MEASUREMENT EQUIVALENCE OF A SAFETY CLIMATE MEASURE ACROSS NATIONAL CULTURES, LANGUAGES, AND WORK ENVIRONMENTS

A Dissertation

by

XIAOHONG XU

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Stephanie C. Payne
Committee Members,	Mindy E. Bergman
	Winfred Arthur, Jr.
	Myeongsun Yoon
Head of Department,	Douglas W. Woods

December 2015

Major Subject: Psychology

ABSTRACT

Given the relevance and importance of safety climate to workplace safety in organizations worldwide, researchers and practitioners recognize the utility of measuring and tracking safety climate, especially in high-reliability organizations. However, sample characteristics or faultlines including national culture, language, hierarchical position, employment arrangement, and work environment create meaningful differences between groups of respondents that may make it inappropriate to compare safety climate scores across groups. Differences were expected to emerge for numerous reasons including the construct relevance of item content, response sets or tendencies to use response scales in a particular manner, the relative strength of item endorsement, and/or the frame of reference used. The purpose of this study was to examine the measurement equivalence of a safety climate measure across five faultlines within an archival dataset containing survey responses from 8,790 multinational chemical processing and manufacturing employees. In order to take the multilevel nature of the data into account, the measurement equivalence of the safety climate measure was examined by using multilevel multi-group CFAs for the Level-3 (national culture) faultline and multilevel factor mixture model for the Level-1 faultlines. Scalar (intercept) equivalence was not established across all five faultlines, whereas metric equivalence held for hierarchical position and employment arrangement. These results suggest that the same safety climate instrument may not be used across different contexts. Results have important practical implications for benchmarking safety climate ratings across studied faultlines.

ACKNOWLEDGMENTS

I would like to thank my committee chair, Dr. Stephanie Payne, and my committee members, Dr. Mindy Bergman, Dr. Winfred Arthur Jr., and Dr. Myeongsun Yoon, for their guidance and support for this research. Their comments and suggestions on my earlier drafts have been very helpful for the completion of my dissertation.

I am thankful to Dr. M. Sam Mannan and T. Michael O'Connor at the Mary Kay O'Connor Process Safety Center for their support of my research and for facilitating the collaboration that lead to the collection of the data examined in the project.

Finally, I want to thank my husband for his encouragement, patience, and love. He has been supporting me in every aspect and has been with me amid all difficulties through the years while we studied together.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

CHAPTER I

INTRODUCTION AND LITERATURE REVIEW

The rapid globalization of organizations and immigration has contributed to the need to assess psychological constructs across multiple faultlines or boundaries (Schmitt & Kuljanin, 2008). That is, researchers and practitioners need valid instruments when collecting data from multinational organizations composed of individuals from various cultures who speak many languages (Bartram & Coyne, 1998; Hu & Oakland, 1991; Oakland, 1997, 2004; Oakland & Hu, 1992). This involves more than simply addressing the issue of the adequacy of translation across languages, but also a demonstration of the measurement equivalence of the measures across different groups (Hui & Triandis, 1985; Ryan, Chan, Ployhart, & Slade, 1999; Vandenberg & Lance, 2000). Measurement equivalence (also referred to as measurement invariance) exists when items measuring the latent construct(s) are interpreted the same way across different groups (Vandenberg & Lance, 2000). Mathematically, measurement invariance "holds" (exists) if the conditional probability of an observed score given a latent construct is independent of the group membership (Meredith & Millsap, 1992; Millsap, 1997). That is, measurement invariance holds when individuals who have the same standing on a latent construct have the same probability of obtaining an observed score regardless of the group membership (Meredith & Millsap, 1992; Millsap, 1997).

Psychologists use a variety of instruments to assess individuals and groups on latent psychological constructs. Before meaningful comparisons between groups can be made, any observed differences that emerge should reflect true differences on the underlying construct of interest. There are a number of measurement-related reasons why different scores may emerge for the groups that do not reflect differences on the latent construct. It is critical to

differentiate true differences from measurement-related differences. Only after measurement equivalence is established can true differences be revealed. Indeed, researchers have recognized the importance of the measurement equivalence (e.g., Millsap & Kwok, 2004; Schmitt & Kuljanin, 2008; Yoon & Millsap, 2007). Based on a literature review on measurement invariance, Schmitt and Kuljanin (2008) noted increased interest and practice in conducting measurement invariance tests, advocating for it as a prerequisite for the use of a measure.

Individuals are defined in part by their demographic and personality characteristics. Individuals who share demographic characteristics are likely to be more psychologically similar to one another than individuals who do not share demographic characteristics (Lau & Murnighan, 1998). Organizational scientists refer to hypothetical dividing lines that split individuals into groups based on one or more attributes (e.g., age, tenure, values) as group faultlines (Lau & Murnighan, 1998). For instance, demographic group faultlines, such as age, divide individuals into groups based on age. Depending on the similarity and salience of individuals' attributes, many different potential faultlines may exist within groups, and each of these faultlines may increase or activate the potential for particular subgroupings (Lau & Murnighan, 1998). The existence of faultlines presents a need to test for measurement equivalence to rule out measurement explanations for any differences that emerge.

There are an infinite number of faultlines for measurement equivalence tests. The decision to examine a given faultline should be driven by theory. For instance, theories (e.g., social comparison theory, response bias, item response theory) suggest that measurement equivalence of the safety climate measures may not hold across various faultlines. Thus, only theoretically-defendable faultlines are examined in the present study. Further, as the present

study used archival data, some theoretically-defendable faultlines (e.g., laws) are not considered due to data unavailability.

In sum, the purpose of this study is to empirically test whether the following faultlines pose threats to the measurement equivalence of a safety climate measure: language, national culture, hierarchical position, work environment, and employment arrangement. The results of this study have practical implications for the organization, such as the extent to which it is appropriate for them to benchmark safety climate scores across studied faultlines within and across organizations.

Measurement Equivalence

Measurement equivalence is a concern across a number of different settings and possible groupings. Vandenberg (2002) identified four specific situations in which it would be important to test for measurement equivalence, including (a) comparing survey responses generated by individuals from different cultures, (b) comparing two raters' evaluations of the same person on the same performance dimensions, (c) comparing people with different demographic characteristics on survey responses, and (d) comparing responses before and after an organizational intervention.

Ignoring group differences is like mixing "apples and oranges." Measurement equivalence is particularly important when the assessor intends to compare groups. Inferences based on observed scores are valid only if the observed scores have the same relationship with the corresponding latent constructs across groups and only if the latent constructs have the same meaning across groups. In this study, independent groups will be examined. Vandenberg and Lance (2000) stated, "if not tested, violations of measurement equivalence assumptions are as threatening to substantive interpretations as is an inability to

demonstrate reliability and validity" (p. 6). In other words, a lack of measurement equivalence between groups indicates that the measure or instrument is not functioning the same way across groups and any substantive interpretation of similarities or differences is meaningless without supporting evidence of measurement invariance. Measurement equivalence is not a property of a measure (Van de Vijver & Leung, 1997). It is specific to or bounded by the groups compared. In other words, measurement equivalence must be interpreted in the context of the interaction between the measure, the sample characteristics, and the characteristics of the administration (e.g., Carter, Kotrba, & Lake, 2014; Robert, Lee, & Chan, 2006; Vandenberg & Lance, 2000; Vandenberg, 2002). For instance, a specific safety climate measure that is invariant between men and women may or may not be invariant for groups that work in different environments or between groups that vary on individualism. Likewise, a safety climate measure that is invariant between males and females, and between individualistic and collectivistic cultures, may or may not be invariant between collectivistic females and individualistic males, or between individualistic females and collectivistic males (e.g., the interaction between sex and individualism). In sum, various potential threats to the interpretation of psychological constructs exist when individuals from different groups are surveyed and compared.

Testing Measurement Equivalence

Measurement equivalence is frequently examined using the differential item functioning (DIF) analysis based on item-response theory or by the confirmatory factor analytic mean and covariance structure analysis. The analysis of measurement equivalence in this study will focus on the equivalence of slopes (i.e., metric invariance) and intercepts (i.e.,

scalar invariance), because they are important for correlational and mean-level inferential analyses (Robert et al., 2006; Vandenberg & Lance, 2000).

Vandenberg and Lance (2000) observed that there are eight primary tests of measurement invariance. Among these tests, there are four common models or measurement invariance tests that should be tested in the following order: configural, metric, scalar, and error variance invariance (Vandenberg & Lance, 2000). Each of these tests provides a subsequently more rigorous level of measurement invariance and will be described in turn.

Configural invariance reveals the extent to which individuals from different groups conceptualize the latent construct the same way in terms of dimensionality (i.e., uni- or multidimensional). The configural properties of a measure consist of the number of dimensions and their meaning which is based on the content of the items. Configural invariance is the first test to conduct when examining measurement invariance and it is satisfied if the basic model/factor structure is invariant across groups (Horn & McArdle, 1992). The test of configural invariance has been referred to as a "baseline" model (Bagozzi & Edwards, 1998; Marsh, 1994; Reise et al., 1993), a test of "equality of factor structures" (Cole & Maxwell, 1985), as well as a test of "equal number of factors and factor pattern" (Taris et al., 1998).

Tests of configural invariance determine whether the number and kinds of factors underlying the item responses are the same across groups. It is tested by constraining the pattern of fixed and free loading items of the measure to be equivalent across groups. Thus, configural invariance requires a demonstration of the same factors and patterns of factor loadings across groups. If configural invariance does not hold, the latent construct is not

conceptualized in the same way across groups. Once configural invariance is established, one can proceed to examine metric invariance.

Metric invariance (i.e., slope equivalence) refers to the equivalence of the slopes in the regression of an item on the latent construct variable or the quality of scaling units across groups (Joreskog, 1969; Schmitt, 1982; Vandenberg & Self, 1993). Slope equivalence is important for establishing the validity of both correlational and mean-level analyses.

Metric invariance is examined by constraining the factor loadings of items to be equal across groups. Factor loadings are the regression slopes relating the items to their corresponding latent dimensions/variables, and represent the expected change in the observed score on the item per unit change on the latent dimension/variable. Metric invariance holds if the strength of the relationships between specific items and the underlying construct are the same across groups. At least partial metric invariance (relaxed invariance constraints) must be established before continuing the sequence of measurement equivalence tests (Vandenberg & Lance, 2000).

Scalar invariance (i.e., intercept equivalence) indicates that individuals from different groups, with equal standing on the latent construct, interpret the scale anchors the same way. For example, two individuals from different groups assign the same meaning to the scale anchor "strongly agree." Some scholars interpret the test of scalar invariance as a test for systematic response bias (e.g., leniency) between groups (Bollen, 1989) when latent group mean differences are not expected. On the other hand, intercept differences could also reflect response threshold differences that might be predicted based on known group differences. An examination of scalar invariance is critical before drawing conclusions about observed mean differences on the latent construct between groups.

Operationally, scalar invariance indicates the intercepts of the items' regressions on the latent construct(s) are invariant across groups. Intercept equivalence is potentially relevant for mean-level comparisons. Thus, this is the last model necessary to make meaningful comparisons of observed scores across groups.

Strict invariance, or error variance invariance, means factor loadings, intercepts, and unique variance (i.e., the amount of measurement error) are invariant across groups. The strict invariance test should only be conducted if (at least partial) metric and scalar invariance has been established first (Vandenberg & Lance, 2000). When strict invariance holds, observed group differences on means or covariances are due to true group differences on the latent construct(s). Strict invariance is a highly constrained model and often does not hold. Even if all samples come from a common population with given error variances, it would be expected that error variances would vary from one sample to another.

Testing Measurement Equivalence in Multilevel Data

Multilevel data have a hierarchical structure in which individual observations are nested within clusters. In multilevel data, Level 1 refers to the lowest level in the nested structure, Level 2 refers to the next level within which Level-1 observations are nested. As group membership can exist at the individual level (i.e., Level 1; e.g., hierarchical position), a higher level (i.e., Level 2; work site/location), or an even higher level (e.g., Level 3; country), measurement invariance tests can be conducted at these corresponding levels (Mehta & Neale, 2005). Neglecting the dependence of observations leads to underestimated standard errors in statistical significance tests resulting in incorrect statistical inferences (i.e., Type I error or an increased likelihood of finding measurement non-equivalence when there is equivalence). For instance, Kim et al. (2012) demonstrated that there is substantial Type I

error rate inflation when using the single-level multi-group ordinary confirmatory factor analysis (CFA) for measurement invariance tests in multilevel data, which ignores the hierarchical data structure and assumes independent observations. In other words, the measurement invariant model is more likely to be rejected and misleadingly interpreted as being non-equivalent when the non-independent observations are not taken into account.

Although several methodological studies have discussed and explored measurement invariance issues in multilevel data (e.g., Curran, 2003; Jones-Farmer, 2010; Kim, Kwok, & Yoon, 2012; Kim, Yoon, Wen, Luo, & Kwok, in press; Mehta & Neale, 2005; Selig, Card, & Little, 2008; Ryu, 2014; Zyphur, Kaplan, & Christian, 2008), the influence of the dependence of observations on measurement invariance tests have been ignored in organizational research. Although researchers have started to examine the measurement invariance of different constructs of interests across different faultlines, few researchers have considered the nested nature of the data (e.g., Oreg et al., 2008; Woehr, Sheehan, & Bennett Jr., 2005). For instance, Woehr et al. (2005) examined the measurement invariance of performance ratings across rating sources but ignored the nested nature of their data, which consisted of 1,028 airmen from seven different Air Force job categories. As a result, their conclusions of measurement equivalence may not be valid.

In the present study, individuals are nested in organizational sites, which are nested in countries. This hierarchical nesting will be taken into account when examining the measurement invariance of the safety climate measure.

Safety Climate

Occupational safety is an issue associated with significant financial and societal consequences. For instance, in 2013, there were over four million non-fatal work injuries,

and more than 4,405 work fatalities reported in the United States (Bureau of Labor Statistics, 2013). The most disabling workplace injuries and illnesses in 2011 amounted to $55.4 billion in direct U.S. workers compensation costs (Liberty Mutual Research Institute for Safety, 2013). A review of the research literature suggests that safety climate is one of the most important factors that contribute to workplace safety. Safety climate is defined as shared employee perceptions of organizational policies, practices, and procedures regarding safety (Zohar, 2003). Empirical studies have supported that safety climate predicts safety behaviors and safety-related outcomes, such as accidents and injuries (e.g., Beus, Payne, Bergman, & Arthur, 2010; Christian, Bradley, Wallace, & Burke, 2009; Nahrgang, Morgeson, & Hofmann, 2011).

Measurement of Safety Climate

Although workplace safety researchers agree that safety climate plays an important role in safety behaviors and safety-related outcomes in the workplace, researchers have not always agreed on the definition and measurement of the construct. As a result, numerous measures and articles exist about the measurement of safety climate. For instance, Flin, Mearns, O'Connor, and Bryden (2000) identified more than 20 empirically tested safety climate measures capturing more than 50 different conceptual themes (Guldenmund, 2000) reflecting the state of development of this construct (Flin et al., 2000).

One of the measurement issues raised about safety climate is its multidimensional structure and the lack of consensus over the number and names of the dimensions (Flin et al., 2000; Guldenmund, 2000). In a review of the studies published at the time, Flin et al. (2000) identified management commitment to safety, safety systems, risk, work pressure, and competence as the most common safety climate dimensions. Beus et al. (2010) argued that

risk and competence are contaminated dimensions that should not be included in a safety climate measure. The one dimension that most safety climate researchers agree upon is management commitment to safety (Flin et al., 2000).

Despite the proliferation of measures, safety climate researchers have called for the development of even more measures. For example, in a review of 30 years of safety climate research, Zohar (2010) noted the development of new scales "should also be encouraged as it is likely to identify new, context-dependent targets of climate perceptions" (p. 1521). This suggests that the manifestation of safety climate varies across contexts, or employees from different contexts may not interpret a safety climate measure in the same way. This raises a very important measurement issue: can the same safety climate instrument be used across different contexts?

Measurement Equivalence of Safety Climate

Among the various measurement issues raised about safety climate, measurement equivalence has received a fair amount of recent research attention. Six publications have examined the measurement equivalence of a safety climate measure across four datasets (Beus, Jarrett, Bergman, & Payne, 2012; Cheyne, Thomas, Cox, & Oliver, 2003; Cigularov, Adams, Gittleman, Haile, & Chen, 2013; Cigularov, Lancaster, Chen, Gittleman, & Haile, 2013; Huang, Robertson, Lee, Murphy, Garabet, & Dainoff, 2014; Lee, Huang, Murphy, Robertson, & Garabet, in press). First, Cheyne et al. (2003) examined measurement equivalence across hierarchical groups (i.e., managers, supervisors, and general employees), and revealed that metric invariance, as well as the scalar invariance of a safety climate measure did not hold across these three hierarchical groups.

