error. Model fit was compared using a Likelihood Ratio Test. A $LRT = 8.806$, $p > .05$, indicated that the reduced 17-predictor model provided as good of a fit as the full 21-predictor model. The simpler model was retained. The individual contribution and significance of each variable is summarized in Table 11 and ranked by Odds Ratio. The model summarized in Table 11 was the final model used to develop recommendations for reducing human errors in this automotive assembly context.

Table 11

Results for the individual variables in the reduced final prediction equation for Study 3.

Variable	B	S.E.	Wald	df	Sig.	Exp(B)
Auditory Workload	1.257	.066	359.093	1	<.001*	3.514
Auditory Workload^2	-.238	.019	153.217	1	<.001*	.788
Cognitive Workload	1.220	.077	247.960	1	<.001*	3.386
Cognitive Workload^2	-.402	.026	245.577	1	<.001*	.669
Equipment Feedback	2.395	.081	865.793	1	<.001*	10.971
Information Flow	-.918	.050	337.615	1	<.001*	.399
Information Presentation	-1.555	.108	205.408	1	<.001*	.211
Psychomotor Workload	1.651	.066	633.102	1	<.001*	5.211
Psychomotor Workload^2	-.238	.017	201.294	1	<.001*	.788
Task Dependency	1.031	.086	144.282	1	<.001*	2.803
Task Frequency	.863	.236	13.412	1	<.001*	2.371
Task Frequency^2	-2.896	.287	101.726	1	<.001*	.055
Teamwork	.459	.035	173.875	1	<.001*	1.583
Time Pressure	-7.007	.418	280.630	1	<.001*	.001
Time Pressure^2	3.072	.188	267.087	1	<.001*	21.581
Visual Workload	-1.716	.091	352.839	1	<.001*	.180
Visual Workload^2	.415	.026	259.548	1	<.001*	1.515

Note. *$p < .05$.

CHAPTER 4:

INDIVIDUAL PREDICTOR RESULTS

The individual predictor results discussed in this section are proposed predictors that were developed based on situational variables from Sharit's (2006) model. The specific variables were originally chosen using findings from the literature that indicated that these variables may be useful predictors of human error. The variables included in the models were selected based on their applicability to the automotive assembly context. For example, time constraints related to the amount of time each vehicle is in a workstation may have be a great indicator of human error, but training may not have been since all associates received the same amount and type of training before being allowed to work on the assembly line.

Each variable is accompanied by a hypothesis proposed in the introduction. The results are discussed in terms of testing each hypothesis. All statistical χ^2 results stem from the binary logistic regression tests covered in Chapter 3. This chapter simply discusses the results of each individual predictor variable separately. The graphs presented in this section display the exact predictions made by each model as solid lines and the actual error probabilities observed in each data set as corresponding bars on the graph. The solid lines are an indicator of how well each model variable predicted human error and was not directly based on the actual error data represented by the bars on the graph. All continuous predictor variables are also accompanied by an ROC curve that assess their accuracy in predicting human errors. All ROC curves were statistically tested

against 0.5, representing chance. Each predictor is discussed in a separate subsection containing results from all 3 models (Model A, Model B, Study 3).

4.1 Time Pressure

Time pressure was a significant predictor of human error as a first-order effect in both Model A and Model B; $\chi^2 = 5.964$, $p = .015$, and $\chi^2 = 111.748$, $p < .001$, respectively. Time pressure was also a significant predictor of human error as a second-order effect in both models; $\chi^2 = 4.320$, $p < .001$, in Model A and $\chi^2 = 100.368$, $p < .001$, in Model B. The data for the effect of Time Pressure in Model A and Model B are graphed in Figures 5 and 6, respectively.

Hypothesis 1 proposed that there is an optimum percentage of time utilized, such that the probability of human error would be higher when time utilized was low or high. Hypothesis 1 was partially supported by Study 1. As Figure 6 shows, the probability of human error in Model B was indeed significantly higher when the time utilized was low or high. The odds ratio for this effect was exceptionally high, with a value of 326.244 indicating a strong association between high or low time pressure and human error. This effect was also significant in Model A; however, it is worth noting that there were no instances in Figure 5 where the time utilized fell below 60% in Model A. Although we do not get to see the spike in actual error probability in Model A which occurred with low utilization times in Model B, the overall pattern of the second-order effect in Model A matches Model B and Hypothesis 1. Nonetheless, the conclusion drawn based on significance testing is that Hypothesis 1 was only partially supported by Model B.

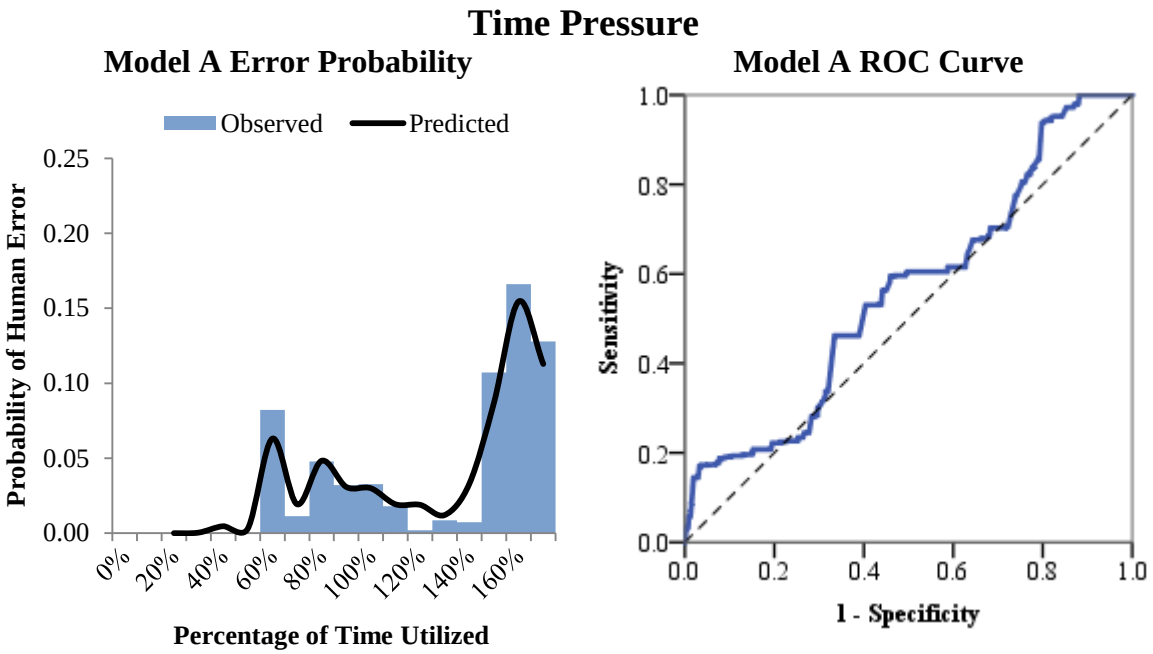

Figure 5. The effect of time pressure on the predicted and observed probability of human error in Model A (left graph), and ROC curve of human error prediction in Model A (right graph).

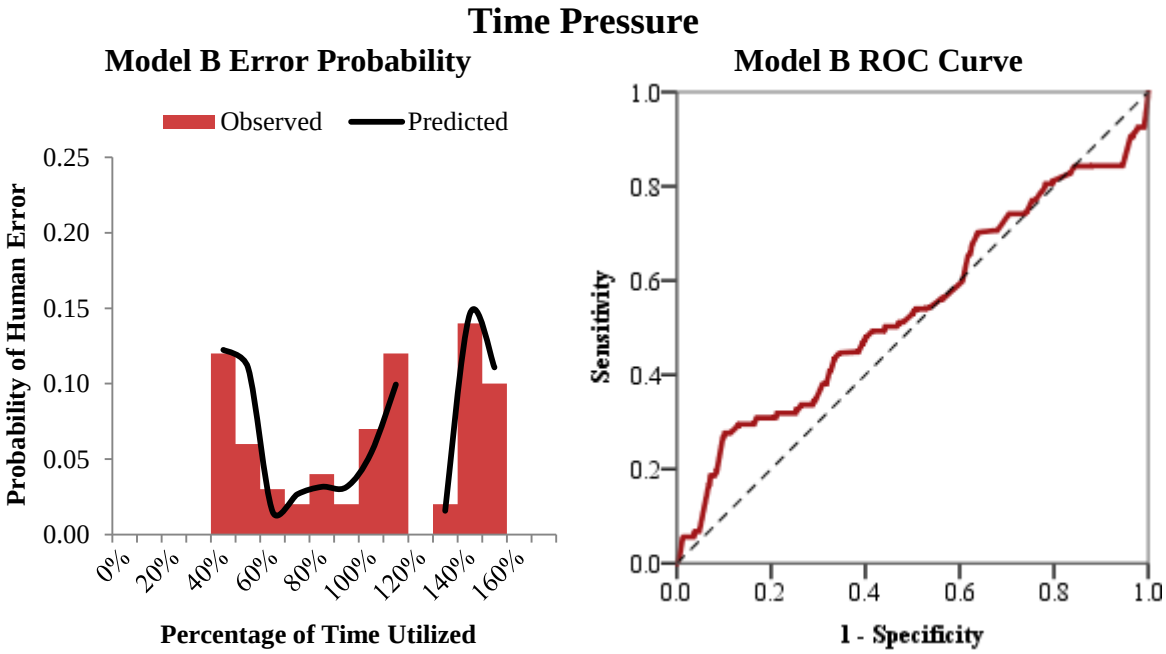

Figure 6. The effect of time pressure on the predicted and observed probability of human error in Model B (left graph), and ROC curve of human error prediction in Model B (right graph).

Time pressure was also a significant first and second-order effect in Study 3; χ^2 = 281.058, $p < .001$, and χ^2 = 267.527, $p < .001$, respectively. The data for the effect of time pressure in Study 3 are graphed in Figure 7. As Figure 7 shows, the probability of human error was higher when time utilized was high, but not low. The larger sample of vehicles analyzed for this study included many more observations where time utilized fell below 60%, yet the spike seen in Model B was not replicated. Based on this finding in the larger and more complete sample, the overall conclusion was that Hypothesis 1 was not supported. It appears as if the second order terms in the models are significantly predicting the curvilinear effect of time pressure on the probability of error, rather than the true ∪-shaped effect hypothesized.

The measure of percentage of time utilized was a significant, yet rather poor predictor of the occurrence of human error. The accuracy of time pressure as a predictor of human error was measured by computing the area under the ROC curve (*AUC*) in each model. Model A *AUC* was .561, $p < .001$, Model B *AUC* was .539, $p < .001$, and Study 3 *AUC* was .548, $p < .001$. All three *AUCs* were significant, but practically not much better than an *AUC* of .500 representing chance. Nevertheless, time pressure still accounted for a significant portion of the variance in human error and thus was further investigated as part of Study 3 for the purposes of human error reduction. The results indicated that time pressure contributed to a 5.3% increase in human error. This increase occurred when the percentage of time utilized was exceptionally high. Looking back at Figure 7, a large and sharp increase in the probability of human error occurred when the percentage of time utilized exceeded 140%.

The recommended intervention was to move tasks between stations or workers so that time utilized no longer crossed above 140%.

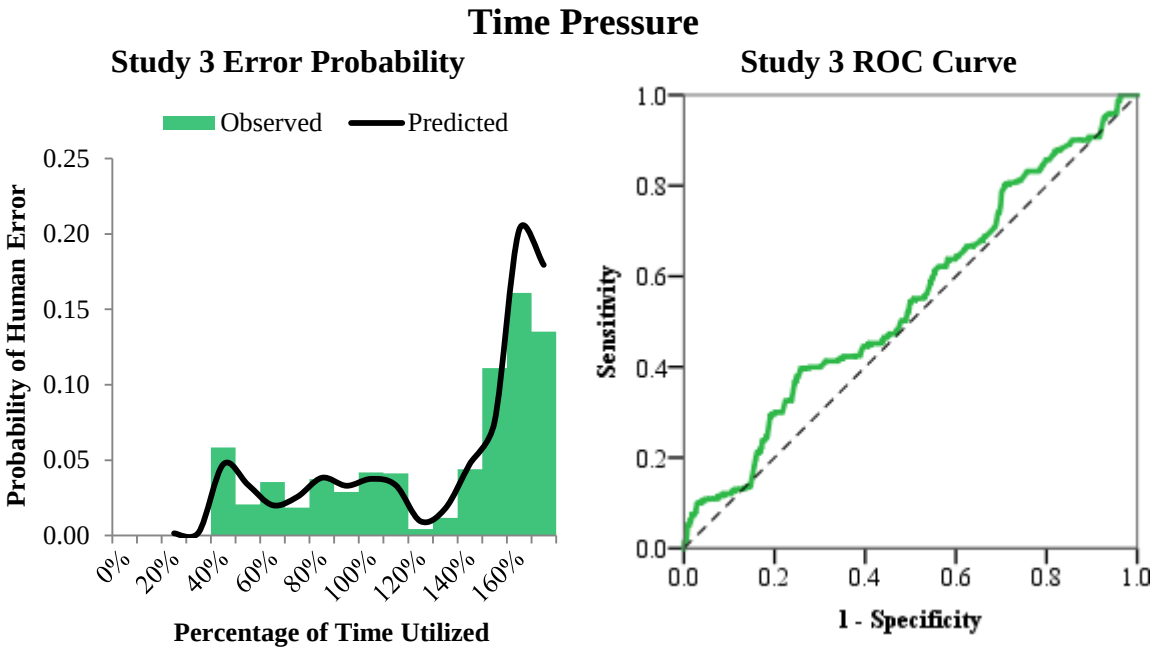

Figure 7. The effect of time pressure on the predicted and observed probability of human error in Study 3 (left graph), and ROC curve of human error prediction in Study 3 (right graph).

4.2 Workload

Workload was analyzed according to the separate demands placed on visual, auditory, cognitive, and psychomotor resources. The results are presented separately by component in sections 4.2.1 – 4.2.4. The implication of the results on Hypothesis 2 is discussed at the end, in section 4.2.5.

4.2.1 Visual Workload

Visual workload was a significant predictor of human error as a first-order (χ^2 = 9.979, $p < .001$ in Model A, and $\chi^2 = 5.865$, $p = .015$ in Model B) and a second-order effect in Model A but not Model B ($\chi^2 = 8.940$, $p < .001$ in Model A, and $\chi^2 = 1.601$, $p =$

.206 in Model B). The data for the effect of visual workload in Model A are graphed in Figure 8 and the data for visual workload in Model B are graphed in Figure 9.

As Figures 8 and 9 show, the overall probability of human error increased as visual workload increased, in both models. Looking at Figure 8 for Model A, the probability of human error increased more sharply after visual workload demands surpassed a certain level. The second order effect in Model A significantly predicted this curvilinear sharp increase. Looking at Figure 9 for Model B, there was also a distinct sharp increase in the probability of human error after visual workload demands surpassed a certain level; however, there were also no observed instances where visual workload was between 250 and 300, and the second order effect in Model B was not significant.

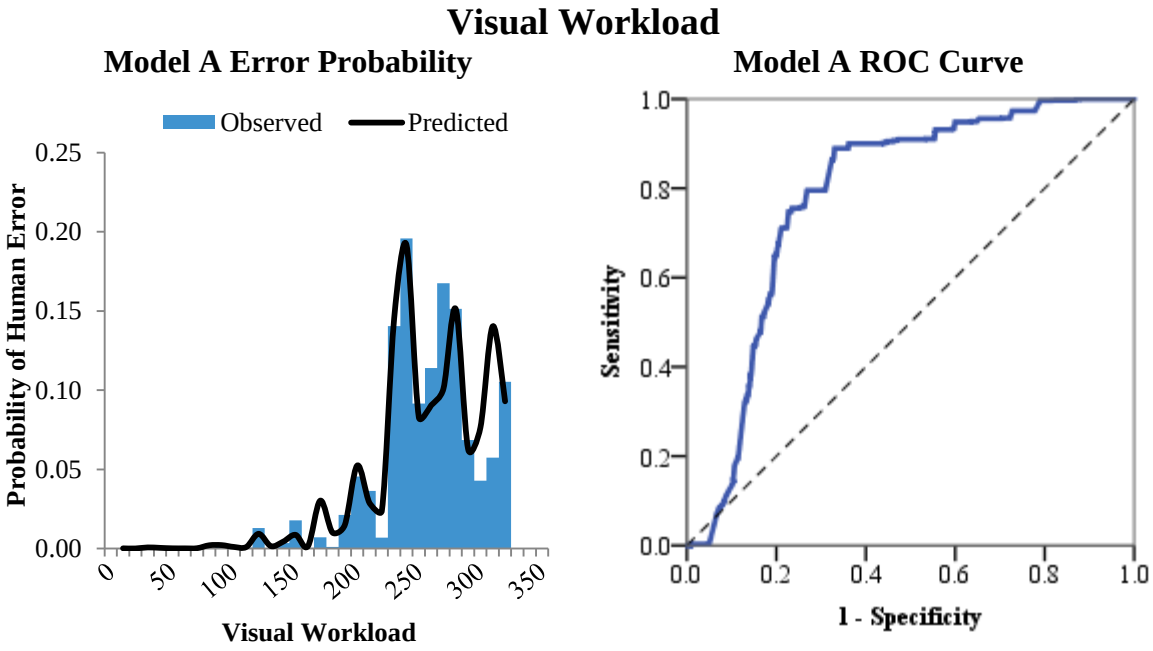

Figure 8. The effect of visual workload demands on predicted and observed probability of human error in Model A (left graph), and ROC curve of human error prediction in Model A (right graph).

