2016-04-11

An Alternative Method for Evaluating the Risks Resulting from the Use of Procedures

Gregory T. Praino
University of Miami, prainog@gmail.com

UNIVERSITY OF MIAMI

AN ALTERNATIVE METHOD FOR EVALUATING THE RISKS RESULTING FROM THE USE OF PROCEDURES

By

Gregory T Praino

A DISSERTATION

Submitted to the Faculty
of the University of Miami
in partial fulfillment of the requirements for
the degree of Doctor of Philosophy

Coral Gables, Florida

May 2016

UNIVERSITY OF MIAMI

A dissertation submitted in partial fulfillment of
the requirements of for the degree of
Doctor of Philosophy

AN ALTERNATIVE METHOD FOR EVALUATING THE RISKS RESULTING FROM THE USE OF PROCEDURES

Gregory T Praino

Approved:

Joseph Sharit, Ph.D. Professor of Industrial Engineering	Shihab Asfour, Ph.D. Professor of Industrial Engineering
Murat Erkoc, Ph.D. Assistant Professor of Industrial Engineering	Guillermo Prado, Ph.D. Dean of the Graduate School
Arzu Onar-Thomas, Ph.D. Department of Biostatistics St. Jude Children's Research Hospital	

PRAINO, GREGORY T (Ph.D., Industrial Engineering)
(May 2016)

An Alternative Method for Evaluating the Risks Resulting from the Use of Procedures.

Abstract of a dissertation at the University of Miami.

Dissertation supervised by Professor Joseph Sharit
No. of pages in the text. (136)

Organizations create procedures as a way of reducing risks by influencing worker behavior but an ineffective procedure may fail to reduce risks and may create new risks. This paper presents a method of pseudo-quantifying those risks as an alternative to the labor intensive conventional computation of risk. The model considers scores for the value and failure likelihood of procedural controls as a surrogate for the consequence and likelihood measures normally associated with risk. Scores were provided by experts in space shuttle processing regarding a selected set of procedures in place for shuttle ground processing in place prior to the *Columbia* disaster. It was concluded that the effectiveness of some portions of the model were dependent on the professional background of the evaluator. A recommendation is enclosed for further study of the model using a cross-disciplinary team and for using correlations from observed failure rates of procedures during the *Columbia* investigation as a basis for creating procedure improvement guidelines.

DEDICATION

This work is dedicated to the memory of Rick D. Husband, William C. McCool, David M. Brown, Kalpana Chawla, Michael P. Anderson, Laurel B. Clark and Ilan Ramon, without whose sacrifice it would never have been conceived of or been recognized as necessary.

ACKNOWLEDGMENTS

First, I would like to thank my parents or the roots and the wings that made everything since possible. I look forward to finally hearing "My son the doctor."

I also want to thank the bosses who accommodated the time off that was necessary to make this happen: Keith Jones, Mike Holman and Anne Ramsey. Critical thanks also goes to the original architects of the idea that became DATOM: Keith, Dave and Keith and to the team that originally put it to practical use: Keith, Carrie, Karen, Brock and Mike.

Next, I would like to thank my children Tori, Beth and Alex. I hope that seeing me work on this for all of their lives so far has given them an appreciation for lifelong learning.

Finally, I want to thank my wife, Jenn, for the patience needed to see this through for the 14-years it's taken.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1 INTRODUCTION

Every person or organization faces risks each day that can potentially cause them some loss. To cope with the existence of these risks, risk reduction tools have been developed to prevent the losses and to reduce the severity of the losses. One such tool, which is found in almost all high-risk industries, is the written work procedure. Unfortunately, the use of any such tool can never completely eliminate uncertainty, so some risk remains.

The inherent assumption in the use of procedures is that they will not increase the risk faced by the person or organization. However, this is not necessarily true because deficient design and poorly implemented procedures can lead personnel into undesired behavior, increasing risks to themselves or to the goals of their organization.

Traditional Quantified Risk Analysis (QRA) techniques can be used to evaluate the risks that remain, or result, when procedures are implemented, but the goal of this research is to develop a less resource intensive and more intuitive approach to quantifying such risks. This proposed method will be targeted towards processes where a human operator has a role that is governed by procedures in the form of documented instructions.

Rather than directly using the conventional risk components of consequence and likelihood, the method presented here uses characteristics of a procedure of interest to gauge the associated risk. By using a framework associated with the effectiveness of procedures, this method is intended to avoid the need for long observation periods or expensive data collection to determine likelihoods, and to similarly avoid the need for converting consequence magnitudes across measurement scales.

Managing the risks that arise from the use of procedures requires two fundamental concepts to be explicitly defined: *procedure* and *risk*. While these terms are commonly used in everyday situations, the meanings of both are flavored by the context of the particular usage. Once it is understood exactly what is meant by procedures and risks, a technique is necessary for identifying risks, ranking those risks, and prioritizing potential preventive actions. The end result of assessing procedure risks will be a prioritized list of preventive actions, though implementation of the identified actions is outside the scope of the assessment.

Prioritization of preventive actions will allow an organization to focus limited resources on the most effective preventive actions. To determine which actions are the most effective to implement, the reduction in risk associated with each action will be determined. This is done by comparing the risk before a change with the projected risk after the preventive action.

While additional risks to the organization invariably exist, by limiting the assessment to risks caused or allowed by procedures, tools that exclusively address procedure content can be developed. Items that cannot be controlled by procedure, such as environmental or design limitations, fall outside the scope of the intended analysis even though they may be recognized in the risk identification portion of the assessment.

Ranking of the risks will be based on characteristics of risk that are observable for individual portions of a procedure. Thus, an assessor will be able to draw conclusions about specific risks by examining the procedures themselves and the context in which they are expected to be performed. A method of assessing procedures without necessarily needing to observe their execution provides insight that is useful during the design of those procedures. Therefore, the goal of this research is to develop a technique for estimating the net effect of written work procedures on the magnitude of risk based on the characteristics of the procedure clauses themselves. For scenarios where such estimates cannot be reasonably produced, the aim is to provide a relative ranking of the procedures by their impact on risk.

Risk Defined

To explore the relationship between procedures and risk, one of the many definitions of risk must be selected from among the many in use in daily life and academic literature. A universal definition of risk is an elusive goal that continues to plague the fields of risk analysis and risk management. In fact, the Risk Definition Committee of the Society for Risk analysis identified at least 13 different definitions prior to its first meeting in 1981 (Haimes, 2009) with additional nuanced definitions published in the years since. One simple definition from the early days of formal risk analysis is that “Risk is a measure of the probability and severity of adverse effects.” (Lowrance, 1976) While this does agree with one of the common language uses of the term, it includes the idea of measurement and would be more appropriately a definition of the term ‘risk exposure.’ In an effort to maintain precision and accuracy in the basic terminology, this paper will use a simpler definition for the abstract concept of risk provided in the ISO 31000 (2009) standard: *Risk is the effect of uncertainty on objectives*. As Lowrance’s definitions shows, some definitions only address the possibility of adverse effects, but the more general definition is well suited to risks related to procedures because procedures are as easily intended to cause desired effects as they are intended to prevent adverse effects. Thus, the benefits of a procedure can be weighed accurately against both the fixed cost of implementing, for new procedures, and the potential adverse effects allowed or caused by the procedure.

While this is somewhat inconsistent with the traditional view of risk, considering ‘upside risk’ in risk management activities is an increasingly common practice among project management (Hillson, 2002) (PMI, 2008) and investment industry practitioners, where risk management is primarily a financial activity. In considering risks associated with procedures, the underlying assumption is that an assessment of a procedure will lead to a choice of retaining the existing procedure or replacing it. To effectively make such a decision though, the costs and benefits must both be considered. For maintaining simplicity of vocabulary however, ‘loss’ will be used for describing the combined gain/loss and ‘cost’ will be used for the combined benefit/cost, though the net value for either quantity could be positive.

Since there is always a potential for some uncertain effect, risk is an unavoidable circumstance. As a practical matter though, a working definition must consider just how much effect comes from specific uncertain events. For example, when considering the safety of their workers, most organizations would not attempt to provide additional protection to prevent a fatality from a meteor strike; while it is a possibility, the likelihood of a meteor striking a particular target is so small that it is trivial compared to the myriad of other events that are also possibilities.

In a landmark paper in the inaugural issue of the journal *Risk Analysis*, Kaplan and Garrick (1981) proposed that risk could be described as the answers to three questions: “What can go wrong?” “What are the consequences?” “What is the likelihood?” The use of this ‘risk triplet’ to describe risk continues to be the foundation of the practical definitions of risk used by many researchers today. In later refinements (Kaplan, 2001), they clarify that the set of possible scenarios is generally nondenumerable and infinite, but that a practical analysis can involve partitioning the set into a finite number of disjoint scenarios that encompasses all of the combinations of what can go wrong, the likelihoods and the consequences.

Even with the refined definition, however, there remain multiple interpretations. One aspect that is responsible for many of the differences is the concept of likelihood: some definitions consider it the likelihood of the consequence itself; the likelihood of the event causing the loss (Haimes, 2009); or the likelihood that a particular probability distribution will describe the distribution of the consequences (Kaplan 1993). The ambiguity regarding likelihood as a probability of occurrence has even led to the opinion that a fourth question is appropriate: "Over what timeframe?" (Haimes, 2009)

While the nuances of the different definitions are significant in their own contexts, the practical issue faced by an organization when considering risk is optimizing the response to the possible effects. Dekker addresses this relative to safety (2008), but the point applies to all uncertainty: organizations do not exist to reduce risk, but to provide a service or product, to achieve economic gain or maximize capacity utilization. Thus, when an organization elects to hold a portion of its finite resources in reserve to prepare for uncertainty, there is a corresponding loss of ability to provide the service or product that is the organization's primary goal.

Categorizing Consequences

Since the risks that threaten an organization's ability to satisfy its objective come in a variety of forms, it is useful to cluster similar risks together to facilitate understanding and responses. One useful way to organize risks is to group by the type of consequences.

When considering risks to an organization, the consequences of interest are the effects on the organization's ability to operate effectively in response to an occurrence of the uncertain event. To identify measures of how the organization's effectiveness is impacted, the goal of the organization must be considered to ensure that improvement activities are focusing on what is truly important and not finding ways to waste resources more efficiently. The goal of an organization is not to provide a product or service at an arbitrary moment and cost, but to *safely* provide a *sufficient* product or service at a *timely* moment for an *economical* cost. Consequently, it is the safety, sufficiency, timeliness and economy of the product or service that are the success measures against which risks can be evaluated.

As a discipline, systems engineering takes a holistic view of all risks, using a similar categorization of cost, schedule, technical and programmatic risks (INCOSE, 2006) (NASA 2007). Other practitioners tend to focus on the measures they have the most control over or are most affected by. Project managers are typically concerned with the ways risk affect the cost or schedule of the project (PMI 2008) while quality managers tend to focus on the sufficiency of the product or service relative to the requirements and needs they are seeking to satisfy. Procurement functions will often split attention between aspects of all three of those factors (DoD, 2015), with system safety professionals prioritizing on prevention of mishaps that impact the safety of personnel or equipment (DoD, 2012). These four areas where the effects of risk appear—safety, performance, schedule and financial cost—provide a high-level set of categories that are convenient for categorizing risks.

Performance, schedule and financial cost represent aspects of operating the organization's system, which can be improved or degraded by the choices made by management, by designers and by workers. Safety is fundamentally distinct in that the upper limit of the measure is fixed, regardless of the scenario, at successful operation with no injuries or other mishaps. While the existence of an upper limit on safety sets it apart in theory, for systems that remain below that limit safety may be treated as an aspect of the organization's system that is affected by the choices made by the participants. Often, these choices are a tradeoff between improving one at the expense of one of the others: working overtime brings added cost, but may improve the schedule margin; choosing a less expensive part with a shorter lead time may require sacrificing performance by omitting some functional capability.

Uncertainty and Risk Costs

To develop a means of measuring the costs of procedural risk, it is essential to understand where the conventional method of determining risk costs is inefficient. The idealized method of identifying the cost of the uncertainty is to consider risk as an expected loss (Kumamoto, 1996), in which each potential consequence is multiplied by the likelihood of that consequence occurring, yielding the expected net cost of the uncertainty. Under this method, the consequence value is one to be directly measured, such as number of fatalities or monetary costs, or it can be qualitative, based on the utility of the different outcomes. Risk as an expected value can be generalized to include beneficial as well as adverse effects.

The likelihood used in determining expected cost is the total probability of that specific consequence occurring during the timeframe being examined. Whether the probability is a constant rate or some distribution that varies over time, proper accounting for the likelihood includes the timeframe over which the consequence can occur, showing that the fourth question proposed by Haimes is superfluous.

Integrating the probability over the relevant timeframe (T), and summing up the expected losses associated with each of the possibilities considered yields a magnitude measure for the total risk faced by the subject. (Figure 1-1)

$$TotalRisk = \sum ExpectedLoss_i = \sum \left(consequence_i \times \int_T probability(t)_i \, dt \right)$$

Figure 1-1 - Total risk as the sum of expected losses from uncertainty

Scenarios with extremely low likelihoods will have correspondingly low expected losses, and will not measurably affect the total risk by being included or omitted unless the consequences are extremely large. While there might be 10 or even 100 fatalities caused by a meteor striking a factory, the probability is so low that the contribution to overall risk would be a fraction of that from a single fatality from a worker falling off a 12-foot ladder.

As noted above, the idea of risk normally includes only outcomes that may occur because of the uncertainty about the future. By that definition, outcomes that are certain to occur pose no risk. However, for the purposes of considering risks associated with procedures, the definition must be expanded to include sure losses as well since procedures aim to improve outcomes. To measure the magnitude of improvement associated with those outcomes, costs of implementing new procedures must be properly accounted for along with any sure costs of working to existing procedures regardless of the final uncertainty.

To assess effectiveness of the procedure, it is the *net expected cost*, not just the *expected cost from future uncertainty* that must be considered. Without this shift in definition, the risks of two pure improvement scenarios are shown inaccurately. In the first case, a consequence otherwise certain to occur may have a potential preventive measure introduced. The calculated risk after the procedure is the product of the consequence and the likelihood that it will not be prevented—a term in the summed risk equation that did not exist when the probability of occurrence was 100%, artificially inflating the total risk. The second case is the scenario where the procedure has the effect of purchasing insurance. By implementing the procedure, the likelihood of occurrence for the consequence is eliminated and the risk is removed. The financial cost of the insurance, or the implementation of the procedure in this case is neglected, when it must be treated as a sure loss to accurately represent the tradeoff. Figure 1-2 shows the additional term added to the total risk equation to address the sure losses surrounding the procedure-related operations, though the additional term is mathematically trivial. The addition of this term is meaningful only in that it explicitly partitions the set of consequences into two subsets: *j*, representing those consequences with a probability of occurrence below 100% and *k* for consequences where the probability equals 100%.

$$\boxed{TotalRisk = \sum ExpectedLoss_i = \sum consequence_j \cdot probability_j + \sum SureLoss_k}$$

Figure 1-2 - Total risk as the sum of all expected losses

This second scenario above touches on a related subject that will be excluded from further discussion by assumption: the economics of risk aversion. The option to buy insurance at a financial cost equal to the expected loss of the consequence is economically equivalent to accepting the risk and an organization that is neither risk-seeking nor risk-averse would have no preference for one over the other because the expected cost is identical.

In actuality though, behavioral economics shows that the responses of individuals are not consistent with classical economic theory (Fredrickson 1993) and that organizations often prefer to pay a premium that is marginally more than the expected risk loss to protect their cash flow from a large loss at an inconvenient time. This leads to the following assumption: that the total risk accounts for the risk aversion factors in the selection of the consequences being considered. For an organization, the protection of the cash flow represents a sure gain in utility that is factored into the net expected loss. Similarly, the choice to select an economically sub-optimal decision comes with the sure gain of reduced cognitive or managerial effort.

Risks with Incomparable Consequences

With the inclusion of sure losses, the summed-risk method can help understand the expected loss from all undesirable outcomes; however it requires that the consequences all be measurable on a single scale. Standardization of this type presents minor annoyance in some situations and significant difficulty in others. If overtime rates and personnel availability schedules are known, then equating a schedule slippage in days to a net financial cost impact would be relatively simple. However there is no uniformly recognized method for equating a financial loss with an injury, or an injury with a fatality so it is unrealistic to equate any financial loss with a single fatality or to set the price of a lost limb as a quarter as much.

While quantifying consequences on a single scale is often difficult and sometimes impossible, from the standpoint of an organization, catastrophic losses that lead to the demise of the organization have an effectively equal consequence. Regardless of the type of consequence, the maximum realized loss can be considered equivalent to the total value of the organization. By focusing analysis activities on these risks, only hazards that can lead to catastrophic outcomes are considered and the management efforts are aimed at reducing the likelihood of the catastrophe, eliminating the vulnerability that may allow it to occur or improving the recovery after an event to a non-catastrophic state.

Where none of these is possible, the organization must choose between engaging in the activity or not. In some scenarios, the short term profit stream could be more dear than the present value of the organization, leading to the logical decision to engage in an activity that could ultimately ruin the organization. The scenarios where this is valid depend on complex market factors that determine the value of the organization, which are not addressed here. Instead, the assumption is made that the organization has decided to engage in the relevant activity and the goal of the risk management activity is to minimize the likelihood of a terminal event.

To that end, the organization responds to potential events through eliminating vulnerability, limiting the magnitude of consequences, reducing the likelihood of a consequence or effectively recovering to an operational state. These four actions are intended to reduce vulnerability and improve resilience, two concepts that have been increasingly tied to the field of risk analysis in the last decade (Haimes, 2009) (Woods, 2006). In the absence of a standard definition of risk however; the exact meaning and role of these ideas is highly variable from use to use.

In managing risks, it is important to note that a single hazard can be associated with multiple risks, as shown in Figure 1-3. This is significant because the existence of multiple risks associated with each hazard is relevant when considering the impact of preventive procedure changes. Specifically, while the elimination of the hazard could eliminate multiple risks, controls that only reduce likelihood or mitigate a consequence will have a smaller, more localized impact.

$$
\begin{array}{llllll}
\text{Total Risk:} & \text{Hazard}_{1} & \text{Consequence}_{1\,1} & \text{Likelihood}_{1\,1} & \rightarrow & \text{Risk}_{1\,1} \\
 & & \vdots & \vdots & & \vdots \\
 & & \text{Consequence}_{1\,a} & \text{Likelihood}_{1\,a} & \rightarrow & \text{Risk}_{1\,a} \\
 & \text{Hazard}_{2} & \text{Consequence}_{2\,1} & \text{Likelihood}_{2\,1} & \rightarrow & \text{Risk}_{2\,1} \\
 & & \vdots & \vdots & & \vdots \\
 & & \text{Consequence}_{2\,b} & \text{Likelihood}_{2\,b} & \rightarrow & \text{Risk}_{2\,b} \\
 & \text{Hazard}_{n} & \text{Consequence}_{n\,1} & \text{Likelihood}_{n\,1} & \rightarrow & \text{Risk}_{n\,1} \\
 & & \vdots & \vdots & & \vdots \\
 & & \text{Consequence}_{n\,m} & \text{Likelihood}_{n\,m} & \rightarrow & \text{Risk}_{n\,m}
\end{array}
$$

Figure 1-3 - Total risk as the answers to the three questions

This can be seen using the example of the meteor strike. If a facility is staffed for eight hours a day, the single hazard to consider is the meteor strike. Two risks with distinct consequences and likelihoods are readily identified: the fatalities that are possible while the building is occupied and the facility damage. The probability of the damage is three times as great because it can happen at any time while the fatalities will only occur during the eight hours that the building is occupied. As noted earlier though, the risk of the meteor striking the building is trivially small, but it is also beyond the organization's influence. Consequently it would be beneficial for reviewers to screen out this hazard and not to concern themselves with the associated risks.

Procedures Defined

At their basic level, procedures are simply a means of communicating task knowledge to personnel who will perform the task. The goal of that communication is to influence their behavior so they accomplish the task as effectively as possible without the personnel having to determine for themselves what actions are required or the best way to perform those actions. The primary sources of that knowledge are the workers who have performed the task before and developed a method that worked, or the designers who planned the task or designed the system with which the worker will be working.

Without procedures to restrict or guide them, workers are faced with the jobs of interpreting what they believe are the organization's goals, and determining an effective series of tasks to reach those goals and remembering the details and order of that sequence—all in addition to their primary function of actually performing the tasks. When procedures exist, workers are provided with information describing what is expected, reducing the need for the worker to make decisions based on local, often incomplete, information. By reducing the reliance on incomplete judgment and on limited memories, significant variation in task execution is prevented.

Efficiency improvements and standardization of the task are possible because the workers benefit by knowing how the task is performed, without the need for lengthy study to determine how best to operate the system. The standardization arising from procedure use allows the same tools and techniques to be used each time the task is performed. Efficiency and task standardization also lead to stabilization of the output, allowing the customer to more accurately anticipate and prepare for the product or service generated by the task.

The content of procedures and the method of communicating them vary according to the intent of the procedures and the system they are intended to control. A procedure could be locked into hardware, as the actions of specialized machines on an assembly line are a function of their design, or could be revisable instructions used by a human worker or a programmable machine.

This paper uses the term *control* to describe these individual items contained in a procedure, though they are a subset of the broader definition of control provided by ISO 31000 (2009) as any measures that modify risk. In written procedures, the controls are easily recognizable as the clauses in the document, but they are sometimes difficult to identify in other contexts. An example of a straightforward non-written control is a mechanical interlock on a cutting machine that requires both hands to be placed on handles. Prominent placement of an emergency stop, on the other hand, is a subtler example. In essence, controls are any means by which undesired states are made more difficult to achieve or desired states are made easier to achieve.

Documented procedures are a special case of non-physical controls, where task knowledge can be communicated explicitly but without the need for the worker and the expert to interact directly. The use of documented procedures allows the organization to benefit from procedure use when it would be impractical or impossible to have the knowledgeable personnel present. Capturing and communicating the knowledge of experts that are remote in distance, time or both allows workers to take advantage of the experts' experience without the cost of having them at the worksite at the time of task execution.

With advances in multimedia technology, it is important to note that non-written procedures, such as audible and pictographic instructions provided by automobile navigation systems, are becoming much more widespread. Interactive procedures such as expert systems are similarly becoming more common as technological advances reduce the costs of implementing them. For this research concerning the assessment of risks associated with procedures, the ubiquity of static, written procedures across most industries led to the choice to focus on these types of procedures. However, it is believed that the lessons learned about written procedures are likely to apply to procedures implemented using other media as well, as there are many issues concerning procedure risks that are independent of the form in which information is presented.

Characterization of Procedures

A working list of seven independent scales describing the characteristics of procedures has evolved during the development of this research. This taxonomy characterizes procedures to describe how the control works and the effect it is intended to elicit from the system. Two characteristics shown in the table in Figure 1-4, Structure and Level of Detail, each represents a continuum of possible values, while the remaining characteristics are discrete descriptors.

<table>
<tr><td><b>Purpose</b></td><td colspan="2">Operating</td><td colspan="2">Protecting</td><td colspan="2">Restoring</td></tr>
<tr><td><b>Nature</b></td><td colspan="3">Inherent</td><td colspan="3">Imposed</td></tr>
<tr><td><b>Structure</b></td><td colspan="3">Comprehensive</td><td colspan="3">Limited</td></tr>
<tr><td><b>Target</b></td><td colspan="3">Process</td><td colspan="3">Output</td></tr>
<tr><td><b>Level of Detail</b></td><td colspan="3">Goal-oriented</td><td colspan="3">Rule-oriented</td></tr>
<tr><td><b>Method</b></td><td colspan="3">Directing</td><td colspan="3">Limiting</td></tr>
<tr><td><b>Duration</b></td><td colspan="3">Discrete</td><td colspan="3">Sustained</td></tr>
</table>

Figure 1-4 - Control Characteristics Scales

The influence a control has on risk under different circumstances is related in large part to the mix of these characteristics and the circumstances.

- Purpose (operating vs. protecting vs. restoring): describes the way in which a control influences the system. Operating procedures are success-oriented and provide workers with information about activities that comprise the direct path to the output. Protecting and restoring controls are the items that are not explicitly necessary for an output but are necessary for assuring continued ability to provide the output. Examples of protective controls include cross training, physical barriers, quality inspection and even risk analysis. Restoring controls include corrective maintenance, emergency activities to secure an out-of-control system, and replenishment of safety stock, for example. In essence, protecting and restoring controls are activities that respond to the uncertainty inherent to a non-ideal system.

- Nature (inherent vs. imposed): addresses the means by which the procedure is implemented. The nature of an inherent control is that it is a feature of the system configuration and system states. An example of an inherent control is the order of operations in a production line. The alternative, an imposed control is the communication of a decision to the workers to resolve an ambiguity in the process flow or system states. Fabrication instructions for a machined part are an imposed set of controls—if multiple tasks could be performed, but an optimal sequence has been identified, the local operator would only know the sequence by relying on the procedure.
- Structure (comprehensive vs. limited): addresses how thoroughly the procedure addresses the associated tasks and the possible variations. A limited procedure may only have controls related to a sub-process or a portion of the time a system will be running, whereas a thorough procedure will contain controls that relay task information for all possible settings on a machine. Since individual

controls are most effectively parsed to small portions of a system's activities, this characteristic is relevant only to the scope of the entire procedure.

- Target (process vs. output): points to the subject of the control—if it is aimed at controlling the processes of the system or the output of the system. A process-targeted control inherently assumes that an effectively managed process is necessary to sustain acceptable outputs. Output-targeted controls make no assumptions and aim to directly control the resulting product or service. This categorization is roughly equivalent to the protection/production objectives identified by Reason (1997).
- Level of Detail (goal-oriented vs. rule-oriented): describes how specifically an individual control is presented. Goal-oriented controls typically leave flexibility in how they will be achieved. Specific controls will provide the step-by-step details necessary to accomplish the task under the anticipated conditions.
- Method (directing vs. limiting): describes whether the control establishes a desired state or prevents an undesired one. Conventional physical or procedural safety barriers that seek to ensure that personnel and sources of stored energy are not in close proximity would be considered limiting controls, while production activities would be typically be directing controls.
- Duration (momentary vs. sustained): the timeframe over which the worker exerts effort to comply with the procedure. Momentary tasks have an inherent end, such as flipping a switch or entering an input to a computer. In contrast, sustained tasks are ones that are executed until a procedural cue ends the need

for the task. An example of a sustained task would be monitoring a temperature gauge to ensure the system remains within the defined limits until the operation is complete.

Purpose in Relation to Resilience and Vulnerability

These descriptive characteristics of procedures describe factors that influence an organization's risk, depending on the conditions. When considering risk however, 'purpose' stands out as unique among them because the need for protective and restorative procedures exists directly in response to the existence of risk. In a deterministic environment, operating procedures would represent the full set of procedures needed because variation in materials, processes and performance would not exist—all activities performed would add value to the output product or service.