Beus et al. (2012) examined the configural and metric invariance of a safety climate measure across hierarchical positions (i.e., front-line employee, supervisor/manager) and organizational heritage (i.e., Company X, Company Y, direct hire, contractor) in the same data used in the present study, and the results indicated that the safety climate measure achieved configural but not metric invariance in the majority of comparisons they made. Cigularov, Adams et al. (2013) demonstrated scalar invariance of a safety climate measure between ten construction trade groups. Cigularov, Lancaster, et al. (2013) revealed metric invariance but not scalar invariance of a safety climate measure for groups based on race and language, specifically between White English-speaking, Hispanic English-speaking, and Hispanic Spanish-speaking respondents.

In the most recent publications, Huang et al. (2014) provided evidence for scalar invariance of safety climate between supervisors and employees; Lee et al. (in press) examined the measurement invariance of a safety climate measure across multiple companies using samples of truck drivers.

Whereas these studies are informative, they provide an incomplete assessment of the measurement equivalence of a safety climate measure as existing theories (e.g., social comparison theory, item response theory) suggest that there are a number of other important faultlines that may theoretically threaten the measurement invariance of the safety climate measures and that have not been empirically tested within the research literature. Further, they do not take into consideration the hierarchical nature of the data examined (e.g., Beus et al., 2012). Thus, the purpose of this study is to test the measurement equivalence of a safety climate measure across the following four faultlines in addition to the previously examined hierarchical position: language, national culture, work arrangement, and work environment,

while taking into consideration the hierarchical nature of the data, further contributing to the empirical research on the measurement of safety climate. The following sections describe the possible threats for each faultline and the underlying theory that supports its relevance.

National Culture

Within the organizational culture and climate literatures, national culture has been proposed as a key determinant of organizational culture or climate, as well as the nature and effectiveness of the human resource management practices in the organizations (Schneider, 1988). Organizations function within a cultural context, regardless of whether the context is defined in terms of shared meanings, values, and assumptions or observable rites and rituals (Kopelman, Brief, & Guzzo, 1989). Although regional differences may exist within nations, national culture has been shown to uniquely influence organizations and human resource management practices (Huang & van de Vliert, 2004; Ryan, McFarland, Baron, & Page, 1999). Thus, organizational climate is expected to be influenced by national culture (Riordan & Vandenberg, 1994). Organizational climate per se is meaningless without attaching a referent (Schneider & Reichers, 1983). Organizations have numerous climates and these climates are all "for something." Safety climate is one of the most studied organizational climates in the research literature (Zohar, 1980, 2010).

The establishment of new constructs and measures is typically done in the context of a specific culture. For instance, the establishment of the safety climate construct and the first measure of safety climate was done in Israel (Zohar, 1980). Specifically, Zohar's 40-item measure of safety climate was constructed and validated in a stratified sample of 20 organizations in Israel (Zohar, 1980). Since then, safety climate has been assessed in numerous other organizations in a wide range of countries. However, there do not appear to

be any studies that have examined whether the safety climate measure operationalizes the same safety climate construct across more than one national culture.

Theoretically, safety climate is a universal phenomenon that applies to any organization in any national culture or country. However, the equivalence of the meaning of the construct across cultures should not be taken for granted. Cultural differences can have a signficant impact on environment, health, and safety, and safety climate (Huang & Fu, 1999). For instance, different cultures may assign a different amount of importance to safety, have different attitudes toward risk taking, different religious beliefs that influence their locus of control, and have different norms when it comes to adhering to safety rules and procedures (Huang & Fu, 1999). Therefore, the meaning and manifestation of safety climate may not be equivalent across national cultures.

One of the most widely used frameworks for examining cultural differences was developed by Hofstede (1980, 1991; Triandis, 2004). Other researchers have proposed similar or related frameworks for studying cultural differences (e.g., House et al., 2004; Inglehart & Baker, 2000; Schwartz, 1999). However, Hofstede's dimensions have generally been shown to more stable over time than other frameworks (Barkema & Vermeulen, 1997; Sondergaard, 1994). Thus, the present study used Hofstede's dimensions to operationalize national culture.

Hofstede (1980) proposed four dimensions (i.e., collectivism-individualism, power distance, masculinity-femininity, and uncertainty avoidance) to classify societies. Subsequently, Hofstede (1991) added a fifth dimension, long-versus short-term orientation, to his earlier four dimensions. Among the five dimensions, the individualism dimension can be conceptually and empirically linked to many other identified cultural dimensions (e.g.,

power, femininity; for a review, see Blondel & Inoguchi, 2006) and has been empirically tested the most (e.g., Oyserman Coon, & Kemmelmeier, 2002; Oyserman & Lee, 2008; Triandis, 2004; Voronov & Singer, 2002). Studies of the individualism dimension have generated important insights into psychological processes (for a review, see Oyserman et al., 2002), and psychological outcomes of interest (e.g., values, self-concept, relationality, cognitive processes) (Oyserman & Lee, 2008).

The present study examines individualism and operationalizes it based on Hofstede's (1980, 1983) cultural dimension index values (see Table 1, the index values for other cultural dimensions are also provided) assigned to the country that the respondents indicated they were in when they completed the measure. It is important to acknowledge that cultural boundaries do not perfectly correspond to the geographical boundaries of nations, and some nations may comprise several subcultures (Schwartz, 1999).

Strong forces towards integration may produce substantial sharing of the culture within the nations (Hofstede, 1990; Schwartz, 1999). Nations typically have one single dominant official language; educational, military, and political system; and shared mass media, markets, services, as well as national symbols (e.g., flags, sports teams) such that each nation typically has one dominant culture, with core attributes that are shared among its subcultures (Liu, Borg, & Spector, 2004; Schwartz, 1999). Accordingly, the measurement invariance of the safety climate may also be detectable at the country level.

Table 1 Cultural Values for Individualism, Uncertainty, Power Distance, Masculinity, and Long/Short-Term Orientation from Hofestede (1980)

Country	*N*	Individualism	Uncertainty	Power Distance	Masculinity	Long/Short-term Orientation
Brazil	644	38	76	69	49	65
Canada	597	80	48	39	52	NA
China	777	20	40	80	66	118
Germany (F.R.)	572	67	65	35	66	31
Mexico	1306	30	82	81	69	NA
Netherlands	295	80	53	38	14	44
Singapore	312	20	8	74	48	48
Taiwan	209	17	69	58	45	87
United Kingdom	488	89	35	35	66	25
United States	3361	91	46	40	62	29

Individualism

The individualism dimension refers to the degree to which individuals are integrated into a group (Hofstede, 1980; Triandis, Leung, Villareal, & Clack, 1985). Triandis (1989) defined individualists as the individuals who "give priority to personal goals over the goals of collectives" and collectivists as individuals who "either make no distinctions between personal and collective goals, or if they do make such distinctions, they subordinate their personal goals to the collective goals" (p. 509). In individualistic societies, the ties between individuals are loose such that individuals tend to focus on themselves and their immediate families, with emphasis on individual initiative, self-sufficiency as well as individual accomplishment. In contrast, in collectivistic societies, individuals are integrated into strong and cohesive groups, emphasizing cooperation, group welfare, duty, security as well as stable social relationships (e.g., Hofstede, 1991; Triandis et al., 1985).

National culture influences the implementation of organizational policies, practices, and procedures, including those concerned with safety (e.g., Cigularov et al., 2013b; Janssens, Brett, & Smith, 1995) thus it is likely to affect safety climate (Flin et al., 2000; Janssens et al., 1995). Research supports that the individualism dimension exerts strong influences on management practices (e.g., Adler, 1986; Hofstede, 1992; Janssens et al., 1995). For instance, in individualistic societies, the same standards are applied to all employees, as employees are all viewed as potential resources such that tasks are considered as more important than relationships. That is, the relationship between employer and employee is "calculative" in individualistic societies (Hofstede, 1992) and the corresponding managerial style is likely to be directive or authoritarian (Kerr, Dunlop, Harbison, & Myers, 1960). In contrast, in collectivistic societies, employees are viewed as members of a group in

the organization such that the standards applied to in-group members are different from the standards applied to out-group members. Relationships are more important than tasks in collectivistic societies (Hofstede, 1992). The managerial style in such countries can be described as directive but welfare-oriented or paternalistic (Kerr et al., 1960).

Hofstede (1980, 1983) conducted pioneering research to systematically map 53 countries onto the individualism dimension. Meta-analytic studies on cross-cultural research have supported Hofstede's work indicating relatively higher scores of individualism in Western European nations (e.g., France and Spain) relative to Northern and Eastern European countries (e.g., Norway and Finland; e.g., Oyserman et al., 2002).

The individualism dimension of the national culture is expected to influence the measurement of safety climate. Specifically, the individualism dimension may influence the construct relevance of item content, drive response sets or tendencies to use response scales in a particular manner, or influence the relative strength of item endorsement. This is also consistent with the research finding that culture plays an important role in occupational safety and health (Burke, Chan-Serafin, Salvador, Smith, & Sarpy, 2008) and the development and perception of safety climate (Peckitt et al., 2004; Rochlin & von Meier, 1994).

It is anticipated that the measurement invariance of the safety climate measure will not hold across the individualism faultline for numerous reasons elaborated upon below.

Slope Equivalence Threats

Metric invariance (i.e., slope equivalence) of the safety climate measure is not expected to hold, because items may vary in their relevance to the construct across national cultures and extreme response styles differ across national cultures.

Item relevance. One reason why slopes of safety climate items are not expected to be equivalent across cultural groups is because of a lack of item relevance. The etic approach to cross-cultural studies involves administering a measurement tool that measures a specific latent construct derived in one culture to a different culture and assuming the measure is universally applicable to all cultures (Berry, 1969; Church, 2001; Hulin, 1987; Triandis & Marin, 1983). However, researchers have long criticized the etic approach to the cross-cultural measurement of latent construct(s) (e.g., Berry, 1969; Church, 2001; Triandis & Marin, 1983), as research has demonstrated that the same measure may actually assess different constructs across national cultures (e.g., Hambleton, 2005; Poortinga, 1995; van de Vijver & Leung, 1997). This may be due to the variation in the definitions of the constructs or the differential appropriateness of the behaviors or indicators associated with the latent construct (e.g., Byrne et al., 2009; Poortinga, 1995; van de Vijver & Leung, 1997; van de Vijver & Tanzer, 1997). As a result, the items developed in one culture may not tap the relevant indicators of the latent construct, thus important culture-specific indicators may be missed for the new cultural setting. For instance, Hoshmand and Ho (1995) found that social aspects are more important aspects of the Chinese conception of "self" compared to the conceptualization of "self" in Western cultures.

Some safety climate researchers have speculated that variation in the concept of safety climate across cultures could be one factor that contributes to the lack of consensus concerning the number of factors in the measures of safety climate (Lin, Tang, Miao, Wang & Wang, 2008). Lin et al. (2008) found that the concept of safety climate in China (a collectivistic country) emphasizes safety awareness and competence which explained the largest variance in safety climate, whereas in the U.S. safety climates emphasize manager

commitment to safety (e.g., Beus et al., 2010; Flin et al., 2000). Similarly, Ma and Yuan (2009) found that Chinese employees had strong perceptions of safety competence and employee (rather than management) commitment to safety.

Further, empirical studies support the influence of national culture on item relevance to the latent construct (e.g., Bryne & Campbell, 1999; Tanzer, 1995). This has been demonstrated with measures of depression (Bryne & Campbell, 1999) and self-concept (Tanzer, 1995). Some safety climate items may not be relevant to all employees. For instance, safety committee meetings are a common practice in Mainland China (Zhou, Fang, & Wang, 2008). When the construct relevance of item content differs across national cultures, the slope of the item will not be equivalent across national cultures as the specific item will not discriminate equally between different levels of the latent construct (i.e., the factor loadings will vary across national cultures) (Robert et al., 2006).

Extreme response style[1]. Another reason why slopes of safety climate items are not expected to be equivalent across cultural groups is because of extreme responding. Cronbach (1946) defined response styles as "any tendency causing a person to consistently give a different response to test items than he (sic) would when the same content is presented in a different form" (p. 475). Among various response styles, extreme response style, and acquiescent response style are the most common (e.g., Schwarz & Oyserman, 2001).

Extreme response style is the tendency for respondents to use the extreme ends of the rating scales. For instance, individuals with a high extreme response style will tend to select

[1] It is important to note that a researcher's perspective on faultlines may be influenced by his/her own culture, where he/she was educated, and by the empirical research to date which has been largely conducted in North American and Europe (Triandis, 1994). For instance, collectivists might interpret social desirability as an effort to be" harmonious" and acquiescence as "humility".

either one (strongly disagree) or five (strongly agree), when responding to items using a 5-point Likert scale (Cronbach, 1950). Between-group differences in extreme response style can be either non-uniform or uniform. Only a subset of items is affected in a non-uniform extreme response style difference, whereas in uniform extreme response style, all items are affected. Some researchers propose that extreme response style can be quantified by the scale's standard deviation, as extreme response style is highly correlated with the standard deviation (Greenleaf, 1992; Hui & Triandis, 1985). However, extreme response style is not identical to the standard deviation (Greenleaf, 1992). Cheung and Rensvold (2000) illustrated that extreme response style results in nonequivalence of slope (i.e., factor loading). Because empirical studies have documented that ERS varies across cultures and nations (e.g., Bachman & O'Malley, 1984; Greenleaf, 1992; Shulruf, Hattie, & Dixon, 2011; Triandis, 1994), it is expected that the slope of the safety climate measure will not be equivalent across national cultures. In summary, because researchers use the imposed etic approach to develop safety climate measures and culture is related to specific response sets and/or response styles (e.g., Cheung & Rensvold, 2000), it is anticipated that slope equivalence will not hold across national cultures, operationalized as individualistic and collectivistic countries.

Intercept Equivalence Threats

Scalar invariance (i.e., intercept equivalence) of the safety climate measure is not expected to hold because multiple response styles (extreme response style, acquiescence), social desirability bias, and frames of reference are likely to differ across national cultures.

Extreme response style. In addition to slope nonequivalence, extreme response style is also expected to result in nonequivalent intercepts (Cheung & Rensvold, 2000). Empirical studies have documented cross-cultural differences in extreme response style (e.g., Bachman

& O'Malley, 1984; Greenleaf, 1992; Shulruf, Hattie, & Dixon, 2011). For instance, several empirical studies have demonstrated that individuals in collectivistic societies tend to avoid extreme responses and use the midpoint on the scales compared to individuals in the individualistic societies (e.g., Lee & Green, 1991; Triandis, 1995). That is, culturally-based response sets lead to a different degree of endorsement in responding to the items of a scale (Guptara, Murray, Razak, & Sheehan, 1990; Hui & Triandis, 1989; Marin, Gamba, & Marin, 1992; Zax & Takahashi, 1967). This results in the nonequivalence of intercepts across cultural groups and a threat to scalar invariance.

Acquiescence response style. Another response style that varies by culture and expected to influence intercept equivalence is the acquiescence response style (also known as acquiescence bias or agreement bias) which refers to the tendency to agree with an item regardless of the content (Billiet & McClendon, 2000; Couch & Keniston, 1960; Cronbach, 1960). Like extreme response style, differences in ARS across cultures can be either nonuniform (affecting responses to some items) or uniform (affecting responses to all items). ARS threatens the validity of the measurement scores, as it is a source of "correlated errors" that bias the measurement scores. ARS marks the true relationships among the items by falsely increasing the strength of the intercorrelations among the items that are worded in the same direction (e.g., Winkler et al., 1982). Nonequivalence due to ARS occurs when one group systematically gives more acquiescence responses than another group regardless of the item content, resulting in a scale displacement (Mullen, 1995).

Studies have documented ARS differences across cultures and nations (Bachman & O'Malley, 1984; Cunningham, Cunningham, & Green, 1977; England & Harpaz, 1983; Grimm & Church, 1999; Marin et al., 1992; Morris & Pavett, 1992; van Herk, Poortinga, &

Verhallen, 2004). For instance, several studies support that collectivists tend to acquiesce more than individualists (Johnson, Kulesa, Cho, & Shavitt, 2005; van Herk et al., 2004). In measurement invariance terms, ARS differences across national cultures lead to nonequivalence of item intercepts.

Social desirability bias. Another construct expected to vary by culture and influence intercept equivalence is the social desirability bias which refers to the tendency to both consciously and unconsciously respond in a way that is socially acceptable based on cultural norms (e.g., Paulhus & Reid, 1991). Johnson and van de Vijver (2003) defined social desirability bias "as the tendency of individuals to 'manage' social interactions by projecting favorable images of themselves, thereby maximizing conformity to others" (p. 194). Crowne and Marlowe (1960) found that individuals' need for approval was associated with their conformity, sensitivity to norms, as well as social influence. Social desirability bias is a serious concern in the measurement of the latent construct because of its potential to introduce response bias (Johnson & van de Vijver, 2002; Paulhus, 1991).

Social desirability consists of two dimensions: impression management and self-deception (e.g., Paulhus, 1984). Impression management refers to a tendency to intentionally distort one's response in order to be viewed favorably by others. In contrast, self-deception refers to an unintentional propensity to portray oneself in a favorable light. The distinction between impression management and self-deception is that impression management involves intentional manipulations of one's image in the eyes of outsider beholders, whereas self-deception involves some unconscious attempts to maintain a positive self-image. Both dimensions are relevant to the national culture faultline (i.e., individualism), because empirical studies support that collectivists are more likely to manage impressions as well as

engage in self-deception compared to individualists, which in turn may influence the measurement equivalence of any construct across cultural groups (e.g., Heine & Lehman, 1997; Johnson, 1998; Lalwani et al., 2006).