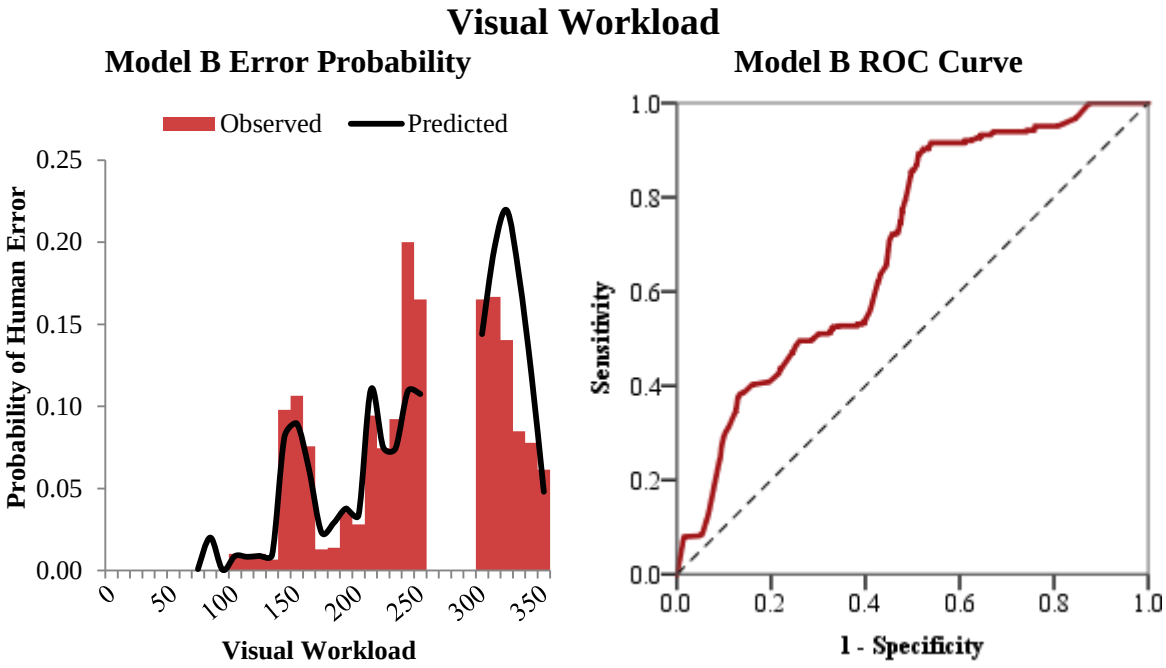

Figure 9. The effect of visual workload demands on predicted and observed probability of human error in Model B (left graph), and ROC curve of human error prediction in Model B (right graph).

Visual workload was also a significant first and second-order effect in Study 3; $\chi^2 = 352.758$, $p < .001$, and $\chi^2 = 259.327$, $p < .001$, respectively. The data for the effect of visual workload in Study 3 are graphed in Figure 10. As Figure 10 shows, once again, the overall probability of human error increased as visual workload increased. Like in Model A, the probability of human error increased more sharply after visual workload demands surpassed a certain level, which was significantly predicted by the second order effect.

Overall, visual workload was a fairly accurate predictor of human error in all three models. The accuracy of visual workload as a predictor of human error was measured by computing the area under the ROC curve (*AUC*) in each model, and testing them statistically against .5 (chance). Model A *AUC* was .782, $p < .001$, Model B *AUC* was .693, $p < .001$, and Study 3 *AUC* was .724, $p < .001$. The results from Study 3

indicated that visual workload contributed to a 9.4% increase in human error. This increase occurred when visual workload was exceptionally high. Looking at Figure 10 again, the probability of human error increased more sharply after visual workload demands surpassed a certain level, in this case 235. The probability of human error in the sample of application vehicles from Study 3 was broken down as a function of visual workload and revealed that the probability of human error was 0.024 when visual workload was below 235, and 0.119 when visual workload was above 235.

The recommended intervention was to move tasks between workers or stations or provide necessary task information auditorally to keep visual workload from becoming excessively high (in this case above 235). A visual workload score of 235 or above indicated that the tasks required to be performed on a vehicle at a station added up to greater than 235 for the duration of time that vehicle was in that station. The visual workload score could be lowered by reducing the visual demands of enough tasks until the score is below 235. For example, if a station required a worker to visually read vehicle option information as well as visually align the orientation of the part being installed, the visual workload score could be lowered by presenting vehicle option information auditorally and changing the mounting tabs on the part so it could only be installed in one orientation.

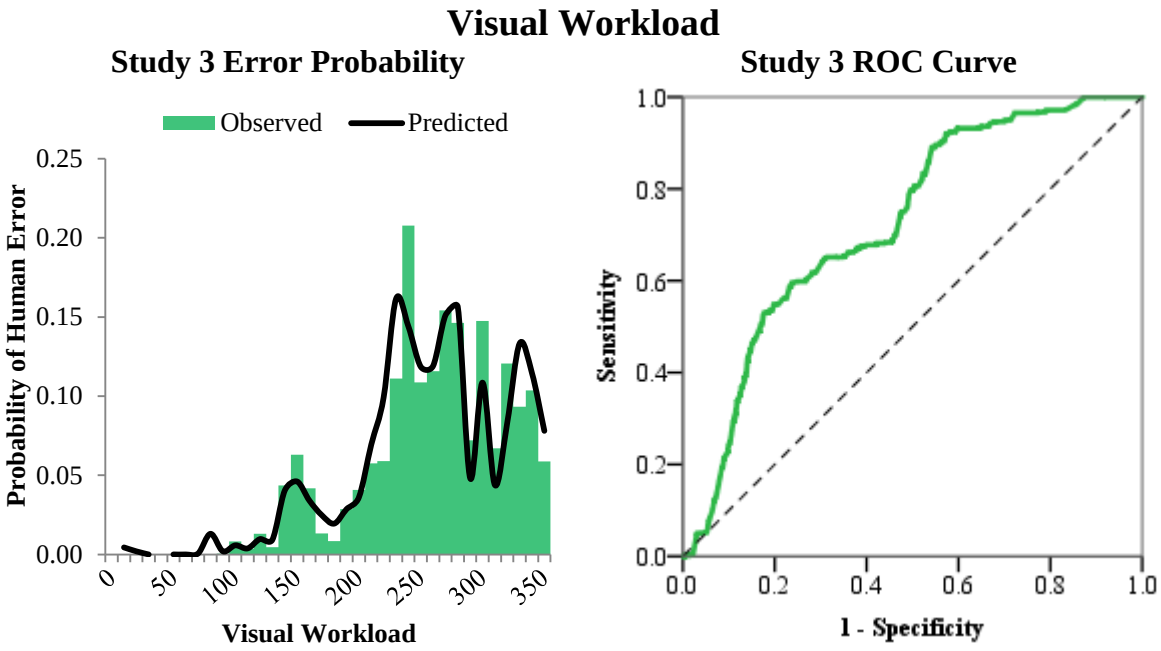

Figure 10. The effect of visual workload demands on predicted and observed probability of human error in Study 3 (left graph), and ROC curve of human error prediction in Study 3 (right graph).

4.2.2 Auditory Workload

The auditory workload of assembly tasks was comparatively much lower than the workload of the other resource components. Nonetheless, auditory workload was a significant predictor of human error as a first-order effect in both Model A, $\chi^2 = 7.816$, $p = .005$, and Model B, $\chi^2 = 4.832$, $p = .028$. The second-order effect of auditory workload was not significant in Model A, $\chi^2 = 3.209$, $p = .073$, but was significant in Model B, $\chi^2 = 30.223$, $p < .001$. The auditory workload data for Model A are graphed in Figure 11 and the data for Model B are graphed in Figure 12. As seen in Figures 11 and 12, the probability of human error generally increased as auditory workload increased. The significant second-order effect present in Figure 12 indicated that the increase in the probability of human error as auditory workload increased may be curvilinear.

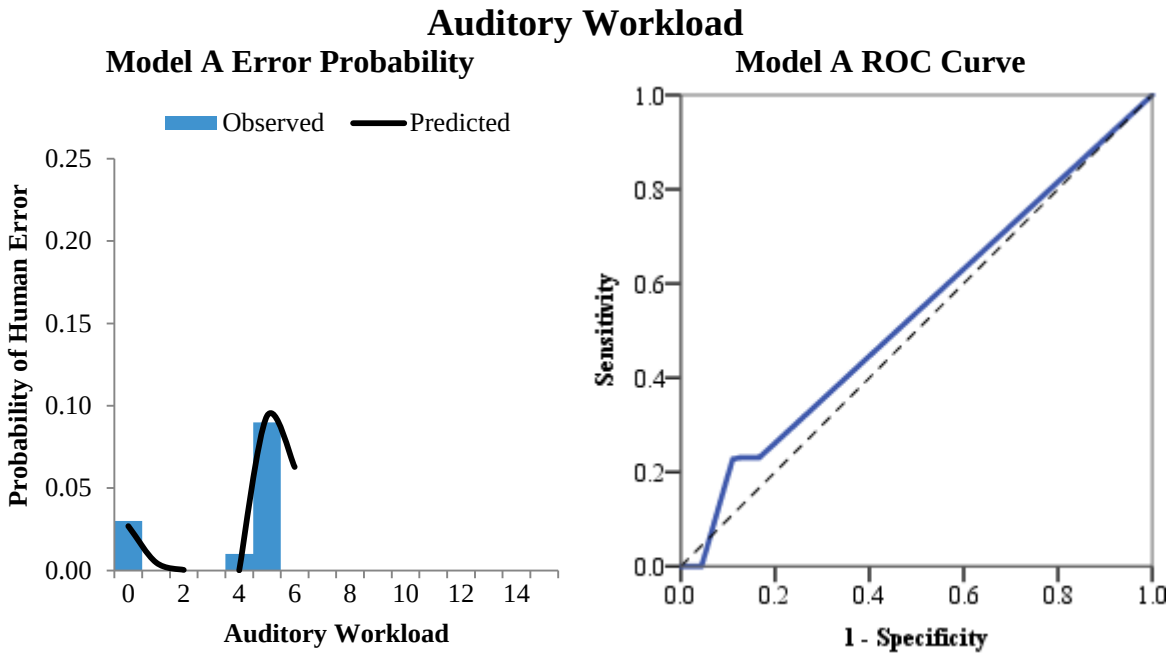

Figure 11. The effect of auditory workload demands on the predicted and observed probability of human error in Model A (left graph), and ROC curve of human error prediction in Model A (right graph).

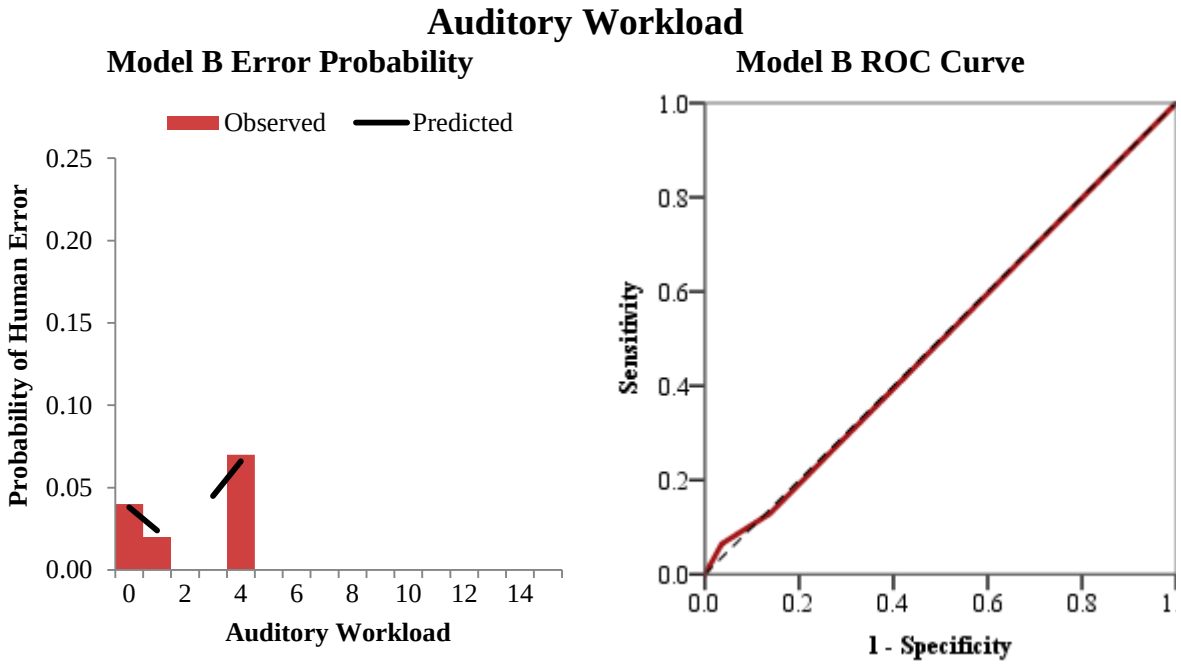

Figure 12. The effect of auditory workload demands on the predicted and observed probability of human error in Model B (left graph), and ROC curve of human error prediction in Model B (right graph).

Auditory workload was also a significant predictor of human error in Study 3, both as a first-order effect, $\chi^2 = 359.026$, $p < .001$, and a second-order effect, $\chi^2 = 153.227$, $p < .001$. The auditory workload data from Study 3 are presented in Figure 13. As Figure 13 shows, once again, the probability of human error generally increased as auditory workload increased and the second-order effect indicated that this increase may be curvilinear. However, the nature of the effect of auditory workload on the probability of human error cannot truly be specified from these data. Even in the larger 4,188 vehicle sample in Study 3, there were many gaps in the values of auditory workload observed. Looking at Figure 13, there were cases where the auditory workload was 15 but no cases where the auditory workload rsnged between 6 and 14, thus we could not truly observe the probability of human error as auditory workload increased.

The accuracy of auditory workload as a predictor of human error was quite low and sometimes no better than chance. Once again, accuracy was measured by computing the area under the ROC curve (*AUC*) in each model and testing it statistically against .5. Model A *AUC* was .533, $p < .001$, Model B *AUC* was .498, $p = .799$, and Study 3 *AUC* was .509, $p = .025$. The results from all three models indicated that although auditory workload may be a significant predictor of human error, it could not account for a significant contribution to human error in the current context. As a result, no recommendation is made for auditory workload in the reduction of human error in this automotive assembly context.

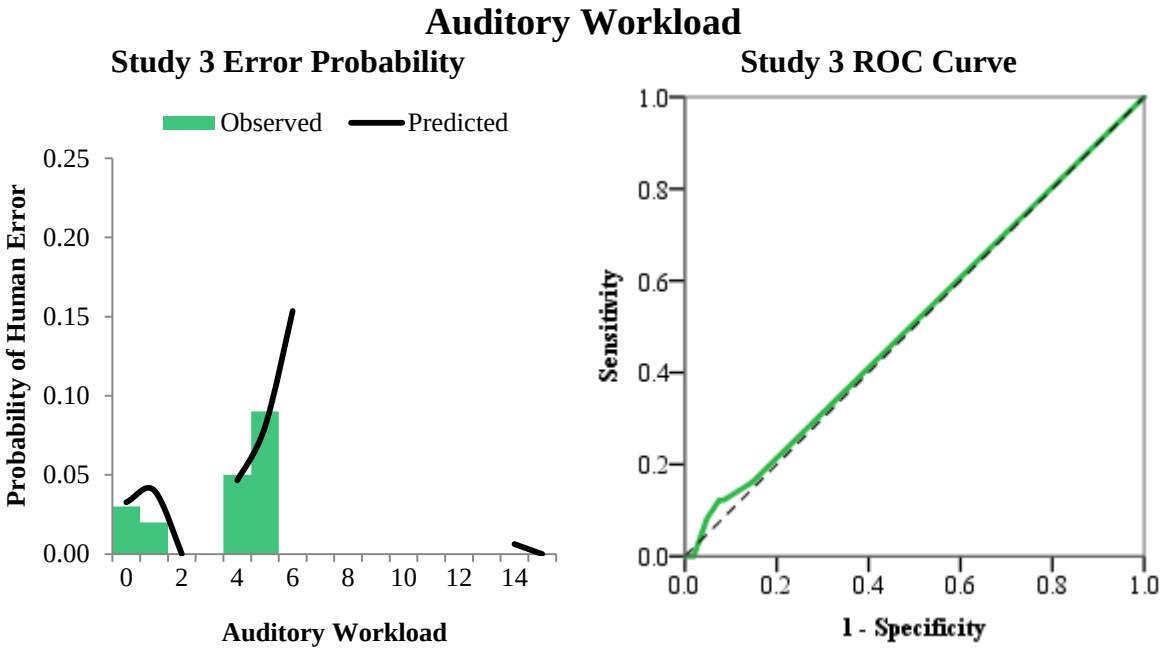

Figure 13. The effect of auditory workload demands on the predicted and observed probability of human error in Study 3 (left graph), and ROC curve of human error prediction in Study 3 (right graph).

