

Technical Paper

Independent Evaluation of INPO's Nuclear Safety Culture Survey and Construct Validation Study

Stephanie Morrow, Ph.D.
Human Factors and Reliability Branch
Valerie Barnes, Ph.D.
Division of Risk Analysis
Office of Nuclear Regulatory Research

June 2012

Preface

This paper presents information that does not currently represent an agreed-upon staff position. The U.S. Nuclear Regulatory Commission has neither approved nor disapproved its technical content.

The safety culture survey data presented in this report are the property of the Institute of Nuclear Power Operations, shared under a Memorandum of Agreement between the Nuclear Regulatory Commission and the Institute of Nuclear Power Operations, dated December 10, 2007.

Acknowledgements

We would like to thank Dr. G. Kenneth Koves and Darrel Maret of the Institute for Nuclear Power Operations (INPO) for the information they provided on the administration of the safety culture survey, explanations of their analyses, and willingness to share raw data and make this evaluation possible. We would also like to acknowledge the work of the Idaho National Laboratory in support of the analyses documented in this paper and Pacific Northwest National Laboratory in support of the literature review documented in this paper.

Abbreviations

ACRS	Advisory Committee on Reactor Safeguards
BI	Barrier Integrity
EP	Emergency Preparedness
HP	Human Performance
IAEA	International Atomic Energy Agency
ICC	Intraclass Correlation
IE	Initiating Event
INPO	Institute for Nuclear Power Operations
INSAG	International Nuclear Safety Advisory Group
IP	Inspection Procedure
ITP	Industry Trends Program
KPI	Key Performance Indicator
LER	Licensee Event Report
MOR	Monthly Operating Report
MS	Mitigating Systems
NEI	Nuclear Energy Institute
NRC	Nuclear Regulatory Commission
OE	NRC Office of Enforcement
OR	Occupational Radiation Safety
PAF	Principal Axis Factoring
PCA	Principal Components Analysis
PI&R	Problem Identification & Resolution
PP	Physical Protection
PR	Public Radiation Safety
RES	NRC Office of Nuclear Regulatory Research
ROP	Regulatory Oversight Process
SCCI	Substantive Cross-Cutting Issue
SCWE	Safety Conscious Work Environment
SPSS	Statistical Package for the Social Sciences
TMI	Three Mile Island
US	United States

Table of Contents

1.	Introduction.....	1
1.1.	Safety Culture and the Nuclear Industry.....	1
1.2.	Safety Culture Policy Statement Definition and Traits.....	1
1.3.	Purpose of Current Paper.....	3
2.	Overview of Safety Culture Research.....	4
2.1.	The Underlying Theory of Organizational Safety Culture.....	4
2.2.	Distinguishing Safety Culture and Safety Climate.....	6
2.3.	Definitions and Dimensions of Safety Culture.....	6
2.4.	Relationships between Safety Culture and Safety Performance.....	7
2.5.	Safety Culture Interventions.....	9
3.	Development of the Safety Culture Survey.....	11
3.1.	Survey Item Development.....	11
3.2.	Survey Administration.....	13
4.	Analysis of Safety Culture Survey.....	14
4.1.	Exploratory Factor Analysis.....	14
4.2.	Reliability Analysis.....	17
4.3.	Consistency of Survey Factors with Other Research.....	19
4.4.	Within-Group Reliability Analysis.....	19
4.5.	Descriptive Analysis of Safety Culture Factors.....	21
5.	Criterion-Related Validity of the Safety Culture Survey.....	23
5.1.	Description of NRC Performance Metrics.....	23
5.2.	Concurrent Validity of the Survey Data with NRC Performance Metrics.....	28
5.3.	Exploratory Predictive Validity of the Survey with NRC Performance Metrics.....	35
5.4.	Relationships among NRC Performance Metrics.....	39
6.	Alignment of Survey with Safety Culture Policy Statement.....	46
6.1.	Mapping of Survey Factors to Policy Statement Traits.....	46
7.	Summary and Discussion.....	48
8.	References.....	50

List of Tables

Table 1 Results of PCA of 110 Items with a 9 Factor Solution.....	15
Table 2 Safety Culture Factors, Sub-Factors, and Example Items	15
Table 3 Results of Reliability Assessment using Cronbach’s Coefficient Alpha	18
Table 4 Range of ICC(1) and ICC(2) Values for Safety Culture and Safety Culture Factors.....	20
Table 5 Descriptive Statistics for Safety Culture and Safety Culture Factors.....	21
Table 6 Intercorrelations among Safety Culture Overall and Safety Culture Factors.....	22
Table 7 Correlations between Safety Culture and 2010 Performance Indicators, ROP Action Matrix, Inspection Findings, and SCCIs.....	31
Table 8 Correlations between Safety Culture and 2010 NRC Allegations, ROP Cross-Cutting Areas and Components.....	32
Table 9 Correlations between Safety Culture and 2011 ROP Cross-Cutting Areas and Components....	36
Table 10 Correlations between Safety Culture and 2011 ROP Action Matrix, Inspection Findings, SCCIs, and Allegations	37
Table 11 Correlations among 2009 and 2010 NRC Performance Metrics.....	40
Table 12 Correlations among 2009 and 2010 Allegations, Inspection Findings, and SCCIs.....	41
Table 13 Correlations between 2009 and 2010 NRC Performance Metrics.....	42
Table 14 Correlations between 2010 and 2011 NRC Performance Metrics.....	45
Table 15 Crosswalk of NRC Policy Statement Traits and INPO Safety Culture Survey Factors.....	46

List of Figures

Figure 1 Layers of Organizational Culture adapted from Schein (1992).....	4
Figure 2 Reason's (2000) Swiss Cheese Model of Accident Causation	5
Figure 3 Overview of NRC Performance Metrics Used in Validity Analyses.....	27
Figure 4 Scatterplot of Relationship between Safety Culture and Unplanned Scrams in 2010.....	33
Figure 5 Scatterplot of Safety Culture and Total ROP Aspects in 2010	33
Figure 6 Scatterplot of Safety Culture and SCWE-Related Allegations in 2010	34
Figure 7 Scatterplot of Safety Culture and Problem Identification and Resolution Area in 2011.....	37
Figure 8 Scatterplot of Safety Culture and Allegations in 2011	38
Figure 9 Scatterplot of SCWE-related Allegations in 2009 and SCCIs in 2010	43
Figure 10 Scatterplot of Resources Component in 2009 and Equipment Outages in 2010.....	44

1. Introduction

1.1. Safety Culture and the Nuclear Industry

The term “safety culture” was first introduced to the nuclear industry as part of the International Atomic Energy Agency (IAEA) assessment of the causes of the 1986 Chernobyl accident. The International Nuclear Safety Advisory Group (INSAG), an advisory group reporting to the Director General of the IAEA, concluded that “The vital conclusion drawn is the importance of placing complete authority and responsibility for the safety of the plant on a senior member of the operational staff of the plant. Formal procedures, properly reviewed and approved, must be supplemented by the creation and maintenance of a ‘nuclear safety culture’” (INSAG, 1986). Safety culture was formally defined by IAEA in a follow-on report (INSAG-4) as the “assembly of characteristics and attitudes in organizations and individuals which establishes that, as an overriding priority, nuclear plant safety issues receive the attention warranted by their significance” (INSAG, 1991).

The Nuclear Regulatory Commission (NRC) recognized that organizational factors have the potential to contribute to accidents well before IAEA introduced the term “safety culture.” The NRC’s investigation of the Three Mile Island (TMI) accident in 1979 stated that “The one theme that runs through the conclusions we have reached is that the principal deficiencies in commercial reactor safety today are not hardware problems, they are management problems” (Rogovin, 1980). Lessons learned from both TMI and Chernobyl led the NRC to formally define its expectations for licensees to promote a strong safety culture in nuclear power plant operations through the issuance of the Conduct of Operations Policy Statement in 1989 and the Safety Conscious Work Environment Policy Statement in 1996 (NRC, 1989; 1996).

The NRC took additional steps to address safety culture within its reactor oversight process (ROP) following the events at Davis Besse in 2002. Changes were made to incorporate safety culture into the NRC’s inspector training program, enhance the ROP to include cross-cutting aspects and components related to safety culture, and develop inspection procedures related to the formal assessment of safety culture for licensees with degraded performance through the assessment guidance in Inspection Procedure (IP) 95003 (NRC, 2011d).

In February 2008, the Commission directed the NRC staff to expand the policy on safety culture to address the unique aspects of security and ensure that the policy was applicable to all licensees and certificate holders. The goal of the policy statement was to make explicit the NRC’s “expectation that all licensees and certificate holders establish and maintain a positive safety culture that protects public health and safety and the common defense and security when carrying out licensed activities” (NRC, 2011b). The NRC staff held a series of public meetings and workshops with stakeholders to inform the development of the policy statement and gather a broad spectrum of views on safety culture and the traits that comprise the safety culture concept.

1.2. Safety Culture Policy Statement Definition and Traits

The primary input to the NRC’s Safety Culture Policy Statement came from a public workshop held in February 2010. The structure of this workshop was unique in that the NRC selected 16 external stakeholders to participate as members of a panel. The workshop panelists represented a wide range of stakeholders regulated by the NRC and/or the Agreement States, including medical,

industrial, and fuel cycle materials users, nuclear power reactor licensees, and members of the public. Panelists were organized into groups by affiliation and interest, given examples of behaviors that exemplify a positive safety culture from previous nuclear and non-nuclear literature, and asked to develop additional behaviors that they would associate with a positive safety culture. Then they used the written behaviors as inputs into an affinity diagramming exercise (Tague, 2004). Affinity diagramming is used to organize large amounts of information into groups by finding relationships in the content. It is particularly useful when the goal is to develop consensus around some concept, such as was the case in developing a definition and set of traits to describe safety culture. Panelists grouped behaviors that seemed to them to be related until all the behaviors were within a group. The panelists discussed the emergent groups and developed labels to capture the meaning of the groups. The panel then reached consensus on language for the definition of safety culture and used the labels developed from the affinity diagramming exercise to characterize the traits of a positive safety culture. Additional public meetings and public comment periods helped to refine the definition and traits developed during the workshop and confirm stakeholder agreement.

The final NRC Safety Culture Policy Statement published in June 2011 adopted the following definition of safety culture:

“Nuclear Safety Culture is defined as the core values and behaviors resulting from a collective commitment by leaders and individuals to emphasize safety over competing goals to ensure protection of people and the environment.”

The list of traits further defining safety culture includes:

- **Leadership Safety Values and Actions:** Leaders demonstrate a commitment to safety in their decisions and behaviors.
- **Problem Identification and Resolution:** Issues potentially impacting safety are promptly identified, fully evaluated, and promptly addressed and corrected commensurate with their significance.
- **Personal Accountability:** All individuals take personal responsibility for safety.
- **Work Processes:** The process of planning and controlling work activities is implemented so that safety is maintained.
- **Continuous Learning:** Opportunities to learn about ways to ensure safety are sought out and implemented.
- **Environment for Raising Concerns:** A safety conscious work environment is maintained where personnel feel free to raise safety concerns without fear of retaliation, intimidation, harassment, or discrimination.
- **Effective Safety Communication:** Communications maintain a focus on safety.
- **Respectful Work Environment:** Trust and respect permeate the organization.
- **Questioning Attitude:** Individuals avoid complacency and continuously challenge existing conditions and activities in order to identify discrepancies that might result in error or inappropriate action.

1.3. Purpose of Current Paper

Concurrent with the development of the NRC's Safety Culture Policy Statement, The Nuclear Energy Institute (NEI) volunteered to sponsor a study, to be conducted by the Institute for Nuclear Power Operations (INPO), assessing safety culture via a survey of employees at nuclear power plants across the United States (US). The primary purposes of the study were to investigate the factors that comprise the concept of safety culture in the nuclear power industry, assess the extent to which they match the traits identified in the NRC's Safety Culture Policy Statement, and evaluate the relationships between the safety culture factors identified from the survey and other measures of organizational and safety performance. The administration of this survey also provided a unique opportunity to explore the relationship between perceptions of safety culture among nuclear power plant personnel and safety performance in the nuclear industry by comparing the results of the safety culture survey to existing performance indicators maintained by INPO and NRC. The Office of Enforcement (OE) requested the Office of Nuclear Regulatory Research (RES) to perform an independent evaluation of INPO's research. Note that because the survey was administered only to nuclear power organizations, the relationships explored in this paper are limited to nuclear power operations. However, the policy statement is intended to be applicable to all individuals and organizations in NRC-regulated communities.

The purpose of this white paper is to present the results of the NRC's independent evaluation of the INPO safety culture survey and its construct validation study. The term "construct" is used to describe a theoretical concept or idea, such as safety culture, intelligence, or personality. Construct validity refers to the extent to which an instrument used to measure a construct – the survey in this case – appears to be measuring what it purports to measure. It is established, in part, by assessing whether the instrument 1) covers the breadth of the construct being examined (content validity), 2) measures the construct consistently (reliability), and 3) demonstrates a relationship with outcomes to which it should theoretically be related (criterion-related validity).

Including this introduction, the paper is organized into seven sections. Section 2 provides an overview of the state of safety culture research, including its theoretical foundations, recent research linking safety culture to safety performance, and studies of safety culture interventions. Section 3 outlines INPO's development and administration of the safety culture survey and lays the foundation for establishing content validity. Section 4 describes INPO's exploratory factor analysis of the survey responses and presents an analysis of the reliability of the factors and robustness of the factor structure. Section 5 introduces the concurrent and predictive criterion-related validity analysis carried out by RES, exploring relationships between the safety culture survey and safety performance data maintained by the NRC. Section 6 compares the safety culture survey factors to the traits included in the NRC's Safety Culture Policy Statement. Finally, section 7 presents a summary of the evaluation of the INPO safety culture survey and construct validation study.

2. Overview of Safety Culture Research

In 2002, J. N. Sorensen, then a member of the staff of the NRC's Advisory Committee on Reactor Safeguards (ACRS), published a critical review of the state-of-the-art of safety culture research. Sorensen asserted that safety culture research cannot progress until safety culture has been defined, the characteristics or attributes of safety culture have been delineated, and a link between safety culture and safe operations has been established. Although the concept of safety culture has existed for decades, researchers and practitioners continue to debate the definition of the concept and its associated characteristics. In addition, quantitative evidence demonstrating a statistically significant relationship between safety culture and concurrent or future safety performance has been historically lacking in the research literature. It is only in the last 10 years that researchers have begun to publish more rigorous studies explicitly testing for relationships between safety culture and safety performance and reviews of the safety culture literature have begun to reach agreement around common themes in safety culture definitions and dimensions.

2.1. The Underlying Theory of Organizational Safety Culture

Edgar Schein's (1992; 2010) model of organizational culture is perhaps the most comprehensive and widely-adopted model in both nuclear and non-nuclear domains. Schein defines organizational culture as "a pattern of shared basic assumptions that the group learned as it solved its problems of external adaptation and internal integration, that has worked well enough to be considered valid, and, therefore to be taught to new members as the correct way to perceive, think, and feel in relation to those problems" (1992, p. 12). Safety culture is generally considered to be a specific aspect of organizational culture regarding the organization's shared beliefs, values, and attitudes that contribute to ensuring safe operations. Schein goes on to describe organizational culture as consisting of three layers: artifacts, espoused values, and underlying assumptions (Figure 1).

Schein's Layers of Organizational Culture

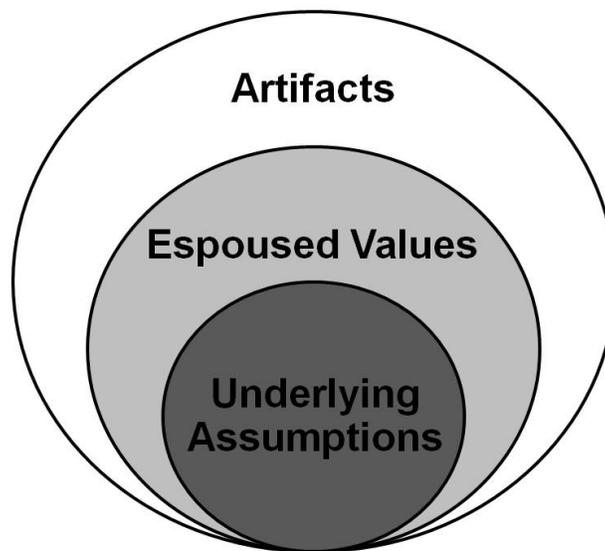


Figure 1 Layers of Organizational Culture adapted from Schein (1992)

Artifacts represent the outer layer of culture, and include visible organizational structures and processes. This layer is easily observed, such as during inspections or behavioral observations, but it is often hard to comprehend how the visible aspects of an organization relate to the underlying culture without more information. The middle layer is termed espoused values, and includes an organization's formal documentation, like policies, procedures, and corrective action plans, but also attitudes expressed by members of the organization. Employee surveys and interviews are often used to capture information at the espoused value layer of safety culture. Schein calls the core layer of organizational culture underlying assumptions. The core layer consists of implicit, basic assumptions that permeate the whole of the organization, such as assumptions about the inherent riskiness of the job. An organization's underlying assumptions are not directly observable, but these assumptions shape the organization's espoused values and artifacts.