Cross-cultural studies provide indirect and direct evidence that individualism is significantly related to the social desirability bias (e.g., Heine & Lehman, 1997; Lalwani, Shavitt, & Johnson, 2006; Triandis, 1995; Triandis et al., 2001; Triandis & Suh, 2002). For instance, empirical studies support that collectivism is significantly related to deception (Triandis et al., 2001), lying (Triandis & Suh, 2002), and face-saving behaviors (Triandis, 1995) that are associated with social desirability. Similarly, Triandis (1995) found that honesty in interactions with strangers is more valued by individualists than collectivists; van Hemert, van de Vijver, Poortinga, and Georgas (2002) found that there is a significant negative relationship between the individualism scores and the Lie scale scores of the Eysenck Personality Inventory (Eysenck & Eysenck, 1964).

There is also direct evidence supporting a significant positive relationship between collectivism and social desirability bias. For instance, Asians tend to be higher on self-enhancement (impression management) than Westerners (e.g., Heine & Lehman, 1997; Lalwani et al., 2006). Likewise, Johnson (1998) found that there is a positive relationship between social desirability bias and collectivism scores (.20) and a negative correlation between social desirability bias and individualism scores (-.19). Taken together, the literature supports that collectivists are more likely to engage in socially desirable responding than individualists resulting in a threat to scalar invariance.

Frame of reference effect. A final threat to intercept equivalence is the frame of reference effect. Social comparison theory proposes that individuals understand themselves

and evaluate their perceptions, attitudes, values, and beliefs by comparing themselves with similar others (Festinger, 1954). Further, the reference group an individual uses affects one's evaluations. This is known as the frame of reference effect (see Robert et al., 2006) or the reference-group effect (see Heine et al., 2002). The frame of reference effect occurs without explicitly asking respondents to compare themselves to others. Individuals use the comparison group with which they are familiar rather than a global comparison group (see Heine et al., 2002). Thus, to the extent that two groups differ in their average level on the dimension/construct under question or differ in the standards/norms by which members of those groups are evaluated, the frame of reference effect will occur and threaten the measurement equivalence of the measures of interest across groups (Heine et al., 2002). Following this logic, individuals from similar cultures should have similar reference groups. That is, individuals from individualistic cultures have highly individualistic reference groups, and their responses may be based on a similar standard (see Heine et al., 2002).

Biernat and colleagues (Biernat & Billings, 2001; Biernat & Kobrynowicz, 1997; Biernat & Manis, 1994; Biernat, Manis, & Nelson, 1991) propose that the frame of reference effect leads to shifting standards when evaluating individuals from different groups. For instance, Hyman (1942) illustrated how one's status on a particular dimension (e.g., intellectual) is determined by comparing oneself to his or her reference group and Sherif (1936) indicated that individuals judge the apparent movement of a light in a room largely based on how their reference group is viewing it. Thus, to the extent that two groups differ on the latent construct, different standards may be applied.

Likert scales are particularly subject to the shifting standards effect (Biernat et al., 1991), as they fail to provide a context-free measure of individuals' absolute standing on the

latent construct. Individuals will respond to Likert response scales relative to similar others or shared norms based on their life experience. In other words, respondents will match the range of the latent construct to what they expect in order to set the endpoints of the Likert rating scale (Volkmann, 1951). For instance, when evaluating whether a man is tall, respondents will likely set the endpoints of the scale to capture a higher range than when evaluating a woman. Thus, the shifting standards individuals use when responding to subjective Likert scales are likely to obscure the true differences on the latent construct between groups. Safety climate is traditionally assessed on a Likert scale (Flin et al., 2000; Guldenmund, 2000; Zohar, 2003), thus it is possible that safety climate measures are subject to a frame of reference effect.

The significance of reference groups or group membership is one of most important conceptual distinctions between individualism and collectivism (e.g., Chen & West, 2008; Triandis, 1995). The boundary between in-group and out-group is sharper among collectivists compared to individualists (Triandis, 1995). Bond and Smith's (1996) meta-analysis supports that collectivists are more likely to conform to in-group members than out-group members. This suggests that collectivists are more likely to be influenced by in-group members or the frame of reference effect. Indeed, empirical studies support that individuals in one culture tend to compare themselves with different referents and standards than individuals in another culture (e.g., Heine, Lehman, Peng, & Greenholtz 2002; Peng et al., 1997).

Likewise, Heine et al. (2002) indicate that cultures can influence the relative strength of item endorsement because one's evaluation or perception of one's standing on an item is interpreted with reference to relevant social groups. Taken together, the literature suggests

that the confounding effect of the frame of reference effect is problematic for the comparisons of subjective constructs across groups (e.g., Heine et al., 2002; Peng et al., 1997).

Language

Translated measures sometimes fail to capture the intended latent construct to the same degree as the original measures (e.g., Schmitt & Kuljanin, 2008; van de Vijver & Tanzer, 1997). Even when a measure is translated by a group of professional translators using the translation-back translation method, it may not share the same psychological meaning as the original measure (e.g., van de Vijver & Tanzer, 1997). Individuals with different mother tongues have different cultural backgrounds. Therefore, measures administered in different languages are almost always multicultural (Liu, Borg, & Spector, 2004; Tanzer, Sim, & Marsh, 1992). That is, language is strongly related to culture. Hence, the mechanisms by which culture is hypothesized to threaten the measurement equivalence of the safety climate measure (e.g., frame of reference effect) are likely to be the same mechanisms by which language threatens measurement equivalence.

Further, translation may result in discrepancies in the meaning of the construct measured between the original and the translated measure. Language is also associated with ambiguities in the original item, low levels of familiarity/appropriateness of the item content in certain cultures, and cultural-specific nuisance factors or connotations associated with the item wording (Robert et al., 2006; van de Vijver & Tanzer, 1997). Therefore, the slope and intercept of items that measure safety climate will not be equivalent across linguistic groups.

Slope Equivalence Threats

The slope equivalence of the safety climate measure (i.e., metric invariance) may not hold due to mistranslation of idioms, colloquialisms, and metaphors (cf. Robert et al., 2006), or due to translators' inability to accurately translate the idioms, colloquialism, metaphors etc.

Intercept Equivalence Threats

The intercept equivalence of the safety climate measure may not hold across languages, as the words with nonspecific meaning may be mistranslated and/or low language proficiency may lead to a misunderstanding of Likert scales.

Mistranslation of ambiguous words. Response anchors often contain words that have nonspecific meaning [e.g., frequency (e.g., rarely), quantity (e.g., many), probability (e.g., likely), and evaluation terms (e.g., good)] (Brislin, 1980). It is difficult to translate these words accurately. Thus, the translations of these words likely fail to accurately reflect the original language version, possibly resulting in more or less extremes in the translated version. Differences in the extremity of the response options can influence the equivalence of item intercepts across groups.

Also for the translated Likert scales, even the closest semantic translation of the rating categories is likely to change their psychological meaning (Tanzer, 1995). For example, it is hard to find a German equivalent for the second and third category of the English intensity ratings "not at all," "somewhat," "moderately," "quite a bit," and "very much" that preserves the same notion of ordering for the respondents with different cultural backgrounds (Tanzer, 1995). Thus, items and response scales containing words with nonspecific meanings will lead

to nonequivalent item intercepts across linguistic groups, as individuals responding to items in different languages will differentially endorse these items and response options.

Hierarchical Position

Safety climate is likely to differ between hierarchical positions within the organization, because daily work demands and experiences are likely to influence individuals' perceptions of safety climate (e.g., Cox & Cheyne, 2000; Glendon & Litherland, 2001; Harvey, Bolam, & Gregory, 1999). In this study, managers are at the top of the hierarchy, followed by supervisors, and then subordinates. Managers usually develop safety policies and procedures, supervisors usually enforce them, and subordinates are required to follow them. Given this, managers and supervisors are more likely to perceive safety climate as it *should* be; that is, as espoused in formal written policy, rather than how it is enacted by supervisors and experienced by subordinates.

A similar pattern is likely for safety training with managers developing training, supervisors sometimes delivering training, and subordinates receiving the training. Thus, supervisors/managers may perceive safety climate differently than their subordinates. Empirical studies have demonstrated differences in safety attitudes and climate across occupational levels (i.e., supervisors/managers and employees) within the same organization (e.g., Cheyne et al., 2003; Cox et al., 1998; Gittleman et al., 2010; Huang et al., 2014). For instance, Harvey, Bolam, and Gregory (1999) found that the conception of safety differs between managers and employees as evidenced by differing factor structures. Additionally, Cheyne et al. found that managers had the most positive perceptions of the safety climate, followed by supervisors, and then subordinates.

Beus et al. (2012) explored the configural invariance and metric invariance between managers/supervisors and subordinates/dependent contractor using the same archival data used in the current study. The present study extends Beus et al.'s (2012) study in three ways. First, the nested nature of the data (i.e., individual observations are nested within worksites/locations which are nested within countries) is modeled to reduce Type I error rate (e.g., Kim et al., 2012; Kim et al., in press). Second, in addition to configural and metric invariance, scalar invariance is also examined which is critical before drawing conclusions about mean differences between groups on the latent construct. Third, the hierarchical position faultline is examined separately from the employment arrangement faultline and the groups within each are differentiated more finely. In the current study, three hierarchical positions are compared (managers, supervisors, and subordinates) and three employment arrangements are compared (employees, dependent contractors, and independent contractors).

Slope Equivalence Threats

The factor loadings of the safety climate items (i.e., slope equivalence) will vary across hierarchical positions, because the relevance of the items to the latent construct of safety climate varies across different hierarchical positions.

Item relevance. Measurement equivalence at the item level influences measurement equivalence at the scale-level (Chan, 2000). Measurement equivalence of item responses holds when the numerical values across groups are on the same measurement scale (Drasgow, 1984; 1987). In other words, an item is not equivalent (i.e., the item functions differentially) across groups when individuals from different groups with equal standing on the latent construct respond differently to that item. Differential item functioning (i.e., item

bias) is due to either differential item difficulty or differential item discrimination (Chan, 2000; Mellenbergh, 1982; 1994). Conceptually, item relevance refers to the extent to which the item is able to distinguish between respondents with high scores and those with low scores on the latent construct. The higher the discrimination parameter, the more the item is able to distinguish where the individual falls on the latent construct. Hierarchical position may influence item discrimination. Chan (2000) proposed that when the concreteness of the items of a scale (i.e., item discrimination/relevance) differs across groups, item slopes will not be equivalent between groups. Some safety climate items seem like they would be less ambiguous to individuals in management positions (i.e., the association between the item and the latent construct [i.e., item factor loading] will be higher for individuals in a management position). For example, for the item "Site management considers health and safety when setting production rates and schedules" the process leading up to setting production rates and schedules is presumably more concrete (i.e., less ambiguous) to individuals in management positions (i.e., managers), because as indicated within the item, managers are the individuals responsible for setting production rates and schedules. As a result, these items should discriminate between different levels of the underlying construct of safety climate better for managers than their subordinates. In measurement invariance terms, this will lead to larger factor loadings for these items for managers relative to supervisors and employees. Thus, item slopes are not expected to be equivalent across different hierarchical levels.

Intercept Equivalence Threats

The intercepts of the items may not be equivalent across hierarchical positions because of differences in social desirable responding, reference groups, and item evocativeness.

Social desirability bias. In addition to different perspectives, individuals in higher level positions may be more inclined to engage in socially desirable responding. Managers and supervisors may be unwilling or afraid for job security reasons to accurately respond to survey items about sensitive topics. As a result, they are more likely to provide responses that are socially acceptable (cf. Huang et al., 2014).

Safety climate consists of employee inferences regarding management commitment to safety (Hofmann & Stetzer, 1996; Zohar, 2002, 2008; Zohar & Luria, 2004). Managers who report poor management commitment to safety are essentially confessing that they do not take their own safety responsibilities seriously (e.g., Huang et al., 2014). For instance, Huang et al. proposed that managers and supervisors identify with upper management in the organization and may alter their responses on sensitive topics (e.g., safety issue) to enhance the image of the organization such that their responses are systematically biased toward what they think is correct or socially desirable. In sum, social desirability bias moves the responses of managers and supervisors up the scale of the safety climate measure as they are the targets of some items (e.g., "Site management provides all necessary safety equipment for workers").

Frame of reference effect. Another reason why hierarchical position is likely to influence measurement equivalence is because people at different hierarchical levels may use different frames of reference, thus resulting in the frame of reference effect discussed earlier. Respondents are most likely to use the same level employees as their referent group.

Managers and supervisors are less likely to have exposure to negative safety referents compared to subordinates, because managers and supervisors' day-to-day responsibilities are to plan and coordinate the organization's strategy and direct subordinates on their tasks. They

tend to be physically away from the safety practice on the ground (Cole & Bruch, 2006). Further, managers and supervisors may not be aware of unreported workplace accidents and injuries (Arthur, Bell, Edwards, Day, & Tubre, 2005; Burns & Wilde, 1995; Probst, Brubaker, & Barsotti, 2008) and less likely to witness and be aware of near misses or close calls (Crowl & Louvar, 2002). On the other hand, managers and supervisors have more opportunities to observe others and have frequent interaction with the higher levels of management (e.g., Lawler, 1967) making it easier for them to discern upper management's true priorities, like safety over competing demands (e.g., production; Huang et al., 2014).

In measurement invariance terms, the frame-of-referent effect will lead to nonequivalence of the intercepts of the safety climate items/measures across individuals in different hierarchical levels.

Item evocativeness. Hierarchical position may also influence the item evocativeness/attractiveness (i.e., the item difficulty or item intercept, Oort, 1998). The item difficulty parameter or the item evocativeness/attractiveness can be interpreted as the location on the latent construct continuum that determines the mean response, and thus the more "evocative" or "salient" the item is in the sense that a higher average response level (i.e., higher intercept) is obtained (Lanning, 1991). The higher the item difficulty parameter, the higher the latent construct level is required for respondents to have a .50 probability of endorsing the particular response (i.e., mean item response) in the context of Item Response Theory (IRT) models or, equivalently, the lower the item mean (or probability of correct response) is obtained for respondents with a value of 0 on the latent construct scale in the context of factor analytic item response models. Therefore, assuming there are two groups of individuals with equal standings on the latent construct, if an item is more salient or

evocative for one group of individuals compared to the other group of individuals, then the former group of individuals will achieve a higher than average score on the latent construct scale compared to the latter group. That is, these two groups with equal standing on the latent construct achieve different observed scores due to the differential item difficulty between groups. Indeed, Chan (2000) demonstrated that different levels of the evocativeness of the items across groups leads to the nonequivalence of item intercepts across groups. An individual's position within an organization determines his or her responsibilities and authority. To the extent that safety climate items focus on these responsibilities, they may be more or less salient/evocative to certain employees. An individuals' position might make certain item content very attractive or salient as a marker of the latent construct of safety climate (Chan, 2000; Robert et al., 2006). For instance, the item "Site management focuses on safety in audits, self-assessments, and inspections" might be very salient for supervisors and managers resulting in this item eliciting higher average responses level (i.e., item difficulty) from supervisors and managers than subordinates despite being the same level on the latent construct because the priority or focus of safety is determined by leaders not employees. Note that both item concreteness (i.e., item discrimination/item factor loading) and item attractiveness (i.e., item difficulty/item intercept) explanations for measurement equivalence suggest that the context of the job or the characteristics (e.g., responsibility, tasks) of the job influence individuals' responses to items (cf. Chan et al., 2002; Robert et al., 2006) but have differential effects on the measurement equivalence, affecting either slope and intercept equivalence.

Employment Arrangement

Contingent work refers to "any job in which an individual does not have an explicit or implicit contract for long-term employment or one in which the minimum hours worked can vary in nonsystematic manner" (Polivka & Nardone, 1989, p.11). Contingent workers are not a homogenous group. Two types of contingent workers, "independent contractors" and "dependent contractors" were investigated in the present study. Independent contractors were hired temporarily to provide services to the client organization on a fixed-term or a project basis (Connelly & Gallagher, 2004). Specifically, in the present study, independent contractors are contractors that are not under the direct day-to-day supervisor of the focal/host company. These contractors perform a specific scope of work (e.g., a construction project, turn around, etc.). An extensive number of these contractors are hired during a "turn around" when the plant is shut down for a few weeks at a time for extensive maintenance. Dependent contractors work daily alongside the regular employees at the plant but they are officially employees of another company contracted to the client organization. Specifically, in the present study, dependent contractors are under the direct day-to-day supervisor of the focal/host company. These contractors have specific roles that do not have a defined termination point. Some example job titles include tubers, loaders, doffers, materials handlers, guards, and cafeteria workers.

Safety climate is likely to differ between permanent employees and contingent workers (i.e., contractors), as empirical studies demonstrate that the employment arrangement with the organization directly influences workers' safety attitudes and behaviors, as well as the development of safety climate (e.g., Clarke, 2003; McDonald & Ryan, 1992; Rousseau & Libuser, 1997).