In the uncertain environment of actual systems, however these variations require activities to preserve the value already invested in the system or the output. Design margins are one means of combating material variation, while maintenance, inspection and rework are techniques for responding to process and performance variability. The operating controls represent the traditional *value-added* activities that benefit a customer. In contrast, protecting and restoring procedures are activities that do not benefit a customer, but instead benefit the organization. This subset of non-value-added activities, the non-wasteful, *value-preserving* activities are the tools that the organization uses to create resilience in their systems.

Vulnerability and resilience, like risk, are terms with multiple definitions, and no standard consensus. In a discussion of the definitions of resilience in systems, Haimes (2009) references a definition for vulnerability and multiple definitions for resilience that are close in meaning to the general definition needed in discussions of controls, but each contains elements specific to its own context so none are quite suited.

One definition describes resilience as "…the ability of the system to withstand a major disruption within acceptable degradation parameters and to recover within an acceptable time and composite costs and risks." (Haimes, Crowther, & Horowitz, 2008) This definition contains the elements necessary for considering resilience relative to controls, but contains certain assumptions that must be relaxed: first, that resilience applies only to "major" disruption; second, that resilience includes the concept of acceptability; and lastly that all disruptions are negative.

The definition used for vulnerability is "the inherent states of the system…that can be exploited by an adversary to adversely affect…that system." This definition too contains assumptions that must be relaxed for a general discussion: the existence of an adversary and the adverse nature of the effect. By relaxing the assumptions, the resulting generalized definitions for vulnerability and resilience are:

- Vulnerability – the system states that expose a system to a disruptive event
- Resilience – the ability of a system to withstand unwanted disruption, accept beneficial disruption and to recover to a nominal state.

Role of Controls in Procedure risk

Since controls are the "mechanisms, techniques and processes that have been consciously and purposefully designed in order to try to control the organizational behavior" (Johnson, 1993), procedural controls are the organization's method for affecting the risks that are dependent on worker behavior.

As an example, consider the process of writing work instructions: a policy clause for the author of a set of work instructions to "number work steps in increments of 10" is a protecting control that specifically describes how the author is expected to act, leaving space for the addition of other steps at a later date and establishing a sequence so that a worker will notice the absence of work steps if all the pages of the work instructions are not present. An example from a process for addressing nonconforming hardware, "Tags on nonconforming items will not degrade the items' functional performance" shows another protecting control, this time identifying a goal to be satisfied without specifically identifying any actions.

As part of the intended sequence of operations, actions and requirements called out in procedures result from formal and informal risk analysis efforts that intend to prevent loss to support commitments to safety and mission goals (Hale and Borys, 2013a).. Risk would then derive from procedural controls that: do not prevent the loss; do not consistently perform their functions; or result in functions that are performed unnecessarily. Failing to prevent the loss or operating consistently do not themselves create a risk, but instead allow existing risks to persist, sometimes when procedure designers believe them to be effectively mitigated. Unnecessary functions on the other hand create new risks in two ways: directly increasing cost or duration of the task; and indirectly by creating additional interactions, leading to opportunities for error (Reason, 1997). One categorical example of these wasteful controls is procedures generated primarily to avoid liability, which are not necessarily intended to be followed but exist primarily to be referred to in a legal suit (Pélegrin, 2013).

This leads to an alternative view of risk that is only indirectly based on the answers to Kaplan and Garrick’s three questions. Where Figure 1-3 shows the total of all risks, procedure risk is concerned only with risks that have had controls enacted to reduce, mitigate, or prevent the hazard. Figure 1-5 shows the procedure risk in terms of a modified set of questions that address how effectively the controls respond to the hazard.

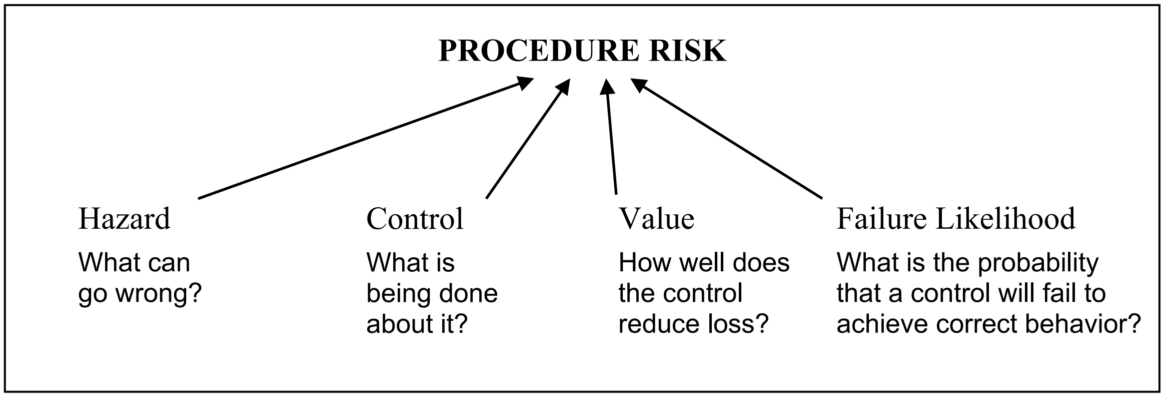

PROCEDURE RISK
Hazard
What can go wrong?
Control
What is being done about it?
Value
How well does the control reduce loss?
Failure Likelihood
What is the probability that a control will fail to achieve correct behavior?

Figure 1-5 - Procedure risk as the answers to the modified questions

How well a control reduces loss is based on two separate elements: that the control provides a necessary function, and that it is successful at accomplishing its function. Neither a control that perfectly performs a trivial role nor a control that fails to fulfill its intended role can be effective at reducing losses. In this schema, the conventional components of risk are replaced with alternatives that give some different information. The substitute for consequence is the Control Value, while likelihood now addresses how likely it is for the control to fail, not how likely a loss would be. The rationale for using Control Value in place of consequence is that procedural controls intended to prevent the greatest consequences would be performing the most necessary functions and therefore be the most valuable controls. Use of failure likelihood as a substitute for the typical likelihood of a loss-event is a considerable difference, but is necessary—when considering procedure risk, this likelihood measure provides information on whether the procedures in place to address a hazard are successful.

Understanding the failure likelihood for a procedural control can be accomplished by directly observing performance and by interviewing the personnel expected to execute the procedure. The observations could include objectively measured failure rates or the reviewers' subjective opinions based on the anecdotal evidence provided by the interviews. The subjectivity of this activity suggests that the reviewers selected for the task should be experts because prior research has shown that experts asked to predict probabilities of undesired events are generally unable to correctly predict outside their field of expertise (Seaver, 1983). Independent of the failure likelihood, though, is the 'value' of a control, which presents a measurement challenge because direct observation is not possible if the control is never called on to perform its function. The Control Assessment technique proposed here seeks to describe the value of the control by considering the consequences of a control failure. Controls that address serious consequences would be more valuable than ones that address lesser consequences. Similarly, a control to prevent a loss that results from a series of events will not be as valuable as one that prevents an inevitable loss.

Though procedural controls are a tool for reducing risk, it is possible to create a risk that did not previously exist. For example, requiring personnel to use protective equipment like a supplied air breathing line during chemical application reduces the consequences of exposure, but can trap personnel in place if they are unable to disconnect the line to evacuate during a fire. Additional procedural controls also have the potential to increase risk by tightening coupling between system elements or increasing complexity of the system. Tighter coupling can allow a series of undesired states to propagate, increasing the likelihood of disturbance to the system. Increased complexity likewise can be responsible for unexpected and vulnerability interactions that contributes to normal failures (Perrow, 1999).

Controls and Worker Behavior

The success or failure of controls is entirely dependent on how well they achieve the desired behavior from the workers who will be assigned to the task. Depending on the culture of the workplace, workers may have different motivations so the differing responses to poor or missing controls could have a wide range of effects on behavior. Therefore, procedures must be evaluated to determine how likely they are to obtain the necessary behavior in real-world conditions.

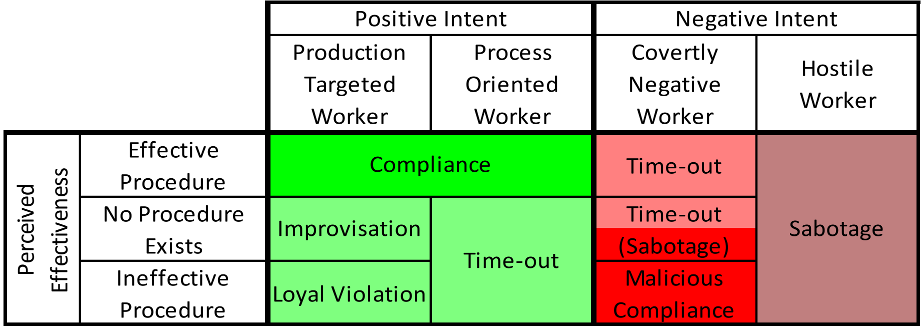
In a workplace culture that values output over process, a missing control can provide a conscientious employee with the flexibility to complete the task quickly, though they may be ignorant of a hazard or eroding design margins in a way their limited role obscures. In a more process-oriented culture, the worker may hold up work while waiting for an update to the procedure to be corrected.

The collapse of the catwalk in the Kansas City Hyatt during a 1981 event is an example of where both situations played out (Petroski, 1992). The construction contractor found a problem with the controls governing the assembly of the catwalk, specifically the installation drawing, stopping catwalk assembly while awaiting updates from the design firm—an ideal example of a time-out scenario.

The construction contractor also provided the design firm with an alternate solution. The design firm relied on the expertise of their licensed engineers to evaluate the design change and did not explicitly require them to recalculate critical loads. In the absence of the load calculations, the design was approved and ultimately proved to be flawed. The lack of the recalculated loads contributed to the approving engineers being convicted of unprofessional conduct in the practice of engineering. By approving the change in the absence of the calculations, the engineers improvised without violating any rules, but the engineering firm failed to obtain the desired behavior.

In the general case of a worker who intends to accomplish the task, they will display different behavior based on the goal of the worker and the perceived effectiveness of the procedure. Table 1-1 shows the various behaviors where the worker's objective is maintaining production or the process in both the absence of procedures and in the absence of procedures perceived as effective by the worker.

Table 1-1 - Nominal Worker Behavior in Response to Procedures



		Positive Intent		Negative Intent	
		Production Targeted Worker	Process Oriented Worker	Covertly Negative Worker	Hostile Worker
Perceived Effectiveness	Effective Procedure	Compliance		Time-out	Sabotage
	No Procedure Exists	Improvisation	Time-out	Time-out (Sabotage)	
	Ineffective Procedure	Loyal Violation		Malicious Compliance	

The table also addresses behavior where workers do not intend to accomplish the assigned task. The behaviors of an openly hostile worker are grayed out to denote that such behavior is independent of the procedures in place and is more appropriately the subject of risk analysis involving an adversary with intent (Reference). Workers who are covertly negative engage in behavior that negatively impacts the organization without giving conclusive evidence of their intent. Whether avoiding the consequences of being caught in detrimental activities or waiting for an opportunity to do greater harm, the lack of effective procedures provides covertly negative workers with system vulnerabilities that can be exploited without significant repercussions to the worker.

The six behaviors identified in Table 1-1 are discussed below in the context of how they relate to risk, vulnerability and resilience. The vulnerabilities identified fall into four categories: safety, performance, cost and schedule, where each vulnerability equates to the possibility of a consequence of that type that is larger than what would exist without the presence of the vulnerability. Safety vulnerabilities translate into a threat of harm to personnel, cost and schedule vulnerabilities address increases in financial cost or time delays respectively and performance vulnerabilities are related to ability of the system or organization to meet goals not related to cost or schedule.

- Compliance – Actions by the worker in accordance with the approved procedure in the absence of knowledge that the procedure is defective. This is the ideal behavioral response to an effective procedure. Regardless of if the worker's objective is maintaining production or the process, a worker intending to perform the assigned task has a clear understanding of what they are to do and the means to perform. In the case of a sufficiently goal-oriented control, it is expected that compliance will involve using judgment and involve freedom-of-action for the worker.

  The unavoidable vulnerability that exists in compliance scenarios is the possibility that the worker fails to recognize an ineffective control, erroneously executing the wrong actions. The responsibility of the procedure designer is to match the structure and detail of the control to the task and worker, not just to improve the effectiveness of the procedure, but also to maximize the ability to recognize ineffective controls.

- Improvisation – Actions taken by workers under their own initiative in cases where the procedure does not identify a sequence of work to be performed. In the absence of a control, improvisation is the logical choice for a worker whose objective is delivering an output. By relying on experience, the worker believes they can execute a sufficient series of actions to obtain the proper output. Improvising reduces schedule and financial cost vulnerability, compared to the alternative Time-out option, but at the expense of possibly creating performance and safety vulnerability if the worker's experience or grasp of the situation is flawed.

  The fundamental difference between improvisation and the freedom to use judgment in complying with a goal-oriented control is the clear intent of the procedure designer. Therefore, the procedure designer is responsible to avoid placing workers in situations where they are expected to exercise their judgment without clearly communicating that expectation.

- Loyal Violation – Work performed by a worker to support organizational goals despite contrary procedural instructions. Like improvisation, loyal violation is an effort to maintain production in the absence of effective procedures; this time forcing the worker to both recognize that the existing control is counterproductive and to determine an appropriate course of action. The performance and safety vulnerabilities created to reduce the financial cost and schedule vulnerabilities are similar to improvisation because both situations represent cases where the worker acts based on their understanding of system performance and organizational goals.

The possible failures of the procedure designer in this case are the failure to provide an effective control, or the failure to communicate the benefits of a genuinely effective control.

- Time-out – Suspending action in response to a procedural conflict or omission so the deficiency can be formally resolved by the procedure designer. In safety-critical situations, Time-out may be preceded by emergency safing activities. For a worker who prioritizes process over production, Time-out is the expected behavior in the absence of an effective control. The wasted time and possible financial costs associated with the delay represent the creation of financial cost and schedule vulnerabilities. As with loyal violation, time-out behavior results from the procedure designer's failure to provide an effective control or to communicate the benefits of the provided control.

  In the case of a covertly negative worker, time-out represents an outlet for exerting their counterproductive aims without the fear of immediate consequences. In the absence of a control, they are protected by the fact that their behavior is identical to some positive-intending workers, so occasionally calling a time-out to falsely claim a control is ineffective could do harm to the organization with no lasting consequence to themselves if done seldom enough.

  Much like the behavior of hostile workers, this covert negativity is not the result of procedures and cannot be easily eliminated by improving procedures though the widespread existence of deficient procedures assists the worker by camouflaging the deceitful time-out. Instead, the responsibility lies with management to end the worker's negative impact by resolving the underlying issues or removing the worker from the system.

- Sabotage – As noted earlier, sabotage is generally the mark of a hostile worker and is not within the realm of procedurally controlled behavior. The exception is in the absence of governing controls, when an antagonistic worker can sabotage the system while claiming to be improvising. Instances of such behavior will typically be limited because workers who engage in this sort of sabotage more than once will be perceived as incompetent and will be removed from positions where they can cause harm.

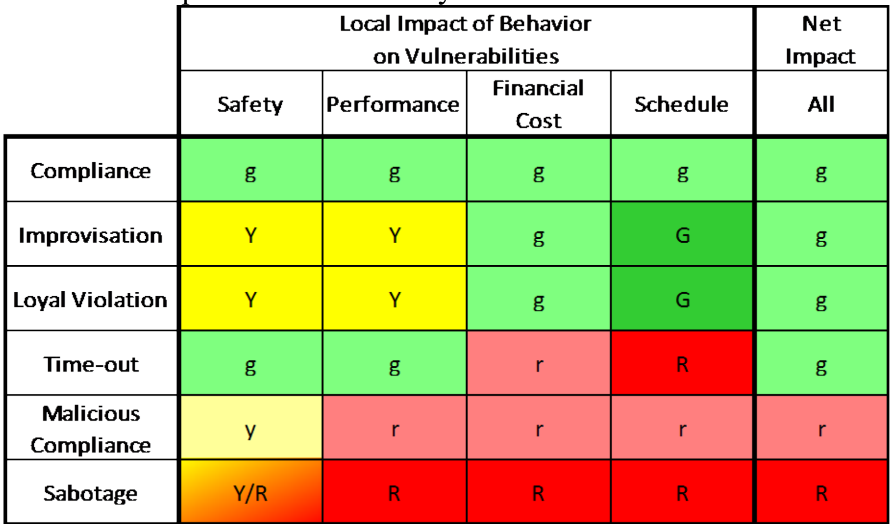
  The vulnerabilities in the system are not categorically limited, as the effects of the sabotage are based on the actions of the worker-as-adversary. Typically though, this form of sabotage does not include added safety vulnerability because workers willing to cause personal harm to themselves or coworkers wouldn't be expected to take efforts to remain covert. As identified earlier, the procedure designer can prevent this form of sabotage by avoiding situations where workers are expected to use their judgment without setting that expectation.

- Malicious Compliance – Malicious compliance is the set of behaviors where a worker recognizes the deficient control and chooses to comply because the destructive effects suit their negative ends. As with sabotage, safety vulnerabilities are generally not created in response to a deficient procedure—an adversary willing to engage in that activity would seek to maximize effect, not wait for the existence of a bad control to minimize reprisal.

  The vulnerability created by maliciously complying with deficient procedures depends on the shortcomings of the controls and is not categorically limited. Resilience against malicious compliance is difficult to achieve because procedure designers would have to implement preventing or recovering controls against the unanticipated failures of their processes—failures, which would be corrected in the original procedure if recognized sufficiently far in advance to plan in the resilience. This leaves only the planning of good procedures as the primary defense against malicious compliance.

To summarize the impact of behavior on vulnerabilities, Table 1-2 shows the relative local and immediate impact on the Safety, Performance, Financial cost and Schedule, and the long term net impact organization-wide for each of the six behaviors. Positive impacts are marked in light and dark green, while negative impacts are in shades of red. Yellow blocks represent neutral or uncertain outcomes. The brighter colors, also marked with capital letters, represent the cases where the impact is independent of the procedures.

Table 1-2 - Behavior Impact on Vulnerability



	Local Impact of Behavior on Vulnerabilities				Net Impact
	Safety	Performance	Financial Cost	Schedule	All
Compliance	g	g	g	g	g
Improvisation	Y	Y	g	G	g
Loyal Violation	Y	Y	g	G	g
Time-out	g	g	r	R	g
Malicious Compliance	y	r	r	r	r
Sabotage	Y/R	R	R	R	R

Compliance is marked as a positive impact, but dependent on the procedures because this behavior is generally positive, but there is the potential for ineffective procedure not to be recognized as such, and by following them, a worker creates vulnerability. Sabotage, on the other hand is near consistently red and also marked as independent of the procedure. This holds true for both hostility-based sabotage and the opportunistic sabotage of missing procedures, with the Safety vulnerability showing some uncertainty because of the supposed reluctance of opportunistic saboteurs to injure other workers and themselves under general circumstances. Malicious compliance is scored similarly to sabotage, but as noted earlier, the effects of this compliance are limited only to the negative actions identified in the procedure.

The remaining three behaviors are significantly different because they involve a trade-off by the worker. The worker's decision affects the impact on the local vulnerabilities though the impact is generally a reduced overall vulnerability when compared to performing a deficient procedure. The tendency towards net positive impact on vulnerability is because workers who understand the system sufficiently to act after recognizing the deficient controls will generally understand the system well enough to choose a course of action that will be an improvement over the deficiency (Lawton 1998). When they choose to seek direction in lieu of acting, even out of malice, the resulting improvement in procedural direction represents a long term gain compared to the transitory nature of the schedule vulnerability created.

Improvisation and Loyal Violation have identical effects because they are the same action with two different causes. In both cases, the worker takes initiative to continue working in the absence of useful direction. The immediate and positive local impact of reducing the schedule vulnerability is associated with dependent financial cost benefits. Using their judgment and experience to identify the sequence of actions, the worker's understanding of the system determines the changes to the performance and safety vulnerabilities and to the extended financial cost and schedule vulnerabilities—a worker who doesn't realize the full implications of an act may introduce unanticipated system states, which could be beyond the worker's ability to control.

Local schedule and financial cost vulnerabilities for time-out on the other hand are increased by the delays from waiting to obtain input from procedure designers, but vulnerability on performance objectives and safety are generally positive because the designers have the opportunity to provide a better solution than a worker likely to have extensive local knowledge but potentially limited scope.

Control Failure as Worker Error

When worker behavior deviates from the expected, the vulnerabilities discussed above are created. In many cases though, the presence of a vulnerability goes unnoticed because the quality of a poorly designed control is unrecognized by the process designer; if the designer was aware of deficiencies, the process would have already been adjusted to compensate. In the case of a missing control, the absence of the control goes unnoticed for the same reason; had the designer anticipated the scenario where the control was needed, the process would have included such a control.

When these deficiencies are observed as contributors during accident investigations, the failure to behave as expected is not attributed to the control. Instead, control failures are often categorized as human error (Sharit, 2012), with the blame placed on the worker even when investigations reveal that the worker could not access the necessary information during the relevant time period. Accidents attributed to "pilot error" reveal examples of these situations. An aircraft in flight is obviously a complex system and a pilot does not have the option of taking time to review complicated procedures. The pressure to maintain 'production,' the safe flight of the aircraft, is often critical, eliminating time-out as a possible behavior.

Emergency checklists are a type of control that is essential in foreseeable undesired situations such as engine fires and wing-icing. When effective checklists exist, they are available and provide the pilot with the necessary information to avoid improvisation or loyal violation via relying on memory. A pilot with an ineffective checklist or in an unforeseen scenario, however, has no usable checklist and is forced to improvise or work without the checklist, developing a response from memories of how the complex system responds under the observed conditions. While successful recoveries can be documented in the FAA's database, possibly resulting in a design change or even an emergency checklist for that scenario, failed recoveries typically lead to investigations where hindsight identifies a clear sequence of actions to avoid the negative outcome. The failure of the pilot to act in this way is often blamed on the pilot though the sequence was developed during an investigation that takes orders of magnitude longer than the event itself.

One example of a failed recovery being classified as a ‘pilot error’ was American Airlines’ flight 965 from Miami to Cali, Columbia in 1995 (Leveson, 2004). While attempting to land in Cali, the pilot lowered the flaps and directed the autopilot to turn towards the non-directional beacon at the airport, named ROZO and marked on the charts as R. By typing in the letter ‘R’ into the flight management computer, rather than the full name of the beacon, the pilot erroneously selected the ROMEO beacon near the Bogotá airport, causing the plane to turn not towards the Cali airport but towards Bogotá and the mountainous region between (Leiden, 2002).

The pilot recovered control by disengaging the autopilot and returning to the original heading that had brought the plane to the Cali area from the coast. Once again, the pilot attempted to direct the autopilot to steer the airplane towards Cali’s ROZO beacon by entering the ‘R’ designator marked on the chart but failed for a second time to note that the computer incorrectly interpreted the input as the ROMEO beacon in Bogotá. With the autopilot already disengaged however, the autopilot did not turn towards either beacon, further distracting the pilot by not acting as expected. Instead, the airplane proceeded towards the mountains ahead as the pilot attempted to understand and respond to the problem. When the proximity warning alerted the pilot of an impending collision with a mountain peak, there was insufficient time to react because the plane was still configured for landing and the lowered flaps reduced the airplane’s maneuverability.

The pilot errors identified by investigations after the fact include the pilot’s failure to notice that entering ‘R’ into the flight management computer selected a beacon other than the intended one and forgetting that the flaps were down so aircraft performance would be sluggish. Identified as a contributing factor was the unfortunate coincidence that another beacon was within range and was the default chosen by the flight management computer for the input that matched the published charts.

As an alternative view of this case, consider the failed controls:

- a beacon identification system where ambiguous single-letter identifiers were allowed
- a data entry system that offered an alternate beacon as the default selection for such an ambiguous entry, though complete and correct
- an auto pilot computer that accepted instructions while deactivated but provided no feedback of that status, even after multiple duplicate inputs
- that American Airlines did not use the portion of Boeing's B-757 Flight Crew Training Manual, which identifies that "The Captain should keep his right hand on the speedbrake lever whenever they are used in-flight." (NTSB 1996)

The associated vulnerabilities were created when the failure occurred, not in the instance of execution. Despite the prior existence of the vulnerability, the pilots were erroneously ruled guilty of willful misconduct by a federal judge (New York Times, 1997). Had the judge considered the failures of the controls and underlying failures of the designers, the otherwise unnecessary cost of the appeal that overturned the verdict (Associated Press, 1999) could have been avoided.

Controls as a Risk Reduction Technique

Since the use of controls are the means by which managers and procedure designers manage risks, it is their responsibility that must be acknowledged when considering the role of controls in reducing risks. Rather than focusing on workers at the 'sharp end' when evaluating risks, and especially realized consequences, it is the managers and procedure designers at the 'blunt end' (Dekker, 2006) who should be accountable for the strategic response to risks.

Ideal controls provide the worker with the knowledge and tools necessary to eliminate the uncertainty that cannot be designed out of the task. Deficient controls, on the other hand represent situations where workers are exposed to uncertainty they are not equipped to handle. These uncertainties can be conflicts in the organizational goals where the designers do not remove the ambiguity for workers. They also take the form of conflicts created by the controls enacted by management and procedure designers, such as when the procedures call for time-out in the face of recognized uncertainty but the culture does not tolerate the associated delays (Dekker, 2005)

The uncertainty allowed or created by managers and procedure designers creates a personal vulnerability for the workers. In these situations, they become exposed to the possibility of being blamed for a negative outcome, sometimes by the same managers who created the conditions for that outcome. By forcing the choice on the workers when no sufficient procedure exists, managers and procedure designers are negligent because they entice workers to develop informal work systems, which hindsight tends to inflate as a causal relationship (Dekker, 2005). The result of allowing this to continue is that the best workers realize their vulnerability in advance and seek alternate employment, taking their experience, while the morale and initiative of other workers is ruined by incidents where any workers are blamed for being “the inheritors of system defects created by poor design, incorrect installation, faulty maintenance and bad management decisions.” (Reason, 1990)

The alternative is to identify deficient or missing controls by turning a critical eye on the uncertainty that is flowed down to the workers by the organization. Dekker notes (2005) that a constant investment in trying to monitor and understand the gap between procedure and practice is a distinguishing characteristic of High Reliability Organizations (HROs). The systematic review of controls described in this research is one method of enabling an organization to formalize that monitoring and of identifying appropriate changes with the goal of moving an organization towards being a HRO.

CHAPTER 2 ASSESSING PROCEDURAL CONTROLS

Each action or requirement placed in the procedure by the process designer or the procedure author corresponds to a part of the process where an undesirable state could develop or where a process gone wrong could be recovered to a desired state. These procedural controls can be examined to systematically identify the inherent hazards that the designer and author considered during design of the process and development of the procedure. By simply asking “what was this intended to prevent?” an analyst without extensive background in the specific process can identify the hazards that concerned the experts during their participation.