Applied to safety culture assessments, Schein's model suggests that data collected from an organization can tap into the artifact and espoused value layers of safety culture, but the raw data must be carefully interpreted to reach conclusions about the underlying assumptions of the organization, and thus the core of the organization's safety culture.

The practical utility of assessing an organization's safety culture is that the assessment may be used as a performance indicator, in addition to more established indicators like safety management audits or analyses of events and near-misses (Guldenmund, 2000). Further, safety culture assessments may be used as leading indicators of performance, and provide opportunities for intervention before negative events occur. Post-event investigations like TMI, Chernobyl, and Davis Besse have repeatedly shown that weaknesses in an organization's safety culture can create opportunities for significant adverse events. Reason's (2000) Swiss Cheese Model of Accident Causation provides a good illustration of how safety culture can function as a contributor to event occurrence (Figure 2).

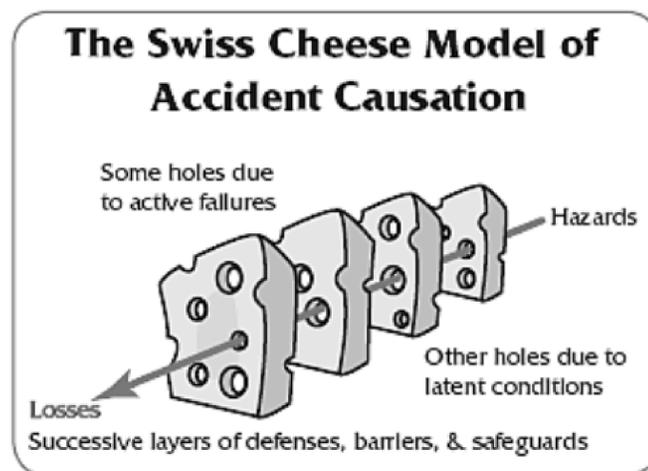


Figure 2 Reason's (2000) Swiss Cheese Model of Accident Causation

In Reason's model, multiple layers of defenses, barriers, and safeguards are characterized as having weaknesses, or holes. Some of those holes may be due to active failures, such as equipment malfunctions or instrumentation errors, but others are due to latent conditions, such as having a management team that stresses productivity over safety, a maintenance department that allows backlogs to add up, supervisors who do not provide adequate oversight of safety-significant actions, or employees who proceed even when uncertain—all potentially signifying a weak safety culture.

When the holes line up, adverse events may occur. As such, the probability of having an event may increase when there are more holes due to latent, organizational conditions because there are more opportunities for the holes to align when active failures occur.

2.2. Distinguishing Safety Culture and Safety Climate

A large proportion of the quantitative research on safety culture, particularly those studies emerging from the organizational psychology discipline, use the term safety climate rather than safety culture in their research. There has been extensive debate about the definition and measurement of organizational culture and climate that is outside the scope of this paper (c.f., Denison, 1996). Generally, organizational culture is characterized as a global construct referring to an organization's deeply-rooted and strongly-held beliefs and assumptions, whereas climate refers to more overt attitudes and perceptions shared by members of an organization. Schein (1992) describes climate as a "reflection and manifestation of cultural assumptions" (p. 230). In Schein's model of organizational culture, the middle layer, espoused values, is most closely related to safety climate. Consequently, safety climate can be considered a component of safety culture, but the full construct of safety culture encompasses more than just the perceptions of its members.

The distinction between safety culture and climate is negligible when it comes to quantitative research. Most safety culture assessments rely on employee surveys for quantitative data, with qualitative data from interviews, focus groups, and behavioral observations used to supplement the quantitative information, whereas assessments of safety climate typically use employee surveys as the sole source of data. As a result, the bulk of published studies exploring safety culture quantitatively are comparable to studies of safety climate because they both use data derived from employee surveys. When safety culture is measured exclusively using a survey to capture employee perceptions and attitudes towards their organization's emphasis on safety, then the construct being measured may be more aptly labeled safety climate. However, to avoid confusion in this paper, we use the term safety culture to refer to the entire body of literature devoted to safety culture and safety climate research.

2.3. Definitions and Dimensions of Safety Culture

One of the common criticisms of safety culture research is that the concept of culture is so abstract and all-encompassing that it loses practical value. The prevalence of research seeking to understand the causes and consequences of safety culture stems from a desire to do more than just describe the culture, but to also influence and change the culture (Guldenmund, 2000). As a result, researchers have devoted a significant amount of time to debating the definition and characteristics that make up an organization's safety culture. Indeed, one of the goals of the workshop held in February 2010 was for "panelists representing a broad range of stakeholders to reach alignment, using common terminology, on a definition of safety culture and a high-level set of traits that describe areas important to a positive safety culture" (NRC, 2011). Arriving at an agreed-upon definition is critical to the advancement of safety culture research because the definition of a construct sets the boundaries for what it is and what it is not, and provides focus for ensuing research and intervention.

Although many different definitions of safety culture exist, most share several commonalities (Wiegmann et al., 2002):

- Safety culture is *shared* by groups of people

- Safety culture is relatively *stable* and enduring within an organization
- Safety culture is comprised of *multiple dimensions*
- Safety culture *shapes behaviors* of members of the organization

In addition, reviews of the various dimensions that comprise the concept of safety culture note that, although different research teams use different terminologies, there is a small group of dimensions that consistently emerge. Wiegmann and colleagues (2002) identify at least five common dimensions:

- *Organizational commitment*: Upper management's values and actions demonstrate an enduring commitment to safety.
- *Management involvement*: Upper and middle-level management demonstrate commitment to safety in day-to-day operations, including their communication with employees about safety issues.
- *Employee empowerment*: Employees in the organization hold themselves and others accountable for safety, have a substantial voice in safety decisions, and feel empowered to express a questioning attitude.
- *Reward systems*: The reward system is structured so that it is perceived as fair and transparent by promoting safe behavior and discouraging or correcting unsafe behavior.
- *Reporting systems*: An effective and systematic reporting system exists that identifies vulnerabilities before serious events occur, enables the organization to proactively learn from past experience, and ensures that employees will not experience reprisals or negative outcomes as a result of using the system.

2.4. Relationships between Safety Culture and Safety Performance

A series of meta-analytic studies published between 2006 and 2010 significantly advanced the state of safety culture research by providing comprehensive analyses of past safety culture studies (Christian et al., 2009; Clarke, 2006; Beus et al., 2010). Meta-analysis is a technique that statistically combines the results of several studies to test shared research hypotheses. This technique allows researchers to develop more accurate and credible conclusions than can be arrived at from a single primary study or a narrative review of research studies (Rosenthal & DiMatteo, 2001). The safety culture meta-analytic studies sought to test the hypothesis that safety culture is related to safety performance by analyzing the collective results from past studies that measured aspects of safety culture and safety performance.

The studies included in the meta-analyses measured safety culture using surveys where employees were asked various questions regarding their perceptions of the extent to which their organization valued safety. The use of self-report surveys to assess safety culture is quite common, primarily because surveys are "relatively easy to use and inexpensive and often are the most plausible alternative for measuring unobservable constructs such as the attitudes of organizational participants (e.g., job satisfaction), individuals' values and preferences, their intentions (e.g., to quit their jobs), and their personalities (e.g., needs and traits)" (Ganster et al., 1983, p. 321). Although these studies did not use the same questionnaire, most had similar questions intending to assess employees' perceptions of management's commitment to safety, supervisor support for safety, safety communication, safety policies, and pressure to work safely.

Safety performance is used as an umbrella term to refer to the various types of safety outcomes that have been used as dependent variables in safety culture studies, ranging from observed or self-reported employee safety behaviors (e.g., following procedures, wearing personal protective equipment, participating in safety meetings) to organization-level safety outcomes like accident and injury rates. The studies hypothesized that there should be a relationship between measures of safety culture and other indicators of safety performance.

A common statistic used to evaluate whether there is a relationship between two measures is a correlation coefficient (Myers & Well, 2003). For example, correlations can be used to assess whether people who work longer hours have more car accidents after work than people who work fewer hours. Correlations do not imply that one thing causes the other, but just that there is an association between two variables. A correlation coefficient also indicates whether the relationship is positive or negative and how strong the relationship is on a scale ranging from -1.0 to +1.0. The closer the value of the correlation is to either -1.0 or +1.0, the stronger the association between the two variables, and values at 0 indicate no relationship. A negative correlation indicates that as one variable increases, the other decreases, whereas a positive correlation indicates that both variables increase together. A positive correlation of .80 using the example above would indicate a very strong relationship between work hours and car accidents, suggesting that people who work longer hours tend to have more car accidents after work. Correlation coefficients can be interpreted based on their magnitude, also called the effect size, and their statistical significance.

The magnitude of a correlation is also called its effect size because it indicates the strength of the effect between two measures. In social science research, correlations around .10 are considered small effects, correlations of .30 are considered medium effects and correlations of .50 and greater are large effects (Cohen, 1988). Effect sizes represent the magnitude of a relationship without making any statement about whether the apparent relationship in the data is statistically significant.

Statistical significance denotes the probability of obtaining the observed correlation assuming that no real effect exists (Myers & Well, 2003). That is, correlations between variables can occur simply by chance, so researchers require a means to assess how likely it is that an observed correlation is a “real” relationship, rather than an error. In the example above we would want to determine the probability of obtaining the correlation of .80 if there is no actual relationship between work hours and car accidents. The p-value of a correlation is a statistic that provides an estimate of how probable it is that the correlation has occurred by chance. Researchers pre-determine the threshold for acceptable p-values to assess whether observed correlations seem to reflect an actual relationship. These thresholds are called significance levels. The most common significance levels used in social science research are .05 or .01 (denoted as α), meaning that there is a 5% or 1% chance of obtaining the observed correlation when no real effect exists in the population. When the p-value of the correlation is equal to or lower than the pre-determined α , the result is said to be statistically significant. Therefore, a statistically significant correlation is one that we are confident did not occur by chance alone.

The results from the meta-analytic studies found consistent evidence of a statistically significant linear relationship between safety culture and accidents/injuries, ranging from a correlation of -.22 to -.39 ($p < .05$), and even larger statistically significant correlations between safety culture and employees' safety behaviors, ranging from .43 to .61 ($p < .05$). Using Cohen's labels, the relationship between safety culture and safety performance appears to be a medium effect, and the relationship

between safety culture and safety behaviors appears to be a large effect. Effect sizes can also be interpreted in terms of the percent of variance shared by two variables. In the case of correlation analyses, the square of the correlation coefficient represents the percent of shared variance. The results of the meta-analyses suggest that, overall, safety culture may account for 5-15% of the variance in an organization's accident and injury rates, and 18-37% of the variance in employees' safety behaviors.

The results of these studies also identified important gaps in the state of the safety culture literature. Most importantly, there are still only a small proportion of studies that have used prospective research designs to test for relationships between safety culture and safety performance. Clarke (2006) identified only six studies that used accident and injury data that were recorded after the administration of a safety culture survey, and the most recent meta-analysis by Beus and colleagues (2010) found only 11 studies that tested the relationship between safety culture and injuries recorded after the safety culture survey was administered. Prospective research designs are important because they help to establish a causal relationship between safety culture and safety performance. As stated previously, the ultimate value of the safety culture construct is as a potential leading indicator of safety performance. This hypothesis cannot be adequately tested unless safety culture is measured before safety performance and can be shown to predict future performance.

2.5. Safety Culture Interventions

Another approach to evaluating the relationship between safety culture and safety performance is through an intervention study. Interventions are actions taken to change an individual, group, or organization, frequently in response to an identified problem, deficiency, or need for improvement. Interventions can vary in intensity, duration, scope, focus, purpose, and method. In a safety culture intervention study, safety culture and safety performance are typically assessed before and after an intervention designed to change some aspect of an organization's safety culture. Improvements in the safety culture and safety performance at an organization following the intervention is said to demonstrate evidence that safety culture is related to safety performance because improvements in the organization's safety culture consequently led to improvements in the organization's safety performance. Of course, a caveat to the utility of safety culture interventions is that estimates of the failure rate of specific change efforts generally range from 50% to 75%. Those efforts aimed at overall cultural change fail at even higher rates (Hale et al., 2010).

One of the few published safety culture intervention studies supported by empirical evidence was conducted by Hale and colleagues (2010). This study reports the evaluation of 17 projects in 29 companies. Each of the projects employed a before-after study design with interventions aimed at changing the organization's safety culture and safety management systems with the goal of ultimately reducing accidents. Most organizations in the study implemented a range of changes directed at different levels and positions within the organization (ranging from directors to new and temporary workers). The specific interventions were chosen primarily based on a review of the organization's existing safety management system and were often framed as a way to fill identified gaps in the system.

Eight of the 17 companies demonstrated positive results on at least one safety performance measure, and three companies demonstrated positive results on multiple safety performance measures (e.g., decreased frequency and severity of injuries, fewer observed unsafe behaviors,

increased reporting of dangerous situations/near-miss events). The successful companies also demonstrated corresponding positive changes on measures of safety culture after the intervention. The characteristics of interventions that seemed to differentiate successful and unsuccessful companies were the number of interventions (i.e., successful companies employed almost twice as many independent interventions targeted at different areas or aspects of the company), the presence of a champion for the project to create enthusiasm and overcome resistance, and the support and participation of upper-level management. Also, interventions targeted at improving communications between the workforce and line management, and setting up a reporting system for dangerous situations or near miss events demonstrated the most consistent success.

Some of the intervention characteristics that did not discriminate between successful and unsuccessful companies included attempts to improve the company's safety management system, training and publications directed at line workers and line management, and changes in the organizational structure (e.g., change in top leadership, reorganization, mergers, and layoffs). These characteristics were present in both successful and unsuccessful interventions, and were not enough on their own to stimulate successful change. Safety culture training, reorganizing, or changing the safety management system often appear only to change safety culture on the surface of the organization (i.e., the artifact layer of the culture), and do not result in deep, lasting changes that ultimately transform an organization's core underlying assumptions.

3. Development of the Safety Culture Survey

In support of the NRC's efforts to develop a common definition and list of traits describing a positive safety culture, NEI and INPO volunteered to conduct a safety culture study by developing and administering a survey to employees at nuclear power plants across the United States. The study had several purposes:

- Confirm that the safety culture construct is multidimensional in the nuclear domain, as has been found in other organizational settings
- Identify the dimensions of the safety culture construct in U.S. nuclear power plants
- Determine the extent to which the safety culture factors derived from the study correlate with concurrent measures of safety performance
- Determine the extent to which the safety culture factors correlate with the same measures of safety performance one year after administration

3.1. Survey Item Development

The first step in investigating the construct validity of a new survey instrument begins by establishing the content validity of the survey items. Content validity is the degree to which the items in a survey cover the breadth of a construct and is the minimum psychometric requirement for establishing measurement adequacy (Schriesheim, 1993). Content validity should be built into the process of generating survey items through a thorough review of the research literature and a clear definition of the construct of interest. The initial survey was drafted using 73 items from a survey developed by the Utilities Service Alliance based on INPO's Principles for a Strong Nuclear Safety Culture (INPO, 2004). NRC staff reviewed the draft items, provided comments, and suggested an additional 37 items based on the IAEA's safety culture characteristics and attributes (IAEA, 2006), the NRC's cross-cutting components and aspects used in the ROP (NRC, 2011a), relevant items from published surveys in the safety culture research literature, and behaviors characterizing a positive safety culture developed by panelists at the February 2010 public workshop. Although many of the original 73 items covered concepts that were similar across the various safety culture publications, the additional items ensured that the survey effectively covered the breadth of the literature on safety culture in the nuclear and other domains. As a result, the items generated appear to demonstrate adequate content validity.

One of the drawbacks of the process used to develop the items for the survey is that the developers did not use a single, specific classification scheme to guide the development process. Instead, items from different classification schemes (i.e., INPO, IAEA, NRC, etc.) were mixed together and item groupings were not hypothesized before the survey was administered. For example, the survey developers could have maximized the likelihood of replicating the traits in the Safety Culture Policy Statement by specifically developing or selecting items that are most representative of the policy statement traits. However, the approach INPO used was an appropriate method for survey development, particularly because the study was exploratory and there is continued debate regarding the characteristics that make up the safety culture construct. Still, it is important to note that INPO's chosen approach relies heavily on exploratory factor analysis techniques, and the resulting factors will be largely based on statistical relationships between items rather than similar themes (Hinkin, 1998).

In addition to content validity, there are a number of guidelines and best practices used in the social sciences for writing survey items (Hinkin, 1998). The items should be as simple as possible, with particular attention devoted to ensuring that the terms used will be familiar to the target respondents. Items should address a single topic and avoid statements that mix two different subjects, like “safety *and* security,” or “management *and* employees ensure safety,” which are called double-barreled statements. Researchers with survey development experience and familiarity with the nuclear power industry from INPO and NRC reviewed the draft survey items to ensure writing clarity and provided editing suggestions to simplify items and ensure familiar terms were used. Although some double-barreled statements were retained as items, efforts were made to ensure that the subjects were not dissimilar, to the extent practical.