Intercept Equivalence Threats

The intercepts of the items may not be equivalent between contractors and employees because of differences in reference groups.

Frame of reference effect. Social comparison theory proposes that individuals understand themselves and evaluate their perceptions, attitudes, values, and beliefs by comparing themselves with similar others (Festinger, 1954). Thus, the employment relationship with the organization is likely to influence measurement equivalence, because contingent workers (i.e., independent and dependent contractors) and employees may use different frames of reference, thus resulting in the frame of reference effect discussed earlier. To the extent to that contingent workers and permanent employees differ in their average level on safety climate, intercept equivalence will not hold between these two groups.

Empirical studies suggest that contingent workers and permanent employees may have different average levels of safety climate. First, contingent workers tend to be less experienced and subject to lower levels of safety training, have lower levels of familiarity with the host company's practices and procedures, and have higher level of injuries and incidents (Clarke, 2003; Rousseau & Libuser, 1997). As such, contingent workers/contractors tend to develop more negative safety climate perceptions. The causal nature of these relationships has yet to be determined. Second, McDonald and Ryan (1992) argued that the development of safety climate is constrained by the control over the work process/tasks. Contractors have less control over their work and may blur the responsibilities in the case of accidents that involve more than one company (Clarke, 2003). They are often contracted and evaluated based on productivity (e.g., meeting a deadline) rather than safety, making safety less salient to them. Indeed, Mearns, Flin, Fleming, and Gordon (1998) found that contractors

had significantly more negative safety attitudes concerning management commitment to safety and incident and accident reporting.

Work Environment

Adverse job characteristics and conditions (e.g., specific tasks, the physical work environment) are critical factors that influence work-related injuries (e.g., Frone, 1998). For instance, there is increasing evidence that excessive noise (e.g., Picard, Girard, Simard, Larocque, Leroux, & Turcotte, 2008; Rabinowitz, 2000), heat (e.g., Ramsey, Burford, Beshir, & Jensen, 1983), poor lighting (Smith, 2001), high physical effort, overcrowding, cognitive demands (e.g., a need for sustained attention), and exposure to chemicals (Nahrgang et al., 2011) lead to occupational injuries.

Work settings differ with regard to the types of hazards and risks and whether or not those hazards pose a risk to one's self and/or others. The work environment influences employees' interpretation of safety as well as safety climate. Employees working in an environment with a high number of hazards may have different expectations and standards related to workplace safety, which may influence the interpretation of and responses to safety climate items (cf. Cigularov et al., 2013b). Employees in work environments with low levels of hazards are not routinely exposed to adverse job conditions and organizations are less likely to provide health and safety training to those employees. Thus, employees in such environments may be accustomed to work situations where hazards are not expected; therefore, they would have few expectations about management taking responsibility for safety. Correspondingly, some safety climate items may not make sense or are ambiguous to those employees who seldom experience hazards in their work setting, such as "Site management provides all necessary safety equipment for workers," and "My supervisor

insists we wear our protective equipment even if it is uncomfortable." As such, employees in different work environments may use different reference points to respond to items regarding the extent to which safety is a priority to their managers or supervisors. If employees have low expectations for the role of managers or supervisors in creating a safe work environment, they may respond to items pertaining to management commitment to safety differently than employees working in environment with high hazards and risks (Cigularov et al., 2013b).

The present study operationalized the work environment on a risk continuum based on the hazards present in the location that employees spend the majority of their time: manufacturing plant, research and development laboratory (R&D lab), or office.

Slope Equivalence Threats

The slope equivalence of the safety climate measure may not hold because the item relevance may vary across different work environments. That is, some items may be more effective at differentiating safety climate for employees working in a higher risk environment compared to those working in less risky environments**.**

Item relevance. People work in a wide variety of work environments, even within the same organization. Employees in different work environments may differ in the way they interpret and rate some items of the safety climate measure due to relevance of the item to their working context. For example, the items "My supervisor insists we wear our protective equipment even if it is uncomfortable," and the item "Site management provides all necessary safety equipment for workers" would be more relevant to employees who are working in the plant than those working in the office. Wearing protective equipment or assessing the safety hazards are common safety practices for employees in the plant. Also, employees in the R&D lab need to handle chemicals with particularly hazardous properties.

Wearing protective equipment (e.g., goggles and gloves) is also a common safety practice for those employees. In contrast, employees in the office do not have routine demands for such behaviors which in turn lead to more ambiguous interpretations of the notion of wearing protective equipment and assessing safety hazards at work as well as less relevance of these items to the latent construct safety climate for these employees. Hence, these items are more effective at differentiating safety climate for the employees in the plant and R&D lab than those for employees in the office.

In sum, because the item relevance varies across different work environments, the slope of the safety climate measure will not be equivalent across different work environments. The scalar invariance (i.e., slope equivalence) of the safety climate measure may not hold because employees in different environments may use different reference groups to evaluate their perception of safety climate, and the differential item evocativeness across work environments may elicit stronger or weaker responses from employees from different work environments.

Intercept Equivalence Threats

The intercepts of the items may not be equivalent across different work environments because of differences in reference groups and item evocativeness.

Frame of reference effect. Employees within the same work environment are likely to serve as an employees' reference group when completing work-related items (Heine et al., 2002). For instance, an average individual in the plant is likely to have a higher true score on the item "we do a good job of routine housekeeping at this site" item than an average individual whose normal work environment is the office. When an employee in the plant makes his or her rating on the item "we do a good job of routine housekeeping at this site," it

is made with respect to people with whom he or she is familiar with, like coworkers in the plant. As a result, it is likely that the employees in the plant will respond more similarly on the item than the employees in the office with equivalent standing on the latent variable, because of the higher or lower average of the safety climate perception established by the comparison group.

Item evocativeness. The work environment of one's job makes certain item content very salient (i.e., higher item difficulty values: Chan, 2000) as a marker of the underlying construct of safety climate. For instance, the items "my supervisor insists we wear our protective equipment even if it is uncomfortable," might be very salient in the plant and elicit more extreme responses from employees in the plant than those in the office, as wearing protective equipment is a routine practice for employees in the plant and R&D lab but not for the employees in the office. In measurement invariance terms, different degrees of item evocativeness across groups will lead to nonequivalence of the item intercepts across groups. Thus, the intercept of the safety climate measure may not be equivalent across different work environments.

CHAPTER II

METHOD

Participants, Design, and Procedure

Archival data were used to test the research questions. A health and safety survey was administered to an international chemical processing and manufacturing organization in 2007. The online questionnaire was sent to 20,260 employees and contractors, of which 8,790 individuals (77.1% male) participated, providing a response rate of 43%. Respondents were from 76 work sites/locations (ranging from 3-1063 employees, $M = 219$, $SD = 248$) in at least 19 countries (see Table 2). Employees' ages ranged from 16 to 77 years ($M = 41.73$; $SD = 10.72$) and organizational tenure ranged from 1 to 45 years ($M = 10.11$; $SD = 9.45$).

Within the sample, 5,366 employees were from individualistic countries, and 3,424 employees were from collectivistic countries (see Table 3). These employees completed the survey in 9 different languages (see Table 4). However, only seven language groups were analyzed, as the sample size for the French and Japanese was too small to draw reliable conclusions. A majority of respondents were subordinates ($n = 6,238$), followed by managers ($n = 1,058$), and then supervisors ($n = 902$), and finally contingent workers (i.e., 362 dependent contractors and 230 independent contractors). A majority of respondents worked in the plant ($n = 5,517$), 2922 employees worked in the office, and 351 employees worked in a R&D lab.

Table 2 Responses by Countries

Country	Frequency	Percent	National Culture[a]
United States	3361	38.2	Individualistic (91)
Mexico	1306	14.9	Collectivistic (30)
China	777	8.8	Collectivistic (20)
Brazil	644	7.3	Collectivistic (38)
Canada	597	6.8	Individualistic (80)
Germany	572	6.5	Individualistic (67)
United Kingdom	488	5.6	Individualistic (89)
Singapore	312	3.5	Collectivistic (20)
Netherlands	295	3.4	Individualistic (80)
Taiwan	209	2.4	Collectivistic (17)
Argentina	83	0.9	Collectivistic (46)
Switzerland	36	0.4	Individualistic (68)
Colombia	32	0.4	Collectivistic (13)
Japan	30	0.3	Collectivistic (46)
Korea	13	0.1	Collectivistic (18)
Italy/Spain	13	0.1	Individualistic (76)
Australia	9	0.1	Individualistic (90)
France	8	0.1	Individualistic (71)
Thailand	5	0.1	Collectivistic (20)

Note. [a] The individualism scores in parentheses in the National Culture column are based on Hofstede's (1980) index.

Table 3 Responses by Languages

Language	Frequency	Percent
Simplified Chinese	757	8.6
Traditional Chinese	215	2.4
Dutch	270	3.1
English	4962	56.5
French[a]	10	0.1
German	534	6.1
Japanese[a]	24	0.3
Portuguese	628	7.1
Spanish	1390	15.8

Note. [a]When examining measurement equivalence across the language faultline, Japanese and French were dropped as the sample sizes for these two groups were too small for reliable results.

Table 4 Responses by Hierarchical Positions and Employment Arrangements

Position	Frequency	Percent
Employees	8189	93.3
Employee/Individual Contributor	6238	76.1
Supervisor	902	11.0
Manager/Leader	1058	12.9
Non-Employees	592	6.7
Dependent Contractor	362	4.1
Independent Contractor	230	2.6

Table 5 Responses by Work Environments

Work Environment	Frequency	Percent
Office	2922	33.2
Plant	5517	62.8
Research &Development Lab	351	4.0

Measures

Safety Climate

Safety climate was assessed with eight items adapted from Zohar and Luria (2005). Professional external translators translated the survey items into nine languages (i.e., simple Chinese, traditional Chinese, Dutch, English, French, German, Japanese, Portuguese, Spanish). All items were administered on a 5-point agreement scale (1 = strongly disagree, 5 = strongly agree, NA). The percentage of NA responses ranged from 0.8% - 1% and were treated as missing data. The Cronbach's alpha coefficient was .91 for the safety climate measure. A complete listing of the items appears in the Appendix.

Prior to the administration of the survey, five items were deemed irrelevant to the office employees (item 1 to item 5; e.g., "Site management focuses on process safety in audits, self-assessments, and inspections.") and one item was deemed irrelevant to the R&D employees (i.e., "Site management focuses on process safety in audits, self-assessments, and inspections."). Therefore, skip logic was embedded into the survey so that employees who identified themselves as office or R&D workers skipped these items.

Faultlines

Respondents were given the option to complete the survey in one of nine languages. By choosing a language, they entered into the corresponding translated survey. Because the survey was only translated one time for each language, language and translation cannot be teased apart. Respondents also indicated the country in which they worked, their hierarchical position/employment arrangement, and their work environment all from multiple choice lists that appear in Tables 2, 4, and 5. The majority of respondents completed their survey in the official language for the country in which they worked (see Table 6).

Table 6 Responses by Country and Language

Country	Language						
	Simplified Chinese	Traditional Chinese	Dutch	English	German	Portuguese	Spanish
China	702	9	0	63	2	1	0
Japan	0	0	0	6	0	0	0
Korea	0	0	0	13	0	0	0
Singapore	54	4	0	254	0	0	0
Taiwan	0	201	0	8	0	0	0
Thailand	0	0	0	4	1	0	0
England	0	0	0	9	0	0	0
France	0	0	0	3	0	0	0
Germany	0	0	0	44	527	0	0
Netherlands	0	0	270	24	0	0	0
Switzerland	0	0	0	31	2	0	0
United Kingdom	0	0	0	487	0	0	1
Italy/Spain	0	0	0	11	0	0	2
Mexico	0	0	0	27	0	1	1278
Canada	0	0	0	596	1	0	0
United States	0	1	0	3350	1	0	9
Argentina	1	0	0	3	0	2	77
Brazil	0	0	0	20	0	624	0
Colombia	0	0	0	9	0	0	23

Data Analysis

All models were estimated with Mplus 7.0 (Muthén & Muthén, 1998-2012).

Multilevel Confirmatory Factor Analysis

Multilevel confirmatory factor analyses were conducted to reveal how well the scale items loaded onto their respective factor.

Measurement Equivalence Tests

The nature of the data examined are multilevel. This means that individuals are nested or grouped within larger mutually exclusive groups. These groups may be nested further into larger mutually exclusive groups. A hierarchical data structure violates the assumption of independent observations; therefore, these groupings need to be accounted for in the analyses. In the current data set, individual employees are nested within worksites/locations which are nested within countries. Thus, country (n = 19) is modeled as a Level-3 variable, worksite/location as a Level-2 variable, and individuals as a Level-1 variable. Language, hierarchical position, employment arrangement, and work environment are all Level-1 variables[2]. Level 1 can also be referred to as a within-level variable, whereas Level-2 and Level-3 are between-level variables.

When testing measurement invariance with multilevel data, it is critical to distinguish the level of group membership (i.e., faultlines). "Measurement invariance must be established at the corresponding level at which an inference is made" (Ryu, 2014, p. 191). In the present study, multi-group multilevel confirmatory factor analyses (Kim et al., 2012)

---

[2] In this dataset, language is not nested within country, as individuals within the same country completed the survey in different languages (see Table 6). A cross-classified data structure (e.g., languages are cross-classified into different countries) was not considered, because measurement equivalence tests in a cross-classified structure have not been developed yet.

were conducted for the Level-3 faultine (i.e., national culture), whereas multilevel factor mixture models for known classes (Kim et al., in press) were conducted for Level-1 faultlines, including language (i.e., simplified Chinese, traditional Chinese, Dutch, English, German, Portuguese, and Spanish), hierarchical positions (i.e., managers, supervisors, and subordinates), employment arrangement (i.e., employees, independent contractors and dependent contractors), and work environments (i.e., office, plant, as well as R&D lab)..

Multi-group Multilevel Confirmatory Factor Analysis for the Level-3 Faultline

Empirical studies have demonstrated the standard procedure for testing measurement invariance with single-level data can be applied to between-level faultlines within multilevel data (e.g., Jones-Farmer, 2010; Kim et al., 2012; Ryu, 2014). Measurement equivalence tests within multilevel data can be conducted using the design-based multi-group multilevel CFA or the model-based multi-group multilevel CFA. For instance, for two level data (e.g., employees nested within locations), the design-based approach analyzes the data with only one overall model and corrects the underestimated standard errors of parameter estimates based on the sampling design, whereas the model-based approach analyzes the data by specifying a within-level (i.e., individual-level) model and a between-level (e.g., location) model, respectively. The design-based multi-group multilevel CFA (i.e., using "Type= COMPLEX" in Mplus) can only adjust one level of clustering; therefore, it is not appropriate for the three level data in the present study.

In contrast, the model-based multi-group multilevel CFA is better suited for the three level data in the present study (e.g., Kim et al., 2012; Kim et al., in press), because it can handle multiple levels of clustering. However, the model for the Level-3 faultline (i.e., national culture; "Type= THREELEVEL" in Mplus) was not identified, because the number

of clusters (i.e., the number of countries, n = 19) at Level-3 is smaller than the number of parameters to be estimated.

A third option is a combination of design-based and model-based multi-group multilevel CFA (i.e., using "Type= COMPLEX TWOLEVEL" in Mplus). In this approach, the design-based approach is used to take into account the small number of clusters at Level-3 (i.e., country) and the model-based approach is used to specify a level-specific model for the between-level model (i.e., location) and a within-level model (i.e., the individual-level), respectively.

Evaluation of Measurement Equivalence Models for the Level-3 Faultline

It is critical to select the appropriate goodness-of-fit index (GFI) to determine different levels of measurement invariance. According to Hus, Kwok, Lin, and Acosta (2015), fit indices (e.g., CFI, RMSEA) along with the traditional cutoff values can only effectively identify the model misspecification at a particular level within a multilevel dataset. Therefore, in the present study, the model fit was evaluated at each level, respectively.

For the within-level (i.e., the individual-level) model, following the recommendations of Hu and Bentler (1999), Cheung and Rensvold (2002), as well as Hus et al. (2015), a variety of fit indices were examined to evaluate the model misspecification, including the standardized root mean square residual for the within-level model (SRMR-W), the root mean square error of approximation (RMSEA; Steiger & Lind, 1980), as well as the comparative fit index (CFI). SRMR-W, a measure of absolute fit, indicates how well (on average) the correlation matrix has been reproduced by the within-level model. RMSEA, another measure of absolute fit, indicates the absolute fit adjusting for model parsimony (i.e., the magnitude of

the covariance residuals are adjusted for degrees of freedom). CFI reflects the proportion of improvement in fit relative to the null model. Ideally, for the model with the adequate fit, CFI should be greater than .90, RMSEA should be less than .06, and SRMR-W should be less than .08 (Hu & Bentler, 1999).

For the between-level model, the standardized root mean square residual for the between-level model (SRMR-B) was used. Hsu et al. (2015) support that SRMR-B is the only fit index that could effectively detect misspecification for the between-level model. SRMR-B should be less than 0.14 for the between-level model (Hsu et al., 2015).