Once the hazards have been identified, ranking of the risks can be performed at the level of the procedural control by evaluating the consequence and likelihood for each hazard. However, in contrast to the conventional method of examining the possible consequences of the hazard, Control Assessments consider the consequence of a failed control. Similarly, the likelihood determination via Control Assessment does not simply consider the likelihood of the specific consequence, but instead evaluates the likelihood of the control failing. By making these substitutions, the assessment does not examine “risk” as classically defined, but instead examines an analog that may be easier to determine and still reflect the threat to the organization.

Cataloging and Screening Hazards

Using the procedural controls as the focus for identifying and ranking risks allows screening criteria to deselect some hazards and risks from the analysis, making the process responsive to the time available for personnel to perform the assessment tasks. Answers to Kaplan and Garrick’s first question “What can go wrong?” help define the hazard, but depending on the nature of the control what can go wrong will vary.

Sometimes it may be a minor deviation from an arbitrarily chosen sequence, or it could be an important task in a tightly coupled process that is performed improperly or omitted.

To get a full understanding of the procedural risks, all controls would be evaluated to identify hazards. Then, for each of these hazards, the consequences of control failure and the likelihood of each control failure would be determined. Although corrective actions for each item on this comprehensive set of risks could, in principle, be implemented to reduce or eliminate these procedure-related risks, in reality the resources provided by the organization for finding and implementing corrective actions are often limited. Thus, screening criteria often need to be made available to reduce the set of hazards to a size that is manageable under the resource constraints.

One screening technique currently used by NASA that is well suited to perform this role is a criticality assessment, which is used by reviewers evaluating a system's reliability. While the term criticality assessment is also used in industry for the summary reporting the results of a Failure Modes and Effects Analysis (FMEA), NASA uses this technique with the same name as a precursor to performing the Failure Modes and Effects Analysis (FMEA). In its more standard use, a criticality assessment provides a criticality number for each failure mode based on the probability of the failure mode occurrence, severity of the failure effect, and the chance of the failure being undetected (United States Department of Defense, 1980). NASA's criticality assessment, on the other hand, is used to screen out system functions that are non-critical, allowing the FMEA to be performed only where there is potential for a critical outcome (National Aeronautics and Space Administration, 1986).

In this context, the critical outcomes are defined explicitly by NASA as a loss of life, a spacecraft, a non-redundant spacecraft system, or the capability to perform a mission objective (Garrick, 1989). Losses of redundant systems are evaluated to determine if a single failure could eliminate both the primary and redundant functions. Safety systems, such as alarms or fire suppression systems are evaluated, even though the loss of such a function would not itself have any consequences without another failure.

During a Control Assessment the organization performing the assessment would select the criteria to identify what is critical to the organization so that minor hazards could be screened. For example, a retailer may be most concerned by employee theft, while theft would be a remote concern for an aircraft manufacturer that is focused on on-time delivery. An accounting firm might consider the integrity of its audits critical and a venture capital group would be most concerned about return-on-investment.

For criticality to be used as an effective screening criterion, a critical control would be one whose failure can possibly cause an outcome that the organization has explicitly identified as critical to the continued ability of the organization to accomplish its goals. Controls intended for the routine management of a process can often be eliminated from consideration because they would not impact critical functions, so a failure wouldn't stop the organization from reaching its goals or threaten its existence. A non-critical control for one organization may be critical for another, or may even become critical, as the organizations' criteria for determining criticality evolve.

The Space Shuttle program has unfortunately had two high-visibility cases where criticality of a function was adjusted as a result of catastrophic losses. In the case of the

Challenger accident, the function of the solid rocket booster o-ring was ambiguous, and for *Columbia*, it was the process for bonding insulating foam onto the external tank.

With *Challenger,* NASA chose to launch despite evidence that hot gas blow-by around the o-rings on the solid rocket booster was seen on previous flights (Vaughn, 1996). While blow-by was not desired, this indicated that the accepted function of the o-ring was not to prevent all gas flow, but to minimize the blow-by. Redesign of the o-ring joint and tighter temperature restrictions on launches to prevent blow-by shows that by the time the shuttles returned to flying in 1988, NASA considered the critical function of the o-ring to be preventing all hot gas flow.

The loss of *Columbia* was caused by impacts of insulating foam falling from the external tank, which had also been seen in prior space shuttle flights. The *Columbia* Accident Investigation Board report noted that “damage caused by debris has occurred on every Space Shuttle flight…” and that foam falling from the external tank bipod area, the source of the piece that destroyed Columbia, had first been seen on STS-7 in 1983, nearly 20 years before the accident (CAIB). The bond between the foam to the tank was originally considered critical, with a baseline design requirement that “no debris shall emanate from the critical zone of the external tank on the launch pad or during ascent…” Unfortunately, that control became ineffective when NASA decision makers decided that the bond between tank and foam was not critical. This decision was never stated as such, but it was implicit in that they allowed space shuttles to launch even though debris was in fact emanating from the “critical zone.”

In both of these cases, schedule pressure was cited as a factor that contributed to the decision to launch. As noted earlier, NASA’s definition of critical is a problem that results in loss of life, a spacecraft, a non-redundant spacecraft system, or the capability to perform a mission objective. Conspicuously absent are cost and schedule criteria, of which, at least schedule has been a *de-facto* criteria at the time of both shuttle accidents. Assuming NASA had intended for schedule to not be considered a primary factor in determining readiness to launch, screening the hazards in the proposed manner may have explicitly revealed the extent of the schedule pressure and provided a re-evaluation of decisions that appear flawed in hindsight.

Consequence Evaluation

A procedural control, in the form of an action to be performed or a requirement to be met, can potentially prevent the existence of a hazard by preventing the condition, or it can simply reduce the likelihood or the effects of the hazard by creating limits and barriers. When the action is not performed or the requirement is not met, however, the control fails and the result is essentially the same as if the hazard was left uninhibited—a failed procedural control doesn't reliably prevent the hazard or mitigate the consequences.

Comparing the consequences of failed controls involves the same difficulty with arbitrary comparisons between different types of consequences as are seen with other quantitative and pseudo-quantitative risk assessment techniques, such as FMEA and to a lesser extent, the Dow Fire & Explosion Index (American Institute of Chemical Engineers, 1994). To resolve this ambiguity, Control Assessment avoids directly evaluating the consequences and instead considers the strength of a control. In the proposed scheme for assessing the risks of procedural controls, the *Control Value* (CV) Score represents how effectively that specific procedural control will reduce the consequences of the associated hazard.

Control Assessment will use two attribute scores as components to the CV: opportunity for intervention and inevitability of consequence. Loss scenarios that are less tightly coupled will provide some opportunity to intervene and will tend to have a lower CV Score because the intervention can negate the consequences arising from the initiating events. Scenarios where the critical consequences won't be realized unless a string of other contributing events occurs will also have lower CV Scores, while scenarios that require active intervention to prevent critical consequences will tend to have higher CV Scores.

Using an example from space shuttle ground processing, there is a requirement for operations involving hypergolic rocket propellants to have limited access. Procedures require that perimeters are established around these operations so that only personnel in SCAPE (Self Contained Atmosphere Protective Ensemble), a sealed full-body garment that does not react violently with fuel or oxidizer, are allowed into the area. The CV score of this requirement would be different in the Orbiter Processing Facility (OPF) than it would be at the launch pad.

There is a greater opportunity for intervention at the pad because the area surrounding the operations is a cleared field so that anyone violating the perimeter can be seen before they are close enough to be exposed, giving the opportunity to warn them away or suspend work until they are clear. In the OPF, on the other hand, the work platforms surrounding the areas where propellant operations occur will obscure the view, allowing someone who violated the perimeter to approach within a few dozen feet without being seen. As a result, the control requiring the establishment of the perimeter in the OPF would have a higher CV score because there is less opportunity for intervention before an exposure would occur.

The CV score at the pad would also further be lowered because of the series of problems that would need to happen before an accidental exposure would occur. In both locations, small amounts of fuel or oxidizer are released into the air when hoses are disconnected. The ventilation system in the OPF turns over the air slowly enough that chemicals dissipating from a typical release could achieve dangerous concentrations outside the immediate area where the technicians are working. At the pad however, operations are outside and exposed to the elements. Consequently, there is sufficient air to dissipate the chemicals and normal releases do not reach dangerous concentrations more than a few feet away from the source unless a strong wind is pushing the cloud. For a person who violated the perimeter in the OPF, exposure to dangerous concentrations would almost certainly occur if they approached, while exposure at the pad would require that there be a strong wind and that they approached from directly downwind.

The opportunity for intervention also accounts for a variety of factors, such as detectability and coupling. In the case of an undetectable hazard, personnel capable of acting would not know what action was necessary and the CV score would be driven higher despite the existence of potential interventions. For tightly coupled systems, this means that even immediate notice of a critical condition may not provide sufficient time to act, so the intervention might be too late or not happen at all.

Continuing with the examples of propellant operations at the launch pad and processing hanger, a ruptured hose at the pad would be detectable nearly instantly because the hoses are all within sight of the operator. A leak at the OPF may not be seen directly and would only be evident by looking at the pressure or flow rate indicated on an instrument panel. The possible delay could result in a larger spill, higher concentrations, and a greater chance that someone who violated the perimeter would be exposed, thereby raising the CV score of the control that requires a perimeter around the operational area.

Likelihood Evaluation

As discussed above, the proposed Control Assessment approach diverges significantly from other risk assessment techniques in how consequences of hazards are treated. The same can be said for how Control Assessment treats likelihood. The assessment starts with the assumption that the critical outcome will occur at a time that will seriously harm the organization unless there is some control placed to prevent it. Since the outcome is expected whenever it is not prevented by a control, Control Assessment considers the likelihood of the control failure rather than trying to approximate the probability of the outcome/consequence as a component of the expected loss. This distinction is not just a subtle one, as a failed control does not necessarily result in a negative outcome. It should be noted, however, that situations where workers perform the correct action despite a failed control are undesired. Without the repeatable process, an action could not be reliably expected to occur again the next time the process is executed because changing conditions may overcome the factors that caused the worker to select the right action in that instance. Without a successful control, the next instance of that process may fail because of a difference as straightforward as a less experienced worker performing the task or as subtle as a temperature change.

Determining how often controls fail involves looking at the failure mechanisms of the controls. The most obvious case of a control that will not affect a worker's behavior is when the worker has a negative intent. Damage resulting from an employee who intends to do harm to the organization by sabotage is outside the scope of control assessment because it is not the result of a control failure. Malicious compliance, on the other hand, is when an employee with a negative intent complies with a procedure they believe to be ineffective or counter-productive to the goals of the organization. This malicious compliance presents a procedural risk because the flawed procedure contains ineffective or failed controls—a worker who is aware that the procedure is not correct but who nonetheless follows the procedure *would not be* executing the actions desired by the organization but would be safe from reprisal.

Malicious compliance is a special case of the first way controls can fail: by not clearly agreeing with the organization's expectations. A control that is ambiguous or conflicts with expectations will leave a worker unaware of the correct action to perform, or in the case of a malicious worker, provide a plausible excuse for acting against the best interests of the organization.

The second way controls can fail is to instruct the worker to perform an action they are unable to, either by providing insufficient details or identifying actions that cannot be performed under the time or resource constraints. The final way controls can fail is by calling for actions that are harmful to the worker. A worker who is aware of what harm may come will not proceed with the action. Usually, such a situation will also be in conflict with the goals of the organization because the costs associated with the organization's liability in such a case could harm the organization as well.

In each of these situations, the worker performs a different action than expected or refrains from performing any action. An unaware worker may happen to perform the correct action, but it will be treated here as an incorrect action—it is not a desired mode of operating to count on happenstance to ensure that workers act correctly.

Although these three failure scenarios describe how a process fails, they are not practical for facilitating an analysis of procedural risk because the level of specification is too general; that is, failures are specified to be the result of badly selected or incompletely described controls. Further specification is necessary to describe the process in useful terms. To accomplish this objective, it is proposed that Control Assessment consider a set of characteristics to describe a process based on five of six basic questions: what? why? when? how? where? and who? 'Why' is excluded because it does not describe the process, but provides rationale for its existence. Providing this rationale can be helpful in motivating the workers who will be performing the task, but is not strictly necessary for successful task completion.

The answers to the five relevant questions are covered by Control Assessment through considering how well the process is defined, to whom it is assigned, the training provided, how the process is organized and what monitoring is performed to ensure that the organization's expectations are met. As explained below, describing a process in terms of how well it is defined, assigned, trained, organized and monitored provides a comprehensive view of the internal functioning of the process as well as the resulting output. It also leads to a handy mnemonic device, DATOM—to remember the key items in describing a process for either assessment or design.

It is intuitive that a relationship should exist between how well the DATOM characteristics are addressed by a process and how successful that process will be at accomplishing its goals. If so, evaluating specific procedural controls using DATOM may provide a substitute for failure likelihood as a component of risk in the same way as the Control Value score is a substitute for the consequence magnitude.

DATOM for Describing Processes

Whether designing a new process or examining an existing one, the first step in describing the process is to define the actions that are expected to take place. 'What' must be firmly established for the action to be part of a process. Without an overarching scheme, a worker will not reliably perform an action or sequence of actions to provide the needed output. Defining the 'what' involves deciding on the extent of the actions involved with the task, along with choosing or identifying the parameters that control the task actions.

In processes that are not completely automated, the unique skills and limitations of the workers influence the 'how' 'when' and 'where' so 'who' must be addressed before progressing to the other remaining questions. Without clearly identifying 'who' will be assigned to the task, some level of confusion is inevitable because of the assumptions that must be made by the participants. Schrock (1988) sums up this problem in the context of quality:

> "Coordination is needed. If quality is everyone's responsibility in general and no one's responsibility in particular, crises develop. Critical requirements are overlooked when each department thinks someone else is minding the store."

However, problems persist even when there is an implicit assignment. An action may be consistently performed by the same worker under normal circumstances, but a substitution creates opportunities for misunderstanding. A substitute worker who is capable of performing the task may be unaware that a particular action needs to be performed, or may assume that the action is performed by another worker.

Once the task has been defined and a worker has been assigned to perform that task, ‘how’ the worker will perform the task becomes relevant. For the task to be effectively performed, the worker needs training in the process knowledge specific to the task and in the skills required to perform the expected actions.

‘Where’ and ‘when’ the task will be performed are linked together because both are limited by the defined process sequence. Some aspects of ‘how’ are similarly constrained, particularly in the context of tools, equipment, and other supporting resources. These three items together describe how the process is organized and determine the efficiency, quality, and safety of the process if a trained worker is assigned to the task.

The links between the ‘when,’ ‘where’ and ‘how’ demonstrate the shortcomings with simply using the five questions as the criteria for evaluating a process. To address this problem, Control Assessment looks at the main concepts associated with the questions to identify the criteria for evaluating the process. ‘What’ addressed defining the process, ‘who’ involved assigning the appropriate worker, ‘how’ in part described training of the worker and, along with ‘when’ and ‘where,’ described how the process is organized. However, although these four evaluation criteria can describe the details of a process, they do not address the quality and effectiveness of the process; that is, there is still the issue of whether the process actually produces the desired results. A process without some form of check will forever be vulnerable to changes in the inputs, the environment, or interpretations of the wording of the documented rules. Based on this need, a fifth concept is included in the process evaluation criteria—monitoring.

The concept of monitoring includes activities that report on the 'health' of the process to management but are independent of the process itself. Inspection activities may seem at first to fit into this category but are not considered monitoring because inspections address the quality of a specific instance of procedure execution. The difference between monitoring and inspection is subtle but significant. One way to differentiate them is to consider the action taken as a result of an observed failure. A failure found by inspection would require a remedial action to fix the nonconforming item, whereas one found by monitoring would lead to management deciding if a corrective action is needed to adjust the process. It is possible for an inspection task to have aspects of both inspection and monitoring, but the failure of the monitoring function cannot in and of itself cause a failure in a well described process.

It is important to note, however, that monitoring can bring about improvement in a badly described process. The Rock Problem is an example of a poorly defined activity where feedback can improve the process description (Leffingwell, 2000). In this problem, the customer requirements simply call for a rock, but after multiple iterations of rocks being refused for various reasons—too big, not round enough—the worker can use the feedback from the monitoring to define the rock selection task more completely.

Table 2-1 considers the ways procedural controls can fail by using the DATOM model to identify the potential weaknesses in process design that can cause that type of failure. The monitoring component is not included as a potential source of process failure because the monitoring only identifies how the process was performed during the observed period and does not directly impact the activities until the results of the observations are acted upon.

	Process Failure Modes		
Characteristic	**Worker is unaware of the actions to be performed**	**Worker is unable to perform expected actions**	**Worker chooses to perform a different action than expected**
Define	-Multiple actions are possible to accomplish task -Actions necessary for task completion are not apparent		-A “better” means of meeting the intent of the task has been identified by the worker
Assign	-Responsibility for task performance has not been specifically given to the expected worker	-Worker lacks the physical or mental capacity to perform the task	-A more appropriate worker for the task may exist, so the assigned worker avoids the task until it is reassigned
Train		-Worker lacks the experience or knowledge necessary for task performance	-Worker confidence in the necessary skills is low and an alternate technique can meet the known goals
Organize	-Appropriate cues and instructions are unavailable so the worker is unable to recall the necessary actions	-Necessary tools or resources are not provided -Action sequence is confusing -Difficulty tracking progress allows distractions to impact task completion	-Worker chooses an easier way to perform the task

Table 2-1 - Process Failures corresponding to DATOM characteristics

Effectiveness of Procedural Controls

To summarize, the effectiveness of procedural controls rests on how well a control lessens the consequence of the associated hazard and how reliably the control fulfills that role.

The importance of the role of a procedural control can be determined by using the criticality assessment, but the effectiveness of even non-critical controls can be considered using Control Value and Failure Likelihood.

Control Value reflects how well the control lessens the consequence by considering what would happen if the undesired state continued unchecked, the inevitability of the outcome, and if a mitigation opportunity exists with enough time to affect the consequences.

Likelihood that a control will fail to perform its role is determined by considering the five process descriptors: definition, assignment, training, organization and monitoring. These five characteristics of a fully-described process provide insight into how a procedural control operates and what can be done to improve the reliability of the control.

By combining the elements of Control Value and Failure Likelihood, it is believed that a functional relationship can be established that uses the individual components of the CV and FL scores to construct an overall ‘risk’ score. This score, in turn, is envisioned as a tool or index that can provide an organization with a fast and efficient means for identifying the threat their procedures pose to their success.

CHAPTER 3 VALIDATION

To demonstrate the usefulness of the method described earlier, which uses the control value and failure likelihood to describe risks, a validation was necessary. Since the technique is fundamentally different from conventional methods of assessing risks, the validation used data from actual procedural controls in place at the time to govern space shuttle ground processing. Data was collected from respondents who were expert in space shuttle operations, asking them to compare the model parameters with their opinions of how valuable each procedural control was and how likely the control was to fail. Observed failure rates were also available for some controls, based on the findings of the *Columbia* accident investigation efforts at Kennedy Space Center.

The validation activity was originally planned in two parts, with a third part added after the initial data was received to refine details of the model that were not included in the initial phase. Phase A presented a set of scenarios to a group of reviewers, who were asked to score each scenario for how well the control in the scenario reflected the attributes of the model, along with how valuable the control was and how likely it was to fail. In Phase B, a different group of reviewers was presented with similar scenarios for which the *Columbia* accident investigation data provided actual failure rates of the controls. To minimize consideration of other factors, these reviewers were asked only to score how well the controls reflected the attributes of the DATOM model. These questions used a 100-point scale, rather than the 10-point scale of Phase A to improve the resolution of the responses.

The final phase, C, presented reviewers with the scenarios from Phase A, but included additional details about the model attributes. Similar to Phase A, Phase C captured the respondents' perceived value of each control and its perceived likelihood of failing. However, the response for each model attribute was split into two questions: one to capture the extent to which the characteristic was displayed and one to record the importance of that characteristic. Phase C also included multiple questions for some attributes in the attempt to isolate the influence provided by the different aspects of that attribute. Details of the questions are provided in the section describing Phase C below.

As discussed earlier, the benefit of the Control Assessment technique is to provide an effective means of comparing risks associated with two controls without needing to quantify the consequence or likelihood. Therefore a validation of the technique must show that the rankings of a control's value and the likelihood of its failure can be consistently obtained with the technique. Phase A collected data necessary to show if respondents were scoring the controls consistently for the model components, and to perform a regression analysis to determine if the model components could be used as a substitute for the respondents' perception of the control's value and its failure likelihood. Phase B addresses a shortfall in the concept of Phase A by comparing the DATOM scores directly to observed failure rates. The objective failure rates used for Phase B were the result of the *Columbia* accident investigation's review of Space Shuttle processing during the final flights of *Columbia*. In an effort to determine if ground processing had contributed to the accident, the investigation collected information on how often personnel did not comply with the rules in place. Phase B compared DATOM scores for a subset of the investigated scenarios directly to the compliance rates, allowing a calibration of sorts to be performed so the impact of each DATOM element on the failure rate could be determined.

Using failure rates determined by the *Columbia* accident investigation, the subjective nature of the respondents' scores of perceived failure likelihood can be eliminated.

Data Collection

For each of the three phases, the respondents were asked to enter their demographic information and to record scores for scenarios representing procedural controls. Selection of the scenarios was made in conjunction with management of the Process Analysis engineering group, who were already familiar with the DATOM model.

Scenarios for Phase A were selected from then-current rules, some of which had been adjusted to account for lessons learned in the *Columbia* accident investigation. Phase B scenarios were selected to be representative of situations where the *Columbia* investigation showed that personnel did not comply with procedures. Phase C used the same scenarios as Phase A. While there were respondents in common between the groups from Phases A and C, there was no influence expected from the earlier answers for multiple reasons: the question sets asked were different, the scoring scales had been switched from the 10-point scale to a manual unnumbered scale, and there was a 6-month interval between the times when the two scores were collected.

Data from respondents who did not complete the full set of questions were not included in any analysis. Also excluded were data from respondents who appeared to be misrepresenting their answers, such as the two respondent in Phase A who provided answers to questions on the first few screens that appeared reasonable, but began answering all questions with the highest value partway through the response. This behavior began on different screens, so it was apparent that it was not related to the questions or scenarios, but to the respondents.

Phase A

The initial sample of respondents was 100% of a relatively small group of engineers in the corrective action engineering department for a NASA contractor at Kennedy Space Center. These personnel were experts in both the standard procedures for processing a shuttle for launch and with the DATOM model. This sample provided a representative cross-section of the workforce makeup, as personnel in the department had come from the ranks of the company's technicians, inspectors and engineering workforce. This original respondent set was expected to be candid in their responses because they had positive experiences developing and using the DATOM model to frame corrective actions in response to the *Columbia* processing errors. As a resut, a formal validation of the model represented an opportunity for these respondents to better understand a tool they found useful but incomplete.

Unfortunately, this sample was not large enough to provide sufficient confidence in any conclusions, so supplementary respondents were sought. The additional respondents were expert in shuttle processing but had not been previously exposed to the DATOM model. These additional respondents were obtained by presenting the overview of this study to randomly selected departments and asking personnel to answer the questions on the data collection website. It was possible that these additional respondents might have been reluctant to respond honestly; they could have feared that their responses would be used against them or had some trepidation because they might be perceived as negative or troublemakers if their responses were made available to management. To mitigate this potential, the instructions informed potential respondents that the responses would be confidential and that no-one from management would even know which personnel provided answers. Additionally, the request came through the corrective action engineering department, who had an established reputation for successfully improving procedures that had been generally acknowledged to be difficult to follow, so it was anticipated that respondents would be comfortable in believing that this effort would help identify ways to make it easier for them to perform the tasks the procedure-writers determined were necessary.

The departments were selected to retain the existing mix of technicians, inspectors and engineers with space shuttle processing experience. This continued until a minimum sample of 50 personnel were obtained. The sample size of 50 was a limitation placed by the management of the workforce, wanting to minimize the distraction on personnel who, at the time, were engaged in preparing *Atlantis* for the final servicing mission to the Hubble Space Telescope and preparations for International Space Station assembly missions for *Discovery* and *Endeavour*. The set of 5 scenarios however, provided 250 data points for analysis of the model and other factors that potentially affected scoring.

The respondents were asked to enter responses via a website that collected demographic data that are likely to have an impact on their ability to evaluate the rules, such as: age, education, occupation, and length of time they have been working in Space Shuttle ground operations. Once the demographic information was recorded, reviewers were asked to answer four sets of related questions regarding a specific rule regarding shuttle processing.

Rules selected for this phase were chosen from the specific procedural requirements in dozens of policy documents that were in effect at the time. The selection of the specific procedural controls was made in collaboration with the management of the corrective action department, who were familiar with the DATOM concepts and the Intervention and Inevitability ideas as well.

The first group of questions, or Utility questions, were presented on a single screen for each scenario and consisted of two questions: one regarding the respondent's opinion of whether the rule was a strong or weak procedural control, the other to record their opinion of how likely to the control was to fail. These two Utility questions were repeated for all 5 scenarios before moving on to the next question sets. Responses were entered using a slider linked to a numerical indicator, so the score was entered by adjusting the slider until the desired value from one to ten appeared.

The subsequent screen contained the remaining three sets of questions for each scenario: group 2, the Intervention and Inevitability or I&I questions asked about the opportunity to intervene after a failure but before a consequence and the inevitability of the consequence after the failure, assuming no intervention occurs; group 3, the DATOM questions asked about the DATOM measures associated with the control and group 4, the Posterior Utility questions repeated the utility questions, asking respondents to update their score for failure likelihood and control value after having considered the I&I and DATOM questions.

As introduced previously, the opportunity to intervene and the inevitability concepts in the I&I questions are loosely related to the factors that Perrow discusses as relevant to normal accidents: coupling and complexity (1999). The opportunity to intervene once a problem exists relates to Perrow's concept of coupling, one of two key dimensions in his systems theory of accidents, where systems exhibiting tighter coupling present less opportunity to intervene when a response is necessary. Inevitability relates to

Perrow's concept of interactive complexity, the other dimension in his theory of accidents, though with the focus on the effect of the complexity rather than on the degree of complexity in the nature of the system itself. Perrow explains that the results of complex interactions can lead to unanticipated failures.

Inevitability however also attempts to account for the accidents prevented by the system complexity. While Perrow correctly argues that complexity introduces unanticipated failure modes, complexity also often eliminates the simple input-output link that would allow an initiating event to propagate though a simpler system and would otherwise lead to undesired outcomes. These I&I questions will be analyzed against the Control Value (CV) score from the Utility Questions; thus they will serve as the independent variables to substitute for the CV.

The DATOM questions in the third group collect information from the respondents on the DATOM measures of a control's effectiveness. For each of the aspects, Definition, Assignment, Training, Organization and Monitoring, each reviewer identified how well the control displays that characteristic. These scores will be analyzed as independent variables related the Failure Likelihood (FL) score from the Utility questions as the dependent variable.