4.2.3 Cognitive Workload

Cognitive workload was a significant predictor of human error as a first-order effect in both models; $\chi^2 = 8.597$, $p = .003$, in Model A, and $\chi^2 = 29.893$, $p < .001$, in Model B. Cognitive workload was also significant as a second-order effect in Model A, $\chi^2 = 9.224$, $p = .002$, but not in Model B, $\chi^2 = .000$, $p = .997$. The data for the effect of cognitive workload in Model A are graphed in Figure 14 and the data for Model B are graphed in Figure 15. Looking at Figures 14 and 15, the overall probability of human error increased as cognitive workload increased. The second-order effect present in Figure 14 also indicated that the increase in the probability of human error may be curvilinear or parabolic. To better understand this effect, the larger sample in Study 3 was examined. Cognitive workload was once again significant both as a first-order effect, $\chi^2 =$

248.008, $p < .001$, and second-order effect, $\chi^2 = 245.554$, $p < .001$. The data for the effect of cognitive workload in Study 3 are graphed in Figure 16. As Figure 16 shows, high cognitive workload resulted in a much higher probability of human error, and this increase was curvilinear rather than parabolic.

The accuracy of visual workload as a predictor of human error was again measured by computing the area under the ROC curve (*AUC*) for each model, and testing it statistically against .5 (chance). Model A *AUC* was .808, $p < .001$, Model B *AUC* was .689, $p < .001$, and Study 3 *AUC* was .728, $p < .001$. These results indicated that visual workload was an accurate predictor of human error.

The results from Study 3 indicated that high cognitive workload contributed to a 10.1% increase in human error. The probability of human error as a function of cognitive workload graphed in Figure 16 is also broken down in Table 12. As Table 12 shows, low cognitive workload, in this case below 20, and moderate cognitive workload, in this case between 20 and 90, resulted in relatively lower probabilities of human error; 0.0020 and 0.0271, respectively. On the other hand, high cognitive workload, in this case above 90, resulted in a much higher probability of human error of 0.1254. Once again, it appears as if the effect of cognitive workload is curvilinear, not parabolic. Thus, the recommended intervention is to move tasks between stations or workers, add automation, or provide build information explicitly so cognitive workload remains below the top 10% of the score distribution (in this case below 90).

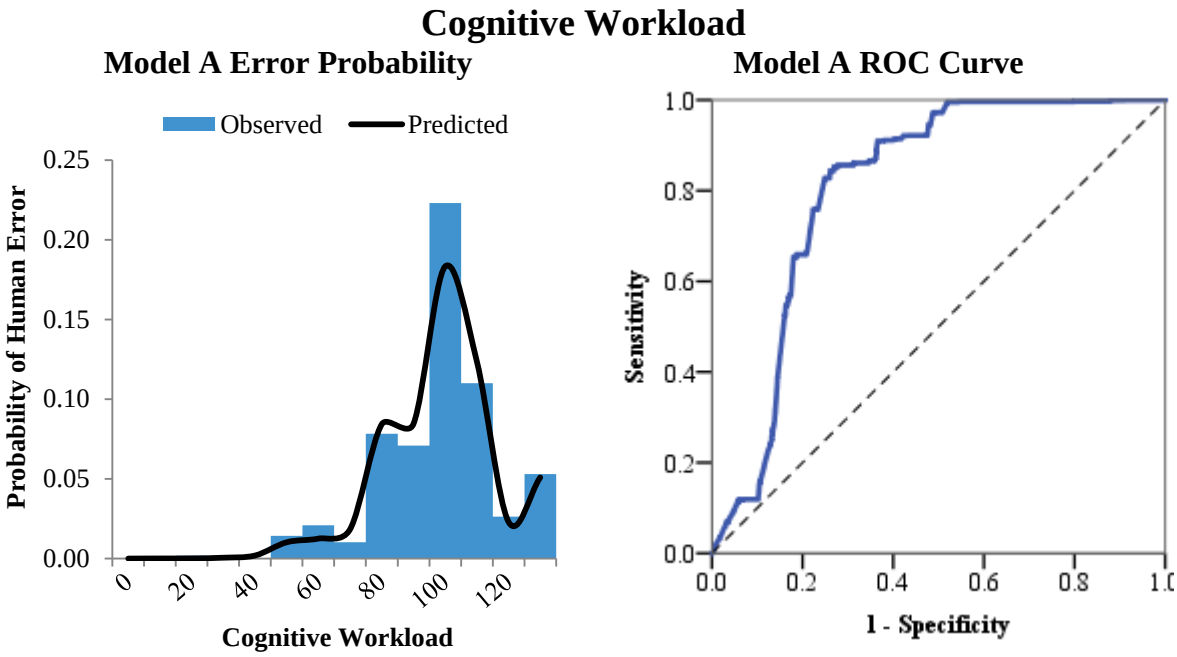

Figure 14. The effect of cognitive workload demands on the predicted and observed probability of human error in Model A (left graph), and ROC curve of human error prediction in Model A (right graph).

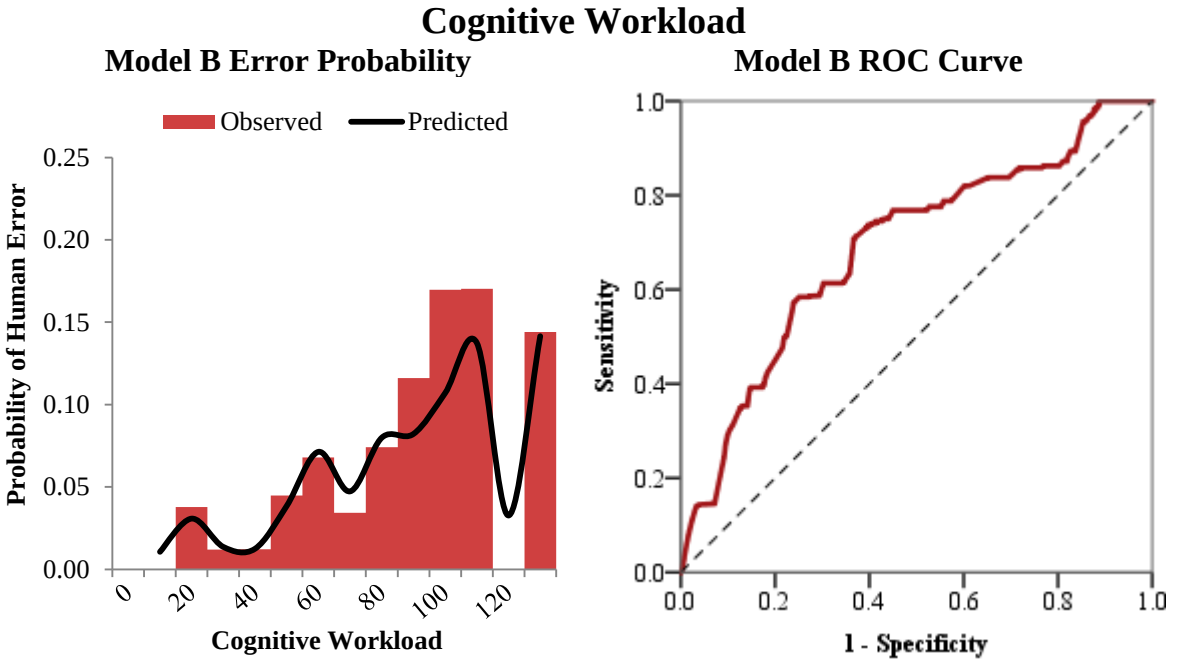

Figure 15. The effect of cognitive workload demands on the predicted and observed probability of human error in Model B (left graph), and ROC curve of human error prediction in Model B (right graph).

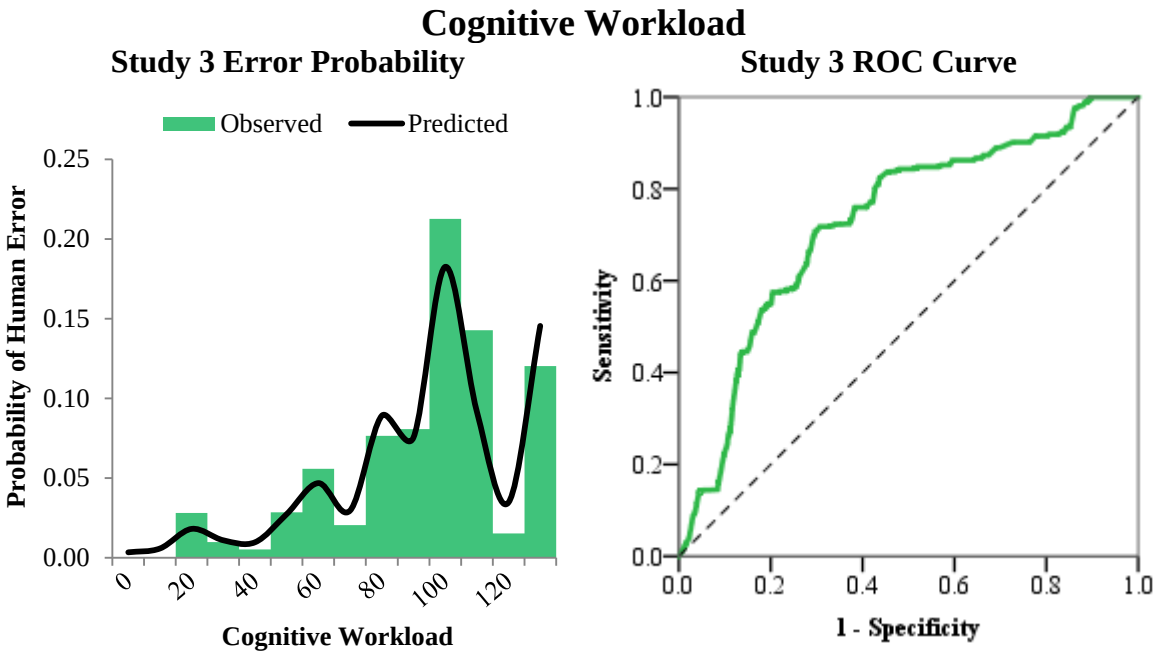

Figure 16. The effect of cognitive workload demands on the predicted and observed probability of human error in Study 3 (left graph), and ROC curve of human error prediction in Study 3 (right graph).

Table 12. *The probability of human error as a function of cognitive workload in the larger (N = 4,188) Study 3 application data set.*

Human Error *P*	Occurrence Condition
.0020	when cognitive workload was 20 or lower
.0271	when cognitive workload was between 20 and 90
.1254	when cognitive workload was higher than 90

4.2.4 Psychomotor Workload

Like the other channels, psychomotor workload was a significant predictor of human error as a first-order effect in both Model A, $\chi^2 = 22.218$ $p < .001$, and Model B, $\chi^2 = 6.158$, $p = .013$. Psychomotor workload was also significant as a second-order effect in Model A, $\chi^2 = 6.794$, $p = .009$, but not significant in Model B, $\chi^2 = 1.988$, $p = .159$.

The data for the effect of psychomotor workload in Model A are graphed in Figure 17 and the data for Model B are graphed in Figure 18. Both Figures 17 and 18 show that the probability of human error increased as psychomotor workload increased. Figure 17 shows that in Model A, the probability of human error increased more sharply after psychomotor workload demands surpassed a certain level; specifically, 100. Figure 18 shows that in Model B, the probability of human error increased more linearly and gradually. Like with cognitive workload, to better understand the effect of psychomotor workload the larger sample in Study 3 was examined. Psychomotor was once again significant as a first-order, $\chi^2 = 633.011$, $p < .001$, and second-order, $\chi^2 = 201.166$, $p < .001$, effect in Study 3. The data for the effect of psychomotor workload in Study 3 are graphed in Figure 19. As Figure 19 shows, high psychomotor workload resulted in a much higher probability of human error, and this increase was curvilinear.

The accuracy of psychomotor workload as a predictor of human error was again measured by computing the area under the ROC curve (*AUC*) for each model and testing against .5. Model A *AUC* was .764, $p < .001$, Model B *AUC* was .661, $p < .001$, and Study 3 *AUC* was .663, $p < .001$. Psychomotor workload was a fairly accurate predictor of human error and the results from Study 3 indicated that high psychomotor workload contributed to a 5.4% increase in human error. The probability of human error in the sample of 4,188 vehicles was further broken down as a function of psychomotor workload. This breakdown revealed that the probability of human error was .028 when psychomotor workload was low or moderate, in this case below 90, and .082 when psychomotor workload was high, in this case above 90. The recommended intervention is to move tasks between workers or stations, add automated tools, or add brackets or

aligning tabs to keep psychomotor workload below the top 10% of the score distribution (in this case below 90).

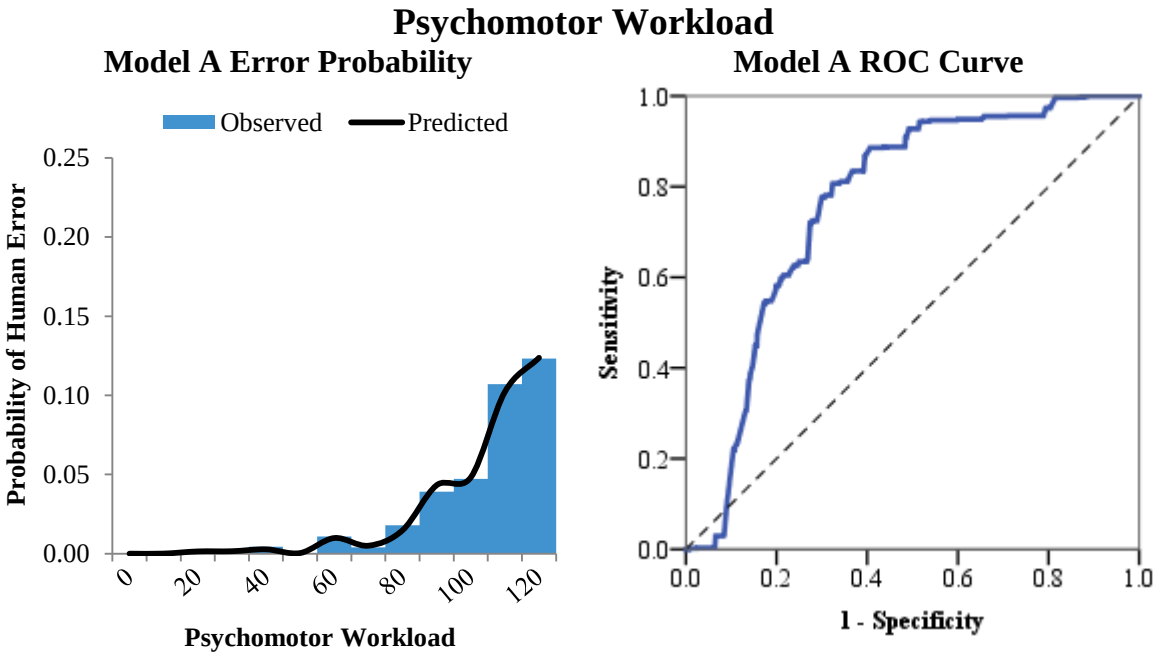

Figure 17. The effect of psychomotor workload demands on the predicted and observed probability of human error in Model A (left), and ROC curve of human error prediction in Model A (right).