Finally, the χ^2 difference test was considered as well. Multilevel Structure Equation Modeling (SEM) uses the maximum likelihood estimation with robust standard errors (MLR). Thus, in the present study, multi-group multilevel CFAs for the Level-3 faultline employed MLR as the estimator. Because the MLR chi-square difference does not follow the chi-square distribution, the Satorra-Bentler scaled chi-square difference test (SBχ^2; Satorra & Bentler, 1994) is recommended for model comparison (Brown, 2006; Heck & Thomas, 2009; Kim et al., in press). A significant decline in fit between models indicates that the more restrictive model has significantly worse model fit relative to the comparison model.

Multilevel Factor Mixture Model for Known Classes for Level-1 Faultlines

Testing measurement invariance with an individual-level grouping variable introduces additional complexities that are beyond a simple extension of the well-established procedures for testing measurement invariance in single-level multiple-group CFA. Multilevel modeling within the SEM framework in which a level-specific model is constructed for each level (the model-based multilevel CFA; Muthen & Satorra, 1995; Wu & Kwok, 2012) is not feasible, as it does not allow for multiple-group analyses specifically at

the individual-level. In other words, the model-based multi-group multilevel CFA and the combination model are not feasible for measurement equivalence tests, when the grouping variable is at the individual level. Although the design-based approach can be used for measurement equivalence tests when the grouping variable is Level-1, this approach can only adjust one level of cluster sampling, which is not appropriate for the three-level data in the present study. Therefore, neither the model-based nor the design-based multi-group multilevel CFAs are feasible with a Level-1 grouping variable. Thus, the multilevel factor mixture model for known classes was conducted when the grouping variable was at the individual level (for the details of model specifications, see Kim et al., in press). Further, based on Kim et al. (in press), the present study combined multilevel factor mixture models with the design-based approach that corrects the Level-3 clustering. Because, as discussed above, it is impossible to specify a Level-3 model with a small number of country clusters, the design-based approach has to be used to correct the standard error of parameter estimates based on the country level of cluster sampling.

Evaluation of Measurement Equivalence Test Models for Level-1 Faultlines

Fit indices that evaluate the fit of the multilevel factor mixture models for known classes or groups are different from those for multi-group multilevel CFAs. Multilevel factor mixture model for known classes also used MLR as the model estimator. When the MLR estimator is used, the Satorra-Bentler scaled likelihood ratio (SBLR: Satorra & Bentler, 1994) is recommended for model comparison (Brown, 2006; Heck & Thomas, 2009; Kim et al., in press). Other fit indices (i.e., information criteria), including Akaike information criterion (AIC; Akaike, 1987), Bayeisian information criterion (BIC; Schwarz, 1978), and sample-size adjusted BIC (SBIC; Sclove, 1987), were also considered. When two models are

compared, the model associated with the smaller AIC, BIC, and SBIC values is considered as a better model (Kim et al., in press). When using the multilevel factor mixture model for known classes for testing measurement equivalence for the Level-1 faultlines, the BIC and SBIC are recommended when the total sample size is sufficiently large (>3,000 for the BIC>2, 000 for SBIC; Kim et al., in press).

CHAPTER III

RESULTS

Descriptive statistics for the safety climate scores for the five faultline groups appear in Table 7.

Table 7 Descriptive Statistics for Safety Climate by Faultline Groups

Faultline	*M*	*SD*
	National Culture	
Individualists	4.05	0.69
Collectivists	4.07	0.60
	Language	
Simplified Chinese	4.06	0.53
Traditional Chinese	4.04	0.62
Dutch	3.89	0.63
English	4.06	0.70
German	4.11	0.64
Portuguese	4.00	0.64
Spanish	4.09	0.60
	Hierarchical Position & Employment Arrangement	
Managers	4.30	0.68
Supervisors	4.28	0.61
Subordinates	4.00	0.65
Dependent Contractors	3.91	0.65
Independent Contractors	3.98	0.65
	Work Environment	
Plant	4.07	0.61
R&D lab	4.17	0.65

Note. Without examining the measurement equivalence of the safety climate measure, it is not clear whether it is meaningful to compare the observed scores across a faultline.

Table 8 presents the estimated intraclass correlations for the eight safety climate items at the country, the location levels of analysis. The intraclass correlation [i.e., ICC(1)] measures the average correlation between observations (e.g., employee) in safety climate item scores within the same cluster (e.g., working group) or it can be conceptualized as the proportion of variance in the safety climate scores that is explained by the group membership (Bliese, 2000). The larger the ICC(1), the more correlated the observations are within a cluster and the more variance that is explained by the group membership. In other words, the larger the ICC(1), the more the assumption of independence between observations is violated (e.g., Muthen & Satorra, 1995).

In general, it is the design effect rather than the ICC(1) that is an issue regarding the multilevel data, as the design effect indicates how much the standard errors of parameter estimates are underestimated (Kish, 1965). This design effect is approximately equal to 1+ (average cluster size-1)*ICC(1) (Muthen & Satorra, 1995). Muthen and Satorra (1995) suggest that a design effect greater than two indicates that the clustering in the data should not be ignored and the clustering will lead to biased estimates. Specifically, the standard errors of the parameter estimates will be negatively biased, which results in spuriously significant effects (cf. de Leeuw & Kreft, 1986; Hox, 2002; Snijders & Bosker, 1999). Thus, measurement equivalence tests will have inflated Type I error rate in terms of rejecting the null hypothesis of measurement equivalence held between groups (Kim et al., 2012), when the design effect indicates that the dependency between observations is an issue. As Table 8 shows, the design effect of the safety climate measure item scores at the country and the location level were larger than two. Therefore, the design effects indicated that examining the

measurement equivalence tests of the safety climate measure across different faultline groups should take into consideration the nested structure of the data.

Table 8 Estimated Intraclass Correlations and Design Effect for Safety Climate Measure Items at the Country Level, and the Location Level

Items	Country ICC (Design[a])	Location ICC (Design)
Item 1	.003(2.364)	.051(6.758)
Item 2	.037(17.824)	.046(6.193)
Item 3	.032(15.550)	.019(3.145)
Item 4	.017(8.730)	.024(3.710)
Item 5	.007(4.183)	.043(5.855)
Item 6	.056(26.463)	.011(2.242)
Item 7	.014(7.366)	.031(4.500)
Item 8	.033(16.005)	.102(12.516)

Note. [a] Design: Design effect = 1 + (average cluster size – 1)* ICC.
ICC = Intraclass correlation.

Multilevel Confirmatory Factor Analysis

Before conducting the substantive analyses, the factor structure of the safety climate measure was examined using a combination model of the design-based and the model-based multilevel CFA. As the number of location clusters is not small (n = 76), the model-based approach was used to take into account the clustering within locations by specifying a model for the location level (and for the individual level, respectively), whereas the design-based approach was used to adjust parameter estimate standard errors for the clustering within countries (n = 19). The results of this multilevel CFA indicated that the safety climate measure had poor model fit: CFI indicated poor model fit (RMSEA = .05; SRMR-W = .05; CFI = .79) for the individual-level model and SRMR-B indicated poor model fit (SMRM-B = .18) for the between-level model (i.e., at the location level). Close examination of modification indices indicate that the residual variance of item 7 (i.e., "My supervisor frequently discusses health and safety issues throughout the work week.") correlates strongly with the residual variance of items 6 and 8, while the residual variances of items 6 and 8 have a small correlation with each other, suggesting that item7 provides redundant information regarding the latent construct of safety climate. Further, if the basic factor structure model did not have good model fit, there is no need to proceed with measurement equivalence tests, as measurement equivalence models would not have good model fit. Therefore, item 7 was dropped from the safety climate measure. The results of the multilevel CFA for the revised 7-item safety climate measure indicated the measurement model had good model fit (RMSEA = .03; SRMR-W = .02; CFI = .92; SRMR-B = .10) at both the individual-level and location level model. Thus, the measurement invariance tests were conducted for this 7-item safety climate measure.

For information purposes (e.g., to reveal the consequence of ignoring the nested structure of the data), a single-level CFA for this 7-item safety climate measure was also conducted, revealing that it had better model fit (RMSEA = .07; SRMR = .04; CFI = .94). The comparison of the single-level CFA with the multilevel CFA revealed that when ignoring the nested structure of the data, the standard errors of the model estimates are biased, which in turn lead to an inaccurate conclusion regarding the model fit.

Results of Measurement Equivalence Tests

Individualism

Table 9 presents the results of the various levels of measure equivalence tests of the safety climate measure between individualists and collectivists, using the multi-group multilevel CFAs, in which the design-based approach was used to adjust the standard errors of the parameter estimates for the country clustering and the model-based approach was used to specify the individual-and the location-level models. The results indicated that the configural equivalence model had acceptable fit, suggesting that the safety climate items evoke the same conceptual framework in defining the latent construct across the individual-level and the location-level for individualists and collectivists. That is, the configural equivalence of the safety climate measure held between respondents from individualist and collectivist countries.

Metric equivalence for the individualism faultline was tested next. The results indicated that the metric equivalence model had good model fit for the individual-level model but bad model fit for the location-level model (see Table 9). Further, the Satorra-Bentler scaled chi-square difference test [$SB\chi^2$ (4) = 94.62, $p < .05$] indicated that the metric equivalence model (i.e., slope equivalence) had significantly worse model fit compared to the

configural equivalence model. Thus, the metric equivalence of safety climate measure did not hold between individualists and collectivists. In other words, the regression slopes associating the manifest safety climate measure to the underlying construct of safety climate were not equivalent between respondents from individualistic and collectivistic countries. These results were consistent with the proposition that the item relevance of the safety climate measure and response styles vary between individualists and collectivists (Bryne & Campbell, 1999; Cheung & Rensvold, 2000; Tanzer, 1995).

Measurement equivalence is established hierarchically. Thus, when the metric invariance of the safety climate measure did not hold between individualists and collectivists, the scalar invariance of the safety climate measure will not hold between individualists and collectivists either. That is, only when the metric invariance is established is scalar invariance even possible. Therefore, there is no need to proceed with the scalar equivalence model when the metric equivalence model was not supported. Nevertheless, the comparison in model fit between these two models was still conducted for the information purposes. As Table 8 shows, the results indicated that the scalar equivalence model had good model fit for the within-level model but poor model fit for the between-level model. Particularly, the Satorra-Bentler scaled chi-square difference test [$SB\chi^2$ (7) = 202.12, $p < .05$] indicated that there was a significant decline in the fit of the scalar invariance model (i.e., the intercept equivalence) compared to the metric invariance model. These results suggested that the regression intercepts associating the manifest safety climate measure to the underlying construct of safety climate are not invariant between respondents from individualist and collectivist countries. That is, the scalar equivalence of the safety climate measure did not hold across the individualism faultline. These results are consistent with the proposition that multiple

response styles, social desirability bias, as well as frames of reference are likely between individualists and collectivists.

In conclusion, responses to the safety climate measure across the individualism faultline were configurally equivalent. This means that the factor structure of the safety climate measure was the same for respondents from individualistic and collectivistic countries. The safety climate measure was neither metric nor scalar invariant for the individualism faultline, indicating that the slopes and intercepts for the safety climate measure were not equivalent for respondents from individualistic and collectivistic countries.

Table 9 Results of Measurement Equivalence Tests for National Culture Operationalized as Individualism[3]

	χ^2 (*df*)	RMSEA	CFI	SRMR-W	SRMR-W
Configural Equivalence	924.22(55)*	.06	.89	.04	.15
Metric Equivalence	1003.13(62)*	.06	.89	.04	.20
Scalar Equivalence	1337.29(68)*	.07	.84	.04	.28

Note. * p <.05.

[3] Measurement equivalence tests were also examined between individualistic countries (Germany, Netherlands, United Kingdom, Canada, and United States), and between collectivistic countries (China, Singapore, Taiwan, Mexico, and Brazil). The results (See Appendix B and C) indicated configural (but not metric or scalar) equivalence between the individualistic countries and between collectivistic countries. These results are not surprising. Although these groupings of countries are similar on the individualism dimension, they are different on many other variables (geography, laws, economy, etc.). That is, conceptually it is unclear what the country faultine represents. In contrast, it is assumed that the sample of countries included in the individualism/collectivism aggregations are representative of the population of individualistic and collectivistic countries.

Language

Multilevel factor mixture models were also conducted for the language faultline. Table 10 presents the results of the various measurement equivalence tests for this faultline.

The lower values of AIC and SBIC for the metric equivalence model suggested that the metric equivalence model provided worse fit than the configural equivalence model, whereas the value of BIC supported that the configural equivalence model (see Table 9). The SBLR produced a negative value. Therefore, additional adjustment was required to ensure the positive chi-square statistic for the comparison between the metric equivalence and the scalar equivalence models (Satorra & Bentler, 2010). Based on Asparouhov and Muthen's approach (2013), the adjusted SBLR(42) = 332.62, p <.05 indicated that significant worsening in fit if the equivalence of the factor loadings of the safety climate items across linguistic groups was imposed. That is, the metric equivalence of the safety climate measurement did not hold across the seven linguistic groups. These results are consistent with the proposition that language is an important facet of the national culture, and that translation may result in discrepancies in the meaning of the construct measured between the original and the translated measure, which in turn may change the relevance of the safety climate items to the latent construct across different linguistic groups.

Although the metric equivalence did not hold across linguistic groups, the comparison between the metric equivalence and the scalar equivalence models is reported for information purposes only. As Table 10 shows, all three information criteria (i.e., AIC, BIC and SBIC) indicated that the metric equivalence model provided better fit than the scalar equivalence model. Further, the adjusted SBLR (36) = 308.24, p < .05 indicated that the scalar equivalence model (i.e., intercept equivalence) had significantly worse model fit compared to

the metric equivalence model (i.e., slope equivalence). In other words, the regression intercepts associating the manifest safety climate measure to the underlying construct of safety climate are not equivalent across the seven linguistic groups. These results were consistent with the proposition that some words with nonspecific meaning may be mistranslated leading to a misunderstanding of Likert scales such that individuals who speak different languages may interpret the response options differently.

In sum, responses to the safety climate measure across the language faultline were configurally equivalent, indicating that the factor structure of the safety climate measure was the same for all languages tested. The safety climate measure was neither metric nor scalar invariant between respondents completing the survey in simplified Chinese, traditional Chinese, Dutch, English, German, Portuguese, and Spanish, indicating that the slopes and intercepts for the safety climate measure were not equivalent across the language faultline.

Table 10 Results of Measurement Equivalence Tests for Language

	Loglikelihood	AIC	BIC	SBIC
Configural Equivalence	-65030.89	130382	131512	131003
Metric Equivalence	-65215.08	130666	131500	131125
Scalar Equivalence	-65583.98	131332	131911	131651

Hierarchical Position

Table 11 presents the results of three measurement equivalence tests for the safety climate measure across the hierarchical position faultline. Contrary to expectation, the metric equivalence model had lower values for all three information criteria (i.e., AIC, BIC, and SBIC) compared to the configural equivalence model, suggesting that the metric equivalence model is a better fitting model. Further, the SBLR (14) = 9.96(14), $p > .05$ indicated that there was no significant decline in the model fit, when the factor loadings of the items were constrained to be equal across managers, supervisors, and subordinates. That is, the results supported that regression slopes associating the manifest safety climate measure to the underlying construct are equivalent across three hierarchical positions.

None of the three information criteria (i.e., AIC, BIC, and SBIC) favored the scalar equivalence model. Particularly, the SBLR (12) = 629.68, $p < .05$ indicated that the scalar equivalence model provided significantly worse model fit compared to the metric equivalence model, indicating that the regression intercepts associating the manifest safety climate measure to the underlying construct of safety climate are not equivalent across managers, supervisors, and subordinates.

In sum, the results indicated that the factor loadings of the safety climate items but not the intercepts of the safety climate items were equivalent across managers, supervisors, and subordinates. Therefore, it is not meaningful to compare safety climate scores across managers, supervisors, and subordinates.

Table 11 Results of Measurement Equivalence Tests for Hierarchical Position[4]

	Loglikelihood	AIC	BIC	SBIC
Configural Equivalence	-56600.80	113346	113849	113621
Metric Equivalence	-56609.35	113335	113741	113556
Scalar Equivalence	-56773.05	113638	113960	113814

[4] The subgroups of the employment arrangement faultline can also be conceptualized as subgroups of the hierarchical position faultline. Thus, an alternative conceptualization of hierarchical position clusters consists of the following five subgroups: managers, supervisors, subordinates, dependent contractors, and independent contractors. The results indicated metric equivalence but not the scalar equivalence of the safety climate measure for this alternative conceptualization of the hierarchical position faultline.

Employment Arrangement

Table 12 presents the fit indices for the configural equivalence, metric equivalence, and scalar equivalence of the safety climate measure for the employment arrangement faultline.

All three information criteria indicated that the metric equivalence model provided better fit than the configural equivalence model for the hierarchical position faultline,. Particularly, the SBLR (14) = 18.30, $p > .05$ indicated that there was no significant difference in model fit between the configural equivalence and the metric equivalence models. In other words, the factor loadings of the safety climate measure items were equivalent across employees, dependent contractors, and independent contractors.