For the Posterior Utility questions, the respondents were shown their answers from the original Utility questions in the first group and were offered the opportunity to adjust the values. These scores were analyzed against the I&I and the DATOM questions in the same way as the anterior Utility scores. This was done to determine if the respondents' scoring of the Utility values were significantly influenced by the explicit consideration of the model factors covered in the I&I and DATOM questions.

Demographic data was collected from the respondents to provide the ability to sort the population into relevant subpopulations in the event that *post hoc* blocking analysis was appropriate. Demographic data collected included the respondents' level of education and their experience in different relevant job functions, specifically, Space Shuttle system engineers, technicians and inspectors.

Table 3-1 summarizes the data collected during Phase A and identifies the intended purpose of that data. Other than the demographic information, all items are subjective ratings provided by the reviewers, based on their expert knowledge of the situations presented. The scale for all quantifications is from 1 to 10, lowest to highest.

Table 3-1 - Phase A data types collected

	DATA TYPE	DESCRIPTION	PURPOSE
	Demographic	Individual characteristics such as age, education, occupation, etc.	Potentially used for differentiating sub-populations of reviewers
Utility	Control Value, CV (Initial)	A judgment of how important the control in a scenario is at preventing loss, as perceived by the reviewer before explicitly considering the coupling and inevitability	Provides the basis of comparison for the control value score
	Failure Likelihood, FL (Initial)	Reviewers' judgment, before explicitly considering the DATOM components, of the probability that the control in a scenario will fail to perform its function	Provides the basis of comparison for the DATOM score
I&I	Intervention	A score representing how tight the coupling is between the failure of a control and the consequence	This is a component of a procedural control's strength, which is part of the model that is expected to correspond to the reviewer's score for Control Value.
	Inevitability	Reviewer's scoring to represent how inevitable a consequence will be once the control fails	This is a component of a procedural control's strength, which is part of the model that is expected to correspond to the reviewer's score for Control Value
DATOM	Define	A score representing the interpretation of how well the control defines the task to be performed or the requirement that must be satisfied.	This is a component of DATOM for a procedural control and is part of the model that is expected to correspond to the reviewer's score for perceived Failure Likelihood of the control.
	Assign	Reviewer's scoring to represent how clearly the task was assigned to the person most appropriate to perform the task	This is a component of DATOM for a procedural control and is part of the model that is expected to correspond to the reviewer's score for perceived Failure Likelihood of the control.
	Train	The reviewers' score to identify the training provided, relative to the skills required for the task. Negative scores correspond to skills required but not trained and positive to training provided in excess of the needed skills.	This is a component of DATOM for a procedural control and is part of the model that is expected to correspond to the reviewer's score for perceived Failure Likelihood of the control.
	Organize	A score representing how well the reviewer believes the control organizes the task and how straightforward the task or requirements are.	This is a component of DATOM for a procedural control and is part of the model that is expected to correspond to the reviewer's score for perceived Failure Likelihood of the control.
	Monitor	Reviewer's scoring to represent how well an independent monitor examines the control to ensure the appropriate actions are executed	This is a component of DATOM for a procedural control and is part of the model that is expected to correspond to the reviewer's score for perceived Failure Likelihood of the control.
Utility (Posterior)	Control Value, CV (Reevaluated)	Revised judgment of how important the control in a scenario is at preventing loss, as perceived by the reviewer after explicitly considering the coupling and inevitability	Provides the basis of comparison for the control value score
	Failure Likelihood, FL (Reevaluated)	Reviewers' revised judgment, after explicitly considering the DATOM components, of the probability that the control in a scenario will fail to perform its function.	Provides the basis of comparison for the DATOM score

Phase B

Phase B used a random sample of 50 respondents from among multiple departments randomly chosen at the same NASA contractor used in Phase A. The department selection was stratified to obtain a mix of the technicians, inspectors and engineers involved with the processing of space shuttles. Respondents were contacted by e-mail and asked to visit a link provided, where the questions were presented.

Similar to the DATOM questions in Phase A's third group of questions, the questions presented in Phase B asked about the DATOM elements for 6 scenarios. The scenarios were associated with procedural controls that had been in place during the final two ground processing flows for *Columbia*. For every scenario, 50 respondents each provided scores for the five DATOM questions, using a 100 point scale that more effectively approximate a continuous scoring range than the 10 point scale from Phase A.

The recorded responses included the DATOM scores identified in Table 3-1 and sufficient demographic data to determine which discipline of employee each respondent represented. The utility questions used in Phase A to analyze control value were not included because the accident investigation data did not include any objective details regarding the value of a control, which would have been comparable against I&I scores.

Phase C

Phase C was constructed to determine if there was a method of separating how important a particular piece of the DATOM model was from the extent to which that aspect was displayed in a scenario. Also, factors that had been implicitly grouped under relatively broad headings were split apart from each other where there were multiple related concepts. Respondents for this phase were asked to consider the same scenarios from Phase A and were selected from the group of corrective action engineers used as the gold-standard reviewers in Phase A.

Concerns had been raised about using the same scenarios in both phases, but there was no expectation that the use in phase A would influence the scoring in Phase C for multiple reasons. Of the hundreds of rules that these engineers were expected to be familiar with, there was no compelling reason why they would remember which had been included in the earlier phase 9-months earlier. Assuming though that a respondent remembered the earlier questions and remembered the responses he or she provided, the questions in the newer phase were significantly different and used a different scoring method.

As previously described, Phase A and B used a website for data collection, but Phase C used a manual method of marking responses. Respondents were provided with a multiple page packet of questions during a time that had been set aside by their management for answering the questions. Below each question was a line where they were asked to mark their answer between a pair of defined extremes, as can be seen in the example in Figure 3-1 below.

In your opinion, how valuable is it to have a control that does what Rule #1 is trying to accomplish?

Unimportant –	*Vital* – critical
no critical losses	losses would be
would result	common without
from the absence	this function

Figure 3-1 - Example answer space for questions from Phase C

The data collection packet included a sheet that listed all five scenarios followed by pages with the same groups of questions asked in Phase A: Utility questions, I&I Questions and DATOM questions. The Utility questions were not repeated at the end because the respondents in this phase were already expected to be using the I&I and DATOM model as part of their daily job functions as the basis for routine evaluation of procedural controls. As a result, there was no assumption that the scores would not be affected by their explicit scoring of those items.

The wording of the questions is provided in Table 3-2, broken out by the question group and the model element with which it is associated. The DATOM questions are organized in pairs that represent how well each measured characteristic is displayed (extent) and how important it is for that characteristic to be present (importance).

This measuring of the different aspects of the model elements resulted in the removal of one of the respondent's answers from the dataset. The respondent was one of the early adopters of the DATOM model and argued that importance was not dependent on the scenario, but was absolute for each of the five measures. This respondent argued that he could not understand how to rank the importance separately for each scenario and would not score the answers according to the instructions.

Table 3-2 - Phase C Questions

	Model Element	Question
Utility	Control Value	In your opinion, how valuable is it to have a control that does what Rule #N is trying to accomplish?
	Failure Likelihood	In your opinion, how likely is it that Rule #N will operate and perform its intended function?
I&I	Inevitability	How likely would it be for a critical incident to develop due to the absence of Rule #N? (Consider redundant rules but do not include contingencies outside the procedure)
	Intervention	How detectable would a failure of Rule #N be to the personnel following the rule?
		How aware would personnel be of the way to prevent a critical incident in the event of a failure of Rule #N?
		How much time would personnel have to act to prevent a critical incident after a failure of Rule #N?
DATOM	Defining	To what extent is Rule #N written in clear language? (e.g. commonly used words, consistent acronyms, no conditional statements, no double negatives, etc…)
		How important is it for Rule #N to be written in clear language? (e.g. commonly used words, consistent acronyms, no conditional statements, no double negatives, etc…)
		To what extent is Rule #N unambiguous and easy to understand?
		How important is it for Rule #N to be unambiguous and easy to understand?
		To what extent are personnel aware of the intent of Rule #N?
		How important is it for personnel to be aware of the intent of Rule #N?
	Assigning	How well does Rule #N identify who is expected to follow the rule?
		How important is it to identify who is expected to follow Rule #N?
	Training	How extensive is the training that personnel receive in the process governed by Rule #N?
		How important is it for personnel to receive extensive training in the process governed by Rule #N?
	Organizing	To what extent are the tools provided to personnel appropriate to the task governed by Rule #N? (this includes software, as well as physical tools)
		How important is it for tools provided to personnel to be appropriate to the task governed by Rule #N? (this includes software, as well as physical tools)
		How memorable are the details of Rule #N under the conditions that personnel would be exposed to? (Consider complexity and reminders – mnemonics, checklists)
		How important is it for the details of Rule #N to be memorable under conditions that personnel would be exposed to? (Consider complexity and reminders – mnemonics, checklists)
		How strongly would time pressure hamper personnel while trying to comply with Rule #N completely?
		How important is it for time pressure not to hamper personnel while trying to completely comply with Rule #N?
		To what extent does Rule #N conflict with other goals personnel are expected to support? (e.g. safety, quality, schedule, etc…)?
		How important is it for Rule #N to not conflict with other goals personnel are expected to support? (e.g. safety, quality, schedule, etc…)?
		To what extent may personnel improvise in response to changing conditions to comply with Rule #N?
		How important is it for personnel to improvise in response to changing conditions to comply with Rule #N?
	Monitoring	To what extent does Rule #N have an alternative means to verify compliance? (Consider difficulty of observing and if evidence persists)
		How important is it for Rule #N to have an alternative means to verify compliance? (Consider difficulty of observing and if evidence persists)
		To what extent does Rule #N have an alternative means to verify success? (Consider difficulty in discerning results of the rule)
		How important is it for Rule #N to have an alternative means to verify success? (Consider difficulty in discerning results of the rule)

Phase A Data Analysis

As part of validating that the model could provide a tool to be an effective replacement for the conventional measurement of risk, the analysis of the Phase A data was intended to support one primary conclusion: to what extent the respondents' perceptions of the controls' value and failure likelihood could be related to the DATOM and I&I question values. A secondary conclusion was also considered, to what extent the respondents perceptions of those values and failure likelihoods changed after explicitly considering the model elements.

The analysis of the data included a Generalized Linear Model (GLM) regression for determining factors of the model that had a significant influence on the dependent variable. Additionally, an assessment of the data was performed within and between respondents using Kendall's Tau to measure the degree of concordance in responses to a question. By considering the relative ranking of respondents' scores to the scenarios, rather than the absolute scores, a mix of linear and nonlinear perceived scoring schemas could be observed. For example, two respondents may agree on the order in which 5 scenarios would be ranked, but one scored linearly with values 2, 4, 6, 8 and 10. If the second respondent scored with a more exponential scheme—1, 2, 4, 7, 10—Kendall's Tau would show the concordance but a linear regression might not.

Ideal results would show that the model terms account for 100% of the Control Value and the Failure Likelihood. Additional terms were considered because of the nature of the sample: the effects of individual respondents and the effects of the scenarios. Because answers from the same respondents were used in the analysis of multiple scenarios, it was likely that some differences existed in how each respondent answered for all scenarios. Similarly, the repeated use of a limited quantity of scenarios might reveal scoring anomalies specific to each scenario. The GLM analysis explicitly included the possibility of these differences, though ideal scoring under the model would show no effect from these factors.

Regression Analysis Using a Generalized Linear Model

The analysis of the data used a GLM regression for determining factors of the model that had a significant influence on the dependent variable. Four sets of analyses were performed for this phase, the first pair of analyses used the responses to the initial utility questions as dependent variables and second pair used the posterior utility questions providing the values for the dependent variables.

Prior to using the analytical model for reviewing the data however, the variables were visually compared against each other to determine if there were important relationships that were initially apparent. The visual comparisons were difficult because the ten possible response values for each variable forced most responses to be plotted identically to at least one other response. Appendix A contains the plots of the data for the CV vs. the Inevitability and the Intervention questions and for FL vs. the DATOM questions. The plots show the datapoints with random offsets from the nominal values to allow visualization of the multiple datapoints of the same value. The CV plots include both the full data set, grouped by scenario and separate charts for each scenario to demonstrate the improvement and remaining diccifulty in the visualization of the data using those techniques.

Where the analyses showed no significant influence from specific model terms, those terms were removed from the regression analysis to determine if the reduction in degrees of freedom provided better insight to the remaining terms. This was particularly useful in some cases where an analysis suggested that a relationships may have existed but could not be established because of the small sample population.

The general form of the regression equation is in Figure 3-2

$$y_{i,j} = C + [ID_i] + [Scenario_j] + \sum_{k=1}^{m} \{Q_k * x_{i,j,k}\} + \varepsilon_{i,j}$$
$$+ \sum_{l=1}^{n} \{[Scenario_j \times Q_k]_l\}$$

Figure 3-2 – Phase A GLM Regression Equation

In this equation, a fitted value for Control Value or Failure Likelihood, along with an associated residual error, is calculated for each dependent variable *y*. The equation for calculating the fitted value as a function of the respondents, the scenarios and the model questions, uses the subscripts *i*, *j* & *k* as indices to represent the respondents, scenarios and included model questions respectively. The subscript *l* is an index to represent the terms that show the interaction between scenario and the model questions.

For each respondent, *i*, the dependent CV or FL score associated with the 5 scenarios (*j*=1 to 5) is compared against the fitted value that includes a constant held fixed across all respondents and scenarios. The next term in the fitted value, $[ID_i]$ is vector containing a calculated offset value for each respondent, which is constant across questions and scenarios. Similarly, $[Scenario_j]$ represents a vector with offset values for each scenario that is independent of respondent or question. In both of these vectors, the values sum to 0, as the net effect across all *i* reviewers and *j* scenarios is accounted for in the constant term.

Q_k is the vector containing the coefficients for each of the questions included in the regression model, which is multiplied by the value of each score. For the analyses using CV as the dependent variable, the index *k* had a maximum of *m*=2, one each for Inevitability and Intervention. The analyses that used FL as the dependent variables had up to 5 Q_k values, one for each of the DATOM questions. In cases where one or more of the model terms were removed, the value of the index was reduced to match. As an example, the coefficients for the model terms Inevitability and Intervention represent a factor, *β* in the simplified regression equation in Figure 3-3 for each respondent *i* and scenario *j*.

$$y_{i,j} = C + \beta_{inev} * x(inev)_{i,j} + \beta_{int} * x(int)_{i,j} + \varepsilon_{i,j}$$

Figure 3-3 - Simplified Control Value Regression Equation

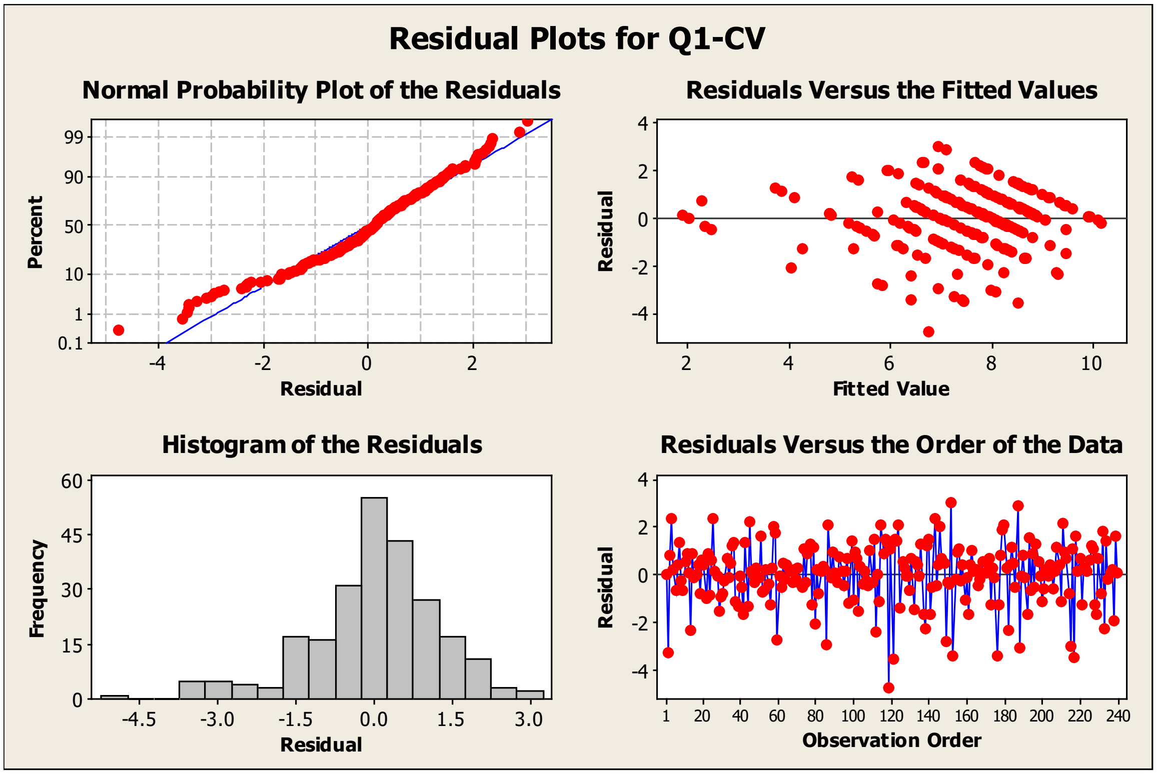
The β factor here represents the Q_k in the generalized case, using *k*=1 for Inevitability and *k*=2 for Intervention. This simplified equation omits the effects of the scenario and the respondent, which are included in the GLM analysis.

The term for $[Scenario_j \times Q_k]$ is a vector of coefficients included to capture the magnitude of interaction between the scenarios and the questions in situations where this is modeled—cases where the responses for one question/scenario combination is significantly different from the others. The term may account for a quantity of scenario/question interactions up to the number of questions, *m*, included in the question term. This interaction term is not used in all of the regressions performed on the model, but is included because some of the subpopulations analyzed appeared to have situationally dependent interactions where the additional term was expected to better describe the relationship between the dependent and independent variables where the responses to the questions were not consistent across scenarios.

Values and coefficients determined in the regression analysis to be non-zero with more than 95% confidence are reported as significant, and items with better than 90% confidence are reported as well. While the lower confidence is by no means conclusive because there is an increased chance that it could be the result of random variation, it suggests areas where a relationship may exist that is masked by the relatively small sample size.

The results of each regression were reviewed to ensure the validity of that analysis, with an emphasis on the distribution of the residuals. The small size of the data set accounts for some departure from a normal distribution, with the effect magnified as analysis the subpopulations used data from only a fraction of the respondents. Figure 3-4shows the 4-in-1 output of the comparisons MiniTab performs on the residuals.

Figure 3-4 - 4-in-1 plot of residuals for the first Control Value regression



This example shows the results from the first regression for Control Value as the dependent variable. The order of the data in the plot on the bottom right quadrant reflects responses for all 5 scenarios in the order in which respondents began answering questions. This order is roughly equivalent to the collection order, but not exactly because of concurrent data collection times, where some respondents had not completed the scenarios before the next respondent began recording responses.

The histogram in the bottom left quadrant shows a peak at zero with a smooth slope on the positive side and an overall downward trend on the negative side. The small rise in frequency near -3 is suggestive of a second peak, though there is insufficient resolution to make a determination due to the size of the data set, though this is supported by the normal probability plot in the upper left quadrant.

The odd stratification in the plot of residuals vs. the fitted value in the top right quadrant is due to the integer nature of the responses. As a result of the integer scores recorded for the dependent variable, there is a functional relationship between the residual/fitted-value pairs that are possible. The plot reflects the constraint, showing plotted points only on the nine lines associated with the integer scores 2 through 10 recorded for the Control Value dependent variable. A tenth line is not evident on the plot because there were no responses in the data set with a recorded 1 as a Control Value.

Phase A Regression on Control Value

The initial regression to determine if the model elements (Inevitability and Intervention) could be used to represent the Control Value (CV) score yielded one coefficient at the 95% significance, the *ID* term. The interpretation of this result is that the most significant source of the differences in CV is the differences between the respondents. However, with a heterogeneous population, such scoring may represent the differences between self-consistent sub-populations, rather than underlying scoring biases from individual respondents.

Before moving on to look at subpopulations though, one additional item deserved further scrutiny. The Inevitability score was significant to the 90% level, so the removal of the Intervention score was considered as a way to determine if the effect of Inevitability would be significant in the absence of the diluting influence of the other term on the power of the model as a substitute.

In the absence of the Intervention term, the follow-up regression analysis revealed that the Inevitability was indeed significant to the 95%, with no appreciable change in the adjusted R^2 value (43.3% with Intervention vs. 43.5% without). The lack of impact on the model's ability to fit the CV was puzzling when one considers the practical foundations of the model being reviewed—that increased opportunities to intervene before occurrence of an effect would decrease the value of a control intended to prevent that effect. Rather than immediately concluding that intervention was irrelevant to the model, an alternate possibility was considered: that the effects of intervention are obscured by a dependence on the scenarios.

To test this alternate interpretation, a third regression was performed to determine if the interaction between the scenarios and the Intervention scores were a significant factor in the model. The results of this analysis showed that the interaction was significant and also that the reduced P-value for the coefficient associated with the scenario suggested that the scenario might directly be a factor with significance if the population was larger. A separate regression was performed to check for a significant interaction between scenario and inevitability, but the interaction was not significant and the significance of all other factors except ID were reduced.

Table 3-3 shows the results of the three regressions performed on the entire sample population to determine the portion of the model best able to represent the respondents' control value scores. The table includes the P-values from the ANOVA table corresponding to the coefficient associated with the respondent ID, the scenario, the Inevitability and Intervention scores and the Intervention interacting with the scenario. Items highlighted in green where they are statistically significant at the 0.05 level and in yellow where P-values are significant to the 0.1 level, suggesting significance but without being conclusive, possibly because of the small sample size. R^2 values are also included to identify how well the model fits the CV as a dependent variable. Care must be taken to avoid drawing conclusions about the model as a suitable substitute from the R^2 values, as the results include the respondent ID as a significant factor, indicating that respondents' scoring biases may be the largest factor affecting the CV score.

Table 3-3 - Control Value regressions performed on entire sample population

CV Regr.	ID	Scenario	Inev.	Int.	Int. x Scen.	R^2	R^2(adj)
1	0.000	0.196	0.077	0.689	-	55.86%	43.28%
2	0.000	0.196	0.047	-	-	55.82%	45.53%
3	0.000	0.076	0.037	0.693	0.029	58.39%	45.36%
Post	0.000	0.026	0.005	0.330	0.007	61.81%	49.85%

An additional regression analysis was performed on the whole sample of respondents, using the post-survey Control Value scores as the dependent variable, rather than the values captured before the respondents recorded scores for the model elements. These results, marked as 'Post' on Table 3-3, indicate that the same relationships exist in the revised scoring of the control value, and the significance of the Inevitability has increased, with a marginal improvement in the adjusted R^2 value.

The subsample of the gold-standard reviewers from the Corrective Action engineering group was considered next, to determine if their job experience led them to differing scoring methods from the rest of the population, as their normal job function included the evaluation of the suitability of procedures. The results are summarized in Table 3-4 similarly to Table 3-3 with the addition of the interaction of the Inevitability score and the scenario.

Table 3-4 - Control Value regressions performed on gold-standard reviewers

CV Regr.	ID	Scenario	Inev.	Int.	Inev. x Scen.	Int. x Scen.	R^2	R^2(adj)
4	0.000	0.033	0.193	0.706	-	-	67.55%	54.85%
5	0.000	0.012	0.088	0.742	0.066	0.018	78.19%	63.26%

Post	0.000	0.007	0.096	0.672	0.044	0.013	79.26%	65.06%

The differences in scoring across the respondents was significant in regression #4, as seen in the prior analysis of the whole sample, but their scoring of Control Value showed that the scenario was also a significant factor, while neither of the model terms was. Based on the significance of the scenario, the analysis was reperformed in regression #5 to include the scenario's interaction with both model terms.

This regression confirmed the significance of the respondent ID and the scenario in the model, as well as the Intervention score's interaction with the scenario seen in the analysis of the whole sample. Unlike the analyses of the whole though, this subpopulation had an inconclusive P-value for the significance of the inevitability alone, and interacting with the scenario as well, though both were low enough to suggest the possibility of a relationship, which might be masked by the small sample size.

An additional regression was performed with the post-survey Control Value scores, identified in the table as 'Post.' The results of this analysis were consistent with the results seen for the whole population—some shifts occurred in the level of significance, but the R^2 improved, indicating a marginally better correlation between the model and the scored Control Values adjusted by the reviewers after having considered the model elements.

Regressions were performed on the other available subpopulations included in the sample: the system engineers, technicians and inspectors. These groups included any of the respondents who identified that they had experience in the relevant position, so some personnel were counted in multiple categories. The summary of these regressions performed on other subpopulations is provided in Table 3-5.

Table 3-5 - Control Value regressions performed on remaining subpopulations

CV	ID	Scenario	Inev.	Int.	Inev.	Int.	R^2	R^2(adj)
Regr.					x Scen.	x Scen.		
Tech7	0.000	0.134	0.725	0.728	-	-	56.68%	42.85%
Tech8	0.000	0.836	0.778	0.903	0.764	0.424	59.77%	41.99%
Insp9	0.000	0.052	0.957	0.949	-	-	66.65%	54.81%
Insp10	0.000	0.540	0.794	0.421	0.214	0.464	73.52%	58.81%
Eng11	0.564	0.180	0.044	0.941	-	-	30.21%	1.97%
Eng12	0.554	0.128	0.095	0.673	-	0.097	42.97%	11.45%
Eng13	-	0.220	0.056	0.851	-	-	14.08%	4.36%
Eng14	-	0.073	0.094	0.571	-	0.062	28.20%	13.54%

Regression *Tech7* for the technicians showed no factors with significance except the ID. As neither Inevitability nor Intervention was close to being significant, there was no reason to expect that a relationship would be clearer by eliminating one or the other from the model. Instead, both factors were checked in regression *Tech8* for significant interactions with the scenario, but neither showed evidence of a relationship. Review of the associated correlation coefficient showed that the R^2 improved slightly but the adjusted R^2 dropped; suggesting the improvement in the capability of the model as a substitute was typical of the minor improvements associated with adding additional terms to a regression that are not necessarily related to the model.

Analysis of the regressions *Insp9* and *Insp10* for the inspectors was similar, except that a weak relationship with scenario did appear before considering the interactions. The disappearance of scenario as a significant factor upon the introduction of the interactions suggests that the interactions are not beneficial to the model, despite the improvement of both the R^2 and the adjusted R^2.

Results for the engineers however, provided a rather different situation. The feature of these results that set the engineers apart from the rest of the sample was that ID was not a significant factor in regression *Eng11.* While the correlation coefficient showed much lower ability of this model to serve as substitute, this is to be expected as the respondent-to-respondent differences had consistently been most significant in all other regressions. In the other groups, the relatively large correlation included respondent-unique effects, suggesting that the behavior of Control Value overwhelmingly represented the scoring biases of the non-engineer respondents, rather than the factors believed to be relevant.