The final survey consisted of 110 items. The survey also included demographic questions asking respondents to indicate their plant, work group (e.g., operations, maintenance, engineering), and work status (permanent or contractor). Example items from the survey are listed below:

- “Our leadership frequently communicates the importance of nuclear safety”
- “Our corrective action program is effective”
- “Management acts decisively when a nuclear safety concern is raised”
- “Continuous learning is expected of everyone”
- “My supervisor discusses safety with me before I start work on a job”
- “Staffing levels are adequate to meet work demands”
- “Station management gives us clear direction”
- “People here are comfortable challenging each other, regardless of level, when they feel something is not correct”

Survey participants were asked to rate their degree of agreement with each statement using a 7-point Likert scale ranging from strongly disagree to strongly agree. Each item also included a “don’t know/no opportunity to observe” response option. The full rating scale is presented below:

- 1 – strongly disagree
- 2 – disagree
- 3 – somewhat disagree
- 4 – neither agree nor disagree
- 5 – somewhat agree
- 6 – agree
- 7 – strongly agree
- 0 – don’t know/no opportunity to observe

This type of rating scale is one of the most commonly used in social science survey research (Trochim, 2006) and is characterized by symmetrical response options along a bipolar scale from positive to negative with a neutral point in the middle. The scoring system from 1-7 enables the researcher to ascribe a quantitative value to the respondent’s qualitative assessment. The data are traditionally used as an approximation of an interval scale, meaning that the response options are inferred to be relatively equidistant from each other on a continuum. Responses to each item are then summed or averaged for each respondent and treated as interval data measuring a latent variable, thereby allowing for the use of parametric statistical tests, like Pearson correlations, regression, and

analysis of variance. The “don’t know” option is typically coded as missing data so as not to skew the average.

3.2. Survey Administration

A vendor for NEI obtained lists of personnel, which included long-term contractors, from each operating nuclear power plant in the U.S., and randomly selected approximately 100 individuals from each site to participate in the survey. The survey was administered through the internet, and e-mail invitations with a link to the survey were sent to the selected participants at each site. Senior management at each site was contacted to request that the survey be announced and participation encouraged. The survey administrator sent an average of two reminder e-mails to people who did not respond to the first invitation, until a minimum of 30 respondents from each site was obtained. Data collection began on June 14 and ended on August 11, 2010.

The total number of people invited to complete the survey over the administration period was 6,333. Of those, 3,031 individuals responded to the invitation for a 48% response rate. 2,876 respondents provided valid answers to the majority of the 110 statements and their responses were retained for subsequent data analysis. The average number of respondents per site was 46. Sixty-three sites were in the sample, or 97% of the operating nuclear power plants in the US. Two sites were not included in the study. One site was participating in another INPO survey in preparation for an organizational effectiveness review and requested not to participate in this survey. Another site was inadvertently omitted from the survey because of an administrative error (no invitations sent).

Response rates for internet-based surveys typically average around 30-40% of the sample population (Cook et al., 2000; Baruch & Holtom, 2008), particularly voluntary surveys administered within organizations for research purposes. The response rate for this survey is consistent with what might be expected in exploratory survey research. The administrators of the survey also demonstrated good practice by ensuring that the organizations under study endorsed the survey and that participation reminders were sent, likely increasing the final tally of respondents.

More important than the overall response rate of the survey is the sample’s representativeness of the population under study (Krosnick, 1999). One of the goals of INPO’s study was to characterize the factors that make up safety culture in the nuclear power industry. Obtaining a sample that represents a cross-section of the industry is therefore of primary importance. At least 30 employees from all but 2 nuclear power plants in the US provided usable data from the survey. The Central Limit Theorem posits that as the size of a random sample increases the sample mean approaches the population mean (Myers & Well, 2003). When the sample size is greater than 30 it tends to approximate a normal distribution and there are negligible differences between the sample mean and population mean. Thus, a sample size of 30 is commonly used as a minimum threshold to establish that the sample is a good estimate for the population. In addition, the demographic information collected from respondents indicates a good distribution of respondents from different workgroups within a nuclear power plant. For example, approximately 16% of respondents worked in operations, 17% in maintenance, 10% in security, 5% in systems engineering, 6% in training, 6% in radiation protection, 3% in chemistry, and 7% were contractors.

4. Analysis of Safety Culture Survey

4.1. Exploratory Factor Analysis

As stated previously, the first two goals of INPO's safety culture study were to confirm that the safety culture construct is multidimensional in the nuclear domain and to then identify those dimensions. INPO's approach to assessing the dimensionality of safety culture, based on the items included in the safety culture survey, was to perform a type of exploratory factor analysis called principal components analysis (PCA; Jolliffe, 2002). PCA is a variable reduction technique that attempts to reduce a large set of highly correlated variables, in this case the participants' responses on the 110 safety culture items, to a smaller number of uncorrelated variables called principal components. PCA is used to reduce "noise" and redundancy in the data, and aids in clarifying the relationships among the items in the data. The defining feature of PCA is that it attempts to account for all of the variance in the data, such that the first principal component accounts for as much of the variability in the data as possible, and each subsequent component accounts for as much remaining variability as possible.

PCA is performed by computing the correlations between all of the items and sorting the items into factors such that 1) the items within each factor have the highest correlations with each other, 2) each factor accounts for as much variance in the data as possible, and 3) the factors are maximally distinct from each other. PCA attempts to account for all of the variance in the data by creating as many factors as there are data points; in this case the 110 survey items. However, the goal of PCA is to reduce the items into a smaller number of interpretable factors, so only those factors that account for a significant amount of variance and consist of items that appear to represent an interpretable theme in the data are retained in the final factor solution. In practice, a statistical software package, such as the Statistical Package for the Social Sciences (SPSS), is used to perform these complex calculations by entering the raw numerical data into a spreadsheet and selecting the PCA function under the data reduction techniques menu. The SPSS software then prints the output of the analysis with a list of the 110 items, the computed principal components or "factors," and the correlations between each item and component, also called the factor loadings.

INPO first performed a PCA using all 110 items in the survey to identify which factors would emerge. The initial PCA resulted in 9 interpretable factors. INPO researchers developed labels for each of the nine factors based on the survey items that demonstrated high factor loadings for each factor. The items with factor loadings of .40 or greater and no major cross-loadings on other factors are judged as meaningful and representative of the construct under examination (Ford et al., 1986). Four items did not have high factor loadings on any of the nine factors and were therefore removed from subsequent analyses. Overall, the nine factors accounted for 58% of the variance in the data. Table 1 displays the results of the PCA, including the factor label, percent of variance in the survey responses accounted for by the factor, and number of items that demonstrated high loadings on each factor (i.e., greater than .40).

Table 1 Results of PCA of 110 Items with a 9 Factor Solution

Factor Label	% Variance Accounted For	# Items
1. Management Responsibility/Commitment to Safety	15.7%	36
2. Willingness to Raise Concerns	6.9%	9
3. Decision-Making	6.3%	10
4. Supervisor Responsibility for Safety	6.2%	11
5. Questioning Attitude	5.9%	9
6. Safety Communication	5.3%	13
7. Personal Responsibility for Safety	4.8%	6
8. Prioritizing Safety	4.0%	6
9. Training Quality	2.7%	6

Several of the factors that emerged from the PCA consisted of such a large number of survey items that it was deemed worthwhile to perform PCAs on those subsets of items to further reduce the factors into sub-factors. The items included in Factors 1 (Management Responsibility/Commitment to Safety), 2 (Willingness to Raise Concerns), 4 (Supervisor Responsibility for Safety), and 5 (Questioning Attitude) were subject to additional, separate PCAs and evidenced interpretable sub-factors. Although additional PCAs were performed with Factor 3 (Decision-Making) and Factor 6 (Safety Communication), the results did not produce any clearly interpretable sub-factors. The full list of factors, sub-factors, and example items are presented in Table 2. Note that when sub-factors are listed, the example items are presented at the sub-factor level.

Table 2 Safety Culture Factors, Sub-Factors, and Example Items

Factor/Sub-Factor Label	Example Item
1. Management Responsibility/Commitment to Safety	
a. Respectful Work Environment	People are treated with dignity and respect by station leadership
b. Continuous Improvement	At this station, we correct problems the first time they appear
c. Performance Indicators	Our performance indicators and trending programs help us to detect problems early
d. Procedure Communication	The procedures at this site are generally up-to-date and easy to use
e. Resources	Staffing levels are adequate to meet work demands
f. Rewards	At this station, people are routinely rewarded for identifying and reporting nuclear safety issues
2. Willingness to Raise Concerns	
a. Informally	When I make a mistake, I'm not afraid to report it to my supervisor
b. Formally	I have confidence in our Employee Concerns Program
3. Decision-Making	Decision-making at this site reflects a conservative approach to nuclear safety

4. Supervisor Responsibility for Safety	
a. Communication	My supervisor is usually available when I have a question or problem
b. Presence	My supervisor periodically observes me working
c. Coaching	My supervisor gives me useful feedback about how to improve my performance
5. Questioning Attitude	
a. Situation/Problem Awareness	Personnel promptly identify and report conditions that can affect nuclear safety
b. Procedure Use	Workers at this station usually follow procedures
c. Plant Knowledge	In general, employees have a basic knowledge of plant fundamentals
6. Safety Communication	There is good communication about nuclear safety issues that affect my job
7. Personal Responsibility for Safety	It is my responsibility to raise nuclear safety concerns
8. Prioritizing Safety	At this station, nuclear safety takes priority over production goals
9. Training Quality	Training at this site provides me with the knowledge I need to perform my job

One of the objectives of any survey development effort is to maximize content validity while maintaining parsimony and simple structure (Hinkin, 1998). Following the exploratory factor analysis, INPO conducted a review of the 110 survey items within the context of the 9 factors identified by the PCA to assess whether items could be removed from the survey. For instance, cases where many respondents did not respond to an item (i.e., missing data), or responded by choosing the “don’t know” option may indicate that respondents were confused by the item or did not find it applicable to their work. Items that have low correlations with all of the other items in the survey, with the rule of thumb being inter-item correlations less than .40 (Kim & Mueller, 1978), may not be good representations of the construct of interest (i.e., safety culture). Extremely similar items that group together under the same factor may also be unnecessarily repetitive. This review resulted in the elimination of 50 items from the survey, bringing the total number of items to 60.

We conducted an independent PCA of the 60 items that were retained in the survey and found that seven of the original nine factors emerged as distinct factors. The items comprising the Safety Communication, Prioritizing Safety, and Decision-Making factors demonstrated high loadings on a single factor, rather than loading on separate factors. These results indicate that the Safety Communication and Prioritizing Safety factors are not as stable as the other factors and are very similar, statistically, to the Decision-Making factor.

Another type of exploratory factor analysis common in social science research is principal axis factoring (PAF). When using PCA, the researcher is making the assumption that all variability in an item should be used in the analysis, whereas PAF only uses the variability that an item has in common with other items, and as such is more sensitive to the existence of error variance. In practice, use of these methods with large samples and large pools of items typically yield very similar results. However, PAF is the preferred method when the goal of the analysis is to detect the structure of some underlying latent construct (Ford et al., 1986), which characterizes the purpose of the current effort

to identify the factors that make up the underlying construct of safety culture. As such, it was important to conduct a sensitivity analysis of the results of the PCA conducted by INPO using PAF as an alternative exploratory factor analysis method. We conducted a series of independent PAFs using all 110 items and various combinations of items by deleting groups of items with low inter-item correlations, excessive missing data, low factors loadings, and items that appeared to be associated with more than one factor. In all cases, the results of the PAF yielded 7 to 9 interpretable factors, and the individual items demonstrated very similar factor loadings as compared to the results of the PCA analyses conducted by INPO. The one factor that did not consistently emerge within the PAF analyses was Safety Communication. The items in this factor tended to group together with items from the Decision-Making or Management Responsibility/Commitment to Safety factors.

The strongest factor resulting from the series of factor analyses was the one labeled Management Responsibility/Commitment to Safety; it consistently accounted for the most variance in the data (15.7%) and the items within this factor demonstrated consistently high factor loadings. The weakest, or least stable, factor to emerge from the analysis was the one labeled Safety Communication. This factor accounted for relatively less variance in the data (5.7%) and items in this factor tended to load on other factors, like Decision-Making or Management Responsibility/Commitment to Safety, when the analyses were run with different combinations of items. Caution should be used when interpreting this factor as it is possible that future administrations of this survey could reveal a different factor structure that does not include Safety Communication.

Overall, the results of the PAF analyses confirm that safety culture is a multidimensional construct, and the first five factors (i.e., Management Responsibility/Commitment to Safety, Willingness to Raise Concerns, Decision-Making, Supervisor Responsibility for Safety, and Questioning Attitude) appear to be the most stable and account for the most variance in the data. However, retaining less stable factors may be justified if they demonstrate adequate reliability (discussed in section 4.2). There also remains the possibility that the less stable factors are uniquely related to safety performance and therefore worthwhile to keep as a distinct factors.

4.2. Reliability Analysis

Another prerequisite for establishing construct validity is demonstrating that a measure is reliable. Reliability refers to the extent to which measurements are repeatable (Nunnally & Bernstein, 1994). One form of reliability is internal consistency, which establishes whether items in a created factor are consistently measuring the same underlying construct. For example, if a respondent expresses agreement with items in a measure such as, "I like running," and, "I have enjoyed running in the past," and disagreement with the item "I dislike running," then a factor created from these items would be said to demonstrate good internal consistency. The most widely used statistical test of internal consistency is Cronbach's coefficient alpha (Cronbach, 1951). Cronbach's alpha is essentially the mean of all possible pair-wise correlations for a set of items. Cronbach's alpha can be used as a confirmatory measure in factor analysis because it measures the strength or precision of a factor once the items in that factor have been identified (Cortina, 1993). Values can range from 0 to 1.00, with higher values indicating better reliability. The minimum criterion for acceptable reliability is considered a value greater or equal to .70 (Nunnally & Bernstein, 1994).

We calculated Cronbach's coefficient alpha for each of the factors and sub-factors identified by INPO using the reduced 60 item survey. An alpha value was also calculated using all 60 items to characterize a single meta-factor representing the safety culture construct as a whole. An alpha value greater or equal to .70 provides confirmation that the factors and sub-factors demonstrate good internal consistency and can be treated as distinct factors within the construct of safety culture. Cronbach's alpha was not calculated for the sub-factors of performance indicators, coaching, and plant knowledge because each of those factors is represented by only one item in the survey. Cronbach's alpha measures consistency in responses among items, and therefore cannot be calculated for single-item factors. The results of the reliability assessment are presented in Table 3.

Table 3 Results of Reliability Assessment using Cronbach's Coefficient Alpha

Meta-factor/Factor/Sub-factor Label	Cronbach's α	# Items
SAFETY CULTURE	0.98	60
1. Management Responsibility/ Commitment to Safety	0.96	20
a. Respectful Work Environment	0.92	7
b. Continuous Improvement	0.89	5
c. Performance Indicators	--	1
d. Procedure Communication	0.64	2
e. Resources	0.72	3
f. Rewards	0.85	2
2. Willingness to Raise Concerns	0.90	6
a. Informally	0.84	3
b. Formally	0.83	3
3. Decision-Making	0.88	5
4. Supervisor Responsibility for Safety	0.88	6
a. Communication	0.86	3
b. Presence	0.69	2
c. Coaching	--	1
5. Questioning Attitude	0.85	6
a. Situation/Problem Awareness	0.81	3
b. Procedure Use	0.63	2
c. Plant Knowledge	--	1
6. Safety Communication	0.87	7
7. Personal Responsibility for Safety	0.77	3
8. Prioritizing Safety	0.83	4
9. Training Quality	0.78	3

The safety culture meta-factor and 9 factors all demonstrate good internal consistency with alpha values greater than .70. Although the results of the exploratory factor analyses suggested that the Safety Communication, Personal Responsibility for Safety, Prioritizing Safety, and Training Quality factors were not as strong when using the reduced 60 item survey, these factors do demonstrate acceptable internal consistency to be used as distinct factors in subsequent data analyses. All but three of the sub-factors also demonstrate adequate internal consistency. The sub-factors labeled

procedure communication, supervisor presence, and procedure use had alpha values below .70. However, each of these sub-factors was represented by only 2 items in the survey, and measures of internal consistency are sensitive to the number of items included in a factor. Because the sub-factors have fewer items than the factors they are likely to show less internal consistency. Sub-factors with fewer than three items should therefore be used with caution. It is more conservative to perform analyses with inferential statistics, like correlations, at the factor level to avoid drawing conclusions based on spurious relationships with unreliable sub-factors.

4.3. Consistency of Survey Factors with Other Research

The exploratory factor analysis results provide support for the contention that safety culture is a multidimensional construct in the nuclear power domain. The most common factor included in various studies of safety culture relates to management's commitment to safety, appearing in approximately 70-75% of the published measures of safety culture (Flin et al., 2000; Guldenmund, 2000). The fact that Management Responsibility/Commitment to Safety emerged as the factor accounting for the most variance in this study is consistent with previous research. Flin and colleagues (2000) also note that it is not uncommon to see a separate factor for supervisor commitment to safety, particularly because first-line supervisors have more direct interaction with line employees and therefore are more likely to influence the immediate work atmosphere and extent to which safety is emphasized in day-to-day operations.