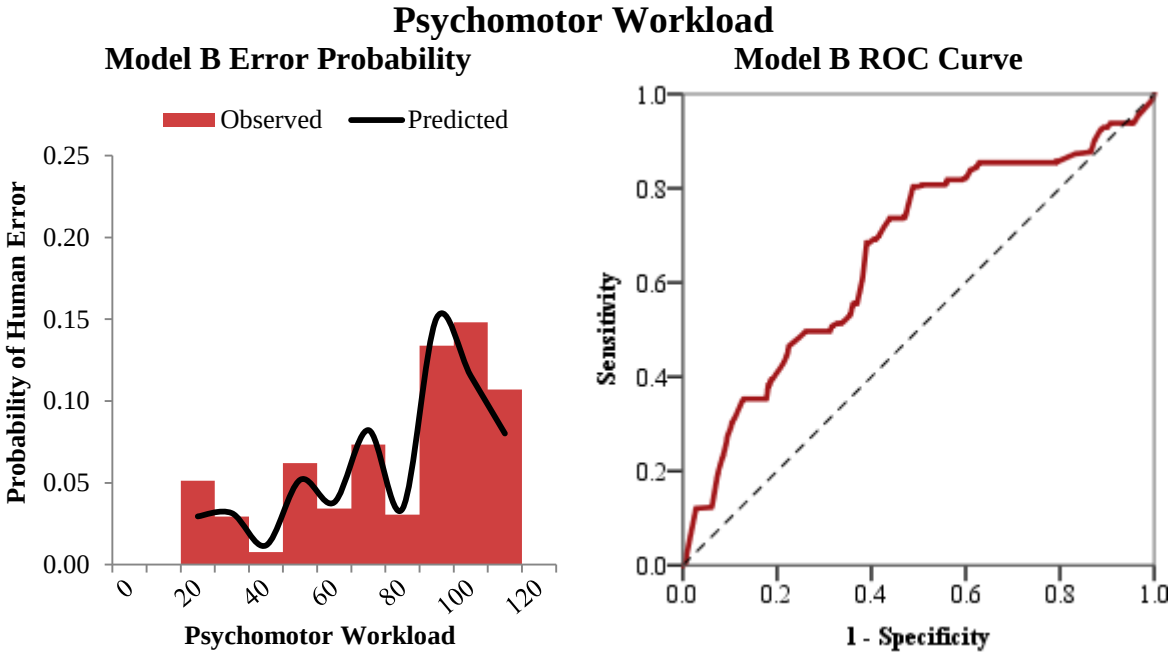

Figure 18. The effect of psychomotor workload demands on the predicted and observed probability of human error in Model B (left), and ROC curve of human error prediction in Model B (right).

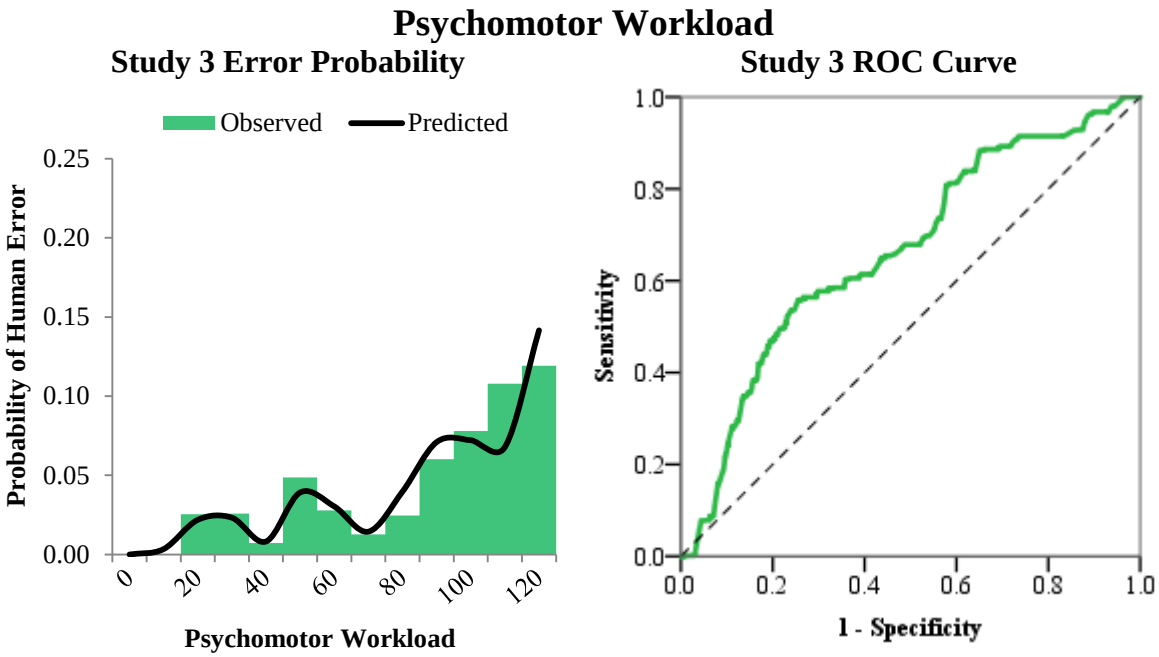

Figure 19. The effect of psychomotor workload demands on the predicted and observed probability of human error in Study 3 (left graph), and ROC curve of human error prediction in Study 3 (right graph).

4.2.5 Workload Discussion

The significant first-order effects for the visual, auditory, cognitive, and psychomotor components indicated that workload is a useful predictor of human error, particularly when broken down by resource component. Hypothesis 2 stated that the probability of human error would be higher when workload was either too low or too high. The results revealed that second-order effects of workload were useful predictors of human error; however, the pattern of the findings did not support Hypothesis 2. The patterns in the visual, cognitive, and psychomotor data indicated that the probability of human error was only higher when workload was high; not also when workload was low as Hypothesis 2 proposed. The pattern in the auditory workload data seemed to match Hypothesis 2, but the results do not offer enough support to overrule the findings of the other components because auditory workload was never truly high. For example, the

auditory workload for a build at a station never exceeded 15, whereas visual workload sometimes exceeded 350, cognitive workload sometimes exceeded 130, and psychomotor workload sometimes exceeded 120. Overall, Hypothesis 2 was not supported.

The typical ∩-shaped function used to depict the impact of workload on human performance (Wickens, 1981) was not observed in the results. Typically, low levels of workload are associated with decreased arousal, situation awareness, and performance, and high levels of workload are associated with missed signals, overload, and task shedding. This was not the case in automotive assembly tasks, specifically for low levels of workload. For example, it is possible that other factors in the automotive assembly context prevent arousal from decreasing in low workload situations. As mentioned before, each vehicle is in a workstation for a precise period of time and the tasks that must be performed depend on the options ordered on each vehicle. Even if the workload for a particular vehicle is low, the assembly line is constantly moving and new vehicles with different levels of workload enter the station every set amount of time. This pace of the assembly line may sustain arousal, even when particular vehicles have low workload demands.

Although the second-order effects of workload did not support Hypothesis 2, the results indicate that these effects are still very useful for predicting human error. Overall, the probability of human error increased as workload increased; however, this increase was not linear. The patterns of data, particularly in the visual and psychomotor components, suggested that the probability of human error sharply increased when workload demands surpassed a certain level. The second-order effects of workload significantly accounted for these non-linear relationships and appear to be useful

predictors of sharp increases in the probability of human error that occur when workload demands exceed the capabilities of workers.

4.3 Task Frequency

Task frequency was a significant predictor of human error as a first-order effect in both Model A, $\chi^2 = 6.456$, $p = .011$, and Model B, and $\chi^2 = 109.646$, $p < .001$. The patterns of this effect are plotted in Figure 20 for Model A and Figure 21 for Model B. Both Figures 20 and 21 show that as the frequency of tasks increases, the probability of human error significantly decreases. In other words, workers are more likely to make an error in tasks that are performed less frequently. The second order-effects of task frequency were also significant; $\chi^2 = 20.701$, $p < .001$ in Model A, and $\chi^2 = 23.755$, $p < .001$ in Model B. These significant effects indicate that the increase in the probability of human error for infrequent tasks may be curvilinear; workers were most likely to make errors in tasks that were infrequent or rare. This finding is consistent with studies in the literature that have found that rare and unexpected tasks are often failed to be noticed even when triggering information is salient, misinterpreted or misdiagnosed, or inappropriately executed (Wickens et al., 2009).

The results from Study 3 matched the results from Model A and Model B; Task frequency was significant as both a first-order, $\chi^2 = 13.026$, $p < .001$, and second-order effect, $\chi^2 = 100.996$, $p < .001$. The effect of task frequency in Study 3 is graphed in Figure 22. Looking at Figure 22, once again there was a curvilinear decrease in the probability of human error as the task frequency increased.

Hypothesis 3 proposed that there is an optimum frequency of tasks, such that the probability of human error is higher when tasks are either performed repeatedly or

infrequently. The findings did not support this hypothesis. The second-order effects indicated that the probability of human error was only higher for tasks that were performed less frequently. The probability of human error was not higher in tasks that were performed more frequently. In fact, tasks that were performed on between 70% to 100% of vehicles had a probability of human error of less than 0.01. Typically, tasks that are performed frequently and repetitively can impact performance by becoming mindless and boring (Rasmussen, 1982). This did not appear to be the case in automotive assembly. It is possible that the high customizability of vehicles and variability in the total number of tasks required keep tasks from becoming monotonous, even when some are performed on every single vehicle. As Figure 22 shows, there was no optimum task frequency; the probability of human error was higher in infrequent tasks but otherwise relatively low, stable, and comparable for other task frequencies. Hypothesis 3 was refuted.

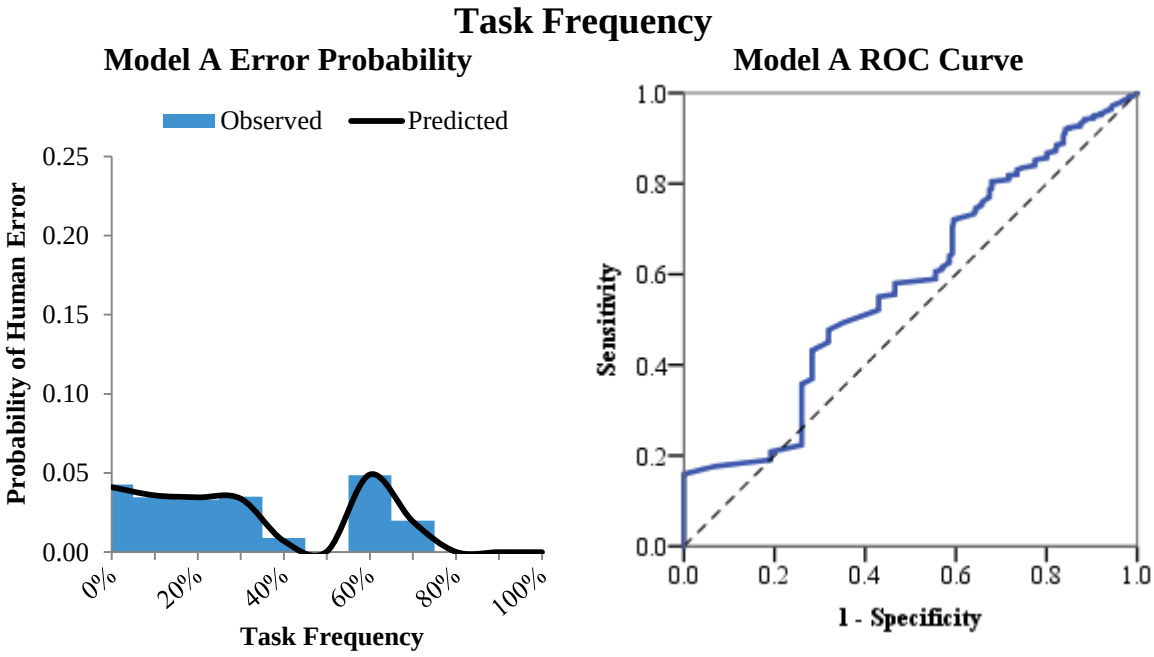

Figure 20. The effect of task frequency on the predicted and observed probability of human error in Model A (left graph), and ROC curve of human error prediction in Model A (right graph).

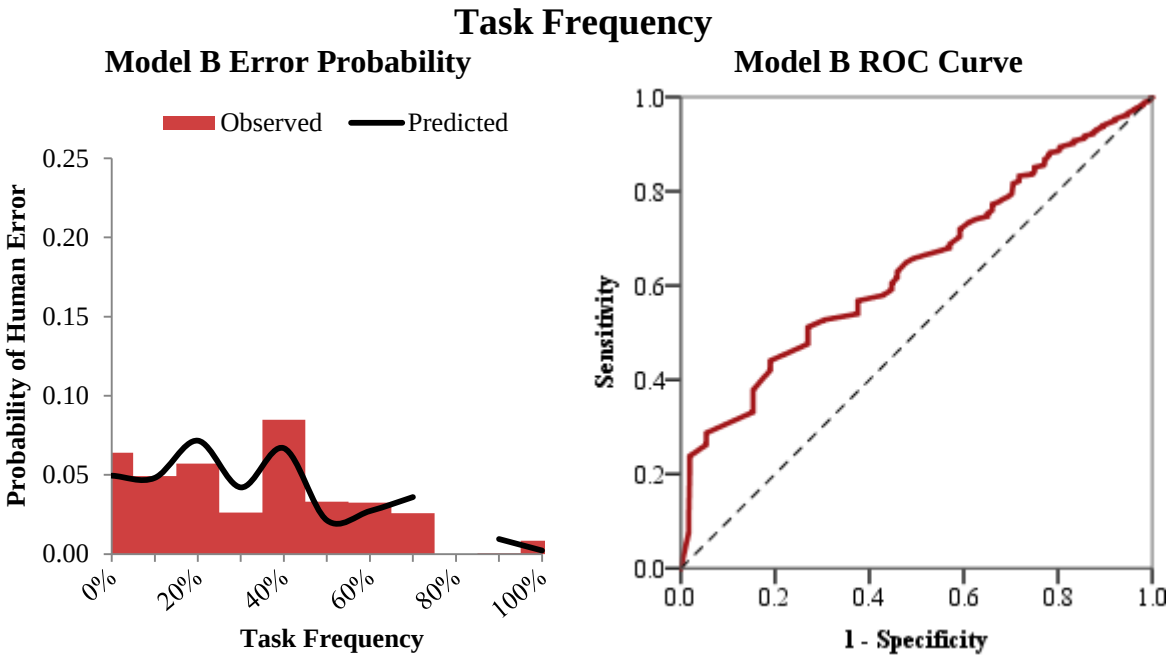

Figure 21. The effect of task frequency on the predicted and observed probability of human error in Model B (left graph), and ROC curve of human error prediction in Model B (right graph).

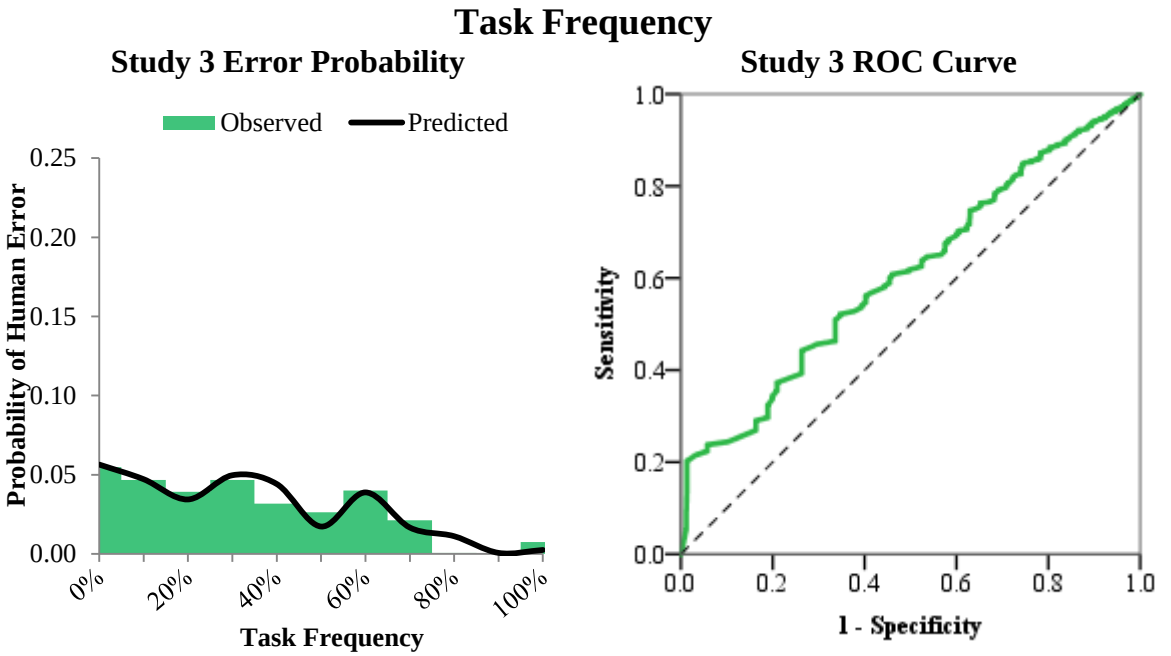

Figure 22. The effect of task frequency on the predicted and observed probability of human error in Study 3 (left graph), and ROC curve of human error prediction in Study 3 (right graph).