Further adjustment of the of the model terms in regression *Eng12* by testing for the interaction of scenario with the intervention score yielded a reduction of significance on the Inevitability score, though still suggesting the existence of a relationship, along with weak evidence of an interaction effect. The noteworthy change stemming from this adjustment is that the correlation coefficients, while still low, increased dramatically.

The regression in *Eng13* eliminated the respondent ID from the model to determine if the suitability of the model would improve with the elimination of a factor that was not significant. The significance of the inevitability did increase as a result, almost to the 0.05 level, though there was a drop in the correlation coefficients.

Regression *Eng14* then incorporated the test for the interaction between scenario and the Intervention score in the absence of the respondent ID. While neither the scenario, the Inevitability nor the scenario/Intervention factors were significant at the 0.05 level, all three were significant at the 0.10 level, suggesting that a functional relationship could exist and that it might be more apparent if a larger sample was used than the 13 system engineers in the respondent set.

Overall, these results show that there are significant differences between the subpopulations. The lack of significance of the ID term for the non-Gold engineers though, suggests the existence of a natural schema understood by this group, while the results for the other groups indicate that individuals in the other subpopulations are not scoring consistently with their group.

In the absence of such a schema at the group level, the cause of the differing scores remains elusive, and is the result of differing individual scoring schemas. The GLM analysis accounts for the possibility of a constant offset between the average answers of respondents, but cannot effectively address nonlinearity in the responses, so the significance of the ID term indicates that the scoring differences are not the result of a simple offset in the scale used by the different respondents.

Phase A Regression on Failure Likelihood

Similar to the analysis of the I&I questions as a substitute for Control Value, a regression analysis was performed for Failure Likelihood using the DATOM questions as the independent variables. This analysis also used the GLM model presented in Figure 3-2, though the index k could go as high as 5 if all the DATOM elements were used. The goal of the regressions was to determine which model terms provided the relationshipwith the best goodness of fit, based on the magnitude of the adjusted R^2 value.

The initial regression using the whole sample of respondents from this phase revealed that ID was significant to better than the 0.0005 level. Similar to the results on the earlier regressions with the Control Value as the dependent variable, this suggests that the respondent-to-respondent variability is a major contributor to the variability in the Failure Likelihood scores. As was postulated when analyzing the Control Value data, this effect could be the result of self-consistent sub-populations

Factors that are found to be significant despite this scoring effect may suggest a natural relationship that could be reinforced by providing training or an objective scale for the respondents to use while scoring. The existence of a subpopulation where the ID is not a significant factor, such as the non-gold engineers in the CV assessment, would support this possibility.

The initial regression is marked as regression #1 on Table 3-6 and shows a significance for the Definition score, in addition to the ID term previously discussed. This was the case in the second regression, which removed Organization. The effect of the removal on the correlation coefficients was so small that regression #3 was performed to check for an interaction between the Organization and scenario, but nothing significant was found.

Table 3-6 - Failure Likelihood regressions performed on entire sample population

FL Regr	ID	Scen.	D	A	T	O	M	Scen. x O	R^2	R^2(adj)
1	0.000	0.130	0.032	0.527	0.567	0.762	0.206	-	50.67%	35.58%
2	0.000	0.117	0.021	0.446	0.551	-	0.175	-	50.65%	35.90%
3	0.000	0.924	0.031	0.663	0.588	0.961	0.194	0.855	51.04%	34.63%
4	0.000	0.091	0.014	0.339	-	-	0.214	-	50.55%	36.12%
5	0.000	0.109	0.000	-	-	-	0.098	-	50.31%	36.15%
6	0.000	-	0.006	0.902	0.363	0.577	0.348	-	48.73%	34.47%
Post	0.000	0.075	0.000	-	-	-	0.138	-	52.75%	39.28%

Further regressions progressively eliminated model terms that did not exhibit noteworthy effects: first Training in regression #4 and Assignment in #5, with marginal improvements in the adjusted R^2 value. The removal of Training did result in scenario being significant at the 0.10 level, suggesting the existence of an effect, but it dropped below that level with the removal of Assignment, as Monitoring became significant to the 0.10 level.

Regression #6 was performed to identify if the removal of scenario from the regression would improve the results by eliminating the source of multiple degrees of freedom in the model. There was no improvement over regression #1, which used the same model terms with the scenario included. To the contrary, along with the expected drop in the R^2 value, the adjusted R^2 also was also slightly reduced, showing that the contribution from the scenario term marginally outweighed the influence of the additional degrees of freedom. The absence of a large change with the removal of the scenario term supports the earlier conclusion that there is no interaction between scenario and the Organization term, and also that there is no interaction between scenario and any of the other model terms.

As with the Control Value regressions earlier, the posterior utility question was used in a regression to determine if the exposure to the model variables improved the relationship between the model variables and the perceived value of the control. In the regression marked as ‘Post,’ the significance values shifted slightly without a change at the 0.05 level, but the adjusted R^2 did increase by a multiple of the changes observed in removing the other model factors.

As in the earlier analysis of the Control Value, analysis of the Failure Likelihood will be broken out by the job functions of the respondents, examining the groups of gold-standard reviewers from corrective action engineering, the system engineers, technicians and inspectors separately.

Analysis of the responses from the gold-standard reviewers was performed using the same method removes the scenario as a factor after regression #7 because of its lack of significance. Other terms removed include Assignment and Organization after regression #8 and Training after regression #9. Removal of Training from the model between regression #9 and #10 resulted in a marginal drop in the adjusted R^2 value.

Table 3-7 - Failure Likelihood regressions performed on gold-standard respondents

FL Regr	ID	Scen.	D	A	T	O	M	R^2	R^2(adj)
7	0.000	0.504	0.244	0.490	0.062	0.552	0.014	68.02%	52.41%
8	0.000	-	0.054	0.321	0.064	0.913	0.007	65.50%	53.03%
9	0.000	-	0.081	-	0.101	-	0.007	64.72%	53.92%
10	0.000	-	0.008	-	-	-	0.001	62.70%	52.26%
Post	0.000	-	0.005	-	-	-	0.001	62.28%	51.72%

The change of failure likelihood scores from the initial responses to the posterior values recorded after participants scored the other model questions marginally reduced the model's suitability, as reflected in the adjusted R^2 value in the table for the run marked *Post*.

It is important to note that the ID term for the gold-standard respondents was significant to at least the 0.0005 level even before taking out any of the other terms, as it was for the whole population. With such overwhelming inter-rater differences influencing the dependent variable's behavior, the significance of any model terms indicates that there is a strongly shared appreciation of the impacts of that element of the model.

Additional regressions were performed on the other subpopulations of reviewers, by occupation, to determine if there was a significant difference between answers by the occupation of the respondents. The results in Table 3-8 show that ID is significant for technicians even before other terms are removed, while Definition also becomes significant and possibly the scenario as well. For inspectors, nothing is initially significant but ID and Definition are, once the extraneous terms are removed. Results for the system engineers are fundamentally different because ID is not significant, even to the 0.10 level, though close.

Table 3-8 - Failure Likelihood regressions performed on remaining subpopulations

FL Regr	ID	Scen.	D	A	T	O	M	R^2	R^2(adj)
Tech 11	0.037	0.114	0.106	0.859	0.242	0.564	0.415	50.91%	33.11%
Tech 12	0.034	0.075	0.005	-	0.181	-	0.356	50.67%	34.22%
Tech 13	0.021	0.064	0.000	-	-	-	-	48.92%	33.33%
Insp 14	0.103	0.390	0.214	0.654	0.847	0.707	0.413	49.86%	28.61%
Insp 15	0.089	0.334	0.007	-	-	-	0.386	49.42%	31.47%
Insp 16	0.032	0.364	0.004	-	-	-	-	48.80%	31.37%
Insp 17	0.027	-	0.000	-	-	-	-	45.22%	31.32%
Eng 18	0.491	0.255	0.028	0.166	0.925	0.766	0.846	47.19%	20.11%
Eng 19	0.071	0.206	0.011	0.105	-	-	-	47.05%	25.62%
Eng 20	0.101	-	0.015	0.424	-	-	-	39.25%	22.08%
Eng 21	0.107	-	0.012	-	-	-	-	38.39%	22.66%
Eng 22	-	-	0.003	-	-	-	-	14.11%	12.63%

The continued significance of Definition through all the groups indicates it is a model term for which all the subpopulations recognized the significance. The significance of Monitoring for the gold-standard respondents suggests that job responsibilities, or possibly the associated training, could be responsible for common scoring behavior.

The absence of significance for the other model terms could be the result of the sampling method. For example, none of the groups selected to participate in this study were responsible for training as a primary job function. It is possible that experts in planning technical training classes or in delivering training materials as instructors would demonstrate a shared scoring methodology for the Training term.

Regression Coefficients

On completion of the regression analyses performed in the previous two sections, the coefficients were tabulated to show the overall influence of the significant factors in Table 3-9 for a subset of the regression runs performed. The runs selected for the table represent the best fit of the models for that subpopulation, as indicated by the highest adjusted R^2 value. Table elements are highlighted in color to represent the level of significance of the coefficient, with green highlighting to represent significance at the 0.05 level and yellow for the 0.10 level. Significance for the array factors on the right hand side represent the significance of the factor as a whole, not the significance of any individual coefficients in the array.

Table 3-9 – Control Value coefficients by subpopulation

Sub Population	Regression	Coefficients			Standard Deviations			
		Constant	Inev.	Interv.	ID	Scenario	Inev* scenario	Inter* scenario
Whole Sample	CV 3	6.783	0.139	-0.026	1.337	1.134	-	0.197
	Post	6.681	0.179	-0.062	1.345	1.260	-	0.223
Gold Std Reviewers	CV 5	6.331	0.207	-0.053	1.725	4.941	0.352	0.497
	Post	6.574	0.191	-0.065	1.686	5.000	0.360	0.486
Technicians	Tech 7	6.977	0.034	0.033	1.355	0.378	-	-
Inspectors	Insp 10	7.394	0.025	0.079	1.207	1.529	0.190	0.174
System Engineers	Eng 12	6.775	0.240	-0.054	0.682	2.631	-	0.408
	Eng 14	7.320	0.160	-0.062	-	2.327	-	0.372

Coefficients are displayed in the table only for the constant and for the model terms, Inevitability and Intervention. Factors that are represented with an array of coefficients are shown on the table with a standard deviation of the values in the coefficient array. This includes the respondent identifier, the scenario and the interactions of model terms with the scenario. These standard deviations provide useful perspective into the variability of the factor in relation to the other components of the Control Value model. In the case of the ID in the CV3 regression as an example, the magnitude of the Inevitability coefficient remained significant even though it was only a fraction of the magnitude of the differences from one respondent to another.

The coefficients for the model terms Inevitability and Intervention in the table represent a factor, β in the simplified regression equation in Figure 3-5 for each respondent i and scenario j. This simplified equation omits the effects of the scenario and the respondent, which are included in the GLM analysis.

$$y_{i,j} = C + \beta_{inev} * x(inev)_{i,j} + \beta_{\text{int}} * x(\text{int})_{i,j} + \varepsilon_{i,j}$$

Figure 3-5 - Simplified Regression Equation

The complete GLM equation shown in Figure 3-6 displays the β for each model term as an array of values for each model term. In the case above for Control Value analyzed against the Inevitability and opportunity to intervene, the summation limit, m, would be 2 for the two model terms included.

$$y_{i,j} = C + [ID_i] + [Scenario_j] + \sum_{k=1}^{m} \{Q_k * x_{i,j,k}\} + \varepsilon_{i,j}$$
$$+ \sum_{i=1}^{n} \{[Scenario \times Q_k] * x_{i,j,k}\}$$

Figure 3-6 - GLM Regression Equation

The additional terms ID and Scenario in the regression equation contain the arrays of values representing the individual effects of the specific respondents or specific scenarios. These two arrays can be thought of as the portion of the error that is specific to the term being considered. The arrays measure the mean residual for each respondent or for each scenario. Each array must sum to 0 because any net effect would increase or decrease the constant value. For these terms, the summary table displays the standard deviation of the values in the array because it is the spread of those values that shows how similarly the respondents score or how closely the scoring for each scenario align.

The significance of the ID term in all regressions except the system engineers shows that the respondent to respondent differences make up a large portion of the differences in the Control Value scores, but that Inevitability shows as a significant factor with the larger population despite those differences. This suggests that the lack of significance of the Inevitability in all the smaller subpopulations may simply be the result of the small sample size.

The similar tabulation of coefficients for the Failure Likelihood is provided in Table 3-10, showing significant differences in scoring between respondents for all groups except the System Engineers. Despite those differences, Definition remained significant for all subpopulations and Monitoring was a significant factor for the Gold Standard respondents.

Table 3-10 – Failure Likelihood coefficients by subpopulation

Sub Population	**Regression**	**Coefficients**						**Standard Deviations**	
		Constant	**Define**	**Assign**	**Train**	**Organize**	**Monitor**	**ID**	**Scenario**
Whole Sample	FL5	5.132	0.237	-	-	-	0.086	0.874	0.270
	Post	5.056	0.255	-	-	-	0.072	0.852	0.275
Gold Std Reviewers	FL10	2.881	0.358	-	-	-	0.256	1.017	-
	Post	2.896	0.361	-	-	-	0.252	0.970	-
Technicians	Tech 12	3.973	0.267	-	0.133	-	0.068	0.783	0.410
Inspectors	Insp 17	4.530	0.364	-	-	-	-	0.809	-
Sys Eng	Eng 21	6.218	0.236	-	-	-	-	0.703	-
	Eng 22	6.049	0.257	-	-	-	-	-	-

Kendall's Tau for Comparing Relative Behavior

While the regression analysis showed some model terms were consistently significant between subpopulations, the differences between respondents unfortunately were almost always significant and often of a magnitude sufficient to dominate the behavior of the scoring. If the scoring differences are typically evidence of differences of opinion between respondents, then the model is not self-consistent and is not effective for capturing the expert judgment in a useful way. If, on the other hand, the differences between respondents are simply the result of the subjective applications of the scoring scale, improving the scoring system by providing objective criteria for each scoring level may improve the applicability of the model.

One characteristic difference between opinion differences and scoring interpretations that can be measured and tested in this circumstance is the degree to which the respondents rank two scenarios on a particular question. When differences in response to a question result from interpretation of the scoring scale, a respondent that scores one scenario higher than a second scenario will be consistent with a second respondent in which of the two scenarios receives the higher score, though the values of the scores may both be higher and may be clustered closer for one respondent than the other. Conversely, when the difference in a pair of scenarios results from differences of opinion, the scenario given the higher score by the second respondent will be independent of which scenario received a higher score from the first respondent.

For pairwise comparisons when there are differences of opinion, one respondent's scoring methodology is independent from another respondent's, so two respondents are as likely as not to rate the same scenario of a pair as the higher. The same pairwise independence exists in comparisons of scenarios with randomly assigned scores, with two pairs being just as likely to agree as disagree. Differences in scoring that are caused purely by subjective perceptions of the scoring scale will result in pairwise concordance. For example, if one respondent scores the Definition on scenario 1 higher than on scenario 2, the magnitude of a second respondent's scores may not be consistent, but the second respondent's Definition for scenario 1 will be scored higher than scenario 2. This agreement can be measured and tested using Kendall's Tau, which measures the degree of concordance between two ranked lists; in this case, the ranking of scenarios from lowest to highest by the scores on a given question. Tau values range from +1 for complete agreement between the lists to -1 for complete disagreement between the lists.

Testing for Opinion Differences with Kendall's Tau

To determine if the scoring differences were the result of scale interpretations, the hypothesis is that nonlinearities in the scoring prevent a relationship from being visible in the regression analyses, and that these are the result of the respondents' perception of the scoring scale, rather than arising from differences of opinion about the measure each question is intended to capture. If the scoring differences reflect fundamentally different perceptions of what is being measured in each question, then the scenario-to-scenario variation is no more likely to increase together for different respondents—a positive τ—than to move in opposite directions—a negative value. Under this situation, the independence of the scores would lead to τ values distributed identically to randomly generated τ values. The null hypothesis then is that τ values calculated for the respondents will be consistent with τ values for a random dataset: H_0 = *Reviewer's order ranking is independent from the question's measure.*

There are tables of values for testing using Kendall’s Tau (Schaeffer & Levitt, 1956) against a random distribution, but the published tables assume the basic case that there are no ties in the rankings and that the 5 scenarios to be compared on each question would have a discrete order. The scoring structure used in collecting the Phase A data however allows the possibility that a respondent would score multiple scenarios with the same value on a single question. By permitting repeat values, the possibility of ties in the ordered list now becomes possible. Conservative adjustments for the tables exist to account for varying levels of potential ties, but adjustments that are overly conservative may obscure results that would otherwise be clear. As a result, the most effective method for obtaining a distribution that was directly applicable to the situation was to develop one by simulation for the specific cases being tested.

Tests of Respondents Against the Group Using Kendall’s Tau

Two tests were performed to determine if the scoring by respondents was different from random. Both required the respondents’ scenario order to be compared against a baseline order. The order selected for each question used the mean of the scores provided for each scenario by the thirteen risk assessment experts who are the gold-standard respondents. In essence, a fourteenth reviewer was artificially generated, where this composite respondent’s answers were a simple average of the scores of the other thirteen reviewers.

The two comparisons started with a comparison of each respondent’s τ value for each of the eleven questions. The first test considered the proportion of the questions where the τ was high enough above the expected τ for random data to be statistically significant. The second test considered the mean τ for the seven model questions: Inevitability, Intervention and the five DATOM elements.

Risk-Assessment Experts vs. Baseline Comparison

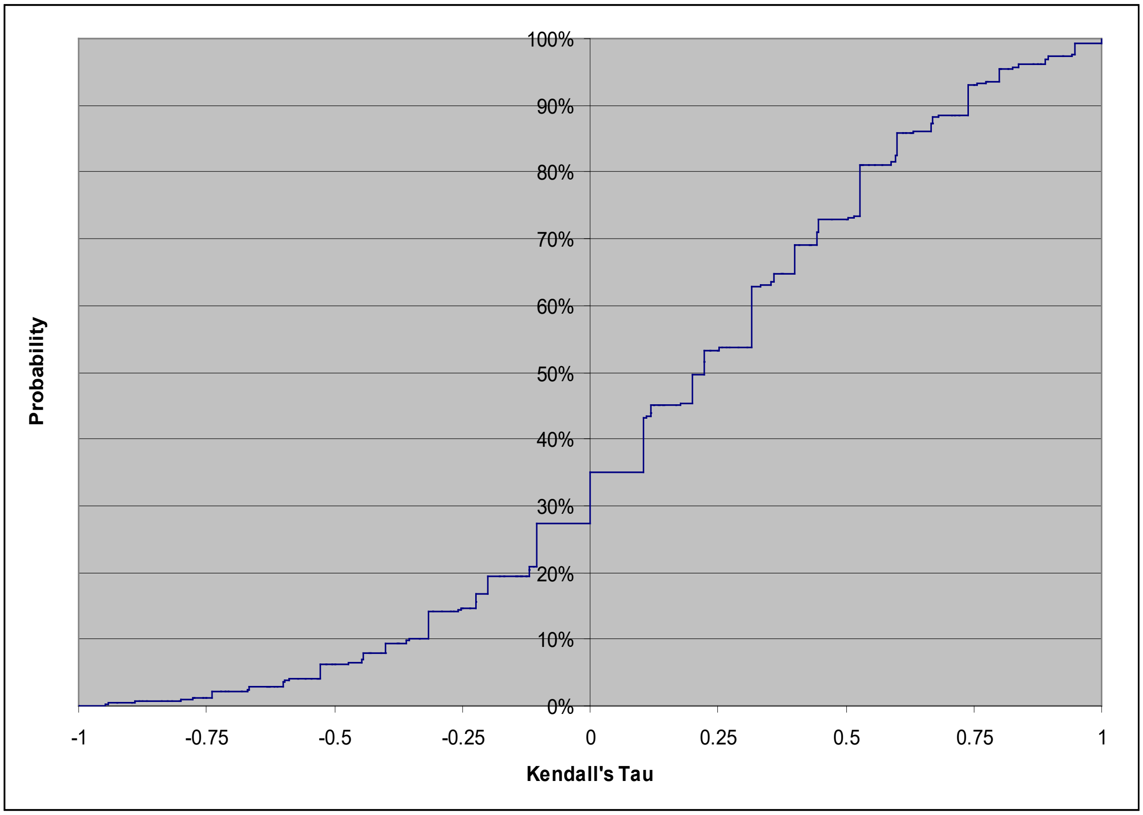
The τ values for the thirteen reviewers, A through M, are shown in Table 3-11 below for each of the eleven questions. The τ values show how closely one reviewer's rankings match the group baseline so a positive score indicates more agreement with the group than disagreement, and a negative score indicates more disagreement. Reviewers' ranking for a question that resulted in a τ of 0.800 or higher highlighted with a gray background. For example, the 0.837 for respondent A on question 4 showed a significant degree of concordance between respondent A's ranking of the five scenarios with how those scenarios were ranked in the composite.

Table 3-11 - τ values comparing risk assessment expert reviewers to baseline

	1	2	3	4	5	6	7	8	9	10	11
A	0.359	0.252	0.671	0.837	0.316	0.447	0.316	0.316	0.316	0.837	0.105
B	0.527	0.825	0.738	0.105	0.447	0.400	0.632	0.527	0.738	0.527	0.598
C	0.105	0.000	0.527	0.224	0.837	0.447	0.000	0.120	-0.316	0.527	0.105
D	0.359	0.825	0.800	0.359	-0.359	0.516	0.000	0.632	0.738	0.359	0.894
E	0.316	-0.816	0.224	0.447	0.000	0.738	0.600	0.949	0.949	0.000	-0.775
F	0.527	0.667	0.316	0.447	0.837	0.400	0.894	0.671	0.224	0.527	0.632
G	0.775	0.680	0.000	-0.120	-0.527	0.000	0.316	0.224	0.120	0.775	0.516
H	-0.120	0.444	-0.837	0.224	-0.224	0.800	0.359	0.316	0.949	0.000	0.447
I	-0.258	-0.667	0.775	-0.671	0.632	-0.316	0.000	0.837	0.527	-0.258	-0.632
J	0.671	0.272	0.000	0.632	0.598	0.598	0.671	0.738	-0.316	0.671	0.258
K	0.447	-0.126	0.120	0.632	0.359	0.316	0.000	0.598	0.359	0.447	-0.120
L	0.400	0.504	0.105	0.527	0.837	0.738	0.949	0.359	0.527	0.400	0.359
M	0.738	0.667	0.527	0.837	0.316	0.200	0.949	0.400	0.316	0.671	0.671

The τ value of 0.800 was determined to be significant because the cumulative distribution function for these τ values, shown in Figure 3-7, indicates the likelihood that random data will yield a τ of 0.800 or higher is only 5%. Therefore, for scores of 0.800, the null assumption, H_0 = *Reviewer's order ranking is independent from the question's measure*, can be rejected with 95% confidence.

Figure 3-7 - CDF of random τ values

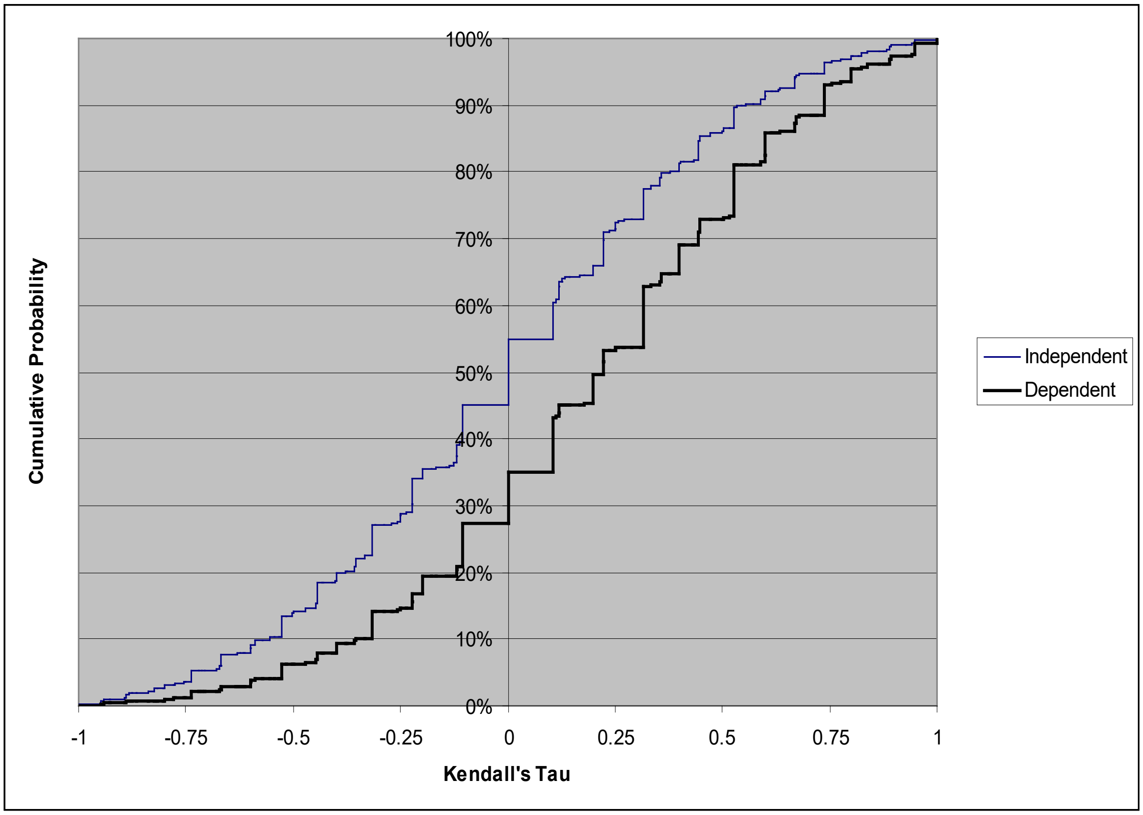


The cumulative distribution function of τ values used to draw this conclusion is based on simulation that compared the ranking of a set of 5 randomly chosen scores to the ranking of a composite of score set that uses the test set and 12 other randomly selected score sets, for a total of 13. This structure mimicked the Gold Standard respondents' scores, where τ compared one respondents ranking of the 5 scenarios to the composite score that was made up of his/her response and the responses of the 12 other Gold Standard respondents.

The simulation was repeated for more than 1.7 million runs, with the results shown in Figure 3-7. The discontinuous shape of the CDF is not the result of a small sample size but the small number of possible τ values, with only 80 possible non-zero τ values above and below the center. The shape of the CDF did not qualitatively change after the initial simulation of 13,000 runs, so the simulation was allowed to run until a sample size of at least 2 orders of magnitude larger was obtained.

The non-symmetrical shape of the plot is the result of the dependency that exists between the group's ranking and the individual reviewer's ranking. Since the composite is partially based on the random set to be tested, the two rankings are more likely to have agreement than if the compared sets were both random. Figure 3-8 shows the CDFs of τ values for dependent and for independent ranks to see how the dependence shifts the distribution. For the independent case, the probability of agreeing more than disagreeing is equal to the probability of disagreeing more than agreeing – 45.1%, which is consistent with the symmetrical appearance of the plot. For the dependent case, however, the probability of having more disagreement than agreement is only 27.4% vs. 65.0% having more agreement.