The factors emerging from the survey also cover the breadth of themes identified by Wiegmann and colleagues (2002) as aspects of safety culture. The organizational commitment theme is reflected in the INPO factors labeled Management Responsibility/Commitment to Safety, Prioritizing Safety, and Decision-Making; the management involvement theme is incorporated in the Supervisor Responsibility for Safety and Safety Communication factors; the employee empowerment theme is reflected in the Personal Responsibility for Safety and Questioning Attitude factors; and aspects of the reward systems and reporting systems themes are captured in the Willingness to Raise Concerns factor and the sub-factors of Respectful Work Environment and Rewards. Although the Training Quality factor does not directly relate to any of the themes identified by Wiegmann et al., it is consistent with the concept of competence, which has been mentioned as an important theme in other safety culture reviews (Flin et al., 2000; Guldenmund, 2007).

4.4. Within-Group Reliability Analysis

A key underlying premise of safety culture is that it is shared among members of an organization. Consequently, assessments of safety culture should seek to reveal underlying beliefs and assumptions that are generally held by all members of the organization. The concept of "relatedness" can be determined by assessing within-group reliability, or the degree to which respondents at the same site had similar responses to items on the safety culture survey. Within-group reliability is necessary to justify aggregating survey data from individuals to the site level and provides support for generalizing data from a sample of employees at a site to the entire site. One method for determining within-group reliability is by using intraclass correlations (ICCs; McGraw & Wong, 1996).

Two types of ICCs are relevant for determining within-group reliability: ICC(1) measures reliability among individuals in a group, and ICC(2) measures the reliability of the group mean (James, 1982). The first, referred to as ICC(1), is similar in interpretation to a traditional correlation and describes the extent to which members within a group had the same responses, in terms of

agreement/disagreement, to the items in the safety culture survey. Some variability in individual responses is expected due to individual differences (e.g., different experiences with supervisors or management), sub-group differences within an organization (e.g., differences in safety culture between operations and maintenance), and measurement variation (e.g., some individuals are more likely to use the ends of the response scale, strongly agree/disagree, whereas others use options toward the middle of the scale, somewhat agree/disagree). Consequently, ICC(1) values are less likely to be close to 1.0 and the criterion for determining adequate reliability is whether the value is statistically significant. The measurement of ICC(2) is very similar in concept to measuring Cronbach's alpha for internal consistency, and uses a cutoff of .70 for acceptability (McGraw & Wong, 1996). However, instead of measuring consistency among items in a factor, the ICC(2) statistic indicates consistency in responses among members in a group and thus the stability of the mean score for the group. The range of ICC(1) and ICC(2) values for the 63 nuclear power stations included in the sample are presented for safety culture overall and the nine safety culture factors (Table 4).

Table 4 Range of ICC(1) and ICC(2) Values for Safety Culture and Safety Culture Factors

	ICC(1)		ICC(2)	
	Range	Sig.	Range	Sig.
SAFETY CULTURE	0.26 - 0.58	p < .01	0.95 - 0.99	p < .01
1. Management Responsibility/ Commitment to Safety	0.33 - 0.63	p < .01	0.91 - 0.97	p < .01
2. Willingness to Raise Concerns	0.38 - 0.81	p < .01	0.79 - 0.96	p < .01
3. Decision-Making	0.32 - 0.79	p < .01	0.70 - 0.95	p < .01
4. Supervisor Responsibility for Safety	0.31 - 0.72	p < .01	0.73 - 0.94	p < .01
5. Questioning Attitude	0.25 - 0.66	p < .01	0.66 - 0.92	p < .01
6. Safety Communication	0.24 - 0.65	p < .01	0.69 - 0.93	p < .01
7. Personal Responsibility for Safety	0.28 - 0.88	p < .01	0.54 - 0.96	p < .01
8. Prioritizing Safety	0.30 - 0.76	p < .01	0.63 - 0.93	p < .01
9. Training Quality	0.24 - 0.73	p < .01	0.49 - 0.89	p < .01

The ICC(1) values were statistically significant for safety culture overall and the nine factors, indicating that at least some portion of the variation in individual responses is shared at the site level. The ICC(2) values for safety culture overall and the first four factors exceeded .70 at all sites, indicating that the group mean scores on each of those factors are relatively stable. All but two sites had ICC(2) values above .70 on factor 5 (Questioning Attitude), and all but one site had ICC(2) values above .70 on factor 6 (Safety Communication). Factors 7 (Personal Responsibility for Safety), 8 (Prioritizing Safety), and 9 (Training Quality) demonstrated less stable group means, and again should be interpreted with caution. Altogether, these measures of within-group reliability provide enough evidence to justify conceptualizing safety culture as an organizational-level construct and aggregating the individual-level survey data to derive a mean score for each site. Aggregating data to the site level is a necessary step to allow for comparisons between the survey data and safety performance metrics, which are traditionally collected by the NRC at the unit or site level. The

analyses discussed in the remainder of the paper use the mean score on the factors and safety culture survey as a whole at each site to represent the station’s safety culture.

4.5. Descriptive Analysis of Safety Culture Factors

We conducted a descriptive analysis of the safety culture survey factors to aid in the interpretation of the results of the inferential statistical analyses presented in section 5 of this paper. As might be expected, the mean value for safety culture overall and the means for the safety culture factors are higher than the mid-point of the response scale (4), suggesting that respondents generally viewed their organizations as having positive safety cultures. Most of the means fall between a value of 5 and 6 on the 7-point scale, which correspond to the response options “somewhat agree” (5) and “agree” (6).

Factor 7, Personal Responsibility for Safety, had a substantially higher mean score than the other safety culture factors at 6.54, and a relatively small standard deviation at 0.52. This suggests that most respondents indicated strong agreement with the items comprising Factor 7, and there was very little variability in those responses. The survey items that make up Factor 7 generally refer to personal attitudes toward safety (e.g., one item is “I understand that I am personally responsible for the behaviors and work practices that support nuclear safety”), rather than perceptions of the organization’s practices. Items in Factor 7 seem to indicate that employees understand that they *should* behave safely, but not whether they engage in safe behaviors on a regular basis, or whether their organization supports and rewards safe behaviors. Table 5 provides the means, standard deviations, and minimum and maximum values for the overall measure of safety culture and each of the safety culture factors.

Table 5 Descriptive Statistics for Safety Culture and Safety Culture Factors

	Mean	Standard Deviation	Minimum	Maximum
SAFETY CULTURE	5.61	0.79	1.72	7.00
1. Management Responsibility/ Commitment to Safety	5.07	1.02	1.30	7.00
2. Willingness to Raise Concerns	5.81	1.02	1.00	7.00
3. Decision-Making	5.98	0.83	1.80	7.00
4. Supervisor Responsibility for Safety	5.63	1.01	1.00	7.00
5. Questioning Attitude	5.93	0.70	2.17	7.00
6. Safety Communication	5.78	0.78	1.14	7.00
7. Personal Responsibility for Safety	6.54	0.52	1.00	7.00
8. Prioritizing Safety	5.98	0.86	1.00	7.00
9. Training Quality	5.75	0.95	1.00	7.00

The correlations among safety culture factors can also provide information to complement the reliability analysis. The various safety culture factors should demonstrate high intercorrelations if they represent different facets of a single underlying construct (i.e., safety culture). However, the factors should not be so highly correlated that they are essentially measuring the exact same thing, as

it would suggest that the safety culture construct is not multidimensional. A general rule of thumb would be to look for intercorrelations that range from .70 to .95 to suggest that the safety culture factors are both related and unique, although these values are not used as rigid cut-off scores. Table 6 presents the intercorrelations among the safety culture factors.

Table 6 Intercorrelations among Safety Culture Overall and Safety Culture Factors

	SAFETY CULTURE	1 Mgmt.	2 Concern	3 Decision	4 Sup.	5 Quest.	6 Comm.	7 Personal	8 Prioritize	9 Training
SAFETY CULTURE	1									
1. Management Responsibility/Commitment to Safety	.98**	1								
2. Willingness to Raise Concerns	.87**	.81**	1							
3. Decision-Making	.96**	.93**	.85**	1						
4. Supervisor Responsibility for Safety	.86**	.81**	.71**	.81**	1					
5. Questioning Attitude	.87**	.83**	.76**	.80**	.69**	1				
6. Safety Communication	.95**	.89**	.83**	.89**	.82**	.85**	1			
7. Personal Responsibility for Safety	.37**	.26*	.45**	.37**	.28*	.37**	.46**	1		
8. Prioritizing Safety	.90**	.86**	.79**	.89**	.75**	.77**	.85**	.28*	1	
9. Training Quality	.80**	.76**	.64**	.75**	.66**	.70**	.80**	.48**	.72**	1

*p < .05; **p < .01

The safety culture factors demonstrated high correlations with the overall measure of safety culture. Factors 1 (Management Responsibility/Commitment to Safety), 3 (Decision-Making), and 6 (Safety Communication) had the highest correlations with the overall safety culture score, suggesting that the scores on those factors are most consistent with the overall mean score for safety culture. Factor 7 (Personal Responsibility for Safety) is concerning because of the considerably lower correlations between it and the other safety culture factors. Recall that Factor 7 did not demonstrate high internal consistency or a stable factor structure as compared to the other factors. Factor 7 also had a substantially higher mean and lower standard deviation than the other factors. Respondents did not seem to respond to items on this factor in the same way that they responded to other items in the survey. Again, caution should be used when interpreting this factor and consideration should be given as to whether this factor, as measured in the survey, is an accurate representation of one of the facets of safety culture. Note that some of the areas of reasonable uniqueness among the factors are between Factor 2 (Willingness to Raise Concerns) and Factor 9 (Training Quality) with a correlation of .64, and between Factor 4 (Supervisor Safety) and Factor 5 (Questioning Attitude) with a correlation of .69.

5. Criterion-Related Validity of the Safety Culture Survey

Next, we sought to evaluate the criterion-related validity of the survey by testing whether the theoretical relationship between safety culture and safety performance is supported by statistical evidence. Some of the metrics the NRC uses to assess the performance of nuclear power plants may be related to aspects of performance that should be influenced by safety culture. In theory, an organization with a weak safety culture will demonstrate behaviors that increase the risk of a negative event. The organization's riskier behaviors may translate into declines in safety performance over time, which can then be detected by performance metrics. Therefore, measures of safety culture (i.e., the safety culture survey) should be statistically related to measures of safety performance (i.e., relevant NRC performance metrics).

However, not all variations in safety performance are likely to be attributable to an organization's safety culture. Natural disasters, "unforced" human error, preventative maintenance, and even safety systems that are functioning properly (e.g., a safety system actuation during a legitimate transient that mitigates a negative event) can adversely affect a plant's performance metrics. It is thus expected that statistical tests of the relationship between safety culture and safety performance will produce, at best, moderate effect sizes. Past research suggests that the overall correlation between safety culture and safety outcomes (e.g., accidents and injuries) is between -.22 to -.39 (Christian et al., 2009; Clarke, 2006; Beus et al., 2010), which is consistent with Cohen's (1988) classification of .3 as a medium effect size. Safety performance variables that are more proximally related to an organization's safety culture should evidence larger effect sizes, whereas variables that are more distally related to safety culture should evidence smaller effect sizes.

The two forms of criterion-related validity of interest in this study are concurrent validity and predictive validity. Concurrent validity uses data from similar points in time to establish whether a measure is statistically related to some criterion. Predictive validity assesses a measure's ability to predict some criterion at a future point in time, and therefore uses data where the criterion (safety performance) is measured after the measurement of the construct of interest (safety culture). Because the safety culture survey was administered in the middle of 2010, performance data from the 2010 calendar year were used to evaluate concurrent validity, and performance data from the 2011 calendar year were used to evaluate predictive validity.

5.1. Description of NRC Performance Metrics

The NRC collects and monitors many different types of data related to the performance of operating nuclear power plants. The staff regularly evaluates these data, primarily through the ROP, to identify indications of degrading performance and determine appropriate regulatory responses. Although there is a wide variety of data available, we determined that some of the data types were more amenable for use in the statistical analyses required for this study than others and could be transformed into variables with reasonably clear conceptual links to both safety performance and safety culture. In this section, we describe the variables used in this study, which were based on performance indicators and inspection reports associated with the ROP, performance indicators from the NRC's Industry Trends Program (ITP), and allegation reports maintained by OE.

The ROP is the NRC's regulatory framework for overseeing the safe operation of commercial nuclear power plants. The ROP is designed to focus on those plant activities that are most important to safety, and uses inspection findings and performance indicators for on-going monitoring of each

plant's performance. The ROP is divided into seven cornerstones representing the essential areas of safe operations: initiating events (IE), mitigating systems (MS), barrier integrity (BI), emergency preparedness (EP), occupational radiation safety (OR), public radiation safety (PR), and physical protection (PP). Performance indicators and findings from risk-informed inspections are intended to provide a broad sample of data to assess licensee performance in the risk-significant areas of each cornerstone, although they are not intended to provide complete coverage of every aspect of plant design and operation (IMC-0308, Attachment 1; NRC, 2007).

The NRC's Inspection Manual chapter 0313 states that one of the objectives of the ITP is to "collect and monitor industry-wide data that can be used to assess whether the nuclear industry is maintaining the safety performance of operating plants" (NRC, 2008). The ITP provides insights about plant performance using information derived from Licensee Event Reports (LERs) and plant Monthly Operating Reports (MORs). The data are consolidated annually by Idaho National Laboratory to create indicators representing aspects of a site's performance.

In high reliability industries like nuclear power, accidents are extremely rare occurrences. As a result, less significant events that occur more frequently are relied upon as indicators of potentially degrading performance. These more frequent events are also more conducive to quantitative data analysis because there are more data points and more variability across organizations. We selected four performance indicators to include in this exploratory analysis. The unplanned scrams variable used in the analysis is based on one of the inputs to the ROP performance indicator program, and the safety systems actuations, force outage hours, and equipment outage variables are based on inputs to the ITP. We selected these indicators because 1) the events they measure seem to occur more frequently than events measured by other NRC performance indicators and 2) they may reflect broader, organization-level processes that are less affected by the functioning of single work groups or departments at a site than some others. Although there are many reasons a plant could have higher counts of the selected performance indicators, some of those reasons could theoretically relate to safety culture. For instance, it is possible that greater numbers of unplanned scrams could be attributable to an emerging pattern of decisions that emphasize productivity over safety or other systematic weaknesses in the organization's safety culture that could be improved. Descriptions of the variables are provided below (with affected cornerstone in parentheses):

- **Unplanned Scrams (IE):** The number of unplanned scrams at a site per year, both manual and automatic, while critical.
- **Safety System Actuations (MS):** The number of safety system actuations at a site per year. Safety system actuations are manual or automatic actuations of the logic or equipment of the Emergency Core Cooling System (ECCS) or Emergency AC Power System.
- **Forced Outage Hours (MS):** The total number of hours a licensee was in a forced outage state per year.
- **Equipment Outages (MS):** The number of forced outages the licensee classifies as equipment-related per year.

In addition to performance indicator data reported by licensees, licensee performance is also evaluated through inspections. The staff documents findings in reports from inspections conducted at each plant throughout the year. Inspection findings are evaluated and given a color designation

based on their risk significance. Green inspection findings indicate a deficiency in licensee performance that has very low safety significance. Licensees are permitted to correct these performance issues without increased regulatory oversight. White, Yellow, or Red inspection findings represent a greater degree of safety significance and therefore trigger increased regulatory attention. However, these “greater than green” findings are rare, compared to the frequency with which inspectors identify green findings. Although green inspection findings are of very low safety significance, they are still indicative of some performance deficiency, and it is possible that small problems could lead to more safety-significant problems in the future if not adequately addressed. Therefore, for the purposes of this study, all inspection findings regardless of color designation were counted in the analysis as one representation of performance to ensure sufficient data were available. The following variable related to ROP inspection findings was used in the analysis:

- **Total ROP Inspection Findings:** The total number of findings reported within ROP inspection reports per year.

The ROP Action Matrix Summary is a high-level indicator reflecting overall plant performance. Each quarter, every plant is assigned a status in the ROP action matrix based on the most recent performance indicators and inspection findings. A plant’s status in the action matrix determines the level of NRC oversight of the plant, including supplemental inspections and pertinent regulatory actions ranging from management meetings up to and including orders for plant shutdown. The columns in the ROP action matrix, ordered by increasing regulatory response and degraded performance, are: 1) licensee response, 2) regulatory response, 3) degraded cornerstone, 4) multiple/repetitive degraded cornerstone, and 5) unacceptable performance. The following variable was used to capture information from the ROP Action Matrix:

- **Action Matrix Elevated:** Indicator of sites that have elevated oversight by the NRC due to degraded performance. Sites in columns 2-5 of the action matrix during the 4th quarter assessment period were coded as 1 (elevated), sites in column 1 were coded as 0 (not elevated).