The accuracy of task frequency as a predictor of human error was again measured by computing the area under the ROC curve (*AUC*) for each model, and testing it statistically against .5. Accuracy was found to be quite poor; Model A *AUC* was .577, $p < .001$, Model B *AUC* was .639, $p < .001$, and Study 3 *AUC* was .609, $p < .001$. Nevertheless, task frequency was still a significant predictor and found to contribute to a 1.4% increase in human error. To better understand this effect, the raw number of human errors that occurred for each task frequency in the in the 4,188 vehicle set is graphed in Figure 23. The percentages of human error that occurred as a function of different task frequencies percentages are also summarized in Table 13. These data indicated that 37.7% of all errors occurred in tasks that were performed less than 20% of the time.

The recommended intervention is to use the existing system on the assembly line to alert workers when vehicles require infrequent processes. The alerts currently on the line to alert for certain tasks, models, or variants should be expanded to include all tasks performed on less than 20% of vehicles that enter a certain station. These alerts should include flashing relevant build information on visual display screens, as well as auditory alerts where possible.

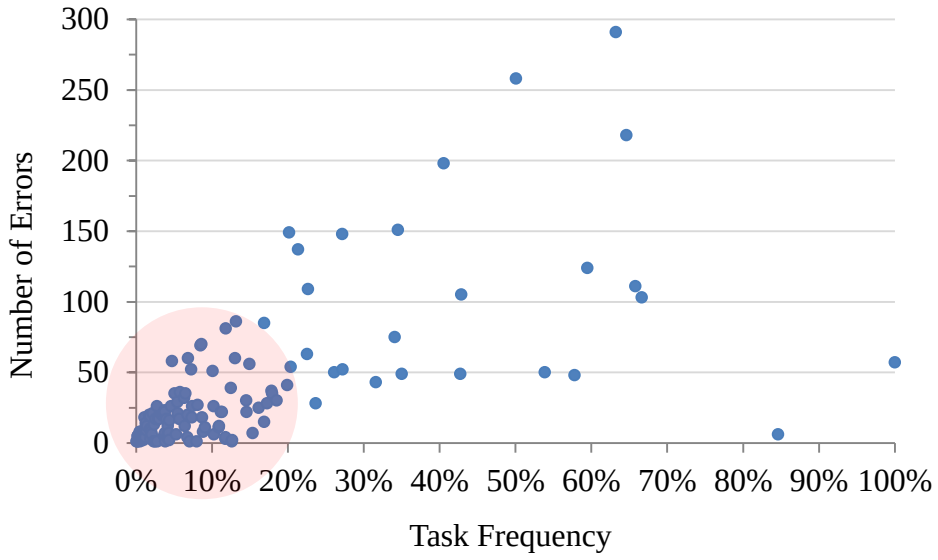

Figure 23. The raw number of human errors that occurred for each task frequency in the larger (*N* = 4,188) application data set in Study 3.

Table 13

The percentages of human error that occurred as a function of different task frequencies in the larger (N = 4,188) application data set in Study 3.

% All Defects	Occurrence Condition
37.7%	Occur when task frequency is less than 20%
20.0%	Occur when task frequency is between 20 - 40%
15.0%	Occur when task frequency is between 40 - 60%
13.1%	Occur when task frequency is between 60 - 80%
01.1%	Occur when task frequency is between 80 - 100%

4.4 Shift

Hypothesis 4 proposed that the probability of human error is higher during night shift than during day shift. The effect of shift was not significant in Model A, $\chi^2 = 0.976$, $p = .323$, Model B, $\chi^2 = 0.114$, $p = .736$, or Study 3 $\chi^2 = 1.257$, $p = .262$. The means and

standard deviations of the probability of human error in each shift for all three models are included in Table 14. The results from all three models indicated that the major impairments in human performance during night shift that have been reported in other studies (Folkard & Monk, 1980) were not found to be predictive of human errors in automotive assembly. The variable of shift does not appear to be a very useful predictor in this context.

Table 14

The means and standard deviations of the probability of human error in each shift (day, night) in each study model.

	Model A		Model B		Study 3	
Shift	*M*	*SD*	*M*	*SD*	*M*	*SD*
Day	.0299	.0604	.0376	.0499	.0344	.0548
Night	.0279	.0574	.0379	.0508	.0336	.0551

4.5 Hours into Shift

Hypothesis 5 proposed that there is a peak period of time in each shift during which the probability of human error is highest. This hypothesis was not supported. In Model A, the effect of hours into shift was not a significant predictor of human error as a first-order, $\chi^2 = 0.039$, $p = .843$, second-order, $\chi^2 = 0.025$, $p = .875$, or third-order effect, $\chi^2 = 0.018$, $p = .894$. The same was true in Model B; the effect of hours into shift was not a significant predictor of human error as a first-order, $\chi^2 = 0.613$, $p = .434$, second-order, $\chi^2 = 1.009$, $p = .315$, or third-order effect, $\chi^2 = 1.115$, $p = .291$. Study 3 found the same results; hours into shift was not a significant as a first-order, $\chi^2 = 0.039$, $p = .843$, second-

order, $\chi^2 = 0.199$, $p = .655$, or third-order effect, $\chi^2 = 0.256$, $p = .613$. The data showing the probability of human error by hours into shift is graphed in Figure 24. As Figure 24 shows, the probability of human error was relatively the same throughout the duration of each shift in each model. The peak in errors that other studies have found between the hours of 2 a.m. and 4 a.m., and 2 p.m. and 4 p.m. (Mitler et al., 1988) was not found in human errors in automotive assembly. Hours into shift does not appear to be a useful predictor of human error in this context.

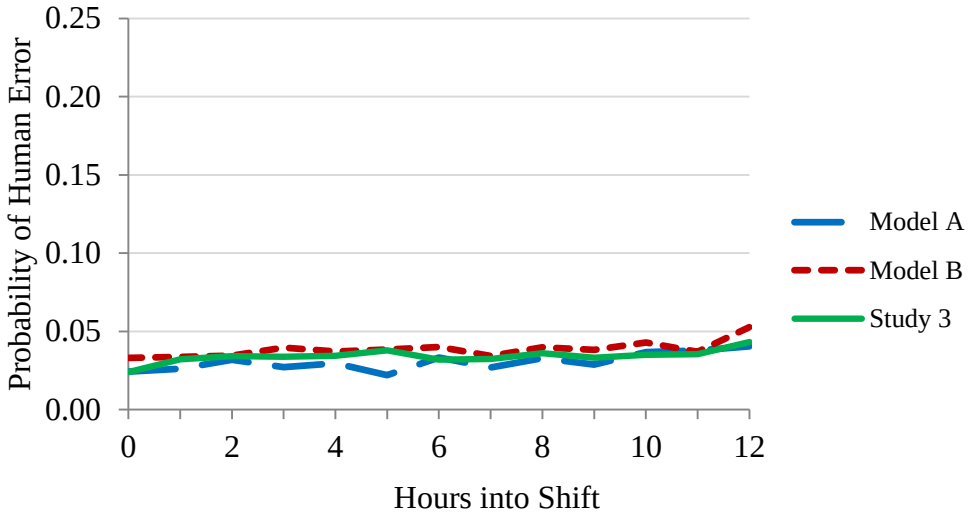

Figure 24. The effects of hours into shift on the probability of human error in all models from Study 1 (Models A and B) and Study 3..

4.6 Information Flow

Hypothesis 6 proposed that the probability of human error would be higher when the flow of information was implicit rather than explicit. The effect of information flow was significant in Model A, $\chi^2 = 23.446$, $p < .001$, Model B, $\chi^2 = 28.749$, $p < .001$, and Study 3 $\chi^2 = 337.858$, $p < .001$. The pattern of data supported Hypothesis 6. In Model A,

the probability of human error was higher when the flow of information was implicit, *M* = 0.0320, *SD* = 0.176, than explicit, *M* = 0.0219, *SD* = 0.146. In Model B, the probability of human error was higher when the flow of information was implicit, *M* = 0.0551, *SD* = 0.228, than explicit, *M* = 0.0215, *SD* = 0.145. In Study 3, again, the probability of human error was higher when the flow of information was implicit, *M* = 0.0422, *SD* = 0.061, than explicit, *M* = 0.0227, *SD* = 0.041. Presumably, the additional mental processing demands required to interpret implicit information led to higher likelihoods of human error (Seminara, Gonzalez, & Parsons, 1976; Welford, 1976). The findings indicate that information flow was a useful predictor of human error. Human error interventions should focus on transforming implicit information into explicit information whenever possible.

4.7 Information Presentation

Hypothesis 7 proposed that the probability of human error would be higher when information was presented visually than auditorily. The effect of information presentation was significant in Model A, $\chi^2 = 25.877$, $p < .001$, and Study 3, $\chi^2 = 205.261$, $p < .001$, but not Model B, $\chi^2 = 0.000$, $p = .983$. The means and standard deviations of the probability of human error for each modality of information presentation are included in Table 15. The results from all three models indicated that the probability of human error was indeed higher when information was presented visually rather than auditorally. These results reflect the findings from Human Reliability Assessments that have demonstrated that the modality of information presentation influences the likelihood of error (Kryter, 1972).

Hypothesis 7 was supported by the findings of two out of the three models. As a results, the suggested intervention for reducing human error in automotive assembly is to, when possible, add auditory information displays for relevant build or task information.

Table 15

The means and standard deviations of the probability of human error for each modality of information presentation in each study model.

	Model A		Model B		Study 3	
	M	*SD*	*M*	*SD*	*M*	*SD*
Visual	.0321	.0577	.0407	.0507	.0379	.0551
Auditory	.0225	.0624	.0030	.0082	.0166	.0502

4.8 Task Dependency

Hypothesis 8 proposed that the probability of human error would be higher when tasks were interdependent and required coordination between workers rather than independent. The effect of task dependency was not significant in Model A, $\chi^2 = 2.293$, $p = .130$, but was significant in Model B, $\chi^2 = 109.646$, $p < .001$, and in Study 3, $\chi^2 = 144.294$, $p < .001$. The pattern of data, however, was the opposite of what was hypothesized. In Model A, the probability of human error was lower when tasks required coordination between workers, $M = 0.01300$, $SD = 0.113$, than when tasks did not, $M = 0.03539$, $SD = 0.185$. The same was true in Model B; the probability of human error was lower when tasks required coordination between workers, $M = 0.03500$, $SD = 0.184$, than when tasks did not, $M = 0.03813$, $SD = 0.192$. In Study 3, the probability of human error was again lower when tasks required coordination between workers, $M = 0.021645$, $SD = 0.038$, than when tasks did not, $M = 0.03739$, $SD = 0.579$.

The hypothesis stemmed from calculations in NUREG-75/014 (Nuclear Regulatory Commission, 1975) that indicated that performance of tasks that require coordination between workers is highly interdependent. Specifically, correct performance by one worker increases the probably of correct performance by the other worker; whereas, incorrect performance by one worker increases the probability of incorrect performance by the other worker (Nuclear Regulatory Commission, 1975). Hypothesis 8 proposed that these interdependent tasks would have a higher probability of human error because incorrect performance by one worker would increase the probability of incorrect performance by the other worker. The results indicated that this was not the case. The analyses from both Model B and Study 3 revealed that interdependent tasks had a significantly lower probability of human error than independent tasks; thus, Hypothesis 8 was refuted. It appears as if interdependent tasks may have a lower probability of human error because correct performance by one worker increases the probably of correct performance by the other worker.

The relationship between task interdependence and performance has been examined in the literature. The pattern of this relationship has been demonstrated to vary based on the behavioral-conceptual dimension of tasks (Stewart & Barrick, 2000). Tasks can be classified according to the extent to which they are either *behavioral*, focusing on the overt execution of manual and psychomotor work; or *conceptual*, focusing on planning, choosing, generating ideas, negotiating, deciding, and problem solving (McGrath, 1984). In conceptual tasks, the relationship between interdependence and performance has followed a ∪-shaped function, and performance was highest when interdependence was either low, and workers operated as individuals; or high, and

workers operated cooperatively as a team (Stewart & Barrick, 2000). In behavioral tasks, the relationship between interdependence and performance follows a ∩-shaped function, and performance is highest when interdependence is moderate and tasks depend on each other yet are clearly defined, dialog is required but intermittent, interactions are necessary but minimal, and information is centralized (Stewart & Barrick, 2000).

In order to better understand why the probability of human error was lower when tasks were interdependent, the specific tasks that required coordination with other workers were examined in both target areas. The tasks fell into three common categories of coordinating with other workers to place and install vehicle components that are large (e.g., carpet, insulation, headliner), align and install parts symmetrically using brackets, or route and secure components of the electrical wiring harness. The tasks in these categories all focus on the execution of manual and psychomotor tasks, and can all be classified as behavioral tasks. The type of coordination required between workers in these tasks can also be classified as moderate interdependence. The ∩-shaped relationship between interdependence and performance for behavioral tasks indicates that this type of moderate interdependence is associated with the highest level of performance. On the other hand, tasks that do not require coordination between workers have low to no interdependence, and as the ∩-shaped relationship indicates, are associated lower levels of performance. This may be why interdependent tasks had significantly lower probabilities of human error than independent tasks.

4.9 Teamwork

Hypothesis 9 proposed that the probability of human error would decrease as the number of workers per station increased. The effect of teamwork was a significant

predictor of human error in Model A, $\chi^2 = 14.330$, $p < .001$, Model B, $\chi^2 = 33.598$, $p < .001$, and Study 3, $\chi^2 = 173.889$, $p < .001$. The data for the effect of teamwork in Model A, Model B, and Study 3 are graphed in Figure 25, Figure 26, and Figure 27, respectively. As these three figures show, it appears as if for stations that had more than one worker, higher numbers of workers at a station had lower probabilities of human error. Hypothesis 9 was supported. The results support the notion that higher numbers of workers at stations put more workers in positions that allow them to observe each other's work, provide performance feedback, distribute high task loads, help each other not fall behind, catch up, and ultimately detect and recover from errors.

The accuracy of teamwork as a predictor of human error was measured by computing the area under the ROC curve (*AUC*) in each model and testing it statistically against .5. The results indicated that teamwork was a very poor predictor of human error, sometimes performing significantly worse than chance: Model A *AUC* was .584, $p < .001$, Model B *AUC* was .428, $p < .001$, and Study 3 *AUC* was .495, $p = .229$. Thus while teamwork may account for a significant portion of the variance in human error, it cannot predict the occurrence of human errors in the current context. As a result, no recommendation is made for teamwork in the reduction of human error in automotive assembly.

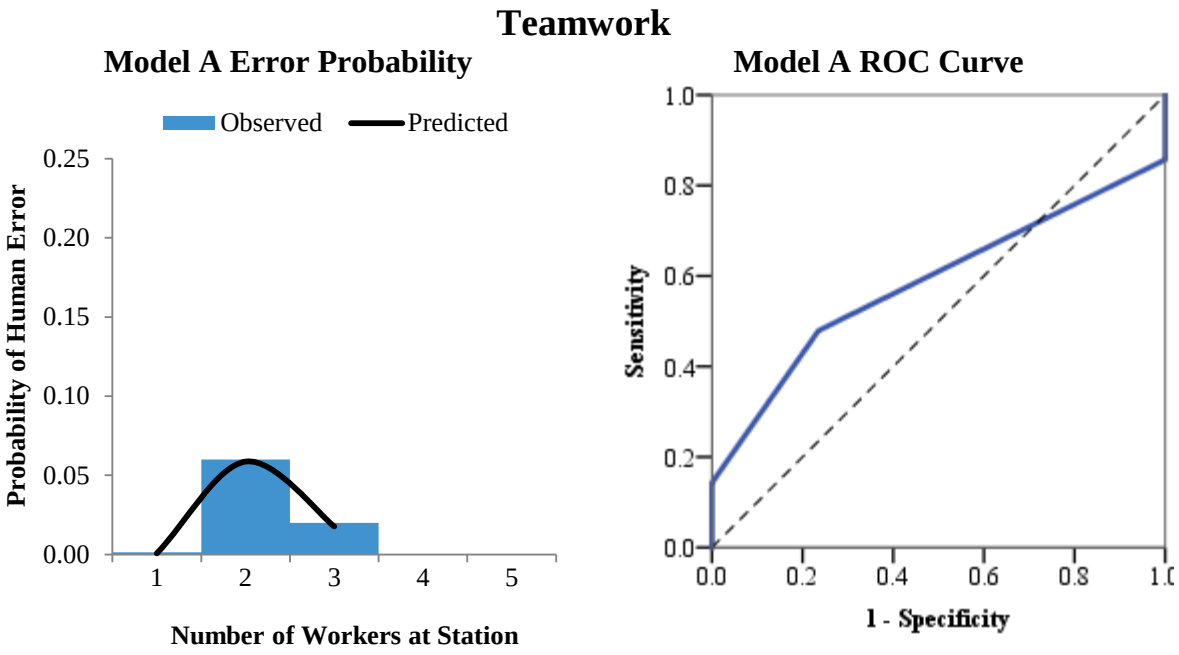

Figure 25. The effect of number of workers at station on the predicted and observed probability of human error in Model A (left graph), and ROC curve of human error prediction in Model A (right graph).