As Table 12 shows, two of the three information criteria (i.e., BIC and SBIC) supported the fit of the scalar equivalence model over the metric equivalence model, whereas the lower value of AIC of the metric equivalence model indicated that the metric equivalence model is the preferred model. However, the SBLR (12) = 33.75, $p < .05$ provided further support that that the scalar equivalence (i.e., intercept equivalence) model provided worse model fit than the metric invariance model, suggesting that the regression intercepts associating the manifest safety climate measure to the underlying construct of safety climate are not equivalent across the employment arrangement groups. These results were consistent with the speculation that contingent workers and employees likely use different frames of reference in responding to safety climate items.

In sum, the factor loadings (i.e., slope equivalence) but not the intercepts (i.e., intercept equivalence) of the safety climate items were equivalent across independent contractors, dependent contractors, and employees. As the scalar equivalence of the safety

climate measure did not hold across the employment arrangement faultline, it is not meaningful to compare the observed safety climate scores between contingent workers and employees.

Table 12 Results of Measurement Equivalence Tests for Employment Arrangement

	Loglikelihood	AIC	BIC	SBIC
Configural Equivalence	-47342.05	94828	95319	95090
Metric Equivalence	-47355.23	94826	95222	95038
Scalar Equivalence	-47375.99	94844	95158	95011

Work Environment

As noted earlier, employees working in the office did not receive five out of the seven items of the safety climate measure, and employees working in the R&D labs did not receive one out of the seven items. Two items is not sufficient for the model identification; therefore, the office were not included in the work environment analyses.

Table 13 shows the results of the work environments faultline. The measurement equivalence tests of the safety climate measure between these two subgroups were limited to six items, because as noted earlier, employees working in R&D lab did not receive one of the seven items.

As shown in Table 13, the lower values of BIC and SBIC of the metric equivalence model indicated that the metric equivalence model was superior to the configural equivalence model, whereas the lower value of the AIC for the configural equivalence model supported that the configural equivalence model as the preferred model. The SBLR (7) = 19.78, $p < .05$ indicated that the metric equivalence model (i.e., slope equivalence) fit the data worse than the configural equivalence model, suggesting that the regression slopes associating the manifest safety climate measure to the underlying construct of safety climate are not equivalent between employees working in the plant and employees working in an R&D lab. These results were consistent with the expectation that the item relevance of the safety climate measure may vary across individuals from different work environments.

Although metric equivalence was not achieved, the comparison between the metric and the scalar equivalence models between the two subgroups of the work environment is reported for the information purposes only. AIC, SBIC, as well as the SBLR (6) = 84.95, $p < .05$ indicated the metric equivalence model is preferred over the scalar equivalence model.

That is, the scalar equivalence of the safety climate measure did not hold between employees working in the plant and employees from the R&D lab, which was consistent with the proposition that the reference groups employees used for responding to safety climate items as well as item evocativeness may vary based on employees' specific work environment.

In sum, neither the scalar equivalence nor the metric equivalence of the 6-item safety climate measure was established between employees working in the plant and employees in the R&D lab. Therefore, any comparison of mean differences on the observed safety climate score between these groups is not meaningful.

Table 13 Results of Measurement Equivalence Tests for Work Environment

	Loglikelihood	AIC	BIC	SBIC
Configural Equivalence	-48619.75	97339	97673	97514
Metric Equivalence	-48633.46	97353	97640	97503
Scalar Equivalence	-48659.54	97393	97640	97522

CHAPTER IV

DISCUSSION AND CONCLUSIONS

Empirical studies support that safety climate is one of leading indicators of safety-related outcomes (Beus et al., 2010; Christian et al, 2009; Nahrgang et al, 2011). To promote a safe environment in the organization, it is critical to establish a valid safety climate measure. However, the existence of various faultiness within the organization (e.g., national culture operationalized as individualism) may threaten the measurement equivalence of the safety climate measure across different subgroups within the organization. If individuals from varying subgroups interpret the safety climate scale differently (i.e., the scale assesses different constructs), then combining subgroup data and/or comparing subgroup safety climate scores is inappropriate (Vandenberg & Lance, 2000). Measurement equivalence is increasingly important due to the diversification of the workforce (Kirchmeyer & McLellan, 1991) and the globalization of business enterprises (e.g., multinational organizations; Schmitt & Kuljanin, 2008).

A few safety climate researchers have started to pay attention to the measurement equivalence of the safety climate measures (Beus et al., 2011; Cheyne, et al., 2003; Huang et al., 2014; Cigularov, Adams, et al., 2013; Cigularov, Lancaster,et al., 2013; Lee et al., in press). However, these studies provide an incomplete assessment of the measurement equivalence of a safety climate measure. Multiple psychological theories (e.g., social comparison theory, item response theory) suggest that there are a number of other important faultlines that may threaten the measurement invariance of a safety climate measure that have not been empirically tested. Thus, extending this line of research, the present study examined

the measurement equivalence of a safety climate measure across four additional faultlines (i.e., language, national culture, employment arrangement, and work environment).

The results indicated that a lack of metric (i.e., slope) equivalence for the safety climate measure across national culture, language, and work environment faultlines. In other words, the slopes of the safety climate items were not equivalent between collectivists and individualists; across the seven linguistic groups examined, or between the plant and R&D lab work environments. Thus, it is not meaningful to compare the safety climate scores across these faultlines. These results are consistent with the speculation that the association of the items with the latent construct is not the same for the respondents within each of these comparisons.

The safety climate measure demonstrated metric (i.e., slope), but not the scalar (i.e., intercept) equivalence for the subgroups based on the hierarchical position faultline and the employment arrangement faultline. This indicates the relationship between the observed safety climate scores and the latent safety climate construct is not the same across subgroups, and thus it is not meaningful to compare manager safety climate scores to supervisor or subordinate safety climate scores. It also means that it is not appropriate to compare plant employee safety climate scores to R&D employee safety climate scores.

National Culture

Cross-national research has become an important trend among organizational researchers and practitioners. Multinational organizations frequently conduct global employee surveys to manage, motivate, and retain employees (Borg, 2003). The ability to meaningfully interpret multinational survey data depends in part on the measurement equivalence of the measures across different national cultures (e.g., Riordan & Vandenberg,

1994; Schmitt & Kuljanin, 2008; Vandenberg & Lance, 2000). Only if measurement equivalence is established can researchers and practitioners feel confident that (a) the conceptualization or definition of the construct that is assessed by the scale or instrument is generalizable to each culture (e.g., Little, 1997; Liu et al., 2004; van de Vijver & Tanzer, 1997); (b) different cultural subgroups interpret the measure in a conceptually similar way (e.g., Liu et al., 2004; Vandenberg & Lance, 2000); (c) respondents from different cultural subgroups calibrate the scalar anchor and/or interpret the response options in the same way (e.g., Liu et al., 2004; Riordan & Vandenberg, 1994; Vandenberg & Lance, 2000); (d) any observed differences between cultural subgroups reflect true differences on the latent construct (e.g., Liu et al., 2004; Raju, Kaffittee, & Byrne, 2002; Vandenberg & Lance, 2000); and (e) sources of bias (e.g., social desirability bias) and error (e.g., translation errors) are minimal (e.g., Little, 1997; Liu et al., 2004). Thus, it is important to establish the measurement equivalence of the safety climate measure across the national culture faultline before researchers and practitioners can confidently make use of the safety climate survey to manage safe behavior. Although safety climate researchers have collected data from multinational organizations (e.g., Cheyne, Cox, Oliver, & Tomas, 1998; Wallace, Popp, & Mondore, 2006), none of these studies have investigated the measurement equivalence of the safety climate measure across national cultures. To begin to address this gap in the literature, the present study provided the first examination of the measurement equivalence of a safety climate measure across national cultures.

Based on a national culture faultline operationalized with Hofstede's (1980, 1983) individualism classification of the country from which each respondent answered the survey, the safety climate measure failed to show metric and slope equivalence. These results are

consistent with empirical findings that survey items may vary in the relevance to the latent construct across national cultures (van de Vijver & Leung, 1997), and that multiple response styles (e.g., extreme response style, acquiescence; Cheung & Rensvold, 2000), social desirability bias (e.g., Heine & Lehman, 1997; Johnson, 1998; Lalwani et al., 2006), and frames of reference (Heine et al., 2002) are likely to differ across cultures, which in turn threaten the equivalence of the safety climate measure between individualistic and collectivistic respondents (Cheung & Rensvold, 2000; Robert et al, 2006; van de Vijver & Leung, 1997).

These findings also provide further support for the importance of examining the degree to which the scale or the instrument measure the same thing across cultures regardless of what constructs are under investigation (Schmitt & Kuljanin, 2008). Because individuals with different cultural backgrounds are likely to engage in different response styles and different amounts of socially desirable responding, cultural differences are not specific to safety climate measures but relevant to all scales and measures. The present findings are also consistent with the proposition that constructs, measures, and theories that are developed in the one culture are likely have limited applicability to another culture (e.g., van de Vijver & Leung, 1997).

Language

Language is an important component and essential indicator of culture (Lenartowicz & Roth, 2001; Liu et al., 2004; Peterson & Smith, 1997). Individuals speaking the same language share many elements of culture (Hofstede, 2001; Inglehart & Baker, 2000). Thus, the mechanisms through which culture is posited to affect the interpretation of a safety climate measure (e.g., social desirability bias, the frame of reference effect) were also

proposed to affect the measurement equivalence of the measure across linguistic groups. For instance, the concept of safety may have a different meaning for Chinese speakers, who may have a different frame of reference for complying with safety rules compared to English speakers. Further, the translation process might result in discrepancies in the meaning of the words used within the original and the translated measure.

Indeed, the results of the current study indicated neither scalar nor metric equivalence for the safety climate measure across the seven linguistic groups (i.e., simplified Chinese, traditional Chinese, Dutch, English, German, Portuguese, and Spanish). These findings also provide further support for the importance of examining the degree to which the scales or instruments measure the same thing across languages (and national culture) for other constructs besides safety climate (Schmitt & Kuljanin, 2008), as the mechanism by which language (and national culture) is hypothesized to affect the measurement equivalence of the scales is not specific to safety climate measures.

Hierarchical Position

Contradicting with the speculation that the relevance of safety climate items may vary across hierarchical positions and result in metric non-equivalence for managers, supervisors, and subordinates, factor loadings of the safety climate items were equal across these groups. This finding contradicts Beus et al.'s[5] (2012) findings which examined hierarchical position

[5] Beus et al. (2012) combined hierarchical position and employment arrangement faultiness to create two subgroups: (a) supervisors and managers and (b) subordinates and dependent contractors. The measurement equivalence of the safety climate measure across these two subgroups was re-examined adjusting for data dependency. Contrary to Beus et al.'s findings, metric equivalence was established between these two groups: supervisors/managers and subordinates/dependent contractors. Scalar equivalence was not supported between supervisors/managers and subordinates/dependent contractors. Beus et al. (2012) also examined the organizational heritage faultline (Company X, Company Y, direct hire, and contractor company), and concluded that the metric equivalence did not hold for this faultline. However, when taking the data dependency into account, the multilevel mixture factor model indicated that the metric equivalence (but not the scalar

in the same data and found metric non-equivalence between supervisors/managers and subordinates/dependent contractors. Extending Beus et al.'s (2012) study, the present study examined the scalar equivalence of the safety climate measure while taking into account data dependency. Consistent with the proposition that managers and supervisors are likely to engage in more socially desirable responding and use different referents regarding safety compared to their subordinates, the present study revealed a lack of scalar equivalence across the hierarchical position faultline. Similarly, Cheyne et al. (2003) found that the metric equivalence of a safety climate measure did not hold between managers, supervisors and subordinates. In contrast, Huang et al. (2014) found that the slopes and the intercepts of a trucking safety climate scale were equivalent between supervisors and subordinates. As noted earlier, measurement equivalence must be interpreted in the context of the interaction between the measure, the sample characteristics, and the characteristics of the administration (e.g., Carter, Kotrba, & Lake, 2014; Robert, Lee, & Chan, 2006; Vandenberg & Lance, 2000; Vandenberg, 2002). As Huang et al. examined a different safety climate measure, it is not surprising that Huang et al. found different results.

Employment Arrangement

Empirical studies demonstrate that the employment relationship with the organization directly and indirectly influences employees' safety attitudes and behaviors, as well as the development of safety climate (e.g., Clarke, 2003; McDonald & Ryan, 1992; Rousseau & Libuser, 1997). Contingent workers and employees were hypothesized to interpret the scale anchor and response options differently because they are likely to use different

---

equivalence) was supported for the organizational heritage faultline. Taken together, these results provided empirical evidence that, ignoring the multilevel nature of the data, can lead to misleading conclusions about measurement equivalence (Kim et al., 2012; Kim et al., in press).

referents/standards (i.e., frame of reference effect). Consistent with this prediction, the employment arrangement subgroups did not demonstrate scalar (intercept) equivalence.

Contrary to expectation, the safety climate measure was metric invariant across the employment arrangement faultline. In other words, the factor loadings of the safety climate items were equivalent across employees, dependent contractors, and independent contractors. This suggests that the safety climate items were equally relevant to the latent safety climate construct across these groups.

Work Environment

Consistent with the proposition that employees in different work environments may interpret the safety climate items in different ways due to the relevance of the items to their working context as well as the frame of reference effect, the present study indicated a lack of metric and scalar invariance between employees working in the in the plant and those working in the R&D lab. This supports the contentions that some safety climate items are more relevant to some work environments than others and/or when employees from different work settings use different referents.

Although some safety practices (e.g., wear personal protective equipment, the practices of handling hazardous material) are similar between employees in the plant and those in the lab, the work setting for R&D lab is also different from that of the plant (e.g., the types of hazards and risks). For instance, employees may have more exposure to dangerous heat and noise in the plant than those in the research lab, resulting in different safety procedures and practices. Thus, it is understandable why the safety climate measure was not equivalent for employees in the plant and employees in the lab; however additional research is needed to test what specifically causes this difference.

Theoretical Implications

The present findings have several theoretical implications. First, faultline theory suggests that the perceptions of organizational climate can be affected by faultlines, as they influence employees' sense-making (Lau & Murnighan, 1998) which is the key process of developing climate perceptions. That is, individuals from different subgroups are likely to have different perceptions of safety climate (Beus et al., 2012). The findings support that faultlines do indeed play an important role in employees' interpretation of the safety climate measure. Incorporating faultline theory into safety climate theory may provide new insights into the mechanisms through which the group-level and the organization-level safety climate emerge from individual-level safety climate. This may lead to the identification of faultline triggers that make the faultlines more or less salient such that safety climate becomes shared by employees from different faultline subgroups, facilitating or inhibiting the emergence of the group-level and the organization-level safety climate (Chrobot-Mason, Ruderman, Weber, & Ernst, 2009; Jehn & Bezrukova, 2010). For instance, Jehn and Bezrukova (2010) found that activated faultlines are more likely to form coalitions and group conflict, which may negatively affect the emergence of the group-level and organization-level climate.

Second, the results supported that national culture, language, and work environment all affect the measurement equivalence of the safety climate measure. These results are consistent with the empirical findings that the conceptualization or meaning of the constructs, the indicators (items) that capture the latent constructs, and the relevance of the items to the latent constructs vary across cultures and languages (e.g., Byrne et al., 2009; Poortinga, 1995; van de Vijver & Leung, 1997; van de Vijver & Tanzer, 1997). In other words, different items may be needed to capture the same latent construct across faultlines and/or the

effectiveness of the items to capture the latent construct may vary across cultures, languages, and work environments (e.g., Byrne et al., 2009; Poortinga, 1995; van de Vijver & Leung, 1997; van de Vijver & Tanzer, 1997). This may be the result of unintentional culture-related biases that influence the way researchers write and select items for inclusion in their safety climate measures.

Third, this study provided the first multilevel examination of the measurement equivalence of a safety climate measure. To the best of my knowledge, despite examining multilevel data, no other studies on the measurement equivalence of a safety climate measure take into account or model data dependency (e.g., multilevel data). In almost every organization employees are nested within groups, departments/units, or locations. Thus, it is not uncommon for safety climate data to be multilevel. When the multilevel structure of the data is not considered, the measurement equivalence tests will have inflated Type I error rate (i.e., higher likelihood of rejecting the null hypothesis of equivalence held between groups; Kim et al., 2012; Kim et al., in press). For instance, Beus et al. (2012) examined the configural and metric equivalence of a safety climate measure in the same data used in the present study, and concluded that the safety climate measure they examined achieved only the configural equivalence across hierarchical positions. However, the present study indicated that the safety climate measure achieved metric equivalence across hierarchical positions, providing further empirical evidence the measurement equivalence model is more likely to be rejected and misleadingly interpreted as non-equivalent.