The ROP also includes three “cross-cutting” areas that extend across all of the ROP cornerstones of safety. These areas are Human Performance (HP), Problem Identification and Resolution (PI&R), and Safety Conscious Work Environment (SCWE). Enhancements to the ROP in 2006 expanded the level of detail in the cross-cutting areas to include components and aspects. Components are nested within one of the three cross-cutting areas, and aspects are nested within components. Inspectors can assign aspects to inspection findings if they determine that the aspect characterizes the most significant contributor to the performance deficiency cited in the finding (IMC -0310; NRC, 2011a). Although the SCWE cross-cutting area is conceptually related to safety culture, the counts of inspection findings assigned to the SCWE area were so low that the inspection data were excluded from the analysis. The following variables related to the ROP cross-cutting aspects from each plant were used in this analysis (note that sub-bullets signify variables that are nested within the variable listed at the higher level):

- **Total ROP Aspects:** The total number of inspection report findings that were assigned as being attributable to one of the aspects in one of the ROP cross-cutting areas per year.
 - **Human Performance Cross-Cutting Area**

- **HP Component 1 - Decision Making**
- **HP Component 2 - Resources**
- **HP Component 3 - Work Control**
- **HP Component 4 - Work Practices**
- **Problem Identification and Resolution (PI&R) Cross-Cutting Area**
 - **PI&R Component 1 - Corrective Action Program (CAP)**
 - **PI&R Component 2 - Operating Experience**
 - **PI&R Component 3 - Assessments**

During mid-cycle and end-of-cycle assessment meetings, the NRC reviews the aspects tagged to inspection findings within each cross-cutting area for substantive cross-cutting issues (SCCIs). According to IMC-0305 (NRC, 2011c), an SCCI is a cross-cutting theme that has been identified in one of the cross-cutting areas (HP, PI&R, or SCWE), about which the NRC staff has a concern with the licensee's scope of efforts or progress in addressing the cross-cutting theme. Cross-cutting themes in the HP and PI&R areas are assigned when multiple inspection findings (i.e., four or more) are assigned the same cross-cutting aspect within a 12 month period. A cross-cutting theme can be assigned in the SCWE area when one or more inspection findings are assigned a cross-cutting aspect within an 18 month period. In all cases, the NRC determines that an SCCI exists only if there is a concern with the licensee's scope of efforts or progress in addressing the cross-cutting theme. The following variable was used to represent the presence of one or more SCCIs at a site:

- **Total SCCIs:** The total number of outstanding SCCIs in the HP or PI&R area at each site during the end-of-cycle assessment. No sites had SCCIs in the SCWE area during the end-of-cycle 2010 or 2011 time period.

The NRC's Office of Enforcement (OE) has an Allegations program that deals with safety concerns related to potential or actual safety issues associated with NRC-regulated activities. Safety concerns may include areas like operations, maintenance, radiation protection, security, harassment, discrimination, wrongdoing, or a work environment that discourages workers from raising safety concerns. Any member of the public or individual who is performing work at a site licensed by the NRC may report an allegation. The NRC maintains and regularly evaluates statistics on allegations from operating power plants to identify significant trends in the data which may indicate a more widespread issue within the organization. For instance, a sharp increase in allegations could suggest that employees are losing confidence in the plant's internal corrective action program or no longer feel safe raising safety concerns within the organization.

Part of OE's evaluation of allegations includes categorizing allegations based on type. For instance, allegations may be categorized as relating to security concerns, the employee concerns program, discrimination in the workplace, fitness for duty, or falsification of information. For the purposes of this study, we grouped together allegations that OE categorized as "discrimination," "chilling effects," "employee concerns program" and "safety culture" and formed a variable called "SCWE-Related Allegations." Allegations within these categories were determined to be most likely to relate to the organization's safety-conscious work environment, and therefore seemed to be most conceptually related to safety culture. Some of the categories not included related to "construction," "security," "health physics," "fitness for duty," and "wrongdoing" because they did not appear to be

allegations related to the overall work environment at the site. The following variables were used to capture information on allegations at a site per year:

- **Allegations from Personnel:** The total number of allegation concerns reported to the NRC by personnel at a site (including contractors).
 - **SCWE-Related Allegations:** The subset of allegation concerns reported by personnel at a site that are related to the site’s Safety Conscious Work Environment.

An overview of the NRC Performance Metrics used to represent plant performance in subsequent data analyses are presented in Figure 3.

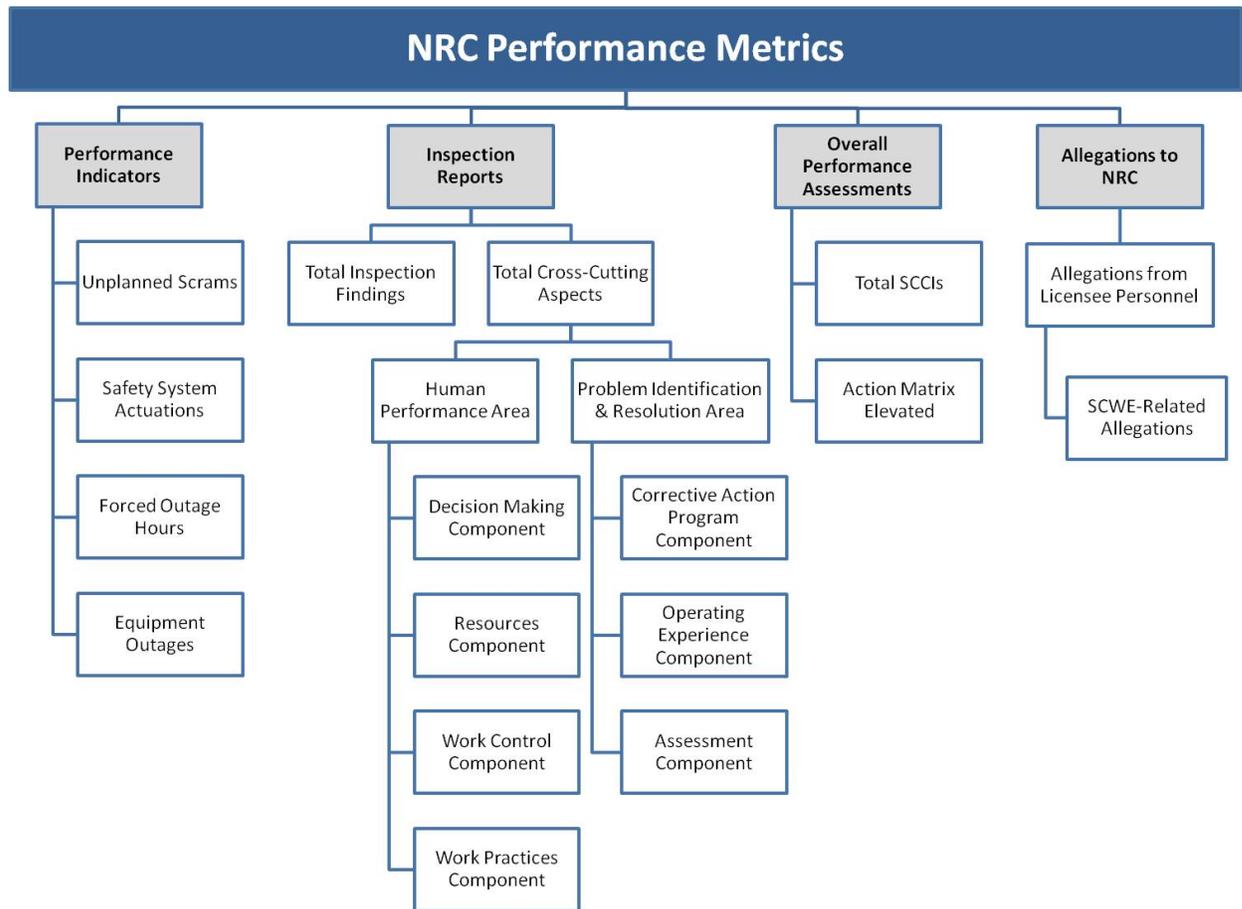


Figure 3 Overview of NRC Performance Metrics Used in Validity Analyses

5.2. Concurrent Validity of the Survey Data with NRC Performance Metrics

The concurrent validity analysis sought to assess whether there was an apparent relationship between safety culture and safety performance using site-level mean scores on the safety culture survey as the indicator of safety culture and aggregated site-level values for NRC performance metrics in 2010 (described in section 5.1) as the indicator(s) of safety performance. The sample size was 63: the total number of sites that participated in the safety culture survey.

The statistic used to test for a relationship between safety culture and safety performance was Pearson's product-moment correlation coefficient. Pearson's correlation is a parametric statistic that tests the linear relationship between two variables. Correlations are often referred to as the foundation for basic and advanced statistics because most statistics seek to describe the relationships among variables of interest (Chen & Popovich, 2002). Pearson's correlation in particular is chosen 95% of the time to describe a relationship in research (Glass & Hopkins, 1996). Pearson's correlation is most appropriately used when the variables being tested are measured on an interval or ratio scale and meet assumptions of normality.

As mentioned previously, the Likert scale used in the safety culture survey uses response anchors (e.g., strongly agree/strongly disagree) that are expected to be equidistant from each other to approximate an interval scale. The survey factors also meet assumptions of normality, as determined by the Shapiro-Wilk test of normality (Myers & Well, 2003). The NRC performance metrics are also considered to be on an interval scale because they are either counts (e.g., number of inspection findings) or ratios (e.g., scrams per 7000 operating hours). However, the NRC performance metrics do not demonstrate normal distributions and fail to meet the Shapiro-Wilk test of normality. This is not surprising because much of the count data, particularly for the performance indicators, have relatively low numbers and zeros are common, resulting in a positive skew to the data. When the normality assumption is violated the maximum possible value of the correlation coefficient is less than 1 because the two variables do not share the same distribution. There is also a greater possibility that a non-linear relationship exists between two variables. Pearson's correlation coefficient can also be artificially high or low when there are outliers present in the data. Interpretations of the correlation coefficient should be accompanied by visual inspection of the data to ensure that the correlation is not being unduly influenced by a small number of outliers. Pearson correlations are less affected by outliers and violations of the normality assumption with larger sample sizes (Chen & Popovich, 2002).

An alternative non-parametric statistic is the Kendall Tau rank-order correlation. Non-parametric statistics do not make assumptions about the underlying distributions of the data and can be used when one or both variables violate assumptions of normality (Chen & Popovich, 2002). The Kendall Tau correlation ranks the data values of both variables from highest to lowest. The rank-ordered values are then compared to assess the extent to which the ranks correspond. For example, does site X, compared to the other sites in the sample, have the same rank on the variable representing safety culture as its rank on the unplanned scrams performance indicator? Kendall Tau correlations are robust when there are outliers in the data because the rank ordering effectively eliminates the outlier effect without removing the data. Outliers are instead treated as the highest or lowest ranking values in the data. However, the Kendall Tau statistic can be significantly influenced by ties in the data (e.g., multiple sites have the same number of unplanned scrams), resulting in spuriously small correlation coefficients.

Studies of safety culture have traditionally used Pearson's correlation coefficient as a starting point for exploring associations between variables, and subsequent meta-analytic studies have relied on the reporting of correlations to calculate effect sizes and estimate composite effects across multiple studies. We calculated correlations between the safety culture survey results and safety performance variables using Pearson's correlation coefficient and the Kendall Tau correlation coefficient. The results of these tests demonstrated similar patterns of statistically significant relationships. Therefore, only Pearson's correlation coefficients are presented in the paper. Use of Pearson's correlation in this analysis is consistent with exploratory research practices in the social sciences and allows for direct comparisons between the current study and past studies documented in the safety culture research literature.

The interpretation of a correlation coefficient depends on the context and purpose. A correlation of .50 may be very low if one is verifying a physical law using high-quality instruments, but may be regarded as high in the social sciences where there may be multiple unknown or unpredictable factors contributing to a complex relationship. Given the results of previous studies, we expected small to medium effect sizes with correlation coefficients of .20 to .30. We also expected the safety culture survey to be negatively related to the NRC performance metrics, such that higher scores on the safety culture survey are associated with lower values on the NRC performance metrics (e.g., fewer scrams, inspection findings, and allegations). In addition, we hypothesized that the NRC metrics related to the ROP cross-cutting aspects and allegations would be more strongly related to safety culture than the NRC performance indicators because they are indicators of aspects of performance that are, conceptually, more likely to be affected by the organization's safety culture than indicators of plant and equipment performance.

It is also important to emphasize that the variables being compared are representations of constructs. The INPO safety culture survey is a representation of the construct of safety culture, and the NRC performance metrics are representations of safety performance. Although they may be valid approximations of their respective constructs, the measurements likely contain error and are therefore not perfectly accurate reflections of reality (Trochim, 2006). All of the variables being investigated are susceptible to both random and systematic measurement error. Random error refers to any factors that randomly affect the measurement of a variable within a sample, such that the error does not have a consistent effect across the entire sample. This results in observed scores that are artificially inflated in some cases and deflated in others. Random error is sometimes considered "noise" because it adds to variability in the data, but does not affect the average for the sample as a whole. Systematic error refers to factors that systematically affect the measurement of a variable, and therefore result in errors that tend to either consistently increase or decrease the observed scores. Systematic error may result in a positive or negative bias in the data, but may not necessarily affect the linear relationships between variables. We compensated for the potential effects of measurement error in this study in two ways: 1) by using a safety culture score that is averaged across items and across people at a site, thereby resulting in a more stable mean score, and 2) by using multiple variables to represent different aspects of plant performance.

Pearson correlations are reported for the overall measure of safety culture and the nine safety culture factors identified by INPO. Relationships between individual safety culture factors and the NRC performance metrics may suggest instances where some aspects of safety culture are more strongly related to safety performance than other aspects.

Table 7 presents the concurrent Pearson's correlations between safety culture, NRC performance indicators, inspection findings, and the ROP Action Matrix in 2010. Table 8 presents the correlations between safety culture, NRC allegations, and ROP cross-cutting areas and components in 2010. Correlations that are statistically significant ($p < .05$ or $p < .01$) are highlighted.

The overall measure of safety culture demonstrated statistically significant correlations with the Unplanned Scrams and Forced Outage Hours performance indicators. The significant correlations were negative, and thus in the expected direction. The scatterplot in Figure 4 shows the linear relationship between safety culture and unplanned scrams. Note that sites with more than two unplanned scrams were generally below the overall mean score for safety culture of 5.61. Sites with higher scores on the safety culture survey were more likely to have fewer unplanned scrams, forced outage hours, and inspection findings in 2010. The effect sizes ranged from .25 to .46, suggesting a medium effect and that safety culture accounts for 6% to 21% of the variance in these performance indicators and inspection findings. Most notably, the correlation between Training Quality and Unplanned Scrams was -.46, indicating that sites with higher quality training, as judged by site personnel, had considerably fewer unplanned scrams in 2010. The correlation between Questioning Attitude and Total ROP Findings was also of a moderate size at -.41, suggesting that sites that were better about encouraging a questioning attitude (as perceived by employees) were more likely to have fewer inspection findings overall. Another interesting result is that sites with elevated NRC oversight (i.e., in columns 2-5 of the action matrix) had lower scores on the factors Willingness to Raise Concerns, Decision-Making, Supervisor Responsibility for Safety, and Personal Responsibility for Safety.

The safety culture survey results evidenced a medium-sized correlation with the total number of ROP cross-cutting aspects (-.41), mainly due to the high correlations between Management Responsibility/Commitment to Safety and Total ROP Aspects (-.44) and Questioning Attitude and Total ROP Aspects (-.45). The scatterplot in Figure 5 shows the negative association between safety culture and counts of ROP cross-cutting aspects. This result suggests that sites with management who had higher safety culture scores (i.e., perceived by survey participants as demonstrating a strong commitment to safety and fostering a questioning attitude among their workforce) were less likely to have inspection findings that inspectors attributed to a cross-cutting aspect. Further, a site's score on the safety culture survey accounted for 15% of the variance in their total number of ROP aspects. Interestingly, Personal Responsibility for Safety was not related to any of the performance indicators or ROP cross-cutting areas. The Decision-Making factor of the safety culture survey was not significantly related to the Decision-Making Component of the ROP, but did have a significant negative correlation with the Corrective Action Program Component of the ROP. In addition, some of the components of the cross-cutting areas (i.e., Work Control and Operating Experience) did not have significant relationships with any of the safety culture survey factors.