Figure 3-8 - Dependent vs. independent distributions of Kendall's τ



Comparing Reviewers by Average τ Value

Where the earlier evaluation of consistency using the τ values only considered one question at a time, a measure that considered multiple questions was needed. τ values for questions 3 through 9, the questions associated with the model attributes, would be used to develop this measure because those are the items that the proposed model considers significant. By averaging the τ values obtained for each reviewer on these 7 questions, the resulting measure provides an indicator with which reviewers could be compared. Table 3-12 shows the average τ for each of the 13 reviewers and the cumulative probability associated with that score from the CDF.

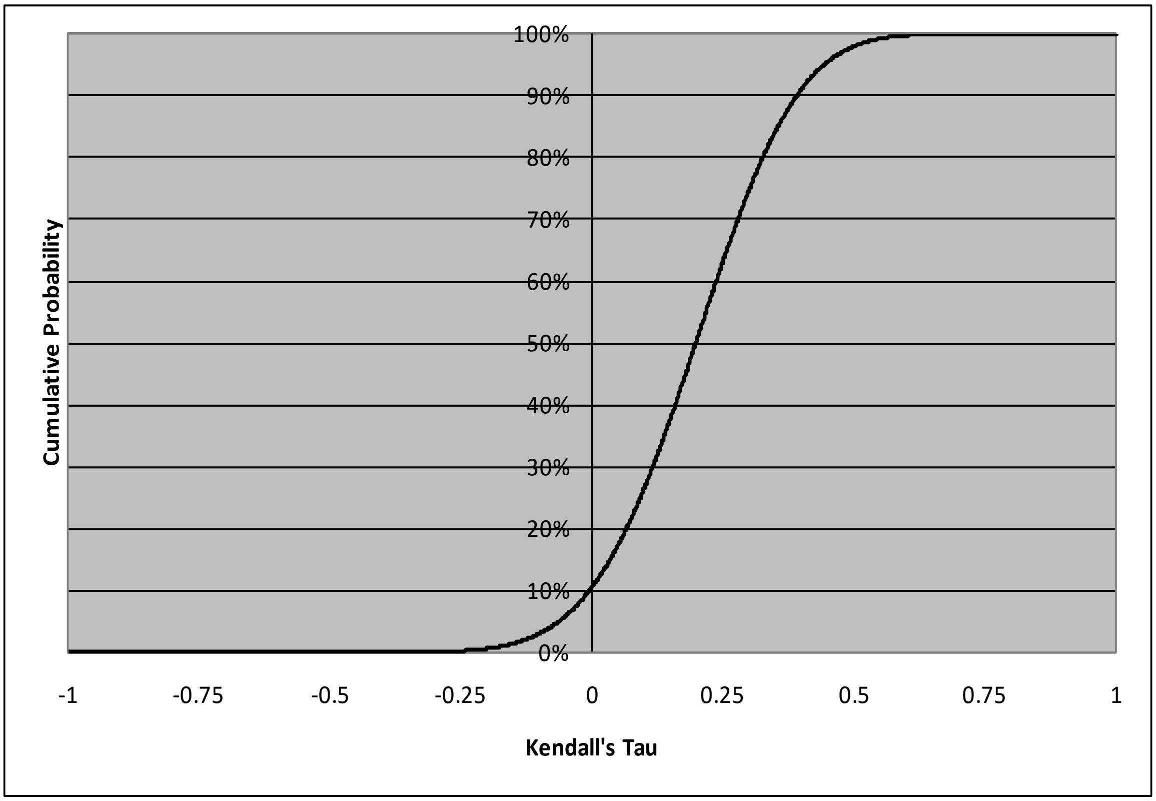
Table 3-12 - Average τ values for expert reviewers

Reviewer	$\tau_{(avg)}$ Q3-Q9	Cumulative Probability (from CDF)
G	0.002	10.8%
H	0.227	57.6%
I	0.255	64.5%
C	0.263	66.5%
K	0.341	82.7%
D	0.384	89.2%
J	0.417	92.9%
A	0.46	96.1%
M	0.506	98.2%
B	0.512	98.4%
F	0.541	99.1%
E	0.558	99.4%
L	0.577	99.6%

At first glance, the results seemed significant because the average τ value for each of the 13 reviewers is positive. For cases where the reviewers' responses were independent from the group's composite response, it would be extremely unlikely for all 13 reviewers to agree with the group more than to disagree. However, the dependent nature of the group response increases the likelihood that a random selection of scores by a reviewer would result in a positive τ value. Averaging together multiple τ values obtained from random selection would further increase the likelihood because the CDF of the averages would be compressed toward 0.222, the 50th percentile, on the CDF of an individual τ value.

Figure 3-9 shows the resulting CDF that is based on the average of 7 τ values that each compare a composite data set based on 13 randomly selected score sets to one of those randomly selected sets. The distribution shows that a positive average τ will result from random scoring with a probability of 89.3%.

Figure 3-9 - CDF for average of 7 dependent τ values



With this information, the probability of all 13 reviewers having a positive average τ value can be calculated from a set of Bernoulli trials with an 89.3% probability of success. Considering 13 attempts with 13 successes:

$$\sum_{k=13}^{13} \binom{13}{k} 89.3\%^{k} 10.7\%^{13-k} = 23.0\%$$

Figure 3-10 - Binomial Distribution for 13 positive average τ values

The probability that random data could yield results similar to the observed case is 23.0%, so it is not significant that average τ values were positive for all 13 respondents. However, considering just the sign of each reviewer's average τ value is insufficient to notice behavior in these results that is statistically significant when considering the magnitudes. The lowest τ value is only marginally positive with a 0.002 value, but the remaining 12 values are 0.227 or above, which will each happen in only 42.4% of random simulations. Using the binomial distribution again in Figure 3-11, the likelihood of having 12 or more τ values at 0.227 or higher at random is only 0.0267%, making it significant to the 0.0005 level

$$\sum_{k=12}^{13}\binom{13}{k}42.4\%^{k}57.6\%^{13-k}=0.0267\%$$

Figure 3-11 - Binomial Distribution for 12 or more τ values at 0.227 or higher

Similar analysis for each of the 11 questions was performed, considering the average of the 13 τ values for each question. The CDF curve for the average was qualitatively similar to the curve in Figure 3-9, but showed that the τ value below which 95% of 13 response averages would fall was 0.277.

Table 3-13 shows the average τ values for each question, averaging the 13 individual values in Table 3-11. Cells are shaded in green to denote the values above the 0.277 threshold where only 5% of values would fall there was no relationship between the respondents' perceptions of the scenarios on the scales measured by the question.

Table 3-13 - Average τ values for each question

1	2	3	4	5	6	7	8	9	10	11
0.373	0.271	0.305	0.345	0.313	0.406	0.437	0.514	0.395	0.422	0.235

It is important to note that Questions 1 and 10, as well as Questions 2 and 11 are related and are not independent. A calculation of the probability that 8 or more average τ values would be above 0.277 in 9 independent questions is shown in Figure 3-12

$$\sum_{k=8}^{9}\binom{9}{k}5\%^{k}95\%^{9-k}=3.60*10^{-10}$$

Figure 3-12 - Binomial Distribution for 8 or more τ values at 0.277 or higher

This apparently posterior analysis necessarily calls into question conclusions reached after reviewing the initial τ values. However, this is not directly an analysis of the τ values, but an analysis of the proportion of the τ values that would behave in this way if there was no relationship between the respondents' perceptions of the scenarios. In essence, the analyses of the averaged τ values addresses the null hypothesis, H_0 = *the proportion of statistically significant average τ values are the result of score selections that are independent of the measurement scales of the questions*.

While the individual τ values were known before the tests were to be performed, the averages were not, so the 0.05 level of significance remains an appropriate standard of proof. An argument could still be made that the independence of the analysis was tainted by the knowledge of the τ values and a stricter burden of proof would be needed for this *ex post facto* analysis. Both averages—the averages of reviewers across all model questions and questions across all reviewers—accomplished this by displaying significance well beyond the 0.05 level: 0.000267 for the reviewers and 3.60x10^{-10} for the questions.

Summary of Phase A

Regressions performed using the entire sample population to determine the suitability of the model for replacing the Control Value using a combination of the scores for Inevitability and the opportunity for Intervention did show that the Inevitability was significant, and the interaction between Intervention and the scenario was also significant. Likewise, when the posterior Control Value was considered, where the respondents updated their CV scores after considering the model elements, Inevitability and the Intervention/scenario interaction were both significant. Unfortunately, the respondent-to-respondent differences were also significant, limiting the use of the model as a substitute for CV when the entire sample population was used.

Similar results were found with the gold-standard reviewers, with the added significance of an interaction between Inevitability and the scenario. Results for the technician and inspector respondents showed significance only for the inter-respondent scoring differences. Results for the system engineer respondents were interesting in that the respondent-to-respondent differences were not significant. Unfortunately, the other model elements significant for the overall sample population were not significant at the 0.05 level for this group, but there were indications of similar behavior at the 0.10 level—not sufficient to draw conclusions, but worthy of attention in such a small sample.

The respondent and scenario dependence of the scoring makes this implementation of the model unsuitable for representing the value of a procedural control. In the case of the front-line technicians and inspectors, the scenario dependence is possibly because the abstract nature of the Inevitability and Intervention concepts were foreign to the mindsets cultivated for their roles, which focus on execution and compliance. The scenario dependence for the sample population as well as for both engineering groups suggests that the model may be incomplete. While the value of the procedural controls across the scenarios were expected to vary, there remained variations unexplained by the model. The unexplained variation represented by the significance of scenario indicates that the original concept of the model did not account for some situational factors associated with the specific scenarios. It is encouraging because the significance indicates that the respondents did share a scoring technique for each of the scenarios; unfortunately whatever factors those subpopulations were treating differently across the scenarios are not currently in the model.

Regressions on the Failure Likelihood, using the DATOM elements as the independent variables, yielded mixed results that were similar to the Control Value regressions performed above: the respondent-to-respondent differences were typically significant, and some scenario-to-scenario effects appeared to be influential as well. For the whole sample population and for all the subpopulations considered, the Definition question showed a significant linear relationship to FL. For the regressions performed on the gold-standard reviewer subpopulation, the Monitor question also showed a significant linear relationship to FL. As was seen in the results of the Control Value regressions, the responses for the system engineering subpopulation were distinctive from the other groups because the respondent-to-respondent differences were not significant. While the R^2 values for this relationship were not very high, this was the only regression relationship that did not require prior understanding of the respondents' individual scoring behavior.

The relative rankings of scenario-to-scenario scores were considered for the gold-standard respondents using the Kendall's τ as a measure of concordance between the respondents. This allowed the existence of relationships to be seen despite a lack of linearity that would have interfered with the regression analysis. The resulting analysis showed that the scoring behavior of the gold-standard respondents was significantly consistent with each other across the 7 model questions (I&I and DATOM) as well as the perceived Control Value and Failure Likelihood. While this cannot be directly applied to the other subpopulations, it does indicate that the overall scoring behavior for at least one subpopulation is consistent beyond just what was found to be significant in the regression analysis.

The consistency visible in the τ analysis for the gold-standard reviewers indicates that there is an underlying agreement in their scoring technique beyond what is shown in the regressions. Though the regression identifies that the ID effect, showing the respondent-to-respondent differences, is significant, the τ shows that the reviewers overwhelmingly agree on which scenarios score highest or lowest on a given question. This qualitative agreement cannot be used directly as a substitute for CV or FL, but it shows that there is a shared perception of the underlying fundamentals of the elements of the model, suggesting that the scoring might become more consistent if the respondents were given training in scoring the scenarios.

Phase B Data Analysis

The second phase of the validation analyzed the responses provided to a set of six scenarios identified during the *Columbia* accident investigation. The investigation considered all the documentation generated during the ground processing for *Columbia's* final two flights and identified failure rates for the controls in place during that timeframe. The DATOM scores provided by each respondent were compared to the observed failure rates reported during the investigation using a GLM regression similar to the method used in Phase A.

One significant difference between Phase A and Phase B, however was the fixed Y values in the regression. Since each scenario had a fixed response variable, including the scenario in the model resulted in a trivial solution. The equation in Figure 3-13 demonstrates the problem with the GLM regression where all values are static.

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{bmatrix} = C + [ID_i] + \begin{bmatrix} Scenario_1 \\ Scenario_2 \\ Scenario_3 \\ Scenario_4 \\ Scenario_5 \\ Scenario_6 \end{bmatrix} + \sum_{k=1}^{m} \{Q_k * x_{i,j,k}\} + \varepsilon_{i,j}$$

Figure 3-13 – Phase B GLM Regression Equation with Scenario

The solution to this case results in the coefficient on the Scenario variable set equal to the y-value for each scenario, with the remaining coefficients— the DATOM coefficients, Q_k; the constant, C; and the ID array—all becoming zero.

To perform the regression in a way that would provide meaningful results, the analysis would have to omit the scenario from the regression model. This change resulted in a model that no longer explicitly included the scenario. However, the scenario-to-scenario differences are still captured implicitly since the value differences in the response variable are entirely due to the effects of the scenarios. Additionally, the results from Phase A showed that Scenario coefficients were not significant across the scenarios included in that phase for most of the subpopulations. The effect of scenario on the regressions for the technicians showed significance at the 0.10 level, which may only have been the result of the small sample size.

The resulting regression model can be seen in the equation in Figure 3-14, which omits the array of coefficients in Scenario term:

$$y_{i,j} = C + [ID_i] + \sum_{k=1}^{m} \{Q_k * x_{i,j,k}\} + \varepsilon_{i,j}$$

Figure 3-14 – Phase B GLM Regression Equation without Scenario

As identified in Phase A, fitted value along with an associated residual error, is calculated for the dependent variable *y*. Regressions in this phase did not evaluate the Control Value and used the Columbia investigation rates as the dependent variable for the comparison against the DATOM model elements as independent variables. The equation for calculating the fitted value as a function of the respondents, the scenarios and the model questions, uses the subscripts *i*, *j* & *k* as indices to represent the respondents, scenarios and included model questions respectively.

For each respondent, *i*, the dependent FL score associated with the 5 scenarios (*j*=1 to 5) is compared against the fitted value that includes a constant held fixed across all respondents and scenarios. The next term in the fitted value, $[ID_i]$ is vector containing a calculated offset value for each respondent, which is constant across questions and scenarios. In this vector, the values sum to 0, as the net effect across all *i* reviewers is accounted for in the constant term. Q_k is the variable containing the coefficients for each question included in the regression model, which is multiplied by the value of each score. The index *k* can range from 1, for a single question being included in the model to 5 when all the DATOM components are included in the regression.

Similar to the removal of the $[Scenario_j]$ vector explained earlier, the $[Scenario_j \times Q_k]$ term—included in Phase A to capture the impact of scenario in the regression—is omitted because of the triviality of the solutions when scenario is explicitly contained in the model.

Values and coefficients determined in the regression analysis to be non-zero with more than 95% confidence are reported as significant, and items with better than 90% confidence are reported as well. While the lower confidence is by no means conclusive because there is an increased chance that it could be the result of random variation, it suggests areas where a relationship may exist that is masked by the relatively small sample size.

Regression on Observed Failure Rate

The observed failure rates presented a problem with calculations because the values of the response variable spanned a large range from approximately 450 failures per million opportunities at the low end to just over 11,500 failures per million opportunities at the high end. The range of values on the model variables was only 100, several orders of magnitude smaller, so the behavior of the scenarios with the highest failure rate tended to dominate the results and the effects of nonlinearities in the scoring were magnified. As a result residual errors were not normally distributed, showing that the regression did not effectively represent the relationship between the model variables.

To mitigate the effects of the disparity between the scales of the response variable and the model variables, the analysis was reperformed using the logarithm of the failure rate as the independent variable, rather than directly using the rate itself.

The first set of regressions were performed using the entire sample of Space Shuttle personnel, which included many of the gold-standard respondents in Phase A from the Corrective Action Engineering organization, along with technicians, inspectors and system engineers. In regression #1, the ID of the respondent was included in the regression equation, as described above, but the associated P-value was calculated as 1.000, indicating it was not significant. This indicated that there were not significant scoring differences between the average values of responses from the different reviewers, once other factors were accounted for.

In the same regression, the model components for the Organization and Monitoring were significant to better than the 0.05 level. The results of this regression are presented in Table 3-14, with the factors that are significant to the 0.05 level or better highlighted in green and the factor significant to the 0.10 level highlighted in yellow.

Table 3-14 – *Columbia* Failure Rate regressions on entire sample population

FR Regr	ID	D	A	T	O	M	R^2	R^2(adj)
1	1.000	0.919	0.147	0.480	0.024	0.000	18.75%	0.77%
2	-	0.934	0.135	0.614	0.026	0.000	13.76%	12.23%
3	-	-	0.065	-	0.002	0.000	13.68%	12.77%

Table 3-14 also shows the remaining regressions performed on the full sample population. Regression #2 was performed without the ID of the respondents included, though the other model terms remained. There was a drop in the R^2 value, which would be expected when removing a single factor that accounts for a considerable fraction of the degrees of freedom in the model. The associated jump in the adjusted R^2, though still low, suggests that the contribution of ID to the fit of the model was primarily due to the relatively large number of degrees of freedom.

Regression #3 was performed without the terms for Definition and Training, and resulted in a small drop in the R^2 value and small but somewhat larger improvement in the adjusted R^2 that further reduced the disparity between eh two values. This shift indicated that those model terms were unnecessary in explaining the behavior of the dependent variable.

As a result of the widely different behavior of the different subpopulations in Phase A, the regressions were performed on the various subpopulations in this phase as well, to determine if there were significant differences between the subpopulations when comparing the DATOM scores to the observed failure rates.

Regression #4 shown in Table 3-15 shows that none of the factors were significant to even the 0.10 level and the adjust R^2 value was 0.0%. Following the example set in the analysis of Regression #1, the ID was removed from the model, as there was nothing in the results to suggest it was significant.

Table 3-15 – *Columbia* Failure Rate regressions on gold-standard respondents

FR Regr	ID	D	A	T	O	M	R^2	R^2(adj)
4	0.855	0.656	0.288	0.907	0.313	0.235	18.85%	0.00%
5	-	0.592	0.369	0.999	0.278	0.396	11.34%	0.79%
6	-	-	0.383	-	0.318	0.236	10.72%	4.64%

Regression #5 showed the expected drop in R^2 associated with the removal of a variable with a large number of degrees of freedom, but there was no significant jump in the adjusted R^2, suggesting that the remaining terms did not materially contribute to the model.

The follow-up in Regression #6 removed the Definition and Training factors from the model because the associated P values were higher than for the other terms, and those factors were not significant in the whole subpopulation. The results of this regression showed an improvement in the adjusted R^2, but did not result in any model terms becoming significant at even the 0.10 level.

Additional regressions were performed on the remaining subpopulations in the sample, and even though the sample sizes were small, some factors were determined to be significant. For all three groups, the system engineers, the technicians and the inspectors, the ID term was not significant and was removed from the model after the first regression. This is seen in Regressions #7, #11 and #15 in Table 3-16.

Table 3-16 – *Columbia* Failure Rate regressions on remaining subpopulations

FR Regr	ID	D	A	T	O	M	R^2	R^2(adj)
Eng7	0.996	0.207	0.948	0.714	0.564	0.000	25.51%	5.95%
Eng8	-	0.151	0.794	0.835	0.379	0.000	21.09%	16.98%
Eng9	-	0.142	-	-	0.249	0.000	21.02%	18.60%
Eng10	-	0.008	-	-	-	0.000	19.93%	18.32%
Tech11	1.000	0.836	0.490	0.819	0.084	0.008	12.10%	0.00%
Tech12	-	0.863	0.656	0.815	0.065	0.006	9.64%	5.88%
Tech13	-	-	-	-	0.001	0.005	9.47%	8.00%
Insp15	1.000	0.574	0.342	0.767	0.450	0.069	15.39%	0.00%
Insp16	-	0.494	0.349	0.543	0.247	0.022	13.61%	8.47%
Insp17	-	-	-	-	0.001	0.005	11.76%	9.73%

Subsequent regressions on these groups showed that Monitoring was significant in all three subpopulations at better than the 0.05 level. The results also showed Definition as significant for the engineers and Organization as significant for both the technicians and inspectors while the remaining factors were not significant. The R^2 values for the engineers were well above the values for the whole sample, and were lower for both the technicians and inspectors.

Regression Coefficients

The coefficients for the regressions performed on the failure rate are presented in Table 3-17, using the notation as above, with the green highlighting to denote significance at the 0.05 level and yellow for the 0.10 level.

Table 3-17 – Failure Rate coefficients by subpopulation

Sub Population	Regr.	Constant	Definition	Assignment	Training	Organization	Monitoring
Whole Sample	3	3.5513	-	-0.0023	-	-0.0048	0.0056
Gold Standard	6	3.7955	-	-0.0032	-	-0.0048	0.0030
System Engineers	10	3.2728	-0.0048	-	-	-	0.0078
Technicians	13	3.5707	-	-	-	-0.0067	0.0050
Inspectors	17	3.5941	-	-	-	-0.0077	0.0058

The primary detail to consider in these results is the relative magnitudes of the model effects—the DATOM elements—compared to the constant term. For all respondent groupings analyzed in these regressions, the magnitude of the constant is overwhelmingly larger than the coefficients of the model terms, indicating that the constant dominates the behavior of the regression.

The secondary point of note is that the Monitoring coefficient is positive, while the remaining model coefficients are negative. The negative coefficients are consistent with a valid model, as the higher DATOM scores were hypothesized in the model to equate to lower failure rates for a control. The positive coefficient though is unexpected for Monitoring since the failure rate being considered in this phase is the failure of the control. The result shows that the control, or rule, fails to effectively create the desired behavior more often as monitoring increases.

Summary of Phase B

As in Phase A, the regressions in Phase B revealed a number of features of interest in the behavior of the scores recorded by the different groupings of respondents that can be divided into three groups: insights into the model itself, insights into the roles of the respondents and insights into the environment in which the respondents were working at the time the data was collected.

Work Environment

Where the theory behind the model predicted that lower control failure rates—higher compliance rates—would coincide with higher Monitoring values under normal circumstances, the opposite was found with significance even in groups having exceedingly small sample sizes. This reflects a reversal of the causality assumed in the original model that resulted from the corrective actions implemented in response to the *Columbia* accident.

The model assumed that poor monitoring would allow rule noncompliance to occur, but the evidence to the contrary does not suggest that increasing monitoring would lead to more violations. Instead, it demonstrates that the monitoring and oversight implemented in the years after the accident was effectively focused on the high failure rate controls identified during the investigation. While the respondents were answering based on their perception of the monitoring, rather than the actual effectiveness over time, it is striking that their perceptions were so significantly related to the initial failure rates that drove improvements in the monitoring.

It is also important to note that the respondents were unlikely to have been exposed to the raw failure rates at the time of the accident investigation so consequently that knowledge could not have influenced their scoring. The one group that was most likely to have had access to the raw failure rates at the time of the investigations was the gold-standard reviewers from the Corrective Action Engineering group—some of them were responsible for implementing process improvements in response to a prioritized list based on the investigation data. On the other hand, they were also the one group that had not shown significance in the positive coefficient on the monitoring term in the regressions.

Workforce Insight

Monitoring's lack of significance in the regressions on the gold-standard Corrective Action (CA) Engineering group may be the result of the small sample size. However, Monitoring was significant in the other subpopulations with similarly sized samples, so it may be an aspect of their job responsibilities that obscures what the other groups found to be a clear relationship.

One fundamental difference between the CA engineers and the rest of the subpopulations is that the other groups directly saw the ongoing monitoring efforts where the CA engineers where only temporarily involved through improvement of the monitoring processes. The other sub-populations were reminded of the ongoing monitoring on a daily basis, while the CA engineers' transient participation in the monitoring makes it likely that their perception is based not on the impact of the change, but on the effort to implement it. Increasing the frequency of random quality inspections that focused on a high failure control would have a large and lasting impact on the monitoring for that failure. In such a case, one CA engineer would have been involved in changing a parameter in a database, but it would be visible and persistent for the other groups.

The significance of the Definition component for the System Engineers and the Organization component for the Technicians and Inspectors also has possible roots in the job responsibilities of those functions. The system engineering role was mostly focused on planning the tasks necessary to restore functionality of a shuttle system or to optimize the performance of the system for a mission—a task that is more closely aligned with

Definition than the other model components. The technicians were more focused on the execution and the inspectors' responsibility was to verify and oversee the execution, so their focus matches the technicians', with the execution of the task being the primary target of their attention. With their attention focused on the execution, the Organization of the task would be their highest priority: ensuring that the necessary tools were available and that they could readily verify successful completion.

The lack of significance in items outside the respondents' area of expertise may be the result of less rigorous scoring of those other model components, or a failure to recognize subtleties in the component's behavior that come as second nature to respondents with more experience focusing in that area.

The absence of significance for Training may also be the result of this phenomenon. None of the respondents were exposed to a daily focus on training, so none may have been sufficiently prepared to recognize the key influencers in the scenarios that would have resulted in significant scoring if a subpopulation with that experience had been included.

Model Insights

Assignment was also not a significant model element in the regressions performed on any of the subpopulations, but most likely for a different reason. It did show significance at the 0.10 level which is insufficient to draw any conclusion, but a weak association would not be surprising because the roles and structure in which the respondents all work had matured through decades of shuttle work and through the Apollo program before. Some respondents may have assumed the Assignment was inherent in the existing structure while others may have scored based only on what was explicitly in the scenarios.

If this is the case, then alternate scenarios without the maturity could show more impact of the Assignment model factor in the results. A follow-up study using scenarios and personnel from SpaceX could be an effective way to accomplish this goal. SpaceX shares a goal with the shuttle team—safely and effectively launching—while sharing the same physical environment and even some of the same personnel. The difference though is that they have the added goal of being smaller, leaner and faster than the prior culture at the launch sites, effectively breaking down the inherent Assignments that existed in the shuttle culture.

Phase C Data Analysis

The final phase of data collection returned to the models considered in Phase A: using the Inevitability and Intervention scores to substitute for the value of a procedural control, and using the DATOM elements—Definition, Assignment, Training, Organization and Monitoring—as a replacement for the likelihood that the control would fail. The same scenarios were used for Phase C as for Phase A, but the question set had been expanded. Where only a single question had been asked regarding each of the model elements, now, there were multiple components to each score that were being asked independently. Table 3-2, repeated below, showed the expanded set of questions, and displayed that the questions on the aspects of the DATOM elements were also expanded into two parts: the respondents' perception of how important that aspect was in determining the likelihood the control would fail, and their perception of the extent the control demonstrated that aspect.