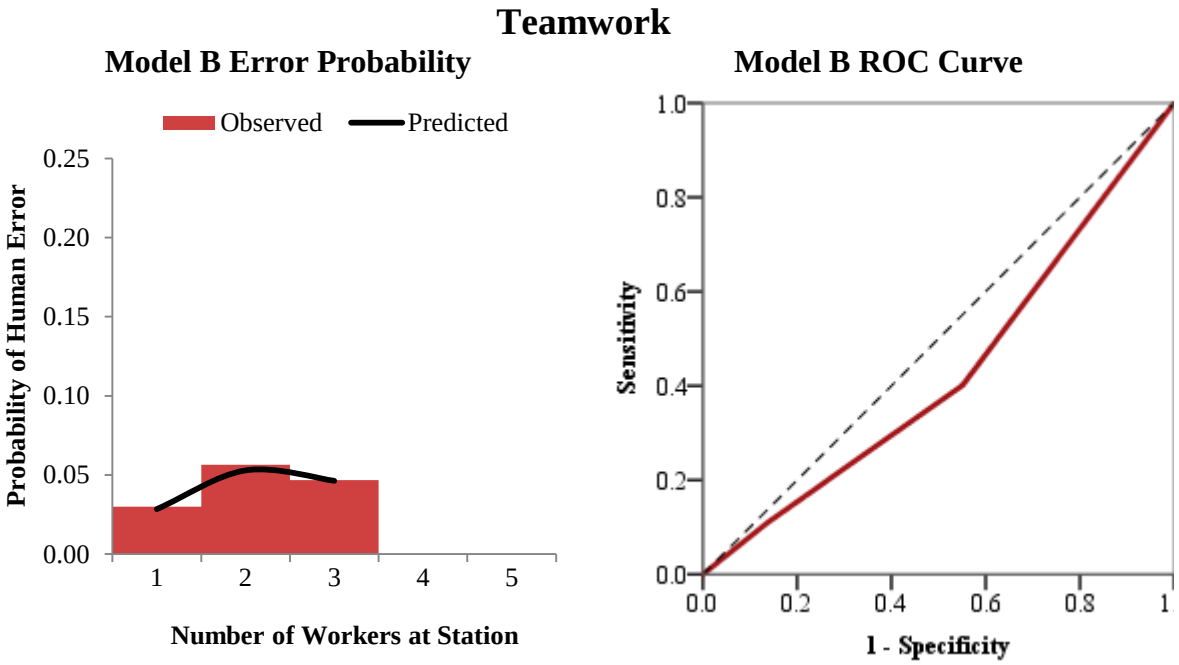

Figure 26. The effect of number of workers at station on the predicted and observed probability of human error in Model B (left graph), and ROC curve of human error prediction in Model B (right graph).

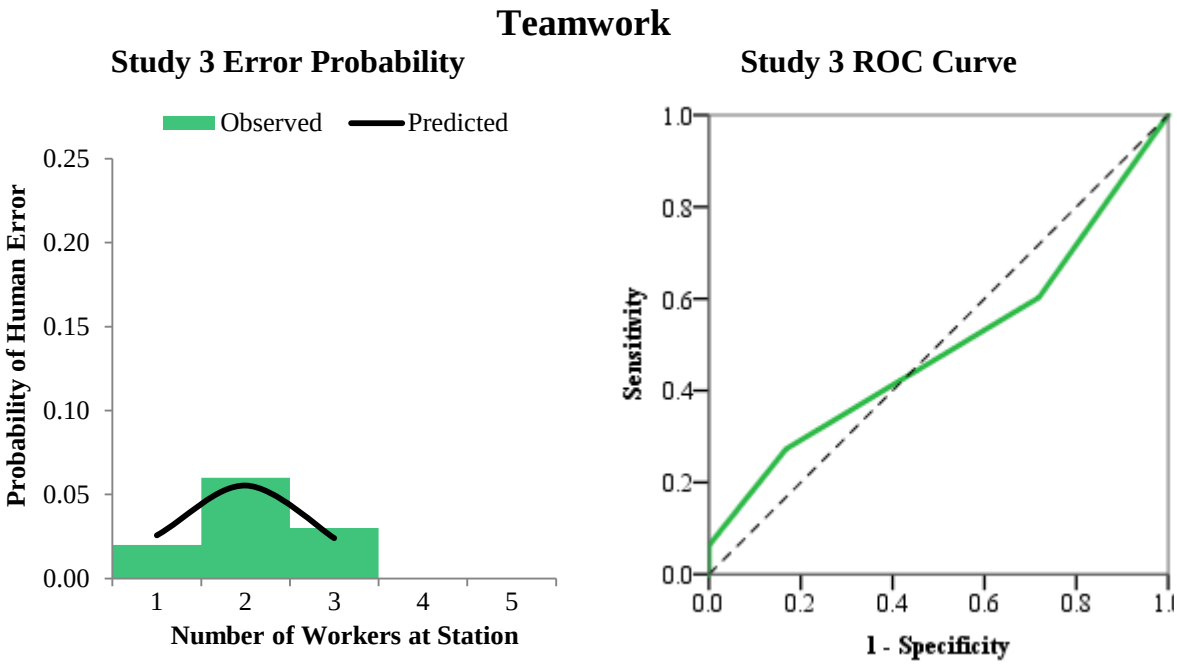

Figure 27. The effect of number of workers at station on the predicted and observed probability of human error in Study 3 (left graph), and ROC curve of human error prediction in Study 3 (right graph).

4.10 Equipment Feedback

Hypothesis 10 proposed that the probability of human error would be lower at workstations with equipment that provided immediate performance feedback, than at stations without such equipment. The effect of equipment feedback was significant in both Model A, $\chi^2 = 159.904$, $p < .001$, Model B, $\chi^2 = 72.710$, $p < .001$, and Study 3, $\chi^2 = 865.527$, $p < .001$. The pattern of data supported Hypothesis 10. In Model A, the probability of human error was lower when stations contained automated tools, fasteners, scanners, lifts, or other equipment that provided feedback about task performance, $M = 0.01494$, $SD = 0.113$, than when stations did not contain any such equipment to provide performance feedback, $M = 0.02175$, $SD = 0.146$. The same was true in Model B: The probability of human error was lower when stations contained automated tools, fasteners,

scanners, lifts, or other equipment that provided feedback about task performance, *M* = 0.01086, *SD* = 0.108, than when stations did not contain any such equipment to provide performance feedback, *M* = 0.05548, *SD* = 0.220. Once again, the same was true in Study 3: The probability of human error was lower when stations contained equipment that provided feedback, *M* = 0.00501, *SD* = 0.011, than when stations did not contain any such equipment, *M* = 0.059755, *SD* = 0.064.

Lack of equipment feedback was the largest influence on the odds of human error occurrence in this context. Of the 4,811 human errors that occurred in the application vehicle set in Study 3, 26.78% of human errors occurred at stations with automated or other equipment that provided feedback about task performance, while 73.22% of human errors occurred at stations that did not contain any equipment to provide performance feedback. The actual probability of human error was .005 in stations with equipment feedback, and .060 in stations without equipment feedback, suggesting that lack of equipment feedback contributed to a 5.5% increase in human error.

The recommended intervention was to expand the automated tools, fasteners, scanners, lifts, or other equipment that provided performance feedback to as many stations as possible.

The connection found between equipment feedback and human error in automotive assembly is consistent with the perspective in the literature of human error as a systems phenomenon (Chemical Process Safety, 1994; Senders & Moray, 1991). The strong association between the presence of equipment feedback and the occurrence of human error that was indicated by comparatively large odds ratios in all three models support the view that human error is systematically connected to the tools and tasks

(Dekker, 2002). In this case, the feedback provided by the tools and equipment fostered the type of learning that has been discussed as leading to the reduction of human error. In order to reduce human error, operators should be provided with knowledge of the results of their actions; otherwise, there is no signal of when to modify their behavior (Reason, 1990; Senders & Moray, 1991). Knowledge of results is most effective in reducing human error when it is immediate, to the point, and nonjudgmental, so operators can correct the actions that were erroneous (Senders & Moray, 1991). Tools and equipment that provide feedback give operators this exact type of knowledge of results: immediate, nonjudgmental, and to the point of specific actions. Operators can then immediately modify their behavior to avoid errors. Operators at stations without tools and equipment that provide feedback can only modify their behavior if they detect and correct their own error, or receive knowledge of results from quality control or their supervisor, which may be delayed until an error occurs several times and may be judgmental.

CHAPTER 5:

CONCLUDING THOUGHTS

The results of this study indicated that an expanded task network architecture has the potential to fulfill the need for better prediction of human error identified by NASA's System-Wide Accident Prevention Program (Leiden et al., 2001). The models developed in this study accounted for 21.9% to 36.5% of the variance in human error. Task network modeling demonstrated to be a promising approach that goes beyond taxonomies and has evidenced to provide accurate and sensitive predictions of human error.

The prediction equations from each model and the application study have important implications for both human error reduction and human reliability analysis. The abundance of performance and error data available for this study made it possible to not only test the significance and fit of the prediction equations, but also rigorously test their validity and stability. The validation procedure in this study involved the computation of 539,000 human error probabilities. These probability computations were used to assess the stability of predictors, as well as the sensitivity and specificity of task network modeling as a probabilistic risk assessment tool. The results were favorable, with accuracy and sensitivity within each area ranging from fair to excellent. The stability of predictors across all three study models is summarized in Table 16. Predictors that were significant, either as a first-order effect, second-order effect, or both are marked with a green checkmark. Predictors that were not significant are marked with a red X. As Table 16 shows, the predictors of Time Pressure, Visual Workload, Auditory Workload,

Cognitive Workload, Psychomotor Workload, Task Frequency, Information Flow, Teamwork, and Equipment Feedback were significant predictors of human error in all three models. The variables of Information Presentation and Task Dependency varied in significance between Models A and B, but both were significant in the larger sample from Study 3. Overall, these two variables were significant in two out of three models. The variables of Shift and Hours into Shift were never significant in any of the three models. The pattern in the significance and stability of model predictors across all three studies is a great indicator of the capability of task network modeling as an error prediction tool. Interestingly, the variables that were greatly stable across studies were all related to the tasks being performed by each worker at each station. For example, what was the time pressure of the tasks, the workload of the tasks, the frequency of the tasks, the information displayed for the tasks, the equipment used for the tasks, the number of workers performing the task, and so on. The variables related to the timing of errors, on the other hand, were never significant. Task network modeling was not useful in predicting the shift and the hour within the shift when errors would occur. Thus it appears as if task network modeling is a great tool for predicting the situations and circumstances in which human errors will occur, but not the timing of when errors will occur.

Table 16

The significance of variables as predictors across all three study models.

Variable	Model A	Model B	Study 3
Time Pressure	✓	✓	✓
Visual Workload	✓	✓	✓
Auditory Workload	✓	✓	✓
Cognitive Workload	✓	✓	✓
Psychomotor Workload	✓	✓	✓
Task Frequency	✓	✓	✓
Shift	X	X	X
Hours into Shift	X	X	X
Information Flow	✓	✓	✓
Information Presentation	✓	X	✓
Task Dependency	X	✓	✓
Teamwork	✓	✓	✓
Equipment Feedback	✓	✓	✓

The results of this study are also beneficial for better understanding the situational variables that can affect the occurrence of errors. Sharit's (2006) modeling framework demonstrated how human error arises from a mismatch between human capabilities and task demands. The models in this study have identified 12 of these factors that significantly predict human error. This is only a beginning. The analyses indicated that deviance was high in both models and the application study, and that model fit could be improved. The contextual component of Sharit's model contains a variety of other situational variables that can affect the occurrence of errors. These variables should be used to further develop the set of predictors that are added to the task network modeling architecture to predict human error.

The odds ratios and AUCs in the analyses of individual predictors were particularly useful for indicating the strength of the relationship between each situational factor and human error. Assessors can use the odds ratios and AUCs to rank the biggest influences to human error in a system and develop more effective interventions for reducing human error. This procedure was used to develop a set of interventions that are currently being implemented in the operational automotive assembly plant from which the process sheets and error data used in this study originated. The influence of these interventions on human error is being tracked over the course of two months and the results will be used to further evaluate the effectiveness of task network modeling as an approach to human error reduction.

The value of task-network modeling human error in general lies in possibility to investigate beyond just individual error cases. Task network modeling makes it possible to take individual error cases and link them with the process during which they occur within the network. This type of link indicates where in the system something went wrong and then allows the investigation of what factors contributed. The major advantage of this type of approach is that it can use error data that already exists. Furthermore, these data allow the investigation of circumstances in which errors did not occur, in the same level of detail as circumstance in which errors occurred. This advantage makes task network modeling especially useful for investigating rare or infrequent events during which human errors occur. These networks, however, require large amounts of error data. Models A and B each had 1,000 vehicles within the sample and some data ranges were still never observed and thus could not be modeled. The application study had 4,188

vehicles and still had a few data ranges that were missing. Thus, care should be taken to examine the range of values within samples to ensure that all predictors can be modeled.

Task-network modeling is also valuable within the specific context of automotive assembly. This type of approach to human error reduction utilizes process sheets, timing data, and error account data that already exist within the assembly plant. Task-network models make it possible to track human errors through the assembly line and pinpoint exactly where these errors occurred. The major advantage to this type of approach is the ability to track errors across time, shifts, workers, and rotation schedules, which was not previously possible because errors were only documented by vehicle, worker, and timestamp. This type of error analysis allows investigators to go beyond just individual errors, and look at patterns that may identify weakness or shortcomings within the system. Human error interventions can then be developed to overcome these weaknesses and reduce the probability of human error. The end result is fewer instances of human error that in the current context had the greatest impacts on customer safety and satisfaction.

The applications of the proposed study can be especially valuable outside the research community. Task network modeling has demonstrated to be an effective method for predicting human error. The task network modeling architecture should be further expanded and applied to other safety critical industries. In the current application, task-network modeling worked well for predicting the circumstances in which errors would occur. Prediction accuracy was also highest when errors were predicting within the tasks that the network models were designed for, which is somewhat expected, especially with regression modeling. Nevertheless, the underlying finding is that task network models in

particular are poor at correctly identifying the situations in which errors *do not* occur when predicting to tasks that were not originally included within the model. Thus, the utility of predictions of expanded task network models is limited in generalization to the area or domain the model is based on.

The task networking modeling architecture if flexible enough to be applied to a wide variety of systems as long as a task analysis is performed and timing data are available or can be collected. Task network modeling can particularly be expanded in the nuclear industry for the development of next generation control room configurations (Boring, 2006). Better prediction of human error and more accurate probabilistic risk assessment are essential for reducing human error and preventing human error disasters in the future. Task network modeling should be an integral part of the approach to human error reduction.

APPENDIX A:

MICROSAINT CODES

To be able to predict human error, 14 additional variables were added to the task modeling architecture in Micro Saint Sharp. This was done by right clicking the 'Variables' node in the tree view at the right of the screen, and selecting 'Add Variable' from the menu that displayed. The name of each variable was then entered using the dialog box that displayed. After the variables were added, their names appeared in the tree view along with the five aforementioned variables that are automatically created by Micro Saint Sharp. The parameters for each variable were then entered by double-clicking the variable name and opening the 'Variable Description' dialog box.

The expanded Micro Saint Sharp architecture was then used to construct separate task network models for each target area based on the data compiled in the Excel worksheet. The model for target area A was constructed first. The point and click graphical interface was used to turn each of the 2,950 tasks identified in the task analysis into separate tasks within the network. The tasks were organized by station and then connected according to the sequence of tasks in the assembly process. For each task, parameters were entered by double clicking the task and opening the 'Task Description' dialog box. These parameters were used to represent how each task influenced the 14 variables being tracked in an effort to predict human error. The specific code for each parameter is discussed by the different tabs in the 'Task Description' dialog box; the 'Main' tab, 'Timing' tab, and 'Paths' tab.