Table 7 Correlations between Safety Culture and 2010 Performance Indicators, ROP Action Matrix, Inspection Findings, and SCCIs

	Unplanned Scrams	Safety System Actuations	Forced Outage Hours	Equipment Outages	Action Matrix Elevated	Total Inspection Findings	Total SCCIs
SAFETY CULTURE	-.35**	.01	-.27*	-.25	-.23	-.37**	-.23
1. Management Responsibility/ Commitment to Safety	-.34**	.03	-.26*	-.22	-.20	-.40**	-.28*
2. Willingness to Raise Concerns	-.29*	.01	-.24	-.30*	-.31*	-.21	-.10
3. Decision-Making	-.33**	.10	-.26*	-.25*	-.27*	-.34**	-.21
4. Supervisor Responsibility for Safety	-.26*	-.07	-.21	-.10	-.27*	-.34**	-.14
5. Questioning Attitude	-.29*	.00	-.22	-.24	-.17	-.41**	-.38**
6. Safety Communication	-.35**	-.04	-.29*	-.22	-.19	-.33**	-.14
7. Personal Responsibility for Safety	-.24	-.05	-.19	-.18	-.28*	.02	.13
8. Prioritizing Safety	-.23	.06	-.16	-.23	-.11	-.27*	-.11
9. Training Quality	-.46**	-.03	-.39**	-.37**	-.23	-.28*	-.10

*p < .05; **p < .01

Table 8 Correlations between Safety Culture and 2010 NRC Allegations, ROP Cross-Cutting Areas and Components

	Allegations From Personnel	SCWE-Related Allegations	Total ROP Aspects	Human Perform. Area	HP 1: Decision Making	HP 2: Resources	HP 3: Work Control	HP 4: Work Practices	Problem ID & Resolution Area	PI&R 1: CAP	PI&R 2: OE	PI&R 3: Assessments
SAFETY CULTURE	-0.21	-0.28*	-0.39**	-0.28*	-0.27*	-0.23	-0.01	-0.21	-0.37**	-0.35**	-0.10	-0.24
1. Management Responsibility/ Commitment to Safety	-0.24	-0.30*	-0.44**	-0.30*	-0.29*	-0.22	-0.05	-0.23	-0.40**	-0.40**	-0.08	-0.23
2. Willingness to Raise Concerns	-0.16	-0.21	-0.21	-0.16	-0.13	-0.15	0.09	-0.16	-0.24	-0.22	-0.14	-0.14
3. Decision-Making	-0.17	-0.25	-0.37**	-0.25*	-0.23	-0.25*	-0.02	-0.17	-0.35**	-0.33**	-0.13	-0.23
4. Supervisor Responsibility for Safety	-0.15	-0.24	-0.28*	-0.27*	-0.25*	-0.25*	0.00	-0.19	-0.24	-0.21	-0.13	-0.30*
5. Questioning Attitude	-0.41**	-0.48**	-0.47**	-0.40**	-0.36**	-0.30*	0.04	-0.37**	-0.37**	-0.36**	-0.15	-0.18
6. Safety Communication	-0.12	-0.19	-0.31*	-0.20	-0.22	-0.20	0.05	-0.14	-0.32*	-0.31*	-0.06	-0.28*
7. Personal Responsibility for Safety	0.17	0.14	0.04	0.05	0.09	-0.12	0.02	0.09	0.00	0.03	-0.17	0.00
8. Prioritizing Safety	-0.09	-0.15	-0.28*	-0.18	-0.17	-0.23	0.01	-0.09	-0.30*	-0.29*	0.00	-0.26*
9. Training Quality	0.04	-0.05	-0.22	-0.14	-0.20	-0.17	0.07	-0.05	-0.25*	-0.23	-0.08	-0.20

*p < .05; **p < .01

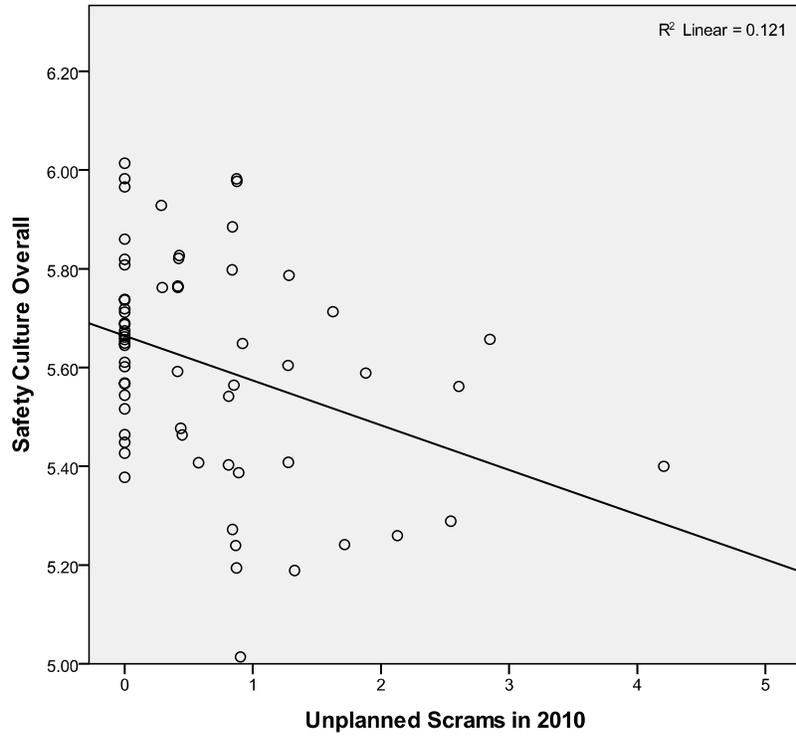


Figure 4 Scatterplot of Relationship between Safety Culture and Unplanned Scrams in 2010

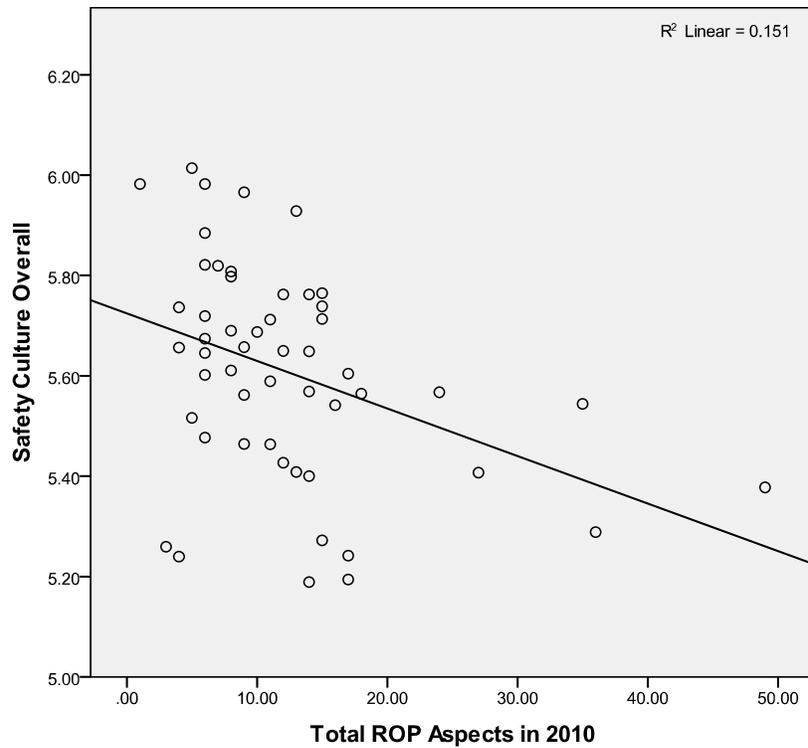


Figure 5 Scatterplot of Safety Culture and Total ROP Aspects in 2010

Although the overall safety culture survey score was not significantly related to whether a site had an SCCI, both Management Responsibility/Commitment to Safety and Questioning Attitude survey factors evidenced significant negative correlations with the SCCI variable. Sites where employees reported better management commitment to safety and more of a questioning attitude were less likely to have a substantive cross-cutting issue at the end of 2010. The highest correlation among all of the concurrent NRC performance metrics was between Questioning Attitude and SCWE-related Allegations. Sites where employees indicated that a questioning attitude was more strongly supported seemed to have fewer allegations related to their safety-conscious work environment. In terms of the effect size, the questioning attitude factor accounted for 23% of the variance in the number of SCWE-related allegations. In addition, although safety culture was not related to the total count of allegations, there was a significant relationship between the overall measure of safety culture and SCWE-related allegations. The scatterplot depicted in Figure 6 shows the relationship between safety culture and SCWE-related allegations. Although most sites had fewer than five SCWE-related allegations in 2010, two of the three sites with more than five allegations had safety culture scores below the mean (5.61).

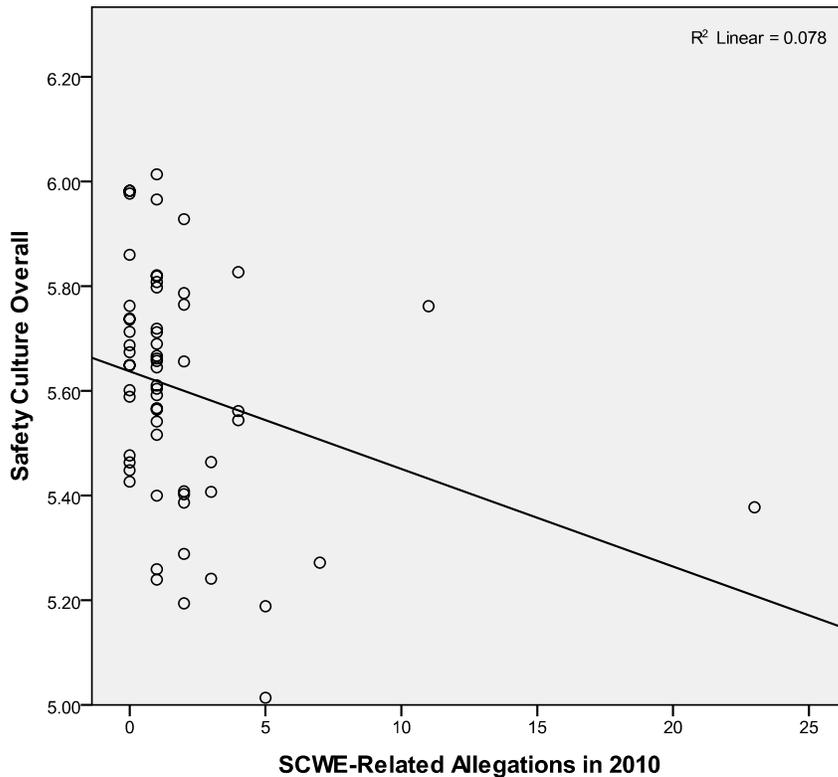


Figure 6 Scatterplot of Safety Culture and SCWE-Related Allegations in 2010

The results of the concurrent validity analysis demonstrate that the safety culture survey is related to some measures of safety performance. Although not related to all of the NRC performance metrics tested, the overall measure of safety culture, derived from INPO's survey, evidenced statistically significant correlations with Unplanned Scrams, Forced Outage Hours, Total Inspection Findings, Total ROP Aspects, Human Performance Cross-cutting Area, Decision-Making Cross-cutting Component, Problem Identification and Resolution Cross-cutting Area, the Corrective Action Program Component, and SCWE-related Allegations. In addition, some of the safety culture survey factors demonstrated unique correlations with the NRC performance metrics. The Management Responsibility/Commitment to Safety, Decision-Making, and Questioning Attitude survey factors seemed to have the most consistent correlations with the performance metrics used in this analysis. The Training Quality factor was significantly related to Unplanned Scrams and Forced Outage Hours, the Questioning Attitude factor was the only factor related to the Work Practices Component of the ROP and total Allegations, and Personal Responsibility for Safety was related to being in an elevated oversight column within the Action Matrix.

5.3. Exploratory Predictive Validity of the Survey with NRC Performance Metrics

The predictive validity analysis consisted of testing the relationship between the safety culture survey administered in 2010 and NRC performance metrics in 2011, the time period one year after the administration of the survey. The purpose of the predictive analysis is to evaluate whether scores on the safety culture survey may relate to future indicators of safety performance and thus could be used as a predictor of safety performance. The same NRC metrics used in the concurrent analysis were used in the predictive analysis, with the exception of the NRC performance indicators (i.e., unplanned scrams, safety system actuations, forced outage hours, equipment outages) and SCWE-related allegations because not all of the 2011 data were available at the time of the analysis.

Table 9 presents the predictive Pearson's correlations between safety culture and the ROP cross-cutting areas and components in 2011. Table 10 presents the correlations between safety culture and NRC allegations, ROP Action Matrix placement, and SCCIs in 2011. Correlations that are statistically significant ($p < .05$ or $p < .01$) are highlighted.

The correlations between the safety culture survey and the Problem Identification and Resolution cross-cutting area remained significant in 2011, in particular the Management Responsibility/Commitment to Safety, Decision-Making, and Prioritizing Safety factors. Other correlations between the safety culture factors and ROP Cross-cutting areas and components did not remain significant using the 2011 data. It appears that the safety culture survey results may be related to future problem identification and resolution issues, such that sites with lower safety culture scores were more likely to have more findings assigned to the PI&R cross-cutting area and CAP component. A scatterplot of the relationship between safety culture and the PI&R cross-cutting area is shown in Figure 7.

Although the safety culture survey results demonstrated fairly strong correlations with the number of ROP inspection findings in 2010, there was no significant relationship between these variables using the 2011 inspection finding data. Safety culture, as measured by the INPO survey, was concurrently related to the number of inspection findings a site receives, but it did not relate to inspection findings in the following year.

Table 9 Correlations between Safety Culture and 2011 ROP Cross-Cutting Areas and Components

	Total ROP Aspects	Human Perform. Area	HP 1: Decision Making	HP 2: Resources	HP 3: Work Control	HP 4: Work Practices	Problem ID & Resolution	PI&R 1: CAP	PI&R 2: Operating Experience	PI&R 3: Assessments
SAFETY CULTURE	-0.24	-0.12	-0.10	-0.14	.01	-0.07	-.27*	-.28*	-0.06	.01
1. Management Responsibility/ Commitment to Safety	-0.27	-0.11	-0.12	-0.14	.05	-0.06	-.30*	-.30*	-0.10	-0.02
2. Willingness to Raise Concerns	-0.11	-0.03	.01	-0.08	.01	-0.01	-.18	-.23	.03	.12
3. Decision-Making	-0.27	-0.13	-0.10	-0.12	.00	-0.09	-.29*	-.29*	-0.12	.03
4. Supervisor Responsibility for Safety	-0.17	-0.12	-0.06	-0.12	-0.01	-0.10	-.20	-.21	-0.07	.06
5. Questioning Attitude	-0.20	-0.15	-0.19	-0.11	-0.07	-0.04	-.19	-.23	.04	.01
6. Safety Communication	-0.21	-0.14	-0.10	-0.13	-0.05	-0.08	-.20	-.23	-0.01	-0.01
7. Personal Responsibility for Safety	-0.02	-0.05	.03	-0.07	.02	-0.10	-.06	-.10	.06	.13
8. Prioritizing Safety	-0.27	-0.14	-0.07	-0.18	-0.04	-0.08	-.25*	-.25*	-0.07	-0.08
9. Training Quality	-0.13	-0.06	-0.01	-0.15	.06	-0.04	-.15	-.16	-0.02	.03

*p < .05; **p < .01

Table 10 Correlations between Safety Culture and 2011 ROP Action Matrix, Inspection Findings, SCCIs, and Allegations

	Action Matrix Elevated	Total Inspection Findings	Total SCCIs	Allegations From Personnel
SAFETY CULTURE	-.30*	.07	-.26*	-.36**
1. Management Responsibility/ Commitment to Safety	-.29*	.07	-.27*	-.38**
2. Willingness to Raise Concerns	-.24	.16	-.18	-.38**
3. Decision-Making	-.32*	.01	-.32*	-.33**
4. Supervisor Responsibility for Safety	-.30*	.01	-.18	-.20
5. Questioning Attitude	-.26*	.02	-.23	-.48**
6. Safety Communication	-.29*	.07	-.21	-.28*
7. Personal Responsibility for Safety	-.05	.19	-.10	-.08
8. Prioritizing Safety	-.27*	.04	-.20	-.21
9. Training Quality	-.23	.02	-.20	-.07

*p < .05; **p < .01

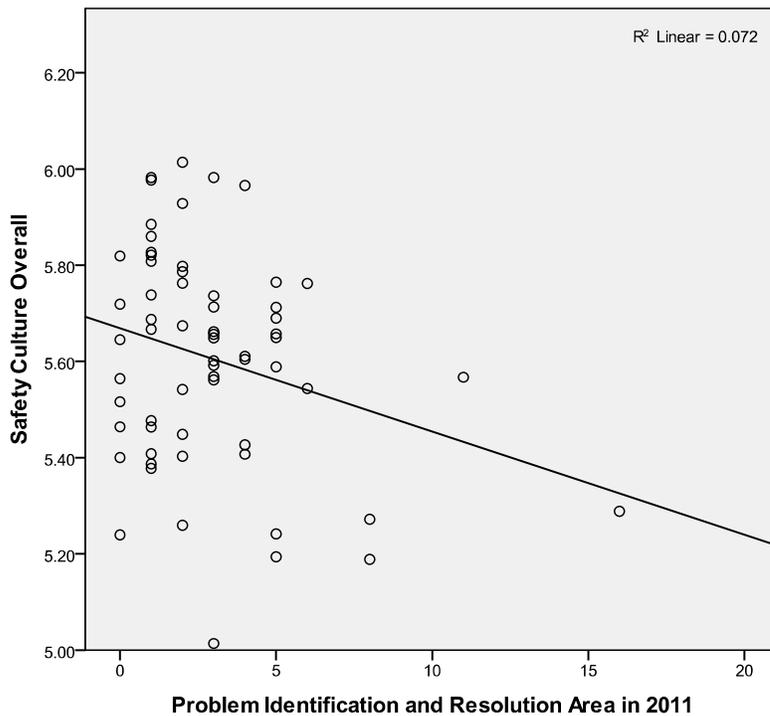


Figure 7 Scatterplot of Safety Culture and Problem Identification and Resolution Area in 2011

The strongest correlations between the safety culture survey and safety performance metrics in the following year relate to the broad performance metrics of SCCIs, elevated oversight in the Action Matrix, and Allegations. The overall measure of safety culture was not significantly related to these three variables using the 2010 data, but demonstrated significant negative correlations with the 2011 data. Sites with lower scores on the safety culture survey, as measured in 2010, were more likely to receive substantive cross-cutting issues, be placed in an elevated oversight condition within the ROP Action Matrix, and receive more allegations in 2011. The scatterplot of the relationship between safety culture and allegations in Figure 8 shows a significant negative trend in the data. Lower scores on the safety culture survey may indicate problems at a site that are starting to add up, and it is only at future points in time that those problems, if not adequately addressed, require NRC response through the issuance of SCCIs and movement in the ROP Action Matrix. In addition, it could be hypothesized that personnel may try to raise safety issues internally at first, and then seek out the NRC for assistance via the allegations process only when they find their organization's internal processes to be ineffective, which is another potential characteristic of weaknesses in safety culture.