Conceptually, safety climate is a multilevel construct (Zohar, 2000, 2003; Zohar & Luria, 2005). Based on Chan's (1998) typology of composition models, safety climate can be described as a referent-shift consensus model in which perceptions at different levels are

aligned. That is, the factor structure at the individual-level and at the group/organization-level should be the same. However, the fact that measurement equivalence of the safety climate measure holds at the within-level (i.e., individual-level) does not mean that the measurement equivalence of the safety climate measure will hold at a higher level (e.g., group-level, organization-level) as well. For instance, as Table 8 shows, the fit index (e.g., SRMR-B) for the between-level model (safety climate at the location level) actually indicated the metric equivalence at the location level was not achieved, whereas the fit indices for the individual-level safety climate model indicated that the metric equivalence model at the individual-level had acceptable model fit. Safety climate theories need to be developed to explain the discrepancy in the measurement equivalence tests of safety climate across different levels of analysis.

Practical Implications

The present findings have several practical implications. First, the present findings indicated that national culture, language, work environment, hierarchical position, and employment arrangement demonstrated important cross-sample differences that researchers and practitioners should consider when collecting data from organizations, as the existence of faultlines in the organization may prevent meaningful comparisons of the observed scores between different groups. Therefore, researchers and practitioners need to establish the equivalence of the safety climate measure before examining group differences. Otherwise, they may erroneously assume measurement equivalence across the compared groups and make misleading conclusions concerning the meaning of differences between groups (i.e., "comparing apples and oranges" (Vandenberg & Lance, 2000).

Second, multilevel researchers advocate confirming there is sufficient agreement across individual-level ratings before aggregating to a higher level (Bliese, 2000). This study provides empirical evidence that agreement within future safety climate assessments should be tested across multiple faultlines.

Finally, the use of different referents/standards in responding to scale items by different groups (the frame of reference effect) may be the reason why intercepts are not equivalent (Heine et al., 2002). To the extent that this is true, researchers and practitioners could use some strategies to avoid the frame of reference effect to ensure that individuals from different faultline groups assign the same meaning to the response options or the same numeric value to the scale anchor (e.g., "strongly agree"). One option would be to use behaviorally anchored rating scales, which provide behavioral descriptions for each rating or response option to ensure that individuals from different groups use the same standard or referent (e.g., Bernardin & Smith, 1981). Another strategy would be to enhance communication. For instance, consistent with the work environment faultline, managers and supervisors do not work side-by-side with front-line employees who engage in safety work practices every day (Cole & Bruch, 2006). As a result, they are less likely to be aware of underreported workplace accidents and injuries (e.g., Arthur et al., 2005; Burns & Wilde, 1995; Probst et al., 2008), giving them less exposure to negative safety referents compared to their subordinates. Encouraging communication (e.g., seeking employee input regarding organizational safety procedures, practices, and facilitating the open reporting of accidents and near misses) across employees from different faultline groups may help to establish the same standard/referents for them resulting in similar interpretation of the safety climate items

(Beus et al., 2012). Communication might also facilitate the emergence of group-level safety climate that is shared by employees from different faultline groups.

Limitations and Future Directions

Despite the numerous strengths to this study including a large, multinational field sample with multiple faultlines, there are some limitations to acknowledge. First, all the subgroup sizes are adequate for the analyses performed ($n > 200$, Kline, 2011), but the size of the subgroups for each faultline is not equal or balanced. The model estimates will be largely driven by the subgroup with the largest sample size (Kline, 2011) for multi-group analysis. For instance, for the employment arrangement faultline, there were significantly less independent contractors ($n = 230$) and dependent contractors ($n = 362$) than employees ($n = 8189$). However, follow-up analyses with matched sample sizes for each subgroup (by randomly drawing an equal number of individuals from each group) for each faultline resulted in identical results. That is, unbalanced sample sizes of the subgroups for each faultline did not appear to be an issue for the present study.

Second, culture was not directly assessed. In the present study, cultural differences were operationalized using the country in which the respondent worked. Using national culture scores assumes culture is homogenous within a country. In reality, culture resides within and is exhibited by individuals. That is, individual culture, such as individual values and beliefs, are not only shaped by the shared meaning system of a culture but also by the unique characteristics of each individual, such as personality (Chao & Moon, 2005). Individual-level culture might be more powerful in explaining the effects of culture on the interpretation of the safety climate measure. Future studies should investigate how individual-level culture influence employees' interpretation and perceptions of safety

climate. However, before researchers can investigate whether individual-level culture affects the measurement equivalence of the safety climate measure, researchers should first address how to meaningfully divide individuals into different groups based on their individual-level continuous cultural scores. In other words, how high should an individual's score be on the scale to be classified into an individualistic person?

Third, the present study only focused on the collectivism versus individualism cultural value dimension. Culture is a multilayered construct represented by values, assumptions, rituals, behaviors, and artifacts (Taras, Rowney, & Steel, 2009). Although researchers have argued that cultural value is the best operationalization of culture (Hofstede, 1980; Javidan, House, Dorfman, Hanges, & De Luque, 2006; Oyserman et al., 2002; Taras et al., 2009), it is very likely that other cultural facets threaten the measurement equivalence of the safety climate measure. For instance, occupational safety practice vary across countries, with different approaches to legislation, enforcement, and incentives for safety compliance. Countries vary on laws and standards with regard to workplace safety, which will influence the measurement equivalence of the safety climate measure (i.e., the frame-of-referent effect). Future studies are needed to assess and quantify variability in safety laws and then examine whether they affect employees' interpretation of safety climate measures.

Fourth, based on the current data, it is impossible to identify the exact source(s) that leads to measurement non-equivalence of a measure between different faultline groups. For instance, the present study indicated a lack of equivalence between respondents from individualist and collectivist countries. It is unclear if these differences are a result of differences in connotations of items and/or in relevance of items to the latent construct (Hulin, 1987), differences in the organizational culture by country (Candell & Hulin, 1986),

or differences in familiarity with surveys (Lonner, 1990). This is true for the other faultlines that indicated a lack of equivalence as well. For instance, language and translation are confounded; therefore, it is not possible to identify whether it is the language or the translation that led to the lack of equivalence between different linguistic groups. Additional research is needed to differentiate all these potential sources of nonequivalence.

Finally, the present study focused on individual faultlines rather than the combinations of different faultlines (e.g., individualistic English-speaking employees versus collectivistic Chinese-speaking employees). However, as the present study indicated that the measurement equivalence of the safety climate measure did not hold across all subgroups of the five faultlines under investigation (i.e., national culture, language, work environment, hierarchical position, and employment arrangement), it is reasonable to expect the measurement equivalence of the same safety climate measure will not exist across different combinations of these five faultlines.

Conclusions

The present study examined the measurement equivalence of a safety climate measure with a sample of 8,790 employees in a multinational chemical processing and manufacturing organization. The multilevel multi-group CFAs indicated that the factor loadings of the safety climate items and the intercepts of the measure were not equivalent between respondents from individualistic and collectivistic countries. The multilevel factor mixture models indicated that the factor loadings of the items were equivalent across hierarchical level and employment arrangement but the intercepts of the safety climate measure were not equivalent across language, hierarchical position, employment

arrangement, or work environment. In other words, it is not meaningful to compare safety climate scores across these five different faultlines.

REFERENCES

Adler, N. J. (1986). *International dimensions of organizational behavior*. Belmont, GA: Kent.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52,* 317-332.

Arthur, W., Jr., Bell, S. T., Edwards, B. D., Day, E. A., Tubre, T. C., & Tubre, A. H. (2005). Convergence of self-report and archival crash involvement data: A two-year longitudinal follow-up. *Human Factors, 47,* 303-313.

Bachman, J. G., & O'Malley, P. M. (1984). Yea-saying, nay-saying, and going to extremes: Black-white differences in response styles. *Public Opinion Quarterly, 48,* 491-509.

Bagozzi, R. P., & Edwards, J. R. (1998). A general approach for representing constructs in organizational research. *Organizational Research Methods, 1,* 45-87.

Barkema, H. G., & Vermeulen, F. (1997). What differences in the cultural backgrounds of partners are detrimental for international joint ventures?. *Journal of International Business Studies, 28,* 845-864.

Bernardin, H. J., & Smith, P. C. (1981). A clarification of some issues regarding the development and use of behaviorally anchored ratings scales (BARS). *Journal of Applied Psychology, 66,* 458-463.

Bartram, D., & Coyne, I. (1998). Variations in national patterns of testing and test use: The ITC/EFPPA International Survey. *European Journal of Psychological Assessment, 14,* 249-260.

Berry, J. W. (1969). On cross-cultural comparability. *International Journal of Psychology, 4*, 119-128.

Beus, J. M., Payne, S. C., Bergman, M. E., & Arthur Jr, W. (2010). Safety climate and injuries: An examination of theoretical and empirical relationships. *Journal of Applied Psychology, 95,* 713-727.

Beus, J. M., Jarrett, S. M., Bergman, M. E., & Payne, S. C. (2012). Perceptual equivalence of psychological climates within groups: When agreement indices do not agree. *Journal of Occupational and Organizational Psychology, 85,* 454-471.

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 349–381). San Francisco: Jossey-Bass.

Blondel, J., & Inoguchi, T. (2006). *Political cultures in Asia and Europe: Citizens, states and societal values*. New York: Routledge.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: John Wiley.

Bureau of Labor Statistics. (2013). Census of fatal occupational injuries (CFOI) – Current and revised data. Retrieved from: http://www.bls.gov/iif/oshcfoi1.htm

Burke, M. J., Chan-Serafin, S., Salvador, R., Smith, A., & Sarpy, S. A. (2008). The role of national culture and organizational climate in safety training effectiveness. *European Journal of Work and Organizational Psychology, 17,* 133-152.

Burns, P. C., & Wilde, G. J. S. (1995). Risk taking in male taxi drivers: Relationships among personality, observational data, and driver records. *Personality and Individual Differences, 18*, 267-278.

Byrne, B. M., Oakland, T., Leong, F. T., van de Vijver, F. J., Hambleton, R. K., Cheung, F. M., & Bartram, D. (2009). A critical analysis of cross-cultural research and testing practices: Implications for improved education and training in psychology. *Training and Education in Professional Psychology, 3,* 94-105.

Biernat, M., & Kobrynowicz, D. (1997). Gender- and race-based standards of competence: Lower minimum standards but higher ability standards for devalued groups. *Journal of Personality and Social Psychology, 72,* 544 –557.

Biernat, M., & Manis, M. (1994). Shifting standards and stereotype-based judgments. *Journal of Personality and Social Psychology, 66,* 5–20.

Biernat, M., Manis, M., & Nelson, T. E. (1991). Stereotypes and standards of judgment. *Journal of Personality and Social Psychology, 60,* 485–499.

Billiet, J. B., & McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling, 7,* 608-628.

Bond, R., & Smith, P. B. (1996). Culture and conformity: A meta-analysis of studies using Asch's (1952b, 1956) line judgment task. *Psychological Bulletin, 119,* 111-137.

Brannick, M. T. (1995). Critical comments on applying covariance structure modeling. *Journal of Organizational Behavior, 16,* 201-213.

Brislin, R. W. (1980). Translation and content analysis of oral and written material. *Handbook of Cross-Cultural Psychology, 2,* 349-444.

Byrne, B. M., & Campbell, T. L. (1999). Cross-cultural comparisons and the presumption of equivalent measurement and theoretical structure: A look beneath the surface. *Journal of Cross-Cultural Psychology, 30,* 555-574.

Candell, G. L., & Hulin, C. L. (1986). Cross-language and cross-cultural comparisons in scale translations independent sources of information about item nonequivalence. *Journal of Cross-Cultural Psychology, 17,* 417-440.

Carmines, E. G., & McIver, J. P. (1981). Analyzing models with unobserved variables: Analysis of covariance structures. In G. W. Bohrnstedt, & E. F. Borgattta (Eds.), *Social measurement: Current issues* (pp. 65–115). Beverly Hills: Sage.

Carter, N. T., Kotrba, L. M., & Lake, C. J. (2014). Null results in assessing survey score comparability: Illustrating measurement invariance using item response theory. *Journal of Business and Psychology, 29,* 205-220.

Chan, D. (1998). Functional relations among constructs in the same content domain at different levels of analysis: A typology of composition models. *Journal of Applied Psychology, 83,* 234-246.

Chan D. (2000). Detection of differential item functioning on the Kirton Adaptation-Innovation Inventory using multiple-group mean and covariance structure analyses. *Multivariate Behavioral Research, 35,* 169–199.

Chao, G. T., & Moon, H. (2005). The cultural mosaic: A meta-theory for understanding the complexity of culture. *Journal of Applied Psychology, 90,* 1128-1140.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14,* 464-504.

Chen, F. F., Sousa, K. H., & West, S. G. (2005). Teacher's corner: Testing measurement invariance of second-order factor models. *Structural Equation Modeling, 12,* 471-492.

Chen, F. F., & West, S. G. (2008). Measuring individualism and collectivism: The importance of considering differential components, reference groups, and measurement invariance. *Journal of Research in Personality, 42,* 259-294.

Cheyne, A., Tomas, J. M., Cox, S., & Oliver, A. (2003). Perceptions of safety climate at different employment levels. *Work & Stress, 17,* 21-37.

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,* 1103-1127.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9,* 233-255.

Cheyne, A., Cox, S., Oliver, A., & Tomás, J. M. (1998). Modeling safety climate in the prediction of levels of safety activity. *Work & Stress, 12,* 255-271.

Cheyne, A., Tomas, J. M., Cox, S., & Oliver, A. (2003). Perceptions of safety climate at different employment levels. *Work & Stress, 17,* 21-37.

Church, A. T. (2001). Personality measurement in cross‐cultural perspective. *Journal of Personality, 69,* 979-1006.

Cigularov, K. P., Adams, S., Gittleman, J. L., Haile, E., & Chen, P. Y. (2013). Measurement equivalence and mean comparisons of a safety climate measure across construction trades. *Accident Analysis & Prevention, 51,* 68-77.

Cigularov, K. P., Lancaster, P. G., Chen, P. Y., Gittleman, J., & Haile, E. (2013). Measurement equivalence of a safety climate measure among Hispanic and White Non-Hispanic construction workers. *Safety Science, 54,* 58-68.

Clarke, S. (2003). The contemporary workforce: Implications for organisational safety culture. *Personnel Review, 32,* 40-57.

Chao, G. T., & Moon, H. (2005). The cultural mosaic: A meta-theory for understanding the complexity of culture. *Journal of Applied Psychology, 90,* 1128.

Chrobot-Mason, D., Ruderman, M. N., Weber, T. J., & Ernst, C. (2009). The challenge of leading on unstable ground: Triggers that activate social identity faultlines. *Human Relations, 62,* 1763-1794.

Cronbach, L.J. (1960), *Essentials of psychological testing*. New York: Harper & Row.

Cole, M. S., & Bruch, H. (2006). Organizational identity strength, identification, and commitment and their relationships to turnover intention: Does organizational hierarchy matter? *Journal of Organizational Behavior, 27,* 585–605.

Cole, D. A., & Maxwell, S. E. (1985). Multitrait-multimethod comparisons across populations: A confirmatory factor analytic approach. *Multivariate Behavioral Research, 20,* 389-417.

Connelly, C. E., & Gallagher, D. G. (2004). Emerging trends in contingent work research. *Journal of Management, 30,* 959-983.

Couch, A., & Keniston, K. (1960). Yeasayers and naysayers: Agreeing response set as a personality variable. *The Journal of Abnormal and Social Psychology, 60,* 151-174.

Cox, S. J., & Cheyne, A. J. T. (2000). Assessing safety culture in offshore environments. *Safety Science, 34,* 111-129.

Cox, S., Tomás, J. M., Cheyne, A., & Oliver, A. (1998). Safety culture: The prediction of commitment to safety in the manufacturing industry. *British Journal of Management, 9,* 3-11.

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement, 6,* 475-494.

Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement, 10,* 3-31.

Crowl, D. A., & Louvar, J. F. (2002). *Chemical process safety: Fundamentals with applications* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24,* 349-354.

Cunningham, W. H., Cunningham, I. C., & Green, R. T. (1977). The ipsative process to reduce response set bias. *Public Opinion Quarterly, 41,* 379-384.

Curran, P. J. (2003). Have multilevel models been structural equation models all along? *Multivariate Behavioral Research, 38,* 529–568.

De Leeuw, J., & Kreft, I. (1986). Random coefficient models for multilevel analysis. *Journal of Educational and Behavioral Statistics, 11,* 57-85.

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are central issues. *Psychological Bulletin, 95,* 134-135.

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology, 72,* 19-29.

England, G. W., & Harpaz, I. (1983). Some methodological and analytic considerations in cross-national comparative research. *Journal of International Business Studies, 14,* 49-59.

Eysenck, H. J. & Eysenck, S. B. G. (1964). *Manual of the Eysenck Personality Inventory*. London: University of London Press.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations, 7,* 117-140.

Flin, R., Mearns, K., O'Connor, P., & Bryden, R. (2000). Measuring safety climate: Identifying the common features. *Safety Science, 34,* 177-192.

Frone, M. R. (1998). Predictors of work injuries among employed adolescents. *Journal of Applied Psychology, 83,* 565-576.