In the 'Main' tab of the 'Task Description' box, the 'Beginning Effects' parameter was used to increment each variable as follows: (Note that 'number' is used in place of the actual numerical value in the corresponding column of the Excel worksheet)

```
\\Beginning Effect
Time_Utilization += (Entity.Duration/1.913); //task time over time available.
Visual_Workload += number;
Auditory_Workload += number;
Cognitive_Workload += number;
Psychomotor_Workload += number;
Task_Frequency = number; //probability of performing task.
Shift = Entity.Error_Shift; //1 day shift, 2 night shift.
Time_into_shift = (Entity.Build_Time + Clock);
Information_Flow = number; //1 implicit, 2 explicit.
Information_Presentation = number; //1 visual, 2 auditory.
Task_Dependency = number; //1 not required, 2 required.
Teamwork = number; //number of workers at station.
Equipment_Feedback = number; //1 not provided, 2 provided.
if  (Entity.Error_Location == Group.ID)
        {
                Human_Error = 1;
        }
else
        {
                Human.Error = 0;
        }
```

In the 'Timing' tab of the 'Task Description' dialog box, the 'Mean' parameter was used to enter the data for the mean time associated with each task. The timing expression was entered as follows:

```
return number; //the time duration of each task to 3 decimal places.
```

In the 'Paths' tab of the 'Task Description' dialog box, the parameters of 'Decision Type' and 'Decision Code' were used to define the decision logic whenever tasks had more than one possible path emerging from them. This occurred under two conditions; either the task was spilt into multiple paths because two or more workers were working on a vehicle at the same time, or the task was split into multiple paths because the next task that needed to be performed depended on the build options of the vehicle. Different parameters were used for each type of decision. For modeling tasks that were executed simultaneously by two or more workers at a station the 'Decision Type' was set as 'Multiple' and the decision code for each path was set as 'return true;'. This allowed Micro Saint Sharp to split each entity into two or more task paths at each station. For modeling tasks that were only executed if the vehicle had a certain option, the 'Decision Type' was set as 'Tactical' and the decision code for the task path was a conditional expression. For example, if a task was only executed if the vehicle build had option ABC, the code was as follows:

```
if (Entity.Build_Information.Contains("ABC")) //Where ABC is option code.
        {
                return true;
        }
else
        {
                return false;
        }
```

The task parameters outlined were used to increment the 14 variables that were proposed as predictors of human error. These 14 variables were tracked throughout the simulation of each vehicle and their values were recorded for each worker at each station on the assembly line. These data were recorded by defining snapshots within the execution of the model to collect the value of each variable at the end of the last task being performed by each worker at each station. These snapshots were added by right clicking the 'Snapshots' node in the tree view at the right of the screen, and selecting 'Add Snapshot' from the menu that displayed. The parameters of each snapshot were edited by double clicking the snapshot that then appeared in the tree view to open the 'Snapshot Description' dialog box. Each snapshot was named according to the worker and the station. For example, A001L, A001R, A002L, or A002R. The 'Trigger ID' for each snapshot was set as the ID number of the last task being performed by the respective worker at the respective section. In the 'Expressions' tab, the 'Add' button was used to add all 14 variables being tracked. The check box for 'Auto Export' was checked, and a file name and save location were entered. The end result was a tab delimited '.res' file for each snapshot that included the value of each of the 14 variables, for each worker, at each station, for each vehicle simulated. In order to be able to track the values of the variables by station, some of the floating point type variables needed to be reset to 0 at the end of each station so they could begin being incremented again by the tasks at the next station. This was done using the following ending effects code entered in the parameters tab of the last task at each station:

```
\\Ending Effect
Time_Utilization = 0; //utilization time is reset for the next station.
Visual_Workload = 0; //visual workload is reset for next station.
Auditory_Workload = 0; //auditory workload is reset for next station.
Cognitive_Workload = 0; //cognitive workload is reset for next station.
Psychomotor_Workload = 0; //psychomotor workload reset for next station.
```

After the model was compiled with all 2,950 tasks from target area A, it was checked for errors using the error checker built into Micro Saint Sharp. After all syntax errors were fixed, the model was run several times to check for logic errors. Logic errors were fixed using the line debugger tool in Micro Saint Sharp. After the model was error free, the values of the 14 variables exported by the model for each worker at each station were compared to manual calculations for the same vehicle. Computational errors were fixed and model accuracy was checked again. After the model was ready for execution, the entire process described in this section was repeated to construct a separate task network model for the 2,027 tasks from target area B.

To enter the human error data into the models, custom entity attributes were added. These attributes were added by right clicking the 'Entity Attributes' node in the tree view at the right of the screen, and selecting 'Add Entity Attribute' from the menu that displayed. The parameters for each entity attribute were then entered by double-clicking the attribute name and opening the 'Attribute Description' dialog box. The error data were entered into each corresponding model by importing the Excel worksheet of errors and their Entity Attributes into Micro Saint Sharp.

APPENDIX B:

VALIDATION RESULTS

Table 17. *Full results of all split-half validation trials for Model A.*

Model A Validation	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5
Model Summary					
Chi-Square	1633.479	1562.479	1632.074	1625.059	1567.933
Significance	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
-2 Log	3194.189	3260.399	3308.702	3308.812	3086.000
Nagelkerke R^2	0.373	0.354	0.361	0.36	0.366
Variable Significance					
Time Pressure	.484	.901	.282	.791	.304
Time Pressure^2	.739	.806	.267	.957	.442
Visual Workload	.001	.000	.000	.000	.000
Visual Workload^2	.014	.000	.000	.003	.000
Auditory Workload	.058	.055	.026	.028	.064
Auditory Workload^2	.238	.182	.117	.135	.239
Cognitive Workload	.000	.000	.000	.000	.000
Cognitive Workload^2	.000	.000	.000	.000	.000
Psychomotor Workload	.000	.000	.000	.000	.000
Psychomotor Workload^2	.177	.000	.000	.038	.005
Task Frequency	.005	.042	.002	.021	.113
Task Frequency^2	.000	.000	.000	.000	.002
Shift	.054	.172	.788	.898	.299
Hours in Shift	.887	.748	.804	.858	.884
Hours in Shift ^{2}	.961	.650	.719	.843	.967
Hours in Shift ^{3}	.995	.619	.692	.860	.976
Information Flow	.000	.160	.014	.005	.055
Information Presentation	.000	.001	.001	.000	.002
Task Dependency	.693	.710	.036	.856	.445
Teamwork	.128	.341	.234	.044	.227
Equipment Feedback	.000	.000	.000	.000	.000
Prediction Results					
Correct Rejections	14633	14112	14130	14803	14611
False Alarms	2858	3374	3373	3419	2851
Misses	54	20	16	18	71
Hits	455	494	481	480	467

Model A Validation	Trial 6	Trial 7	Trial 8	Trial 9	Trial 10
Model Summary					
Chi-Square	1571.62	1554.6	1591.09	1596.86	1537.41
Significance	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
-2 Log	3047.56	3184.39	3217.85	3212.07	3145.34
Nagelkerke R^2	0.369	0.357	0.361	0.362	0.357
Variable Significance					
Time Pressure	.106	.147	.142	.290	.104
Time Pressure^2	.207	.186	.222	.465	.164
Visual Workload	.001	.000	.000	.000	.001
Visual Workload^2	.011	.002	.001	.000	.008
Auditory Workload	.050	.086	.003	.002	.060
Auditory Workload^2	.143	.260	.071	.061	.207
Cognitive Workload	.001	.000	.000	.000	.000
Cognitive Workload^2	.007	.000	.007	.000	.001
Psychomotor Workload	.000	.000	.000	.000	.000
Psychomotor Workload^2	.029	.001	.001	.000	.005
Task Frequency	.007	.047	.180	.027	.067
Task Frequency^2	.000	.001	.005	.000	.001
Shift	.247	.293	.220	.249	.745
Hours in Shift	.761	.479	.440	.722	.536
Hours in Shift ^{2}	.689	.514	.486	.885	.361
Hours in Shift ^{3}	.626	.535	.585	.983	.272
Information Flow	.000	.082	.038	.018	.008
Information Presentation	.003	.002	.002	.002	.001
Task Dependency	.239	.889	.323	.652	.530
Teamwork	.052	.106	.312	.270	.144
Equipment Feedback	.000	.000	.000	.000	.000
Prediction Results					
Correct Rejections	14677	14572	14267	14511	14375
False Alarms	2780	2902	3217	2973	3091
Misses	72	54	37	52	36
Hits	471	472	479	464	498

Model A Validation	Trial 11	Trial 12	Trial 13	Trial 14	Trial 15
Model Summary					
Chi-Square	1606.87	1608.64	1534.63	1637.07	1596.47
Significance	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
-2 Log	3202068	3088.19	3141.08	3296.8	3302.83
Nagelkerke R^2	0.364	0.372	0.357	0.363	0.356
Variable Significance					
Time Pressure	.144	.579	.031	.497	.870
Time Pressure^2	.250	.816	.045	.664	.683
Visual Workload	.000	.000	.000	.000	.000
Visual Workload^2	.000	.001	.000	.002	.000
Auditory Workload	.004	.050	.041	.057	.027
Auditory Workload^2	.078	.100	.181	.232	.140
Cognitive Workload	.000	.000	.000	.000	.000
Cognitive Workload^2	.000	.000	.000	.000	.000
Psychomotor Workload	.000	.002	.000	.000	.000
Psychomotor Workload^2	.002	.197	.003	.000	.000
Task Frequency	.247	.007	.010	.142	.067
Task Frequency^2	.007	.000	.000	.002	.001
Shift	.397	.575	.768	.179	.646
Hours in Shift	.212	.907	.705	.540	.545
Hours in Shift ^{2}	.117	.826	.658	.481	.572
Hours in Shift ^{3}	.082	.796	.655	.466	.583
Information Flow	.007	.001	.008	.026	.004
Information Presentation	.001	.012	.001	.001	.001
Task Dependency	.442	.837	.626	.741	.151
Teamwork	.090	.014	.103	.206	.173
Equipment Feedback	.000	.000	.000	.000	.000
Prediction Results					
Correct Rejections	14343	14122	13981	14434	14216
False Alarms	3141	3346	3484	3068	3281
Misses	42	29	29	38	17
Hits	474	503	506	460	486

Model A Validation	Trial 16	Trial 17	Trial 18	Trial 19	Trial 20
Model Summary					
Chi-Square	1589.36	1636.57	1462.93	1506.1	1605.65
Significance	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
-2 Log	3079.29	3255.79	3042.41	3155.49	3147.36
Nagelkerke R^2	0.37	0.365	0.353	0.352	0.368
Variable Significance					
Time Pressure	.837	.588	.752	.067	.413
Time Pressure^2	.907	.714	.959	.084	.576
Visual Workload	.000	.000	.000	.000	.000
Visual Workload^2	.005	.000	.001	.000	.006
Auditory Workload	.048	.017	.048	.011	.074
Auditory Workload^2	.132	.152	.227	.124	.245
Cognitive Workload	.000	.000	.000	.000	.000
Cognitive Workload^2	.000	.000	.000	.000	.000
Psychomotor Workload	.006	.000	.000	.000	.000
Psychomotor Workload^2	.357	.017	.005	.000	.050
Task Frequency	.005	.107	.065	.092	.430
Task Frequency^2	.000	.003	.001	.002	.018
Shift	.080	.240	.282	.721	.761
Hours in Shift	.721	.724	.192	.536	.578
Hours in Shift ^{2}	.803	.785	.177	.437	.548
Hours in Shift ^{3}	.856	.788	.195	.391	.514
Information Flow	.000	.003	.008	.055	.006
Information Presentation	.004	.000	.001	.000	.001
Task Dependency	.244	.787	.739	.723	.505
Teamwork	.009	.056	.201	.163	.036
Equipment Feedback	.000	.000	.000	.000	.000
Prediction Results					
Correct Rejections	14372	14221	14364	14264	14315
False Alarms	3092	3275	3077	3199	3161
Misses	35	20	38	33	39
Hits	501	484	521	504	485

Table 18. *Full results of all split-half validation trials for Model B.*

Model B Validation	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5
Model Summary					
Chi-Square	763.693	827.909	789.495	768.774	810.513
Significance	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
-2 Log	3362.405	3559.358	3539.588	3495.271	3440.475
Nagelkerke R^2	0.209	0.214	0.207	0.204	0.216
Variable Significance					
Time Pressure	.000	.000	.000	.000	.000
Time Pressure^2	.000	.000	.000	.000	.000
Visual Workload	.453	.121	.552	.286	.266
Visual Workload^2	.021	.105	.026	.049	.123
Auditory Workload	.291	.001	.000	.020	.015
Auditory Workload^2	.000	.000	.000	.000	.000
Cognitive Workload	.000	.000	.000	.000	.000
Cognitive Workload^2	.780	.453	.650	.532	.646
Psychomotor Workload	.571	.078	.287	.138	.066
Psychomotor Workload^2	.985	.324	.645	.385	.219
Task Frequency	.006	.002	.000	.003	.002
Task Frequency^2	.000	.000	.000	.000	.000
Shift	.815	.929	.850	.082	.679
Hours in Shift	.901	.457	.897	.521	.723
Hours in Shift ^{2}	.843	.402	.785	.641	.930
Hours in Shift ^{3}	.707	.387	.779	.685	.968
Information Flow	.000	.000	.002	.000	.000
Information Presentation	1.000	1.000	1.000	1.000	1.000
Task Dependency	.000	.000	.000	.000	.000
Teamwork	.000	.000	.000	.000	.000
Equipment Feedback	.000	.000	.000	.000	.000
Prediction Results					
Correct Rejections	9798	9743	9718	9799	9698
False Alarms	3162	3257	3273	3182	3281
Misses	146	134	126	133	135
Hits	394	366	383	386	386

Model B Validation	Trial 6	Trial 7	Trial 8	Trial 9	Trial 10
Model Summary					
Chi-Square	804.96	851.217	832.544	772.58	841.02
Significance	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
-2 Log	3446.027	3484.346	3593.331	3543.528	3584.855
Nagelkerke R^2	0.214	0.222	0.214	0.203	0.216
Variable Significance					
Time Pressure	.000	.000	.000	.000	.000
Time Pressure^2	.000	.000	.000	.000	.000
Visual Workload	.801	.277	.402	.757	.264
Visual Workload^2	.032	.152	.015	.025	.043
Auditory Workload	.000	.000	.008	.005	.020
Auditory Workload^2	.000	.000	.000	.000	.000
Cognitive Workload	.000	.000	.000	.000	.000
Cognitive Workload^2	.909	.756	.174	.944	.322
Psychomotor Workload	.723	.079	.191	.300	.057
Psychomotor Workload^2	.274	.214	.585	.729	.289
Task Frequency	.000	.000	.000	.038	.006
Task Frequency^2	.000	.000	.000	.000	.000
Shift	.615	.703	.653	.808	.498
Hours in Shift	.799	.218	.205	.445	.161
Hours in Shift ^{2}	.666	.198	.224	.374	.092
Hours in Shift ^{3}	.608	.239	.270	.361	.080
Information Flow	.008	.000	.000	.000	.000
Information Presentation	1.000	1.000	1.000	1.000	1.000
Task Dependency	.000	.000	.000	.000	.000
Teamwork	.000	.000	.000	.000	.000
Equipment Feedback	.000	.000	.000	.000	.000
Prediction Results					
Correct Rejections	9860	10050	9715	9637	9741
False Alarms	3119	2947	3291	3352	3265
Misses	145	162	121	109	126
Hits	376	346	373	402	368

Model B Validation	Trial 11	Trial 12	Trial 13	Trial 14	Trial 15
Model Summary					
Chi-Square	811.175	847.632	811.126	799.774	852.011
Significance	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
-2 Log	3530.866	3623.105	3400.588	3470.793	3444.604
Nagelkerke R^2	0.212	0.216	0.218	0.212	0.224
Variable Significance					
Time Pressure	.000	.000	.000	.000	.000
Time Pressure^2	.000	.000	.000	.000	.000
Visual Workload	.168	.142	.288	.066	.587
Visual Workload^2	.085	.126	.055	.126	.009
Auditory Workload	.004	.000	.006	.003	.001
Auditory Workload^2	.000	.000	.000	.000	.000
Cognitive Workload	.000	.000	.000	.000	.000
Cognitive Workload^2	.325	.703	.778	.387	.398
Psychomotor Workload	.118	.071	.251	.018	.164
Psychomotor Workload^2	.395	.224	.818	.130	.503
Task Frequency	.026	.024	.000	.002	.001
Task Frequency^2	.001	.000	.000	.000	.000
Shift	.982	.166	.839	.667	.380
Hours in Shift	.117	.282	.486	.664	.743
Hours in Shift ^{2}	.066	.130	.288	.435	.726
Hours in Shift ^{3}	.053	.092	.212	.371	.679
Information Flow	.000	.000	.000	.000	.000
Information Presentation	1.000	1.000	1.000	1.000	1.000
Task Dependency	.000	.000	.000	.000	.000
Teamwork	.000	.000	.000	.000	.000
Equipment Feedback	.000	.000	.000	.000	.000
Prediction Results					
Correct Rejections	9787	10050	9818	9937	9779
False Alarms	3206	2963	3155	3045	3207
Misses	127	153	141	148	137
Hits	380	334	386	370	377