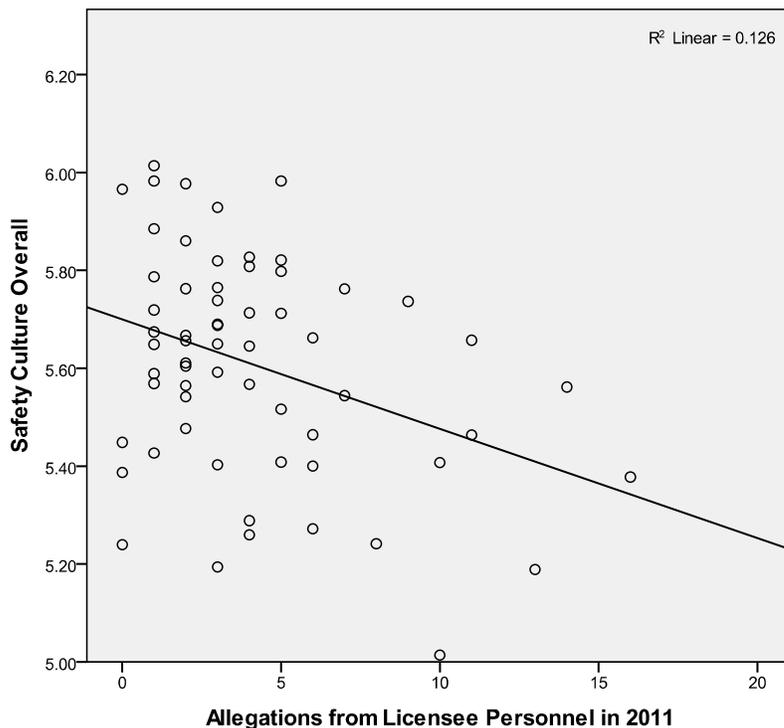


Figure 8 Scatterplot of Safety Culture and Allegations in 2011

It is important to note, however, that the predictive validity findings presented here are exploratory, particularly because the correlational analyses cannot be used to verify causality and the data used represent snapshots of safety culture and safety performance at single points in time. These individual variables do not, by themselves, adequately capture the full scope and dynamic nature of an organization's safety culture or its safety performance. Additional analyses would be necessary to establish confidence that the predictive validity results represent a causal phenomenon that is stable over time. For instance, structural equation modeling using longitudinal data (i.e., multiple administrations of the safety culture survey and annual safety performance metrics over successive

years) would allow for more complex testing of combinations of safety culture factors and performance variables to establish whether safety culture consistently predicts safety performance.

5.4. Relationships among NRC Performance Metrics

The NRC does not collect or maintain data specific to personnel perceptions of the safety culture of its regulated entities, such as may be collected through a safety culture survey. However, some of the data types that the NRC collects may be conceptually related to safety culture. In addition, the NRC collects the data on a continual basis, therefore making longitudinal analyses possible. In this section we explore whether there is statistical evidence of concurrent and predictive relationships among some NRC Performance Metrics and how those relationships compare to the correlations with the safety culture survey.

Of the many NRC performance metrics presented in Section 5.1, the data related to the ROP Cross-Cutting Components and Allegations appear to be, conceptually, the most closely related to an organization's safety culture. The Cross-Cutting Components were added to the ROP to more fully address safety culture. The Allegations represent concerns reported by personnel about their organization, with SCWE-related allegations specifically relating to the work environment. Therefore, we compared the ROP cross-cutting components and allegations to the performance indicators analyzed previously, number of SCCIs, and a site's placement in elevated oversight on the ROP Action Matrix.

We first assessed whether the ROP Cross-Cutting Components and Allegations had similar correlations to other independent safety performance measures as those found between the INPO survey scores and safety performance measures for the same year. Table 11 shows the Pearson's correlations comparing the NRC Allegations and ROP Cross-Cutting Areas and Components to the four equipment performance indicators used in the analyses described above, and placement in the ROP Action Matrix in 2009 and 2010. Table 12 shows the correlations among Allegations, Inspection Findings, and SCCIs in 2009 and the correlations between the same variables in 2010. Correlations between the ROP Cross-Cutting Components, inspection findings, and number of SCCIs are not included because those variables are overlapping. Cross-cutting aspects are assigned to inspection findings, and SCCIs are determined based on themes in cross-cutting areas.

Table 11 Correlations among 2009 and 2010 NRC Performance Metrics

	Unplanned Scrams		Safety System Actuations		Forced Outage Hours		Equipment Outages		Action Matrix Elevated	
	2009 Data	2010 Data	2009 Data	2010 Data	2009 Data	2010 Data	2009 Data	2010 Data	2009 Data	2010 Data
Allegations From Personnel	.07	-.06	-.13	-.10	.02	-.06	.13	-.04	.27*	.04
SCWE-Related Allegations	.12	-.03	-.10	-.13	-.01	-.05	.06	-.01	.24	.00
Total ROP Aspects	.05	.13	-.05	-.13	-.08	.08	.14	-.03	.18	.18
Human Performance	-.02	.09	-.07	-.13	-.06	.01	.11	-.05	.16	.06
HP 1: Decision Making	.07	.07	-.02	-.10	-.09	-.03	.02	.04	.12	-.02
HP 2: Resources	.02	.29*	.11	.12	-.13	.25*	.12	.05	.14	.22
HP 3: Work Control	.09	-.11	-.15	-.17	.21	-.21	.12	-.17	.03	-.02
HP 4: Work Practices	-.16	-.03	-.11	-.19	-.10	-.04	.07	-.09	.13	.00
Problem Identification and Resolution	.14	.14	.01	-.08	-.08	.13	.12	.00	.14	.26*
PI&R 1: CAP	.16	.11	.02	-.13	-.06	.09	.12	-.01	.15	.21
PI&R 2: Operating Experience	.04	.23	-.04	.25*	-.08	.31*	.10	.07	.04	.40**
PI&R 3: Assessments	-.05	.00	.03	-.03	-.06	-.05	-.09	-.06	.04	.01

*p < .05; **p < .01

Note. The correlations presented above are for NRC performance variables measured in the same year. For example, the correlations in the first 2009 column represent the number of allegations from personnel in 2009 compared to the number of unplanned scrams in 2009, and the 2010 column represents the number of allegations from personnel in 2010 compared to the number of unplanned scrams in 2010.

Table 12 Correlations among 2009 and 2010 Allegations, Inspection Findings, and SCCIs

	Total Inspection Findings		Total SCCIs	
	2009 Data	2010 Data	2009 Data	2010 Data
Allegations From Personnel	.38**	.51**	.61**	.70**
SCWE-Related Allegations	.41**	.52**	.69**	.64**

*p < .05; **p < .01

Note. The correlations presented above are for NRC performance variables measured in the same year. For example, the 2009 column under SCCIs represents the number of allegations from personnel in 2009 compared to the number of SCCIs in 2009, and the 2010 column represents the number of allegations from personnel in 2010 compared to the number of SCCIs in 2010.

In the 2009 data, neither the ROP cross-cutting components nor the allegations evidenced any statistically significant relationships with the performance indicators evaluated. The variable representing number of allegations was significantly correlated with a site’s placement in elevated oversight on the action matrix, suggesting that sites with more allegations were more likely to be in an elevated oversight column on the action matrix. However, this relationship was not significant using the 2010 data. The Resources and Operating Experience components were significantly correlated with some of the performance indicators in the 2010 data, and the Operating Experience component was correlated with elevated oversight on the Action Matrix in 2010. Correlations between allegations and inspection findings were statistically significant in the 2009 and 2010 data. In addition, there were very strong correlations between allegations and SCCIs in both the 2009 and 2010 data, suggesting that sites with more allegations from personnel and SCWE-related allegations were more likely to have more SCCIs at the end-of-cycle of that same year.

To assess whether ROP Cross-Cutting Components and SCWE-related Allegations are related to other future measures of safety performance we examined their correlations with the safety performance variables one year later. For instance, the total number of SCWE-related Allegations a site received in 2009 was compared to the site’s performance indicators in 2010.

Table 13 displays the Pearson’s correlations between the 2009 and 2010 NRC performance metrics, and Table 14 displays the correlations between the 2010 and 2011 NRC performance metrics. Note that, as mentioned previously, performance indicator data for 2011 were not available at the time of this analysis.

Table 13 Correlations between 2009 and 2010 NRC Performance Metrics

	Unplanned Scrams (2010)	Safety System Actuations (2010)	Forced Outage Hours (2010)	Equipment Outages (2010)	Action Matrix Elevated (2010)	Total Inspection Findings	Total SCCIs (2010)
Allegations From Personnel (2009)	-0.03	-0.11	-0.06	.07	-0.04	.44**	.51**
SCWE-Related Allegations (2009)	.00	-0.07	-0.01	.07	-0.08	.44**	.57**
Total ROP Aspects (2009)	.02	-0.05	-0.07	.02	-0.07	.62**	.42**
Human Performance (2009)	.01	.05	-0.07	.07	-0.03	.45**	.28*
HP 1: Decision Making (2009)	-0.03	-0.11	-0.09	-0.04	-0.08	.41**	.21
HP 2: Resources (2009)	.06	.15	.01	.27*	-0.04	.15	.03
HP 3: Work Control (2009)	-0.10	-0.01	-0.07	-0.12	-0.02	.25*	.27*
HP 4: Work Practices (2009)	.07	.12	-0.02	.09	.04	.34**	.23
Problem Identification and Resolution (2009)	.03	-0.20	-0.04	-0.06	-0.10	.62**	.45**
PI&R 1: CAP (2009)	.04	-0.16	-0.05	-0.06	-0.06	.68**	.57**
PI&R 2: Operating Experience (2009)	.01	-0.20	.03	.00	-0.11	.14	-0.10
PI&R 3: Assessments (2009)	-0.05	-0.08	-0.06	-0.09	-0.09	.04	-0.06

*p < .05; **p < .01

SCWE-related Allegations in 2009 were significantly correlated with number of inspection findings and SCCIs in 2010. Sites with more SCWE-related allegations were more likely to have higher numbers of inspection findings and SCCIs in the following year. This same relationship held for counts of all allegations from licensee personnel. However, the scatterplot in Figure 8 suggests that an outlier may be driving the high correlation between SCWE-related allegations and SCCIs. When this outlier is restricted to within three standard deviations from the mean (i.e., the outer boundary of a normal distribution) the correlation remains high ($r = .42, p < .01$), but when the outlier is removed altogether the correlation is not significant ($r = .06, p > .05$). Thus, it is not appropriate to draw conclusions based on this relationship.

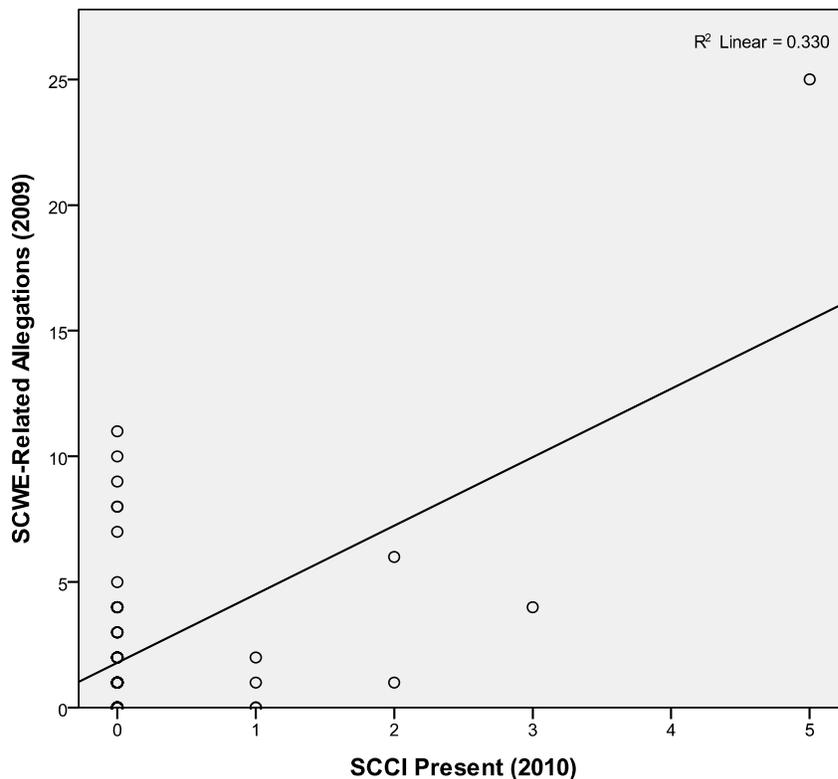


Figure 9 Scatterplot of SCWE-related Allegations in 2009 and SCCIs in 2010

The total number of ROP aspects assigned to findings in 2009, along with the counts of aspects in the Human Performance and Problem Identification and Resolution cross-cutting areas, were significantly related to the number of inspection findings and SCCIs in 2010. Sites with more findings assigned to cross-cutting aspects in 2009 were more likely to have more inspection findings and SCCIs in the following year. This is perhaps not surprising considering that the ROP aspects, inspection findings, and SCCIs are inherently related. The counts of cross-cutting aspects represent a sub-set of inspection findings, and whether a site receives an SCCI is dependent on them having a theme in one of the cross-cutting areas due to counts of cross-cutting aspects assigned to findings within a 12-month period. The 2009 counts of allegations and cross-cutting aspects in the ROP were not significantly correlated with elevated oversight in the ROP Action Matrix in the following year.

The only variable that had a significant correlation with the 2010 performance indicators was the count of aspects assigned to the Resources component of the ROP in 2009. The Resources component

of the ROP was positively correlated with the number of equipment outages in 2010, suggesting that sites with more cross-cutting aspects assigned to Resources in 2009 may be more likely to have more equipment outages in the following year. The scatterplot showing the relationship between Resources and Equipment Outages is presented in Figure 10.

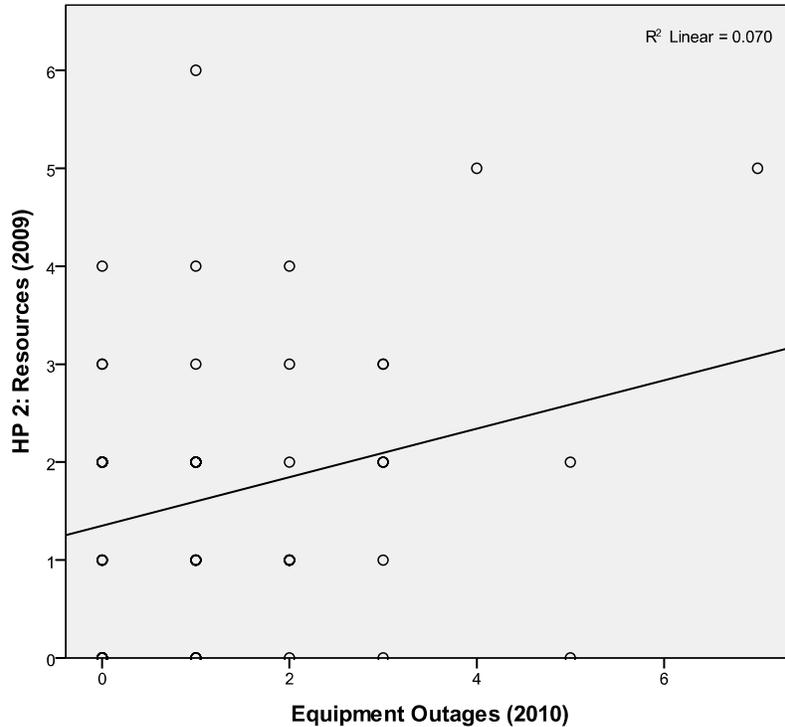


Figure 10 Scatterplot of Resources Component in 2009 and Equipment Outages in 2010

The correlations between the 2010 and 2011 NRC performance metrics were not as strong as the correlations between the 2009 and 2010 data. The total number of cross-cutting aspects assigned to findings was significantly correlated with the total number of inspection findings and SCCIs in the following year. As found previously, there were no statistically significant correlations between any of the NRC performance metrics in 2010 and Action Matrix placement in 2011.

Overall, the only variables that demonstrated consistent correlations were the counts of ROP cross-cutting aspects, inspection findings, and SCCIs. Although these relationships are interesting, the interrelatedness of the variables precludes using counts of ROP aspects as independent leading indicators of performance. Only one variable, the Resources component of the ROP, was significantly correlated with an NRC performance indicator, Equipment Outages.