Table 3-2 - Phase C Questions

	Model Element	Question
Utility	Control Value	In your opinion, how valuable is it to have a control that does what Rule #N is trying to accomplish?
	Failure Likelihood	In your opinion, how likely is it that Rule #N will operate and perform its intended function?
I&I	Inevitability	How likely would it be for a critical incident to develop due to the absence of Rule #N? (Consider redundant rules but do not include contingencies outside the procedure)
	Intervention	How detectable would a failure of Rule #N be to the personnel following the rule?
		How aware would personnel be of the way to prevent a critical incident in the event of a failure of Rule #N?
		How much time would personnel have to act to prevent a critical incident after a failure of Rule #N?
DATOM	Defining	To what extent is Rule #N written in clear language? (e.g. commonly used words, consistent acronyms, no conditional statements, no double negatives, etc…)
		How important is it for Rule #N to be written in clear language? (e.g. commonly used words, consistent acronyms, no conditional statements, no double negatives, etc…)
		To what extent is Rule #N unambiguous and easy to understand?
		How important is it for Rule #N to be unambiguous and easy to understand?
		To what extent are personnel aware of the intent of Rule #N?
		How important is it for personnel to be aware of the intent of Rule #N?
	Assigning	How well does Rule #N identify who is expected to follow the rule?
		How important is it to identify who is expected to follow Rule #N?
	Training	How extensive is the training that personnel receive in the process governed by Rule #N?
		How important is it for personnel to receive extensive training in the process governed by Rule #N?
	Organizing	To what extent are the tools provided to personnel appropriate to the task governed by Rule #N? (this includes software, as well as physical tools)
		How important is it for tools provided to personnel to be appropriate to the task governed by Rule #N? (this includes software, as well as physical tools)
		How memorable are the details of Rule #N under the conditions that personnel would be exposed to? (Consider complexity and reminders – mnemonics, checklists)
		How important is it for the details of Rule #N to be memorable under conditions that personnel would be exposed to? (Consider complexity and reminders – mnemonics, checklists)
		How strongly would time pressure hamper personnel while trying to comply with Rule #N completely?
		How important is it for time pressure not to hamper personnel while trying to completely comply with Rule #N?
		To what extent does Rule #N conflict with other goals personnel are expected to support? (e.g. safety, quality, schedule, etc…)?
		How important is it for Rule #N to not conflict with other goals personnel are expected to support? (e.g. safety, quality, schedule, etc…)?
		To what extent may personnel improvise in response to changing conditions to comply with Rule #N?
		How important is it for personnel to improvise in response to changing conditions to comply with Rule #N?
	Monitoring	To what extent does Rule #N have an alternative means to verify compliance? (Consider difficulty of observing and if evidence persists)
		How important is it for Rule #N to have an alternative means to verify compliance? (Consider difficulty of observing and if evidence persists)
		To what extent does Rule #N have an alternative means to verify success? (Consider difficulty in discerning results of the rule)
		How important is it for Rule #N to have an alternative means to verify success? (Consider difficulty in discerning results of the rule)

As an example, Definition was split into three separate aspects: the clarity of the language in the rule, the ease of understanding and the transparency of the intent of the rule. The six Definition questions were asked as a method of capturing for each of these three aspects the degree to which it is demonstrated in the scenario and the importance that the aspect is displayed in the scenario.

Unfortunately, there was only limited access allowed to the workforce during the collection of data for Phase C, so only the gold standard reviewers from the Corrective Action Engineering department were available to answer the questions. This collection exercise occurred before determining that there were significant differences in the scoring from the various subpopulations of respondents, so the conclusions from Phase C are not generalizable to the entire Space Shuttle team workforce.

Assessing Collinearity in the Responses

Due to the similarity of concepts in the expanded questions, the first step in the analysis was to determine the extent to which the responses for each set of questions were correlated to the others. The two main concerns were that the respondents would not effectively understand the potentially subtle differences between the aspects of each model element, and that they would not differentiate between the extent and the importance measures being collected. Continuing with the Definition questions as an example, respondents may not appreciate that the clarity of the language in the procedure is independent of the ease of understanding. They also may not recognize the difference between how important it is to understand the intent of a rule in this situation and how understandable this rule actually is.

To determine if the respondents, as a group, were able to differentiate on these items, a correlation analysis was performed using Pearson's Correlation Coefficient.

While Kendall's Tau was used in the analysis of the pairwise comparisons performed in Phase A because it was more effective in dealing with the potential ties, Pearson's method could effectively handle the overall correlation across multiple respondents and multiple scenarios.

The correlation between the aspects of Intervention and of the DATOM elements with multiple aspects varied as identified in the tables below. Table 3-18 shows the correlations for the three Intervention questions: how detectable a failure of the control would be, how aware personnel were of the appropriate action and the time available to respond to a failure. The top value in each cell is the Pearson correlation coefficient and the lower value is the associated P-value, with the probability values shaded using the convention of green for 0.05 level of significance and yellow for 0.10 level of significance.

Table 3-18 - Correlation of Intervention aspects

	Int 1	Int 2
Int 2	0.533 0.000	
Int 3	0.272 0.056	0.252 0.077

The correlation between the first two questions shows a significant positive relationship between the questions for detectability and the awareness of the appropriate action. The correlations between the third question—time to respond—and the other questions is weaker, but there are loose indications that relationships may exist, especially after considering that the respondent set was so small.

The correlation between Int 1 and Int 2 are not surprising; both consider the level of a worker's knowledge of the system. The reduced likelihood of significance for Int 3 is similarly unsurprising because the time to respond considers something inherent to the system, the time for a response, which is relatively independent of the worker.

The separate measurements of degree and importance on each aspect of the DATOM elements makes the analysis more complicated for those items. While Assignment and Training only contained one aspect each, Definition considered three, Organization considered five and Monitor considered two. For the elements where more than one aspect was measured, the correlations between each degree measurement and the corresponding importance measure are crucial for understanding the element as a whole. Similarly, the correlations within all the degree measures and within all importance measures provide insight into how the respondents perceived those aspects.

Table 3-19 displays the correlations for all the pairs of Definition variables, with the measures of degree boxed with a darker line, as are the measures of importance. The correlations between degree and importance for each aspect are identified in double-line boxes. The probability measures on the second line of each cell are noted with green and yellow highlighting for significance to the 0.05 level and 0.10 level respectively.

Table 3-19 - Definition Correlations

	D1d	D2d	D3d	D1i	D2i
D2d	0.837 0.000				
D3d	0.458 0.001	0.422 0.002			
D1i	0.468 0.001	0.537 0.000	0.328 0.020		
D2i	0.544 0.000	0.548 0.000	0.388 0.005	0.650 0.000	
D3i	0.122 0.398	0.126 0.383	0.542 0.000	0.245 0.086	0.250 0.080

The correlations on the degree measures show that all three aspects were positively correlated with a high significance. The strong degree correlation for D1 and D2—the clarity of the language and the intelligibility of the rule—shows that the respondents did not score those aspects independently. There is insufficient information to determine if the smaller but still significant correlation of both to D3—the awareness of the rule—is the result of nonlinearities in the relationship, or differences in the aspect itself. Where the first two call on the respondent to consider the rule itself, D3 elicits a response about the perceptions of the workforce in general.

The correlations within the importance measures for all three are similar, with the positive correlation of D3 to both D1 and D2 not being significant to the 0.05 level. Again, this behavior may be the result of nonlinearities or a distinct thought process guiding the D3 importance measure because it considers an opinion-based judgment that can be made directly by the respondent, rather than the second-order judgment of what someone else is aware of or could be expected to be aware of.

Assessing the correlations on Organization was more difficult because of the 45 possible interactions between the five measured aspects. Table 3-20 displays the 45 correlations with the degree measures boxed in the upper left, the importance measures in the lower right and the degree-importance correlations for each of the aspects boxed in the lower left.

Table 3-20 - Organization Correlations

	O1d	O2d	O3d	O4d	O5d	O1i	O2i	O3i	O4i
O2d	0.158 0.273								
O3d	0.107 0.460	0.248 0.082							
O4d	-0.032 0.823	0.074 0.611	0.27 0.058						
O5d	-0.508 0.000	-0.308 0.029	-0.124 0.391	0.051 0.723					
O1i	0.620 0.000	0.108 0.455	0.18 0.211	0.087 0.55	-0.361 0.010				
O2i	0.363 0.009	0.503 0.000	0.256 0.072	-0.08 0.581	-0.590 0.000	0.215 0.133			
O3i	0.498 0.000	0.102 0.483	-0.133 0.357	-0.317 0.025	-0.377 0.007	0.451 0.001	0.376 0.007		
O4i	0.425 0.002	-0.003 0.983	0.191 0.185	0.037 0.797	-0.439 0.001	0.193 0.178	0.522 0.000	0.354 0.012	
O5i	-0.171 0.236	-0.139 0.334	-0.227 0.113	-0.224 0.119	0.339 0.016	-0.169 0.242	-0.096 0.509	-0.200 0.164	-0.073 0.616

Where the Definition aspects were mostly correlated with some exceptions, the Organization aspects show fewer significant correlations. Among the degree vs. degree correlations, only two were significant at the 0.05 level: O5, the ability to improvise, compared to O1, the degree that suitable tools were provided and O2, the degree that task details were memorable. Both correlations were negative, indicating that the respondents correlated both better tools and more memorable rules with reduced presence of improvisation in a scenario. From the available information, it is not possible to determine if the tools and better rules were developed in scenarios where there was a need to standardize the work and avoid improvisation. The opposite may be true, where the improved tools and rules led to better processes and a reduction in improvised work.

Among the comparisons on the importance of the aspects, the importance of O3 was significantly correlated with the importance of O1, O2 and O4. In addition to the correlation with O3, O2 also correlated significantly with O4. Correlation of O3, the importance of time pressure, with the tools, details and goal conflicts in O1, O2 and O4 respectively is not unexpected. These results show that the importance of recognizing time pressure in these scenarios is positively correlated to increased importance of tools, detail memorability and the importance of recognizing goal conflicts—all methods for coping with pressure to perform. The correlation between the importance of memorable details and goal conflicts provides a similar coping mechanism; memorable control details in a scenario will help prioritize actions when goals cause a conflict that may otherwise make the preferred behavior ambiguous.

The significant positive correlation between degree and importance on three of the five aspects suggests that the process for building procedures effectively prioritizes providing tools and memorable details where needed—at least as perceived by these respondents. The lack of correlation on the degree and importance of the time pressure and the goal conflicts suggests that the procedures provided to the workforce neglected these two aspects.

Correlation of the degree of O5—the extent of improvisation in the scenarios—to the importance measures of all five aspects is notable. A positive correlation between degree and importance for O5 is significant, indicating that improvisation is perceived to be allowed where these respondents believe it is appropriate, an ideal situation when trying to create efficient rules for a workforce. The significant negative correlations for all the other importance measures is similar to the negative correlations already addressed above between the degree of O5 and some of the other degree measures; the degree of improvisation increases where the importance of the tools, memorable details, time pressure and conflicting goals increases. This is likely due to a combination of reduced need for improvisation and incentives against it in some scenarios. Improvisation is less necessary where better tools and procedures are needed and provided, and workers' desire to improvise drops when time pressure and goal conflicts drive up the consequences of bad decisions.

Similar significant positive correlations were seen between the degree of O1, the tools provided, and the memorable procedure details, time pressure and goal conflicts. The positive correlation in this case shows that the respondents perceive the tools as being better where the other aspects are important. This too could be a combination both of tool maturity—better tools are provided in response to high pressure scenarios—and tool deficiencies—some scenarios are high pressure because better tools aren't provided.

Correlations on the Monitoring aspects, in Table 3-21, are simpler because there were only 2 aspects: alternative methods of monitoring to verify rule compliance and to verify success. The degree measure for both attributes were significantly correlated to each other, as was the importance measure.

Table 3-21 - Monitor Correlations

	M1d	M2d	M1i
M2d	0.562 0.000		
M1i	-0.275 0.053	-0.310 0.028	
M2i	-0.103 0.476	-0.121 0.402	0.607 0.000

The other significant correlation was a negative relationship between the degree of verifying success, M2, with the importance of verifying compliance, M1. There is nothing surprising in the result that reduced methods of verifying success would be associated with increased importance of measuring compliance. The final comparison that showed signs of correlation, though was less obvious. While only significant to the

0.10 level, a negative relationship is seen between the degree and importance on the M1 aspect, suggesting that the importance of having alternative methods of verifying compliance dropped where those methods existed. Based on the scenarios and the model, there is no apparent explanation for this scoring behavior from the respondents.

Regression Analysis

Performing regression analyses on the data from this phase was unlikely to provide conclusive results about the general use of the model as an alternate for two reasons: the operational limitations on the workforce constrained the sample population to include only the gold standard reviewers and the highly collinear responses would confound relationships in the data. This was verified however by performing regression analyses on both the Control Value and Failure Likelihood.

The regression using Control Value as the dependent variable resulted in appreciably non-normal distribution of residuals with no model factors consistently significant. The results are tabulated in Table 3-22, which show that the respondent-to-respondent differences in the ID variable account for the most significant relationship

Table 3-22 - Control Value regression on Inevitability and Intervention aspects

CV Regr	ID	Scen.	Inev	Int 1	Int 2	Int 3	R^2	R^2(adj)
1	0.014	0.244	0.118	0.109	0.880	0.141	50.29%	23.89%
2	0.010	0.188	0.087	0.097	-	0.094	50.26%	26.14%
3	0.018	-	0.340	0.170	-	0.134	40.39%	21.06%
4	0.020	-	-	0.170	-	0.206	38.89%	21.20%

While the Inevitability and two of the Intervention aspects suggest a relationship in the second run, the loss of significance with the elimination of the non-significant Scenario variable suggests any relationship was not strong or may have been masked by collinearity in the variables.

The regression on the Failure Likelihood considered two different measures. One was a weighted score for each aspect, and the other was the degree score for the aspects. The importance score was not considered independently in the regressions because it measures the respondent's perception of the scenario, rather than the procedural control captured in the rule. In effect, the importance captures the need for the control, rather than anything specific to the control itself. The importance was significant in the determination of the weighted attribute score, as the 1-100 values for importance were used as a percentage multiplier for the degree value. For example one respondent scored the D1 aspect on scenario 1with a 74 for degree and 75 for importance, resulting in a weighted score for D1 of 55.5, 74 x 75%.

For regressions using the degree measures as the independent variables, the residuals were not normally distributed and the ID variable was consistently significant. As non-significant variables were removed, all three Definition aspects remained significant along with O1. The significance of the Definition aspects was not surprising because Phase A showed this to be significant in that phase to the reviewers from this same subpopulation. The significance of the O1 aspect suggests that these reviewers recognized a clear influence of effective tools on reducing failure likelihood. Unfortunately, the significance of both the ID and the Scenario variables did not make this implementation of the model very useful as a substitute.

Regressions using the weighted measures as the independent variables were not appreciably different with only one exception: the D1 measure was not significant. While not conclusive, this lack of significance on the aspect of Definition regarding the clarity of the language could indicate that the degree to which the clear language is related to success is independent of the perceived importance.

Summary of Phase C

The phase C data provided insights into the scoring behavior of the gold-standard respondents because it considered multiple aspects of the DATOM and I&I model elements. The improvement of the scoring scale, by switching from the 10-point to the 100-point scale, eliminated the effects of the discrete differences in scoring value. Unfortunately, the added questions could only be presented to a small respondent set and the additional questions created a large potential number of interactions.

Analysis of the associated data focused on the correlations for the model elements where there was more than one aspect of that element that the respondents were asked to consider: Intervention, Definition, Organization and Monitoring. Intervention questions only looked at a selected set of aspects, where the questions on the DATOM element aspects considered both a measurement of degree and importance of each aspect.

Many of the correlations observed were intuitively obvious but were not explicitly considered by the respondents while answering the questions. This supports the position that the aspects of the model elements can represent the characteristics of procedural controls. The aspects were used only on a limited sample of respondents, so the expansion would have to be assessed across a larger population to determine the extent to which they are applicable. The correlations found are summarized below in general terms based on the detailed analysis provided in sections above.

Within the Intervention aspects, workers were aware of appropriate responses to situations where problems were detectable. They were less likely to recognize problems or know the appropriate actions to take when there was time pressure. While it cannot be directly determined from the responses if the increased time pressure was the cause or result of difficulty detecting problems and identifying responses, this relationship could be the basis for a guideline on designing procedures—when faced with time pressure procedure designers should ensure that workers are provided assistance in detecting anomalies and deciding on responses.

The aspects of definition were highly correlated so it was the non-correlated aspects that were the notable relationships. The importance of rule awareness was unrelated to either the clarity the language or the intelligibility of the rule. However, the degree of rule awareness was highly correlated to both the importance of the rule's clarity and its intelligibility. This suggests that the set of respondents may not explicitly recognize the rules that must be memorable, they do recognize the rules that must be clear and easy to understand, which serves as an indirect indicator of how important it is for that rule to be memorable. Also, regardless of how important it is for a rule to be memorable, the memorable rules are the ones where procedure designers used clear and simple language. This too provides a useful design guideline when creating procedures because procedure designers can improve the memorability of important rules by focusing on simple and clear communication channels.

Within the aspects measured for Organization, as quality of tools dropped, the presence of improvisation increased. Similarly, the improvisation present in executing a task increased with decreasing memorability of task details. The main implication in procedure design is that the designer should provide tools and memorable tasks to avoid improvisation.

Also in the Organization aspects, as time constraints became important, the tools, details and absence of goal conflicts also became important. This suggests that the designer should avoid specific details where the tasks are appropriate to the judgment of the worker, allowing improvisation where it is useful and there is time to allow workers to determine the course of action.

Task monitoring aspects also presented relationships that were unsurprising but useful as guidelines. Where methods for identifying compliance were less effective, methods for identifying success tended to be less effective. Similarly, when compliance was perceived as important, verifications of success were perceived as important. The final observation about the Monitoring aspects was that in situations where ability to verify success increased, the importance of verifying compliance tended to drop. This highlights a shortcoming in the group of respondents used for this aspect analysis. The Corrective Action engineering group that provided the respondents was typically focused on the quality of the output more than the cost, and their perceptions on the monitoring aspects may not be shared by a financial analyst or program manager. Where the respondents were satisfied that quality would be protected by output verification, the cost inefficiencies caused by that behavior would be undesirable if problems could be found earlier than the end of the process.

CHAPTER 4 SIGNIFICANCE

Risk analysis and risk management are fields where exhaustive work has been done on quantifying risks, identifying the hazards that are the source of the risks and finding methods of reducing risks. However, little work has been done regarding procedures, which is one of the main methods an organization has for controlling risks. When procedures and risk are considered together, the main focus in the literature is on the human factors associated with work instructions, and violations are typically treated as an error on the part of the operator, despite evidence that the worker is not entirely to blame.

In the absence of any tools or techniques to systematically gauge the impact of procedures on the risks that an organization faces, the focus of this research has been on providing a tool that can efficiently consider the procedure-related risks faced by an organization without sacrificing the accuracy of the results. Using the control value and DATOM, the method of describing a process in terms of the Definition, Assignment, Training, Organization and Monitoring of the tasks, the idea of deconstructing procedures down into specific controls provides a structure for systematically understanding the effect of each control in the procedure.

The Control Assessment technique identified here is unconventional, not because it attempts to find approximations for the consequence and likelihood components of risk, but because those approximations are only indirectly related to the general ideas of consequence and likelihood that make up risk. The substitution of control value for consequence is based on the idea that the more valuable a control is, the more effective it will be at reducing the effects of the hazard, assuming that the control works when called upon. Substituting DATOM for the likelihood of realizing a particular catastrophic loss assumes that the reliability of the controls put in place to prevent the loss provides insight into how often that loss can be expected to occur.

Using these substitutions, an alternative definition of risk develops, where the risk is described in terms of how well the organization is preventing a loss by responding in its procedures to the hazards. The two main benefits of this alternate method are that it doesn’t require disparate consequences to be equated and that it can provide results for even new systems, where there is no long history of performance on which to base likelihood probabilities.

While this research only concerned itself with written procedures in situations where there was a potential for extreme consequences, there is nothing to indicate that the principles would not apply beyond that boundary, which was chosen only for convenience. In fact, the technique seems well suited to any work domain where failures could result in serious consequences and the existing processes use repeated execution of similar action sequences, such as medicine. Further research to confirm that the technique is portable to other work domains would be useful, but the ability of the technique to obtain results without needing to extrapolate failure rates from a long process performance history suggests that this tool could help gauge risks during the design of a process, providing an augmentation of or alternative to FMEA as a tool for assessing hazards of a system’s design.

As a method of replacing the traditional Quantified Risk Analysis techniques in assessing procedures, the model assessed in this research was not sufficiently validated to be used in the current form. Each of the three phases however did provide valuable insight into improving the model or the effectiveness of possible improvements in creating processes.

While it was not conclusive, Phase A did show some promise for the DATOM model elements as a potential replacement for assessing likelihood in a traditional technique. Inevitability and Intervention elements also showed some potential as proxy for the consequence of a control, though to a more limited extent.

By far though, the most interesting result of this research was the unexpected correlation in Phase B between improved Monitoring and higher failure rates in the *Columbia* data. Because the respondents provided their answers approximately 6-years after the *Columbia* accident, the corrective actions implemented after the investigation had an opportunity to be institutionalized. This showed that areas determined to be error-prone during the accident investigation were perceived by the workforce as being monitored more vigilantly, suggesting that the improvements resulting from the accident were sustained over time

The correlations visible in Phase C have limited applicability due to the limited sample, but they do point to areas where improvement efforts could be targeted for procedure designers to make effective use of limited resources.

In Phase A, the universal significance of the Definition element of the DATOM model, despite significant respondent-to-respondent scoring differences, indicates that to some extent, the DATOM model can be used as a substitute for expert judgment of how likely a procedure is to fail. The model must be refined considerably before DATOM scores could be used as substitutes for CV, but with little modification Definition scores could be used as a validation tool. The Definition question could be posed to personnel who assess likelihood as part of a risk evaluation; responses could be used to screen for validity of the reported likelihood when their answers may be distorted by conflicts of interest.

The Phase B results unrelated to the *Columbia* corrective actions suggest that task organization, as represented by the Organization question provides an opportunity for floor-level personnel (technicians and possibly inspectors as well) to identify where processes are likely to not be successful. By understanding where these personnel believe the task is poorly organized, task designers can focus improvement efforts to ensure they obtain the desired behavior and outcomes.

While Phase C could only be directly applied to a small subpopulation, it demonstrated fundamental relationships between aspects of the DATOM elements that were correlated to the expert perceptions of risk. The results also suggest a general causation that provides a guideline for process designers and procedure authors to develop procedures that are perceived as being effective. By focusing particularly on the relationships dealing with the Organization aspects, creators of processes and procedures can leverage these relationships directly into rules that succeed in obtaining the proper behavior; improving memorability of task details and providing useful tools should lead to higher worker compliance with the expected behavior. The findings regarding goal conflicts also points to the value of avoiding over-specifying tasks when goals are clear and time pressure is low, relying on judgment of the worker to satisfy the goal.

Overall, the use of the model in this implementation as tool that can substitute for quantification of procedure-caused risks is not warranted; the I&I scores are insufficient to capture the differences in Control Value from scenario to scenario and because the DATOM scores have not been shown to completely represent the Failure Likelihood of the controls. On the other hand, there are immediate and practical uses for the model and areas for further development that are promising leads for making the model suitable for the original purpose.

CHAPTER 5 FURTHER RESEARCH

During the course of this research, multiple areas for refinement of the model and further research have been identified. These areas could be the basis of enhancements to the tool that would make it more useful to risk managers or ways to augment the role of process designers.

Objective Scoring Criteria

The most promising improvement to the methodology of collecting expert opinion is to provide respondents with a measure of training or some guidelines for consistently scoring across respondents, at least within a subpopulation. One such guideline would be a set of objective criteria associated with different scores, such as the example rubric in Table 5-1 below for Organization on a 100 point scale. Additional criteria could be added for each model element in an iterative fashion.

Table 5-1 - Example objective scoring criteria for Organization

Score	Criteria
10	No tools are provided and the process is arbitrary
20	
30	
40	
50	Tools are provided but rely heavily on experience or process has an imposed order
60	
70	
80	
90	
100	Intuitive tools are provided with cues on how they can be used in a process with an inherent flow

One benefit of objective scoring criteria to evaluate is if the use of criteria reduces the scoring variability that results in scenarios having a high degree of concordance in the order they are ranked, but still resist attempts to fit a regression line. Any resulting improvement of the regression could provide enough information to show that more portions of the model are effective for more subpopulations than the initial research was able to show.

Model Training

In addition to the objective scoring criteria mentioned above, the respondents could be provided with training in how the model is expected to work. Additional insight into the model might help make respondents more consistent in their scoring, particularly scoring aspects outside their area of expertise, where they might benefit from instruction on the subtleties that they may otherwise miss. A drawback to consider and watch for in this effort is the potential that the model elements could no longer be used to validate an expert's opinion; any biases that would affect the assessment of the likelihood or consequence would likely be applied to items the respondent knows are substitutes for those values. Similar to the objective scoring criteria above, this training could also improve the linearity of the relationships between model variables, leading to broader environments where the resulting tool could be put to use effectively.

Multidisciplinary Combined Scoring

The within-group consistency of scoring seen in Phase A and Phase B for the different subpopulations provides a lead to another of the follow-up areas where additional work can be performed on this subject. If groups can be found who score consistently for an element of the model that could not be verified with the subpopulations considered in the initial research, their scores could be used as substitutes in the portion of the Control Value and Failure Likelihood these groups could not account for. For example, instructors from the training organization could provide insight into the contribution of the Training element in the Failure likelihood.

Combining scores across different groups could be difficult, but there are two options that should be investigated: simple averaging across a diverse respondent set who represents the necessary subpopulations, and scores reached by consensus among the subpopulations. Mean scores across groups would be the simpler and less labor intensive choice, but the consensus scores might be more robust because they will force all respondents to discuss the scoring rationale behind a particular score. Of course, consensus scoring would be subject to the characteristic difficulties of consensus in addition to the added labor costs, because arguments without consensus could occur, or participants with dominant personalities could exert disproportionately large influence on the consensus score.

If the use of mean scores was sufficient to provide insight into the risks, but is not as robust as consensus scoring, perhaps mean scores could be used in the cases where the consequences to the organization are not catastrophic. This way, the lower cost method could still be used to gain moderate improvements where the more expensive would not be possible due to practical constraints of budget and schedule.

Independent Facilitation

Presuming the cost of consensus scoring is found to be more effective, potentially the assistance of a facilitator, skilled in the scoring methodology, would be useful. An independent facilitator's role could include not only coaching the group to ensure the scoring is consistent with the objective criteria; it could also begin as the trainer and conclude with moderating any extreme personalities and ensuring an appropriate pace through the scoring activity.

This role could be similar to the role of a facilitator in the performance of a HAZOP hazard assessment, where an outsider with experience in the technique facilitates without needing any significant amount of experience in the specific process being assessed. In fact, the use of a tool derived from this research may provide an additional option to companies performing OSHA-required Process hazard analysis activities as part of the Process Safety Management (Process safety management of highly hazardous chemicals, 2013).