Model B Validation	16	17	18	19	20
Model Summary					
Chi-Square	798.087	821.98	808.795	810.203	772.517
Significance	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
-2 Log	3459.431	3578.172	3297.441	3466.883	3399.772
Nagelkerke R^2	0.212	0.212	0.222	0.215	0.209
Variable Significance					
Time Pressure	.000	.000	.000	.000	.000
Time Pressure^2	.000	.000	.000	.000	.000
Visual Workload	.829	.504	.640	.658	.992
Visual Workload^2	.007	.044	.007	.000	.014
Auditory Workload	.000	.024	.001	.000	.007
Auditory Workload^2	.000	.000	.000	.000	.000
Cognitive Workload	.001	.000	.000	.001	.000
Cognitive Workload^2	.290	.596	.292	.148	.747
Psychomotor Workload	.225	.434	.863	.420	.583
Psychomotor Workload^2	.704	.802	.465	.910	.822
Task Frequency	.001	.064	.029	.000	.000
Task Frequency^2	.000	.005	.001	.000	.000
Shift	.236	.630	.676	.705	.776
Hours in Shift	.498	.477	.103	.880	.503
Hours in Shift ^{2}	.408	.426	.061	.853	.630
Hours in Shift ^{3}	.377	.438	.050	.776	.644
Information Flow	.000	.000	.000	.000	.001
Information Presentation	1.000	1.000	1.000	1.000	1.000
Task Dependency	.000	.000	.000	.000	.000
Teamwork	.000	.000	.000	.000	.000
Equipment Feedback	.000	.000	.000	.000	.000
Prediction Results					
Correct Rejections	9920	9598	10275	9887	9649
False Alarms	3060	3404	2682	3096	3318
Misses	139	118	192	144	131
Hits	381	380	351	373	402

Table 19. *Results for the individual variables in the cross-group validation trial of using the prediction equation from Model A to predict errors in vehicles from Area B.*

Variable	B	S.E.	Wald	df	Sig.	Exp(B)
Time Pressure	-2.315	1.274	3.303	1	.069	.099
Time Pressure^2	.794	.526	2.281	1	.131	2.213
Visual Workload	-.066	.011	34.745	1	<.001*	.936
Visual Workload^2	.000	.000	23.382	1	<.001*	1.000
Auditory Workload	1.769	.683	6.711	1	.010*	5.868
Auditory Workload^2	-.210	.126	2.798	1	.094	.811
Cognitive Workload	.090	.014	39.582	1	<.001*	1.094
Cognitive Workload^2	.000	.000	27.684	1	<.001*	1.000
Psychomotor Workload	.104	.017	37.481	1	<.001*	1.110
Psychomotor Workload^2	.000	.000	13.866	1	<.001*	1.000
Task Frequency	1.905	.753	6.397	1	.011*	6.721
Task Frequency^2	-4.765	1.047	20.736	1	<.001*	.009
Shift	-.070	.073	.917	1	.338	.933
Hours in Shift	.019	.118	.025	1	.874	1.019
Hours in Shift ^{2}	-.003	.022	.013	1	.908	.997
Hours in Shift ^{3}	.000	.001	.008	1	.931	1.000
Information Flow	-.676	.171	15.738	1	<.001*	.508
Information Presentation	-2.631	.574	21.005	1	<.001*	.072
Task Dependency	-.316	.368	.737	1	.391	.729
Teamwork	-.255	.107	5.657	1	<.001*	.775
Equipment Feedback	5.332	.407	171.944	1	<.001*	206.873

Table 20. *Results for model fit and prediction in the cross-group validation trial of using the prediction equation from Model A to predict errors in vehicles from Area B.*

Model A Cross-Group Validation	Results
Model Summary	
Chi-Square	3169.987
Significance	< 0.001
-2 Log	6322.017
Nagelkerke R^2	.364
Prediction Results	
Correct Rejections	21004
False Alarms	13946
Misses	285
Hits	765

Table 21. *Results for the individual variables in the cross-group validation trial of using the prediction equation from Model B to predict errors in vehicles from Area A.*

Variable	B	S.E.	Wald	df	Sig.	Exp(B)
Time Pressure	-11.871	1.131	110.163	1	<.001*	.000
Time Pressure^2	5.715	.575	98.902	1	<.001*	303.489
Visual Workload	-.033	.009	15.034	1	<.001*	.967
Visual Workload^2	.000	.000	.250	1	.617	1.000
Auditory Workload	-.404	.235	2.961	1	.085	.668
Auditory Workload^2	.286	.055	26.767	1	<.001*	1.332
Cognitive Workload	.104	.014	55.726	1	<.001*	1.110
Cognitive Workload^2	.000	.000	.511	1	.475	1.000
Psychomotor Workload	.043	.015	8.586	1	.003*	1.044
Psychomotor Workload^2	.000	.000	3.832	1	.050*	1.000
Task Frequency	2.619	.569	21.225	1	<.001*	13.724
Task Frequency^2	-4.452	.691	41.529	1	<.001*	.012
Shift	-.027	.074	.134	1	.715	.973
Hours in Shift	-.086	.121	.511	1	.475	.917
Hours in Shift ^{2}	.021	.022	.878	1	.349	1.021
Hours in Shift ^{3}	-.001	.001	.976	1	.323	.999
Information Flow	-1.129	.155	52.874	1	<.001*	.324
Information Presentation	-27.195	123198.8	.000	1	1.000	.000
Task Dependency	2.841	.267	112.896	1	<.001*	17.132
Teamwork	.696	.108	41.401	1	<.001*	2.005
Equipment Feedback	1.646	.194	72.310	1	<.001*	5.184

Table 22. *Results for model fit and prediction in the cross-group validation trial of using the prediction equation from Model B to predict errors in vehicles from Area A.*

Model B Cross-Group Validation	Results
Model Summary	
Chi-Square	1644.527
Significance	< 0.001
-2 Log	7020.121
Nagelkerke R^2	.215
Prediction Results	
Correct Rejections	22422
False Alarms	12528
Misses	336
Hits	714

REFERENCES

Adkins, R., Murphy, W., Hemenway, M., Archer, R., & Bayless (1996). HARDMAN III analysis of the land warrior system (Document Number ARL-CR-291). Aberdeen Proving Ground, MD: U.S. Army Research Laboratory.

Archer, R.D., & Lockett, J.F. III (1997). WinCrew - a tool for analyzing performance, mental workload and function allocation among operators. *Proceedings of the First International Conference on Allocation of Functions*, Galway, Ireland.

Barnes, C.D., & Quiason, J.L. (1997). Success stories in simulation healthcare. *Proceedings of the 1997 Winter Simulation Conference,* Atlanta, GA, December 7-10, 1997.

Beideman, L.R., Munro, I., & Allender, L.E. (2001). IMPRINT modeling for selected crusader research issues. Aberdeen Proving Ground, MD: U.S. Army Research Laboratory.

Bell, B.J., & Swain, A.D. (1983). *A Procedure for Conducting a Human Reliability Analysis for Nuclear Power Plants* (Document Number NUREG/CR-2254). Washington, DC: U.S. Nuclear Regulatory Commission.

Belland, K.M., Olsen, C., & Lawry, R. (2010). Carrier air wing mishap reduction using a human factors classification and risk management system. *Aviation, Space, and Environmental Medicine, 81*(11), 1028-1032.

Boring, R.L. (2006). Modeling human reliability analysis using MIDAS (Document Number INL/CON-06-11170). Idaho Falls, ID: Idaho National Laboratory.

Cain, B. (2007). *A review of the mental workload literature* (Report Number RTO-TR-HFM-121-Part-II). Toronto, Canada: Defence Research and Development, Human System Integration Section

Center for Chemical Process Safety (1994). *Guidelines for Preventing Human Error in Process Safety*. New York: American Institute of Chemical Engineers.

Cohen, J. (1988). *Statistical power analysis for the behavior sciences* (2nd ed.). New Jersey: Lawrence Erlbaum.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Hillsdale: Erlbaum.

Dekker, S.W.A. (2002). The re-invention of human error. Document number 2002-01. Lund University School of Aviation.

Department of Defense (2005). *Department of Defense Human Factors Analysis and Classification System*. Document number DoDI 6055.7. Washington D.C.

Detwiler, C., Hackworth, C., Holcomb, K., Boquet, A., Pfleiderer, E., Wiegmann, D. & Shappell, S.A. (2006). *Beneath the tip of the iceberg: A human factors analysis of general aviation accidents in Alaska verses the rest of the United States* (Report Number DOT/FAA/AM-06/7). Washington DC: Office of Aerospace Medicine.

Folkard, S., & Monk, T.H., (1980). Shiftwork and Performance. In W.P. Colquhoun & J. Rutenfranz (Eds.), Studies of Shiftwork (pp. 263-272). London: Taylor & Francis.

Hill, T., & Lewicki, P. (2007). *Statistics: Methods and Applications*. Tulsa, OK: StatSoft.

Hugo, J., & Gertman, D.I. (2012). The use of computational human performance modeling as task analysis tool. *Proceedings of the Eighth American Nuclear*

Society International Topical Meeting on Nuclear Plant Instrumentation, Control, and Human-Machine Interface Technologies, NPIC&HMIT 2012, San Diego, July 22 - 26, 2012.

Isaac, A., Shorrock, S.T., Kennedy, R., Kirwan, B., Andersen, H., & Bove, T. (2002). *Short report on human performance models and taxonomies of human error in ATM (HERA).* Report number HRS/HSP-002-REP-02. European Organization for the Safety of Air Navigation.

Kryter, K.D. (1972). Speech communication. In H. P. Van Cott and R. G. Kinkade (Eds.), *Human Engineering Guide To Equipment Design.* Washington, DC: U.S. Government Printing Office.

Laughery, R. (1998). Computer simulation as a tool for studying human-centered systems. *Proceedings of the 1998 Winter Simulation Conference,* Washington DC, December 13-16, 1998.

Laughery, R., & Corker, K. (1997). Computer modeling and simulation of human/system performance. In G. Salvendy (Ed.), Handbook of human factors and ergonomics (2nd ed.) (pp. 1375-1408). Hoboken, NJ : John Wiley & Sons.

Leiden, K., Laughery, K. R., Keller, J. W., French, J. W., Warwick, W. & Wood, S.D. (2001). *A Review of Human Performance Models for the Prediction of Human Error*. Boulder, CO: Micro Analysis and Design, Inc.

Lupien, S.J., Maheu, F., Tu, M., Fiocco, A., & Schramek, T.E. (2007). The effects of stress and stress hormones on human cognition: Implications for the field of brain and cognition. *Brain and Cognition, 65*(1), 209-237.

Malkin, F.J., Allender, L.E., Kelley, T.D., O'Brien, P., & Graybill, S. (1997). Joint base station variant 1 MOS-workload-skill requirements analysis (Document Number ARL-TR-1441). Aberdeen Proving Grounds, MD: Army Research Laboratory.

McCrasken, J. H., & Aldrich, T. B. (1984). Analysis of selected LHX mission functions: Implications for operator workload and system automation goals (Technical Note ASI479-024-84). Fort Rucker, AL: Army Research Institute Aviation Research and Development Activity.

McGrath, J.E. (1984). *Group Interaction and Performance.* Englewood Cliffs, NJ: Prentice Hall.

Mitchell, D.K. (2000). *Mental workload and ARL workload modeling tools*. Aberdeen Proving Ground, MD, US Army Research Laboratory, Human Research & Engineering Directorate: 35.

Mitler, M.M., Carskadon, M.A., Czeisler, C.A., Dement, W.C., Dinges, D.F., & Graeber, R.C. (1988). Catastrophes, sleep, and public policy: Consensus report. *Sleep, 11*(1), 100-109.

Myers, J.H., & Forgy, E.W. (1963). The Development of Numerical Credit Evaluation Systems. Journal of the American Statistical Association, 58(303), 799–806.

Nuclear Regulatory Commission (1975). *Reactor Safety Study - An Assessment of Accident Risks in U.S. Commercial Nuclear Power Plants* (Report Number NUREG-75/014, WASH-1400). Washington, DC: U.S. Nuclear Regulatory Commission.

Peacock, B., Savage, E., & Waldock, B. (2009). *Simulation of Emergency Evacuation from Transport Category Aircraft.* Atlanta, GA: Department of Safety Science, Embry-Riddle Aeronautical University.

Rasmussen, J. (1982). Human errors. A taxonomy for describing human malfunction in industrial installations. *Journal of Occupational Accidents, 4*(2–4), 311-333.

Reason, J. (1990). *Human Error.* Cambridge University Press, Cambridge.

Reinhardt, B.M. (1992). Estimating result replicability using double cross-validation and bootstrap methods. *The Annual Meeting of the American Educauonal Research Association,* San Francisco, CA.

Sarno, K.J., & Wickens, C.D. (1995). Role of multiple resources in predicting time-sharing efficiency: Evaluation of three workload models in a multiple-task setting. *The International Journal of Aviation Psychology, 5*(1), 107-130.

Schunk, D., & Plott, B. (2000). Using simulation to analyze supply chains. *Proceedings of the 2000 Winter Simulation Conference,* Orlando, FL, December 10-13, 2000.

Seminara, J.L., Gonzalez, W.R., & Parsons, S.O. (1976). *Human Factors Review of Nuclear Power Plant Control Room Design* (Report Number EPRI NP-309). Palo Alto, CA: Electric Power Research Institute.

Senders, J. W. & Moray, N. P. (1991). *Human error: Cause, prediction, and reduction.* Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Shappell, S.A., Detwiler, C., Holcomb, K., Hackworth, C., Boquet, A. & Wiegmann, D. (2007). Human Error and Commercial Aviation Accidents: An Analysis Using the Human Factors Analysis and Classification System. *Human Factors, 49,* 227-242.

Shappell, S.A. & Wiegmann, D.A. (1997). A human error approach to accident investigation: The Taxonomy of Unsafe Operations. *International Journal of Aviation Psychology*, *7*(4), p. 269-291.

Shappell, S.A. & Wiegmann, D.A. (2000). *The Human Factors Analysis and Classification System (HFACS)* (Report Number DOT/FAA/AM-00/7). Washington, DC: Office of Aerospace Medicine.

Shappell, S.A. & Wiegmann, D. (2009). Developing a methodology for assessing safety programs targeting human error in aviation. *The International Journal of Aviation Psychology*, *19*, 252-269.

Sharit, J. (2006). Human Error. In G. Salvendy (Ed.), Handbook of human factors and ergonomics (3rd ed.) (pp. 708 – 760). Hoboken, NJ : John Wiley & Sons.

Stanton, N. A. & Salmon, P.M. (2009). Human error taxonomies applied to driving: A generic driver error taxonomy and its implications for intelligent transport systems. *Safety Science*, *47*, p. 227–237.

Stewart, G.L., & Barrick, M.R. (2000). Team structure and performance: Assessing the mediating role of intrateam process and the moderating role of task type. *The Academy of Management Journal, 43*(2), p. 135-148.

Swain, A.D. & Guttmann, H.E. (1983). *A handbook of human reliability analysis with emphasis on nuclear power plant applications.* NUREG/CR-1278, USNRC, Washington, DC.

Szumilas, M. (2010). Explaining odds rations. *Journal of the Canadian Academy of Child and Adolescent Psychiatry. 19*(3), p. 227-229.

Walters, B., French, J., & Barnes, M.J. (2000). Modeling the effects of crew size and crew fatigue on the control of tactical unmanned aerial vehicles. *Proceedings of the 2000 Winter Simulation Conference,* Orlando, FL, December 10-13, 2000.

Welford, A.T. (1976). *Skilled Performance: Perceptual and Motor Skills*. Glenview, IL: Scott, Foresman.

Wetteland, C.R., Miller, J.L., French, J., O'Brien, K., & Spooner, D.J. (2000). The human simulation: Resolving manning issues onboard DD21. *Proceedings of the 2000 Winter Simulation Conference,* Orlando, FL, December 10-13, 2000.

Wickens, C.D., Hooey, B.L., Gore, B.F., Sebok, A., Koenecke, C., & Salud, E. (2009). Identifying black swans in NextGen: Predicting human performance in off-nominal conditions. *Human Factors, 51*(5), 638-651.

Wickens, C.D. (1981). *Processing Resources in Attention, Dual Task Performance, and Workload Assessment* (Report Number EPL-81-3/ONR-81-3). Urbana Champagne, IL: Engineering-Psychology Research Laboratory.

Wiegmann, D., Faaborg, T., Boquet, A., Detwiler, C., Halcomb, K. & Shappell, S.A. (2005) *Human error and general aviation accidents: A comprehensive, fine-grained analysis using HFACS* (Report Number DOT/FAA/AM-05/24). Washington DC: Office of Aerospace Medicine.

Wiegmann, D. A. & Shappell, S.A. (2001). *A human error analysis of commercial aviation accidents using the Human Factors Analysis and Classification System (HFACS)* (Report Number DOT/FAA/AM-01/3). Washington DC: Office of Aerospace Medicine.

Yerkes, R.M., & Dodson, J.D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology, 18*(1), 459-482.