Table 14 Correlations between 2010 and 2011 NRC Performance Metrics

	Action Matrix Elevated (2011)	Total Inspection Findings	Total SCCIs (2011)
Allegations From Personnel (2010)	-.03	.12	.00
SCWE-Related Allegations (2010)	.00	.18	.00
Total ROP Aspects (2010)	.08	.26*	.38**
Human Performance (2010)	.01	.26*	.20
HP 1: Decision Making (2010)	-.03	.49**	.04
HP 2: Resources (2010)	.01	.14	.44**
HP 3: Work Control (2010)	.03	.03	-.02
HP 4: Work Practices (2010)	.03	.05	.10
Problem Identification and Resolution (2010)	.14	.17	.46**
PI&R 1: CAP (2010)	.11	.14	.40**
PI&R 2: Operating Experience (2010)	.12	.17	.44**
PI&R 3: Assessments (2010)	.15	.08	.21

*p < .05; **p < .01

6. Alignment of Survey with Safety Culture Policy Statement

6.1. Mapping of Survey Factors to Policy Statement Traits

One of the initial goals of INPO’s safety culture survey was to provide data-driven confirmation for the traits identified in the NRC’s Safety Culture Policy Statement. We reviewed the items comprising each of the factors in INPO’s safety culture survey and compared them to the safety culture traits included in the policy statement. We then created a crosswalk to identify similarities between the policy statement traits and INPO survey factors (see Table 15).

Table 15 Crosswalk of NRC Policy Statement Traits and INPO Safety Culture Survey Factors

NRC Policy Statement Traits	INPO Survey Factors
Leadership Safety Values and Actions: Leaders demonstrate a commitment to safety in their decisions and behaviors.	1. Management Responsibility/Commitment to Safety 3. Decision-Making 4. Supervisor Responsibility for Safety
Problem Identification and Resolution: Issues potentially impacting safety are promptly identified, fully evaluated, and promptly addressed and corrected commensurate with their significance.	1. Management Responsibility/Commitment to Safety (sub-factor b. Continuous Improvement) 3. Decision-Making
Personal Accountability: All individuals take personal responsibility for safety.	7. Personal Responsibility for Safety
Work Processes: The process of planning and controlling work activities is implemented so that safety is maintained.	1. Management Responsibility/Commitment to Safety (sub-factor d. Procedure Communication and e. Resources)
Continuous Learning: Opportunities to learn about ways to ensure safety are sought out and implemented.	1. Management Responsibility/Commitment to Safety (sub-factor b. Continuous Improvement) 9. Training Quality
Environment for Raising Concerns: A safety conscious work environment is maintained where personnel feel free to raise safety concerns without fear of retaliation, intimidation, harassment, or discrimination.	2. Willingness to Raise Concerns
Effective Safety Communication: Communications maintain a focus on safety.	6. Safety Communication
Respectful Work Environment: Trust and respect permeate the organization.	1. Management Responsibility/Commitment to Safety (sub-factor a. Respectful Work Environment)
Questioning Attitude: Individuals avoid complacency and continuously challenge existing conditions and activities in order to identify discrepancies that might result in error or inappropriate action.	5. Questioning Attitude

There was clear correspondence between the Personal Responsibility factor from the survey and the Personal Accountability trait, the Willingness to Raise Concerns factor and the Environment for Raising Concerns trait, the Safety Communication factor and trait, and the Questioning Attitude factor and trait. Some of the factors that emerged from the survey did not have one-to-one correspondence with the Policy Statement traits. The Continuous Learning trait and Training Quality factor were similar, but aspects of continuous learning were also present in the Continuous Improvement sub-factor under Management Responsibility/Commitment to Safety. The factor labeled Prioritizing Safety did not directly relate to any of the traits from the Policy Statement; however, a review of the items suggests that it may be most closely associated with the Personal Accountability trait because the items tend to converge around the idea that individuals in the plant emphasize the importance of nuclear safety. The Management Responsibility/Commitment to Safety factor was the largest with 20 out of the 60 items retained in the survey, and although there were clear similarities between the factor and the Leadership Safety Values and Actions trait, some of the sub-factors were more closely associated with other traits in the policy statement, such as Problem Identification and Resolution, Work Processes, and Respectful Work Environment. Additionally, some of the items comprising the Decision-Making Factor were related to the Leadership Safety Values and Actions trait, whereas others were more similar to the Problem Identification and Resolution trait. The factor analysis results also suggested an important distinction between site management and supervisors with Supervisor Responsibility for Safety emerging as a distinct factor from Management Responsibility/Commitment to Safety. Overall, the results of the exploratory factor analyses provided general support for the Policy Statement traits. Although there was not one-to-one alignment between the survey factors and Policy Statement traits, each trait is represented by one or more of the survey factors.

7. Summary and Discussion

Safety culture has received considerable attention in high reliability industries, including nuclear power operations, because of its potential contribution to ensuring safe performance. Post-event analyses and accident investigations in many different industries have frequently cited weaknesses in an organization's safety culture as significant contributors to event occurrence. Some recent examples of accidents where safety culture was identified as a contributing cause include BP's Texas City refinery explosion in 2005 (Chemical Safety Board, 2007), the Washington Metropolitan Area Transit Authority metrorail collision in 2009 (National Transportation Safety Board, 2010), the Deepwater Horizon oil spill in 2010 (United States Coast Guard, 2011), and the Upper Big Branch mine explosion in 2010 (Mine Safety and Health Administration, 2011).

The primary purposes of INPO's safety culture study were to investigate the factors that comprise the concept of safety culture in the nuclear power industry, assess the extent to which they match the traits identified in the NRC's Safety Culture Policy Statement, and evaluate the relationships between the safety culture factors identified from the survey and other measures of safety performance. Our independent evaluation confirms that the INPO safety culture survey is multidimensional, consists of factors similar to the traits identified in the NRC's Safety Culture Policy Statement, and demonstrates statistically significant relationships with some, but not all, measures of safety performance in the expected directions.

The results of the factor analysis of the INPO survey showed reasonable correspondence between the majority of the survey factors and the policy statement traits. Each trait was represented by at least one factor or sub-factor within the survey. These findings indicate a degree of convergence on what is important about safety culture among the policy statement traits, research theory, and the results of other studies exploring the safety culture construct.

In addition, the INPO safety culture survey factors were significantly correlated with some indicators of safety performance concurrently (in 2010) and some one year after survey administration. When there were significant correlations, the effect sizes were generally consistent with or larger than what has been reported in previous research in other domains. The overall safety culture survey results were significantly correlated with concurrent unplanned scrams, forced outage hours, inspection findings, and cross-cutting aspects. The safety culture survey results were also related to receipt of SCCIs, movement to elevated oversight in the ROP Action Matrix, and number of allegations from licensee personnel in 2011. Based on this study, the INPO safety culture survey factors demonstrated consistent or somewhat stronger correlations with safety performance measures as compared to the data types maintained by NRC that are conceptually related to safety culture. The one exception to this was that, in one instance, the allegations data demonstrated stronger correlations with other measures of safety performance as compared to the safety culture survey data. The number of allegations was moderately correlated with concurrent counts of inspection findings and strongly correlated with concurrent SCCIs. This relationship was also evident when comparing counts of allegations in 2009 to inspection findings and SCCIs in 2010, but did not hold when comparing 2010 and 2011 data.

The INPO's safety culture survey generally demonstrated sound psychometric properties (i.e., reliability and validity) and a factor structure supported by other safety culture research. Construct validity was evaluated by examining the survey's 1) content validity, 2) reliability, and 3) criterion-related validity. There was adequate support for all three criteria. However, these results should be

interpreted within the context of the study, which used data from a cross-sectional survey administered to employees at a single point in time to approximate an organization's safety culture. The cross-sectional survey, combined with the correlational analyses, does not permit conclusions to be drawn regarding the causal relationship between safety culture and safety performance. The results of the INPO study provide some evidence that the survey factors may correlate with concurrent measures of safety performance and may be related to future performance for some measures. Additional longitudinal and experimental research would be required to establish adequate evidence of a causal relationship between safety culture and nuclear power plant outcomes.

Moreover, it is important to recognize that the INPO safety culture survey is only one potential indicator of safety culture and does not constitute a full assessment of a site's safety culture. There are no established thresholds for determining what constitutes a "strong," "nominally acceptable," or "weak" safety culture. Nor do the results presented in this paper attempt to draw any conclusions about appropriate thresholds. Recall that the mean score on the safety culture survey was 5.61, which indicates that respondents generally agreed with the positive statements concerning their organization's safety cultures. There is considerable debate in the research literature about whether a threshold can be defined or is appropriate for evaluating safety culture. It may be that organizations can have significantly different cultures and perform equally well in terms of safety (Reiman, 2007). Edgar Schein echoes this idea, noting that, "In most organizational change efforts, it is much easier to draw on the strengths of the culture than to overcome the constraints by changing the culture" (2010, p. 327).

The INPO study represents a first step in the empirical exploration of relationships between safety culture and other measures of nuclear power plant performance. Additional, ongoing research would be necessary to determine whether the relationships observed are consistent over time, whether the same factors consistently emerge in subsequent survey administrations within the nuclear power industry, and whether different safety culture factors are uniquely related to different aspects of performance. And, finally, the generalizability of this study's results to other sectors within the NRC's regulated communities is unknown. Comparable data, both in terms of a safety culture survey and robust performance metrics, would be necessary to replicate this study in other domains.

8. References

- Baruch, Y. & Holtom, B. (2008). Survey response rate levels and trends in organizational research. *Human Relations, 61*, 1139-1160.
- Beus, J., Payne, S., Berman, M., & Arthur, W. (2010). Safety climate and injuries: An examination of theoretical and empirical relationships. *Journal of Applied Psychology, 95*(4), 713-727.
- Chemical Safety and Hazard Investigation Board. (2007). Investigation Report, Refinery explosion and fire, BP, Texas City, TX, March 23, 2005. [2005-04-I-TX]. Washington, DC.
- Chen, P. & Popovich, P. (2002). *Correlation: Parametric and Non-Parametric Measures*. Thousand Oaks, CA: Sage Publications.
- Christian, M. S., Bradley, J. C., Wallace, J. C., Burke, M. J. (2009). Workplace safety: A meta-analysis of the roles of person and situation factors. *Journal of Applied Psychology, 94* (5):1103–1127.
- Clarke, S. (2006). The relationship between safety climate and safety performance: A meta-analytic review. *Journal of Occupational Health Psychology, 11*(4):315-327.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.
- Cook, C., Heath, F. & Thompson, R. (2000). A meta-analysis of response rates in web- or internet-based surveys. *Educational and Psychological Measurement, 60*, 821–36.
- Cortina, J. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98-104.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 12*, 1-16.
- Denison, D. (1996). What is the difference between organizational culture and organizational climate? A native's point of view on a decade of paradigm wars. *Academy of Management Review, 21*(3), 619-654.
- Flin, R., Mearns, K., O'Connor, P., & Bryden, R. (2000). Measuring Safety Climate: Identifying the Common Features. *Safety Science, 34*, 177-192.
- Ford, J., MacCallum, R., & Tait, M. (1986). The application of exploratory factor analysis in applied psychology: A critical review and analysis. *Personnel Psychology, 39*, 291-314.
- Ganster, D., Hennessey, H. & Luthans, F. (1983). Social desirability response effects: Three alternative models. *Academy of Management Journal, 26*, 321-331.
- Glass, G. & Hopkins, K. (1996). *Statistical methods in education and psychology*, Third Edition. Boston: Allyn & Bacon.
- Guldenmund, F. (2000). The nature of safety culture: A review of theory and research. *Safety Science, 34*(1-3), 215-257.
- Guldenmund, F. (2007). The use of questionnaires in safety culture research – an evaluation. *Safety Science, 45*, 723-743.
- Hinkin, T. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods, 1*(1), 104-121.
- International Atomic Energy Agency. (2006). Application of the management system for facilities and activities: safety guide. IAEA Safety Standards Series No. GS-G-3.1. Vienna: International Atomic Energy Agency.
- International Nuclear Safety Advisory Group. (1986). *Summary Report on the Post-Accident Review Meeting on the Chernobyl Accident*, Safety Series No. 75-INSAG-1, International Atomic Energy Agency, Vienna.

- International Nuclear Safety Advisory Group. (1991). *Safety Culture*, Safety Series No. 75-INSAG-4, International Atomic Energy Agency, Vienna.
- Institute for Nuclear Power Operations. (2004). *Principles for a Strong Nuclear Safety Culture*. Atlanta: Institute of Nuclear Power Operations.
- James, L. R. (1982). Aggregation bias in estimates of perceptual agreement. *Journal of Applied Psychology*, 67, 219-229.
- Jolliffe I. T. (2002) *Principal Component Analysis*, Second edition. Springer Series in Statistics. New York: Springer.
- Kim, J. & Mueller, C. (1978). *Introduction to factor analysis: What it is and how to do it*. Beverly Hills, CA: Sage.
- Krosnick, J. (1999). Survey research. *Annual Review of Psychology*, 50, 537-567.
- McGraw, K. & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1), 30-46.
- Mine Safety and Health Administration. (2011). *Report of Investigation: Fatal Underground Mine Explosion, April 5, 2010*. Coal Mine Safety and Health.
- National Transportation Safety Board. (2010). *Collision of Two Washington Metropolitan Area Transit Authority Metrorail Trains Near Fort Totten Station, Washington, D.C., June 22, 2009*. Railroad Accident Report NTSB/RAR-10/02. Washington, DC.
- Myers, J. & Well, A. (2003). *Research Design and Statistical Analysis*, 2nd Edition, Mahwah, New Jersey: Lawrence Erlbaum Associates.
- Nuclear Regulatory Commission. (1989). Policy Statement on the Conduct of Nuclear Power Plant Operations. *Federal Register*, 54 FR 3424, January 24, 1989.
- Nuclear Regulatory Commission. (1996). Policy Statement: Freedom of Employees in the Nuclear Industry to Raise Safety Concerns Without Fear of Retaliation, *Federal Register*, 61 FR 24336, May 14, 1996.
- Nuclear Regulatory Commission. (2006). Technical Basis for Inspection Program. NRC Inspection Manual, Chapter 0308, Appendix 2. Washington D.C.: United States Nuclear Regulatory Commission.
- Nuclear Regulatory Commission. (2008). Industry Trends Program. NRC Inspection Manual, Chapter 0313. Washington D.C.: United States Nuclear Regulatory Commission.
- Nuclear Regulatory Commission. (2011a). Components within the Cross-Cutting Areas. NRC Inspection Manual, Chapter 0310. Washington D.C.: United States Nuclear Regulatory Commission.
- Nuclear Regulatory Commission. (2011b). Final Safety Culture Policy Statement. *Federal Register*, 76 FR 34733, June 14, 2011.
- Nuclear Regulatory Commission. (2011c). Operating Reactor Assessment Program. NRC Inspection Manual, Chapter 0305. Washington D.C.: United States Nuclear Regulatory Commission.
- Nuclear Regulatory Commission. (2011d). Supplemental inspection for repetitive degraded cornerstones, multiple degraded cornerstones, multiple yellow inputs, or one red input. NRC Inspection Procedure 95003. Washington D.C.: United States Nuclear Regulatory Commission.
- Nunnally, J.C., & Bernstein, I.H. (1994). *Psychometric Theory*, 3rd Edition, New York: McGraw-Hill, Inc.
- Reason, J. (1997). *Managing the Risks of Organizational Accidents*. Ashgate.
- Rogovin, M. (1980). *Three Mile Island – A Report to the Commissioners and the Public*, vol. 1.
- Rosenthal, R. & DiMatteo, M. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology*, 52, 59-82.

- Schein, E. H. (1992). *Organizational Culture and Leadership*, Second edition. San Francisco: Jossey-Bass.
- Schein, E. H. (2010). *Organizational Culture and Leadership*, Fourth edition. San Francisco: Jossey-Bass.
- Schriesheim, C., Power, K., Scandura, T., Gardiner, C., & Lankau, M. (1993). Improving construct measurement in management research: Comments and a quantitative approach for assessing the theoretical content adequacy of paper-and-pencil survey-type instruments. *Journal of Management*, 19, 385-417.
- Sorensen, J. N. (2002). Safety culture: A survey of the state-of-the-art. *Reliability Engineering and System Safety*, 76, 189-204.
- Tague, N. (2004). The Quality Toolbox, Second Edition, Milwaukee, WI: ASQ Quality Press, 96-99. <http://asq.org/learn-about-quality/idea-creation-tools/overview/affinity.html>. Retrieved February 15, 2012.
- Trochim, W. M. (2006). Likert Scaling. *Research Methods Knowledge Base*, Second Edition. <http://www.socialresearchmethods.net/kb/scallik.php>. Retrieved January 24, 2011.
- United States Coast Guard. (2011). *Report of Investigation into the Circumstances Surrounding the Explosion, Fire, Sinking and Loss of Eleven Crew Members Aboard the Mobile Offshore Drilling Unit Deepwater Horizon in the Gulf of Mexico, April 20-22, 2010*. Volume 1. Washington, DC.
- Wiegmann, D., Zhang, H., von Thaden, T., Sharma, G., & Mitchell, A. (2002). *A Synthesis of Safety Culture and Safety Climate Research*. Federal Aviation Administration. [Technical Report FAA-02-2]. Atlantic City, NJ.