Domain Transferrability

All analysis for this research dealt with a narrow subset of the Space Shuttle workforce, itself a highly specialized area of the broader aerospace industry. However, there is nothing to suggest that the concepts cannot be generalized for other industries. Further research could show that a refined model applicable to one particular area may be universal; with the model representing not performance specific to Space Shuttle operations, but perceptions and performance inherent to human nature.

REFERENCE

American Institute of Chemical Engineers. (1994). *Dow's Fire & Explosion Index Hazard Classification Guide*. New York, NY: John Wiley & Sons

Associated Press, "Cali Crash Case Overturned". CBS. June 16, 1999. Retrieved April 10, 2015 http://www.cbsnews.com/stories/1999/06/16/world/main51166.shtml

Columbia Accident Investigation Board. (2003). *Columbia Accident Investigation Board Report*. Washington, DC: United States Government Printing Office

Dekker, S. (2005) *Ten Questions About Human Error: A New View of Human Factors and System Safety*. Mahwah, NJ: Lawrence Erlbaum Associates

Dekker, S. (2006) *The Field Guide to Understanding Human Error*. Burlington, VT: Ashgate Publishing

Fredrickson, B L.; Kahneman, D (1993) Duration Neglect in Retrospective Evaluations of Affective Episodes. *Journal of Personality and Social Psychology*, 65(1), 45-55

Garrick, B. J. (1989). Risk Assessment Practices in the Space Industry: The Move Toward Quantification. *Risk Analysis*, *9*(1), 11-27.

Haimes, Y. Y. (2009), On the Complex Definition of Risk: A Systems-Based Approach. *Risk Analysis*, 29: 1647–1654

Haimes, Y. Y. (2009), On the Definition of Resilience in Systems. *Risk Analysis*, 29: 498–501

Haimes, Y. Y., Crowther, K. G., Horowitz, B. M. (2008), Homeland Security Preparedness: Balancing Protection with Resilience in Emergent Systems. *Systems Engineering*, 11: 287-308

Hale, A., and Borys, D. (2013) Working to Rule, or Working Safely. In: Bieder, C., Bourrier, M. (Eds.), *Trapping Safety into Rules*, pp. 43-68. Aldershot, England. Ashgate Publishing

Hillson, D. (2002) Extending the Risk Process to Manage Opportunities. *International Journal of Project Management*, 20: 235-240

International Council on Systems Engineering. 2006. *INCOSE Systems Engineering Handbook*. San Diego: INCOSE

ISO 31000, 2009. ISO 31000: *Risk Management: Principles and Guidelines = Management Du Risque: Principles Et Lignes Directrices*. Geneva, Switzerland: ISO.

Johnson P. and Gill J. (1993). *Management Control and Organisational Behavior*. London: Paul Chapman Publishing Ltd.

Kaplan, S. (1993) Formalisms for Handling Phenomenological Uncertainties: The Concepts of Probability, Frequency, Variability and Probability of Frequency. *Nuclear Technology*, 102, 137-142

Kaplan, S. and Garrick, B. J. (1981). On the Quantitative Definitionof Risk. *Risk Analysis*, *1*(1), 11-27.

Kaplan, S., Haimes, Y. Y. and Garrick, B. J. (2001), Fitting Hierarchical Holographic Modeling into the Theory of Scenario Structuring and a Resulting Refinement to the Quantitative Definition of Risk. *Risk Analysis*, 21: 807

Kumamoto, H. and Henley, E. J. (1996). *Probabilistic Risk Assessment and Management for Engineers and Scientists*. Piscattaway, NJ: IEEE Press

Leffingwell, D. and Widrig, D. (2000). *Managing Software Requirements: A Unified Approach*. Reading, MA: Addison-Wesley

Leiden, K. Keller, J. and French, J. (2002). *Information to Support the Human Performance Modeling of a B757 Flight Crew during Approach and Landing*. http://humanfactors.arc.nasa.gov/ihi/hcsl/publications/757_ApproachLanding_CTA.pdf

Leveson, N. (2002). *A New Approach To System Safety Engineering*. http://sunnyday.mit.edu/book2.pdf

National Aeronautics and Space Administration. (1986). *Instructions for Preparation of Failure Modes and Effects Analysis (FMEA) and Critical Items List*. NSTS 22206. Washington, DC

National Aeronautics and Space Administration. (2007). *Systems Engineering Handbook*. NASA/SP-2007-6105 Rev1. Washington, DC: NASA

National Transportation Safety Board (1996), Safety Recommendations A-96-90 through -106. http://www.ntsb.gov/doclib/recletters/1996/A96_90_106.pdf

New York Times (1997) “American Airlines Ruled Guilty Of Misconduct in '95 Cali Crash” http://www.nytimes.com/1997/09/12/world/american-airlines-ruled-guilty-of-misconduct-in-95-cali-crash.html

Pélegrin, C. (2013) The Never-Ending Story of Proceduralization in Aviation. In: Bieder, C., Bourrier, M. (Eds.), *Trapping Safety into Rules*. Aldershot, England. Ashgate Publishing

Perrow, C. (1999). *Normal Accidents: Living with High-risk Technologies*. Princeton, NJ: Princeton University Press

Petroski, H.; (1992). *To Engineer Is Human: The Role of Failure in Successful Design*. New York, NY: Vintage Books

Process Safety Management of Highly Hazardous Chemicals, 29 C.F.R. part 1910.119 (2013).

Reason, J.T.; (1990). *Human Error*. Cambridge, UK: Cambridge University Press

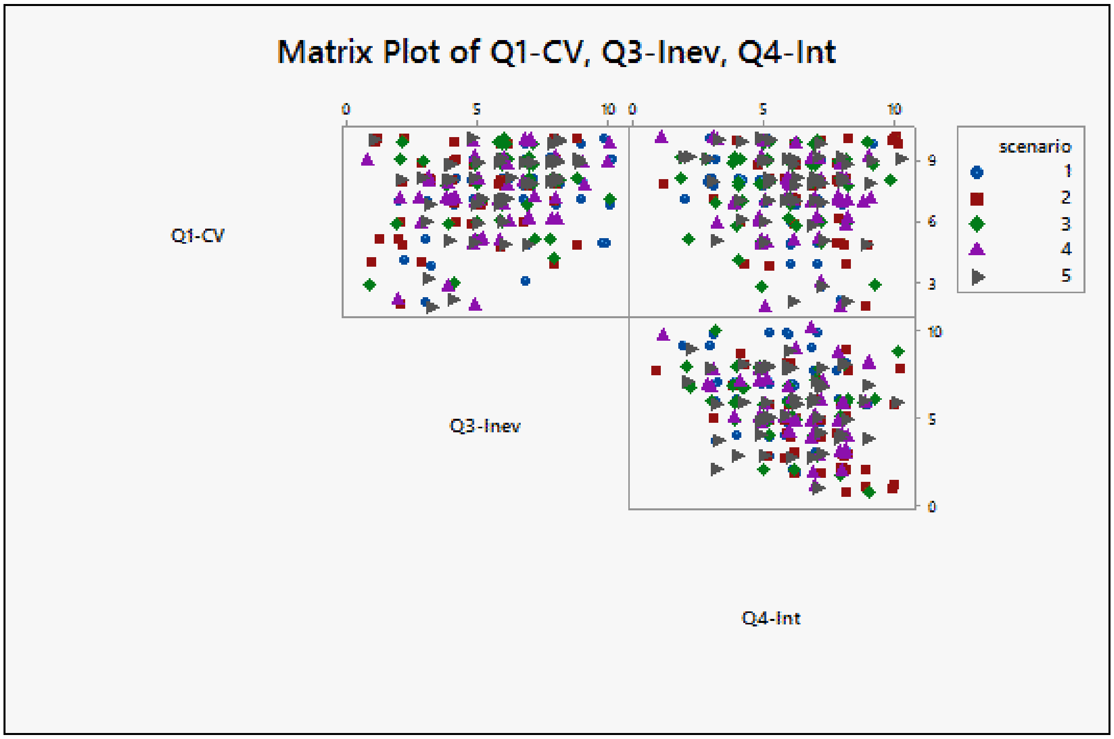
Schrock, E. and Lefevre, H. (1988). *The Good and the Bad News about Quality*. New York, NY: Marcel Dekker, Inc.

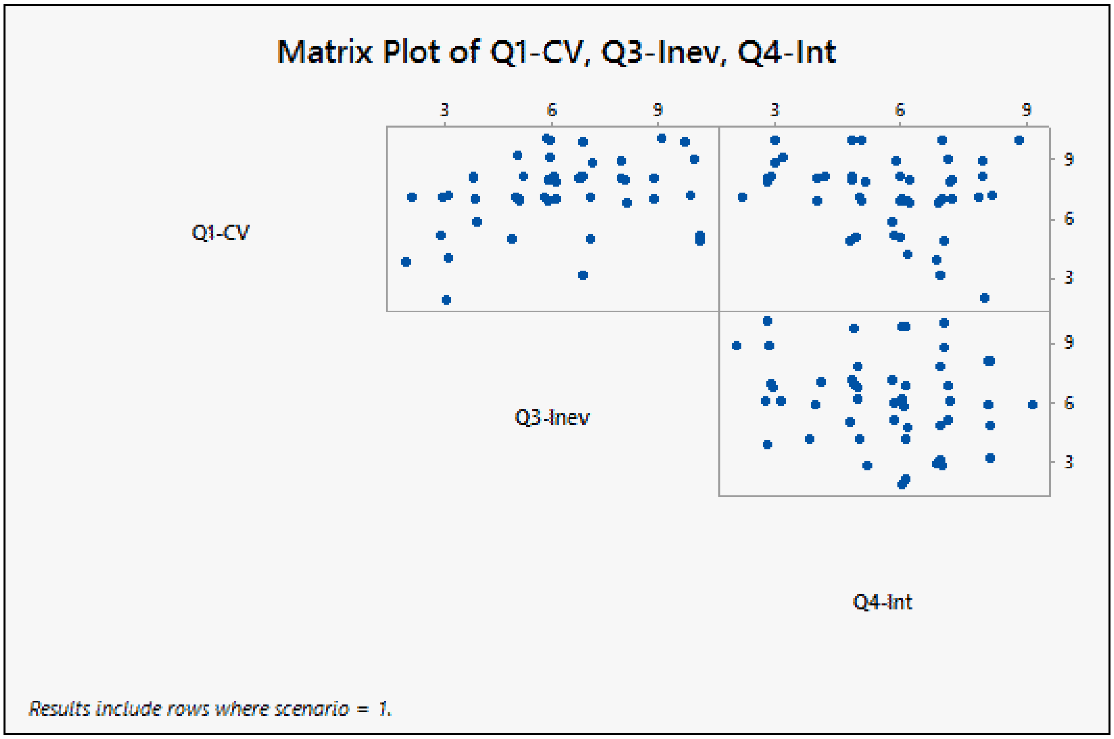
Seaver, D.A.; Stillwell, W.G. (1983). *Procedures for Using Expert Judgment to Estimate Human-Error Probabilities in Nuclear Power Plant Operations*. NUREG/CR-2743. Washington, DC: Nuclear Regulatory Commission

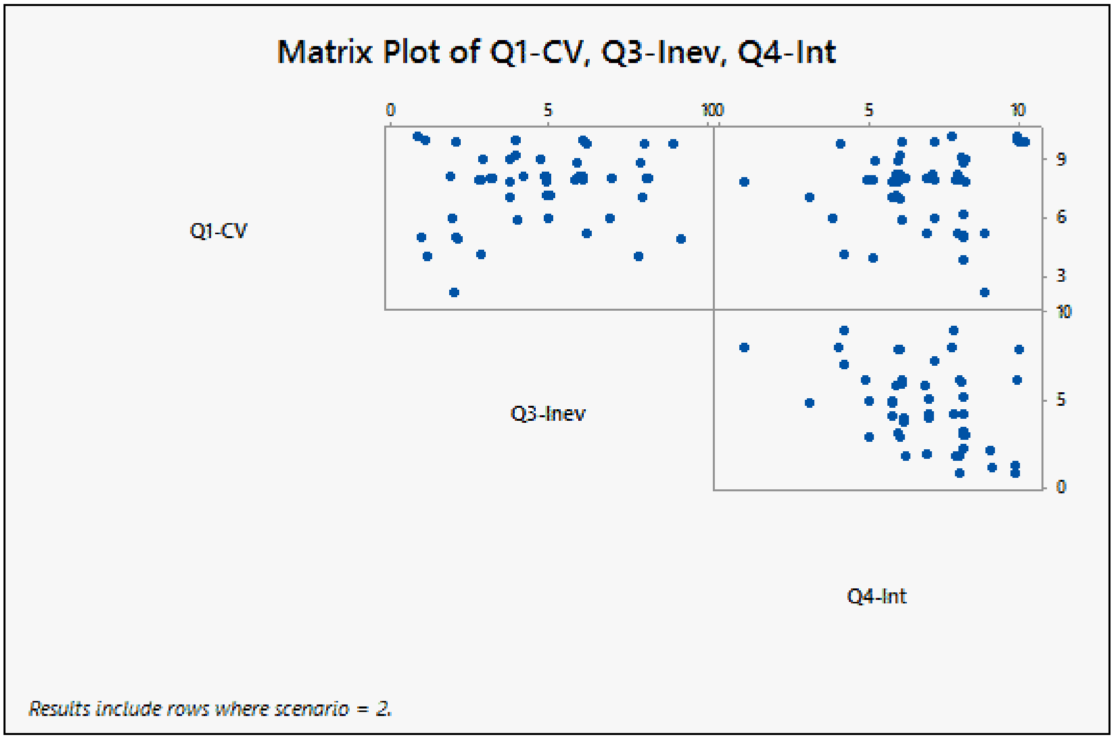
United States Department of Defense. (1980). *Procedures for Performing A Failure Mode, Effects and Criticality Analysis*. MIL-STD-1629A. Lakehurst, NJ: Naval Air Engineering CenterUnited States Department of Defense. (2012). *Standard Practice for System Safety*. MIL-STD-882E. Wright-Patterson Air Force Base, OH: Air Force Materiel CommandUnited States Department of Defense. (2015). *Risk, Issue, and Opportunity Management Guide for Defense Acquisition Programs*. Washington, DC: Office of the Deputy Secretaryof Defense for Systems Engineering

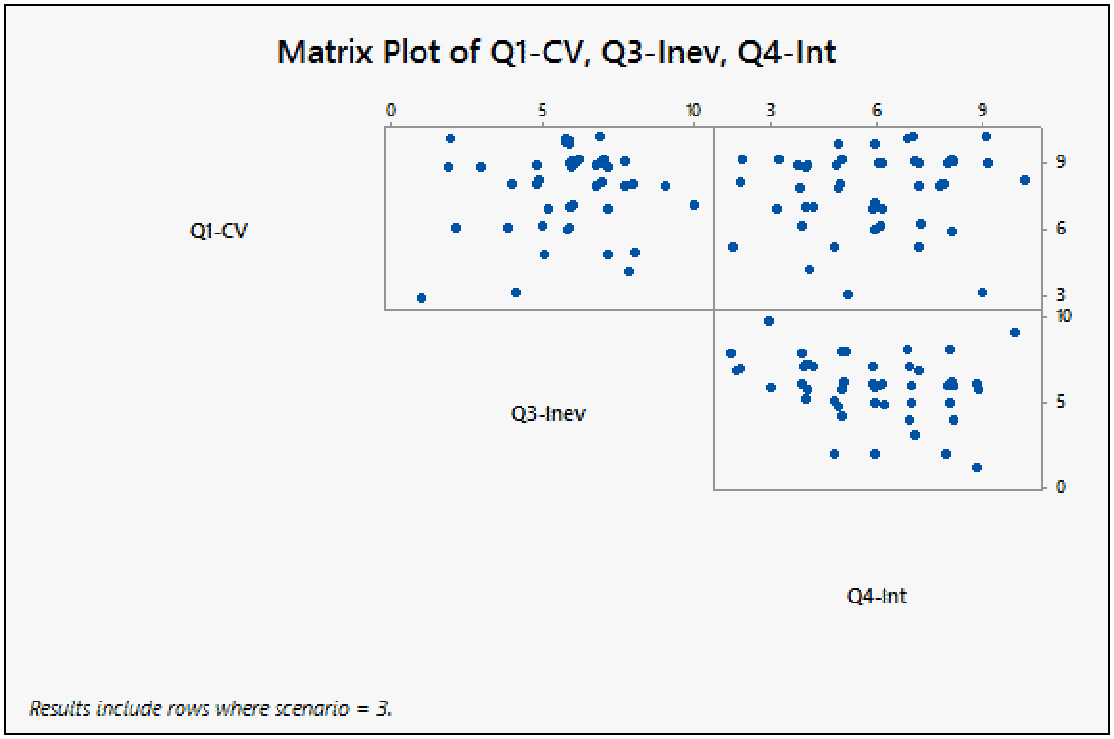
Vaughan, D. (1996). *The* Challenger *Launch Decision - Risky Technology, Culture and Deviance at NASA*. Chicago, IL: University of Chicago PressWoods, D.D., 2006. Essential Characteristics of Resilience. In: Hollnagel, E., Woods, D. D., Leveson, N. (Eds.), *Resilience Engineering: Concepts and Precepts*. Ashgate, Aldershot, England

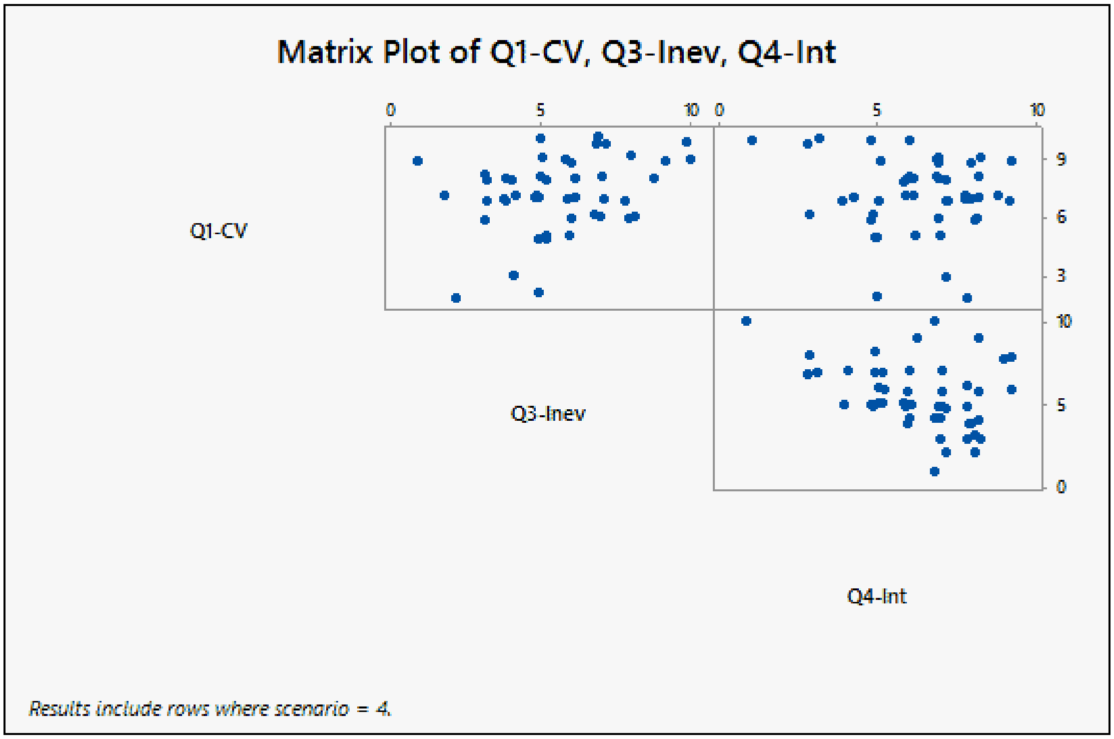
APPENDIX A


Matrix Plot of Q1-CV, Q3-Inev, Q4-Int
0
5
10
0
5
10
9
6
3
10
5
0
Q1-CV
Q3-Inev
Q4-Int
scenario
● 1
■ 2
◆ 3
▲ 4
▶ 5

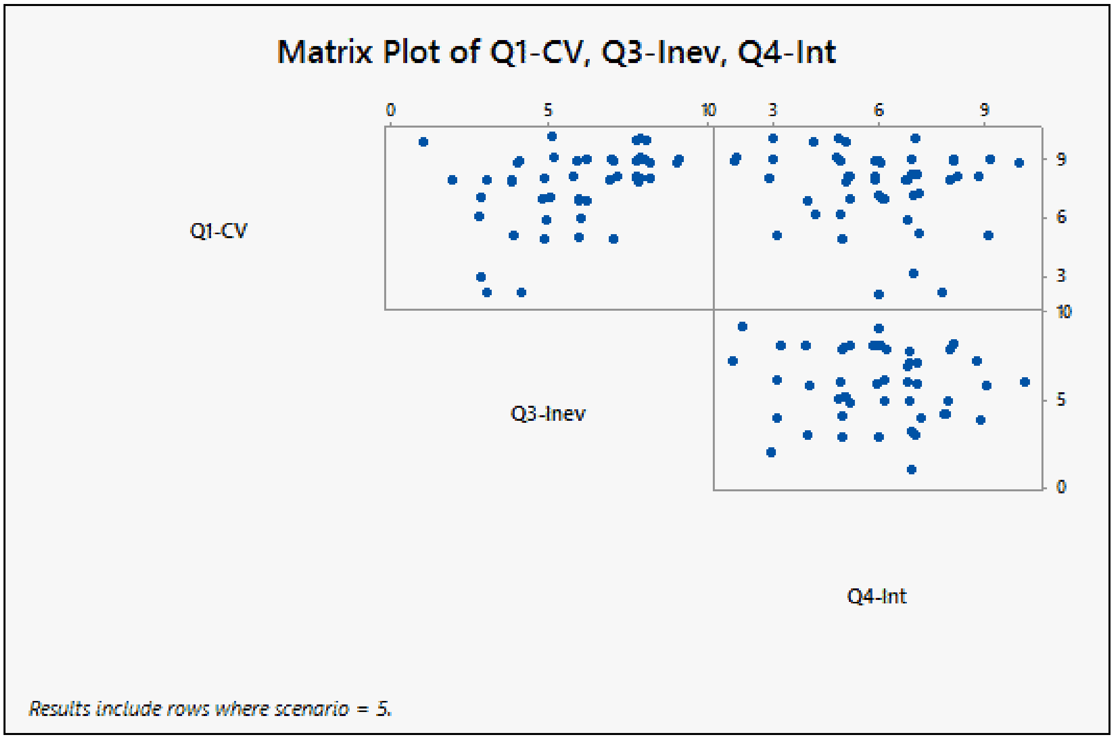

Matrix Plot of Q1-CV, Q3-Inev, Q4-Int
3
6
9
3
6
9
9
6
3
9
6
3
Q1-CV
Q3-Inev
Q4-Int
Results include rows where scenario = 1.


Matrix Plot of Q1-CV, Q3-Inev, Q4-Int
0
5
100
5
10
9
6
3
10
5
0
Q1-CV
Q3-Inev
Q4-Int
Results include rows where scenario = 2.

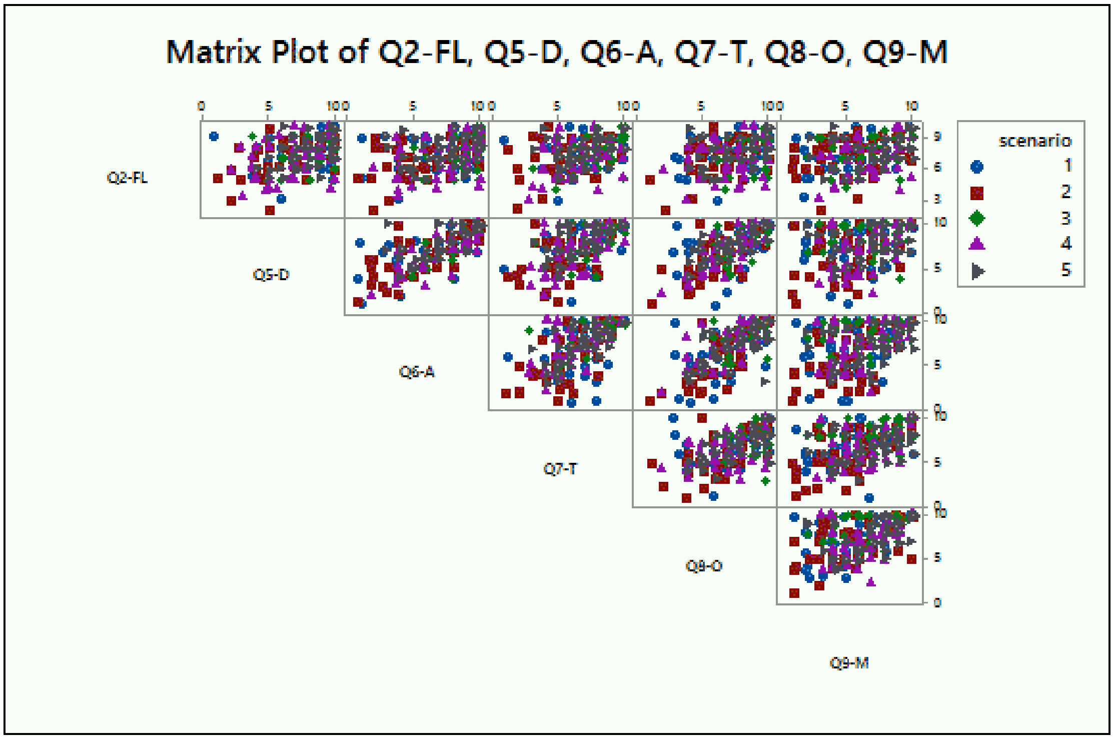

Matrix Plot of Q1-CV, Q3-Inev, Q4-Int
0
5
10
3
6
9
9
6
3
10
5
0
Q1-CV
Q3-Inev
Q4-Int
Results include rows where scenario = 3.


Matrix Plot of Q1-CV, Q3-Inev, Q4-Int

Results include rows where scenario = 4.


Matrix Plot of Q1-CV, Q3-Inev, Q4-Int

Results include rows where scenario = 5.


Matrix Plot of Q2-FL, Q5-D, Q6-A, Q7-T, Q8-O, Q9-M
Q2-FL
Q5-D
Q6-A
Q7-T
Q8-O
Q9-M
0
5
10
9
6
3
scenario
1
2
3
4
5

VITA

Gregory T Praino was born in the Bronx, New York, in 1973. His parents are Joseph Praino and Virginia Praino. He received his secondary education at The Bronx High School of Science. In September 1991, he entered the Florida Institute of Technology from which he graduated with the B.S. degree in Mechanical engineering in December 1995. While working at the Kennedy Space Center as a space shuttle engineer in January 2000 he was admitted to the Graduate School of the University of Miami, where he was granted a M.S degree in Management of Technology in December 2002.