Gittelman, M., McMahon, C. G., Rodríguez‐Rivera, J. A., Beneke, M., Ulbrich, E., & Ewald, S. (2010). The POTENT II randomised trial: Efficacy and safety of an orodispersible vardenafil formulation for the treatment of erectile dysfunction. *International Journal of Clinical Practice, 64,* 594-603.

Glendon, A. I., & Litherland, D. K. (2001). Safety climate factors, group differences and safety behaviour in road construction. *Safety Science, 39,* 157-188.

Greenleaf, E. A. (1992). Measuring extreme response style. *Public Opinion Quarterly, 56,* 328-351.

Grimm, S. D., & Church, A. T. (1999). A cross-cultural study of response biases in personality measures. *Journal of Research in Personality, 33,* 415-441.

Guldenmund, F. W. (2007). The use of questionnaires in safety culture research–An evaluation. *Safety Science, 45,* 723-743.

Guptara, P., Murray, K., Razak, B., & Sheehan, T. (1990). The art of training abroad. *Training & Development Journal, 44,* 13-18.

Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. *Adapting Educational and Psychological Tests for Cross-cultural Assessment, 1,* 3-38.

Heine, S. J., & Lehman, D. R. (1997). The cultural construction of self-enhancement: An examination of group-serving biases. *Journal of Personality and Social Psychology, 72,* 1268-1283.

Heine, S. J., Lehman, D. R., Peng, K., & Greenholtz, J. (2002). What's wrong with cross-cultural comparisons of subjective Likert scales?: The reference-group effect. *Journal of Personality and Social Psychology, 82,* 903-918.

Hsu, H. Y., Kwok, O. M., Lin, J. H., & Acosta, S. (2015). Detecting misspecified multilevel structural equation models with common fit indices: A Monte Carlo Study. *Multivariate Behavioral Research, 50,* 197-215.

Hulin, C. L. (1987). A psychometric theory of evaluations of item and scale translations fidelity across languages. *Journal of Cross-cultural Psychology, 18,* 115-142.

Hofmann, D. A., & Stetzer, A. (1996). A cross‐level investigation of factors influencing unsafe behaviors and accidents. *Personnel Psychology, 49,* 307-339.

Hofstede, G. H. (1980). *Culture's consequences: International differences in work-related values.* Beverly Hills, CA: Sage.

Hofstede, G. (1983). The cultural relativity of organizational practices and theories. *Journal of International Business Studies, 14,* 75-89.

Hofstede, G. (1992). Cultural dimensions in people management: the socialization perspective. In V. Pucik, N.M. Tichy, and C.K. Barnett (Eds.), *Globalizing Management: Creating and Leading the Competitive Organization*. New York, Wiley, 1992, 139-158.

Hofstede, G. H. (2001). *Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations* (2nd ed.). Thousand Oaks, CA: Sage.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18,* 117-144.

Hoshmand, L. T., & Ho, D. Y. F. (1995). Moral dimensions of selfhood: Chinese traditions and cultural change. *World Psychology, 1,* 47-69.

House, R.J., Hanges, P.J., Javidan, M., et al., (2004). *Culture, leadership, and organizations: The GLOBE study of 62 societies* (Eds.). Thousand Oaks, CA: Sage

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6,* 1-55.

Huang, X., & Vliert, E. V. D. (2004). Job level and national culture as joint roots of job satisfaction. *Applied Psychology, 53,* 329-348.

Huang, Y. H., Robertson, M. M., Lee, J., Rineer, J., Murphy, L. A., Garabet, A., & Dainoff, M. J. (2014). Supervisory interpretation of safety climate versus employee safety climate perception: Association with safety behavior and outcomes for lone workers. *Transportation Research Part F: Traffic Psychology and Behavior, 26,* 348-360.

Huang, Y. H., Zohar, D., Robertson, M. M., Garabet, A., Lee, J., & Murphy, L. A. (2013).

Development and validation of safety climate scales for lone workers using truck drivers as exemplar. *Transportation Research Part F: Traffic Psychology and Behaviour*, *17*, 5-19. doi:10.1016/j.trf.2012.08.011

Hui, C. H., & Triandis, H. C. (1985). Measurement in cross-cultural psychology: A review and comparison of strategies. *Journal of Cross-cultural Psychology, 16,* 131-152.

Hui, C. H., & Triandis, H. C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology, 20,* 296-309.

Hulin, C. L. (1987). A psychometric theory of evaluations of item and scale translations fidelity across languages. *Journal of Cross-cultural Psychology, 18,* 115-142.

Hulin, C. L., Drasgow, F., & Komocar, J. (1982). Applications of item response theory to analysis of attitude scale translations. *Journal of Applied Psychology, 67,* 818-825.

Hyman, H. (1942). The psychology of subjective status. *Psychological Bulletin, 39,* 473–474.

Inglehart, R., & Baker, W. E. (2000). Modernization, cultural change, and the persistence of traditional values. *American Sociological Review, 65,* 19-51.

Janssens, M., Brett, J. M., & Smith, F. J. (1995). Confirmatory cross-cultural research: Testing the viability of a corporation-wide safety policy. *Academy of Management Journal, 38,* 364-382.

Javidan, M., House, R. J., Dorfman, P. W., Hanges, P. J., & De Luque, M. S. (2006). Conceptualizing and measuring cultures and their consequences: A comparative review of GLOBE's and Hofstede's approaches. *Journal of International Business Studies, 37,* 897-914.

Jehn, K. A., & Bezrukova, K. (2010). The faultline activation process and the effects of activated faultlines on coalition formation, conflict, and group outcomes. *Organizational Behavior and Human Decision Processes, 112,* 24-42.

Johnson, T., Kulesa, P., Cho, Y. I., & Shavitt, S. (2005). The relation between culture and response styles evidence from 19 countries. *Journal of Cross-cultural psychology, 36,* 264-277.

Johnson, T. P., & Van de Vijver, F. J. (2003). Social desirability in cross-cultural research. *Cross-cultural Survey Methods, 1,* 195-204.

Jones-Farmer, L. A. (2010). The effect of among-group dependence on the invariance likelihood ratio test. *Structural Equation Modeling, 17,* 464–480.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34,* 183-202.

Kelloway, E. K. (1995). Structural equation modeling in perspective. *Journal of Organizational Behavior, 16*, 215-224.

Kerr, C., Dunlop, J. T., & Harbison, F. H. (1960). *Industrialism and industrial man: The problems of labor and management in economic growth*. Cambridge, MA: Harvard University Press.

Kim, E. S., Kwok, O. M., & Yoon, M. (2012). Testing factorial invariance in multilevel data: A Monte Carlo study. *Structural Equation Modeling: A Multidisciplinary Journal, 19,* 250-267.

Kim, E. S., Yoon, M., Wen, Y., Luo, W., & Kwok, O. M. (in press). Within-level group factorial invariance with multilevel data: Multilevel factor mixture and multilevel

MIMIC models. *Structural Equation Modeling: A Multidisciplinary Journal.* DOI: 10.1080/10705511.2014.938217

Kirchmeyer, C., & McLellan, J. (1991). Capitalizing on ethnic diversity: An approach to managing the diverse workgroups of the 1990s. *Canadian Journal of Administrative Sciences, 8,* 72-79.

Kish, L. (1995). *Survey Sampling.* New York: Wiley

Lanning, K. (1991). *Consistency, scalability and personality measurement*. New York: Springer-Verlag.

Lau, D. C., & Murnighan, J. K. (1998). Demographic diversity and faultlines: The compositional dynamics of organizational groups. *Academy of Management Review, 23,* 325–340.

Lalwani, A. K., Shavitt, S., & Johnson, T. (2006). What is the relation between cultural orientation and socially desirable responding?. *Journal of Personality and Social Psychology, 90,* 165-178.

Lee, C., & Green, R. T. (1991). Cross-cultural examination of the Fishbein behavioral intentions model. *Journal of International Business Studies, 22,* 289-305.

Liberty Mutual Research Institute for Safety, (2013). 2013 Liberty mutual workplace safety index. Retrieved from: http://www.libertymutualgroup.com/research

Lin, S. H., Tang, W. J., Miao, J. Y., Wang, Z. M., & Wang, P. X. (2008). Safety climate measurement at workplace in China: A validity and reliability assessment. *Safety Science, 46,* 1037-1046.

Liu, C., Borg, I., & Spector, P. E. (2004). Measurement equivalence of the German Job Satisfaction Survey used in a multinational organization: Implications of Schwartz's culture model. *Journal of Applied Psychology, 89,* 1070-1082.

Ma, Q., & Yuan, J. (2009). Exploratory study on safety climate in Chinese manufacturing enterprises. *Safety Science, 47,* 1043-1046.

Maas, C. J., & Hox, J. J. (2002). Robustness of multilevel parameter estimates against small sample sizes. Unpublished Paper, Utrecht University.

Marin, G., Gamba, R. J., & Marin, B. V. (1992). Extreme response style and acquiescence among Hispanics: The role of acculturation and education. *Journal of Cross-Cultural Psychology, 23,* 498-509.

Marsh, H. W. (1994). Confirmatory factor analysis models of factorial invariance: A multifaceted approach. *Structural Equation Modeling, 1,* 5-34.

McDonald, N., & Ryan, F. (1992). Constraints on the development of safety culture: A preliminary analysis. *The Irish Journal of Psychology, 13,* 273-281.

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods, 10,* 259.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7,* 105-118.

Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin, 115,* 300-307.

Meredith, W., & Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika, 57,* 289-311.

Millsap, R. E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods, 2,* 248.

Millsap, R. E., & Kwok, O. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods, 9,* 93–115.

Morris, T., & Pavett, C. M. (1992). Management style and productivity in two cultures. *Journal of International Business Studies, 23,* 169-179.

Mullen, M. R. (1995). Diagnosing measurement equivalence in cross-national research. *Journal of International Business Studies, 26,* 573-596.

Muthén, L.K. and Muthén, B.O. (1998-2012). *Mplus User's Guide* (7th edition). Los Angeles, CA: Muthén & Muthén.

Nahrgang, J. D., Morgeson, F. P., & Hofmann, D. A. (2011). Safety at work: A meta-analytic investigation of the link between job demands, job resources, burnout, engagement, and safety outcomes. *Journal of Applied Psychology, 96*, 71-94.

Oakland, T., & Hu, S. (1994). International perspectives on tests used with children and youths. *Journal of School Psychology, 31,* 501-517.

Oakland, T. (1997). Test use among school psychologists: Past, current, and emerging practices. *European Journal of Psychological Assessment, 13,* 2-9.

Oakland, T. (2004). Use of educational and psychological tests internationally. *Applied psychology: An International Review, 53,* 157-172.

Oakland, T., & Hu, S. (1992). The top 10 tests used with children and youth worldwide. *Bulletin of the International Test Commission, 19,* 99-120.

Oreg, S., Bayazit, M., Vakola, M., Arciniega, L., Armenakis, A., Barkauskiene, R., ... & van Dam, K. (2008). Dispositional resistance to change: Measurement equivalence and the link to personal values across 17 nations. *Journal of Applied Psychology, 93,* 935-944.

Oyserman, D., Coon, H. M., & Kemmelmeier, M. (2002). Rethinking individualism and collectivism: Evaluation of theoretical assumptions and meta-analyses. *Psychological Bulletin, 128,* 3-72.

Oyserman, D., & Lee, S. W. (2008). Does culture influence what and how we think? Effects of priming individualism and collectivism. *Psychological Bulletin, 134,* 311-342.

Parker, D., Brosseau, L., Samant, Y., Pan, W., Xi, M., Haugan, D. (2007). A comparison of the perceptions and beliefs of workers and owners with regard to workplace safety in small metal fabrication businesses. *American Journal of Industrial Medicine, 50,* 999–1009.

Peng, K., Nisbett, R. E., & Wong, N. Y. C. (1997). Validity problems comparing values across cultures and possible solutions. *Psychological Methods, 2,* 329–344.

Picard, M., Girard, S. A., Simard, M., Larocque, R., Leroux, T., & Turcotte, F. (2008). Association of work-related accidents with noise exposure in the workplace and noise-induced hearing loss based on the experience of some 240,000 person-years of observation. *Accident Analysis & Prevention, 40,* 1644-1652.

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology, 46,* 598-609.

Paulhus, D. L. (1991). Measurement and control of response bias in questionnaires. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). San Diego, CA: Academic.

Paulhus, D. L., & Reid, D. B. (1991). Enhancement and denial in socially desirable responding. *Journal of Personality and Social Psychology, 60,* 307-317.

Poortinga, Y. H. (1995). Cultural bias in assessment: Historical and thematic issues. *European Journal of Psychological Assessment, 11,* 140-146.

Probst, T. M., Brubaker, T. L., & Barsotti, A. (2008). Organizational injury rate underreporting: The moderating effect of organizational safety climate. *Journal of Applied Psychology, 93,* 1147-1154.

Prussia, G. E., Brown, K. A., & Willis, P. G. (2003). Mental models of safety: Do managers and employees see eye to eye?. *Journal of Safety Research, 34,* 143-156.

Rabinowitz, P. M. (2000). Noise-induced hearing loss. *American Family Physician, 61,* 2759-2760.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87,* 517.

Ramsey, J. D., Burford, C. L., Beshir, M. Y., & Jensen, R. C. (1983). Effects of workplace thermal conditions on safe work behavior. *Journal of Safety Research, 14,* 105-114.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114,* 552-566.

Rensvold, R. B., & Cheung, G. W. (1998). Testing measurement models for factorial invariance: A systematic approach. *Educational and Psychological Measurement, 58,* 1017-1034.

Robert, C., Lee, W. C., & Chan, K. Y. (2006). An empirical analysis of measurement equivalence with the INDCOL measure of individualism and collectivism: Implications for valid cross‐cultural inference. *Personnel Psychology, 59,* 65-99.

Rochlin, G. I. (1999). Safe operation as a social construct. *Ergonomics, 42,* 1549-1560.

Rousseau, D. M., & Libuser, C. (1997). Contingent workers in high risk environments. *California Management Review, 39,* 103-123.

Ryan, A., Chan, D., Ployhart, R. E., & Slade, L. A. (1999). Employee attitude surveys in a multinational organization: Considering language and culture in assessing measurement equivalence. *Personnel Psychology, 52,* 37-58.

Ryan, A. N. N., McFarland, L., & Shl, H. B. (1999). An international look at selection practices: Nation and culture as explanations for variability in practice. *Personnel Psychology, 52,* 359-392.

Ryu, E. (2014). Factorial invariance in multilevel confirmatory factor analysis. *British Journal of Mathematical and Statistical Psychology, 67,* 172-194.

Riordan, C. M., & Vandenberg, R. J. (1994). A central question in cross-cultural research: Do employees of different cultures interpret work-related measures in an equivalent manner? *Journal of Management, 20,* 643-671.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variables*

analysis: Applications for developmental research (pp. 399–419). Thousand Oaks, CA: Sage.

Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika, 75,* 243–248.

Satorra, A., & Muthen, B. (1995). Complex sample data in structural equation modeling. *Sociological Methodology, 25,* 267-316.

Schmitt, N. (1982). The use of analysis of covariance structures to assess beta and gamma change. *Multivariate Behavioral Research, 17,* 343-358.

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review, 18,* 210-222.

Schneider, B., & Reichers, A. E. (1983). On the etiology of climates. *Personnel Psychology, 36,* 19-39.

Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6,* 461-464.

Sclove, L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52,* 333-343.

Schwartz, S. (1999). A theory of cultural values and some implications for work. *Applied Psychology: An International Review, 48,* 23–47.

Schwarz, N., & Oyserman, D. (2001). Asking questions about behavior: Cognition, communication, and questionnaire construction. *American Journal of Evaluation, 22,* 127-160.

Selig, J. P., Card, N. A., & Little, T. D. (2008). Latent variable structural equation modeling in cross-cultural research: Multigroup and multilevel approaches. In F. J. R. van de Vijver, D. A. van Hemert, & Y. H. Poortinga (Eds.), *Multilevel analysis of individuals and cultures* (pp. 93–120). New York: Erlbaum.

Sherif, M. A. (1936). *The psychology of social norms*. New York: Harper.

Shulruf, B., Hattie, J., & Dixon, R. (2011). Intertwinement of individualist and collectivist attributes and response sets. *Journal of Social, Evolutionary, and Cultural Psychology, 5,* 51-65.

Søndergaard, M. (1994). Research note: Hofstede's consequences: A study of reviews, citations and replications. *Organization studies, 15,* 447-456.

Steiger, J. H., & Lind, J. C. (1980, May). *Statistically based tests for the number of factors*. Paper presented at the annual spring meeting of the Psychometric Society, Iowa City, IA.

Tanzer, N. K. (1995). Cross-cultural bias in Likert-type inventories: Perfect matching factor structures and still biased?. *European Journal of Psychological Assessment, 11,* 194-201.

Tanzer, N. K., Sim, C. Q. E., & Marsh, H. W. (1992). Test applications over cultures and languages: Theoretical considerations and empirical findings. *Bulletin of the International Test Commission, 19,* 151-171.

Taras, V., Rowney, J., & Steel, P. (2009). Half a century of measuring culture: Review of approaches, challenges, and limitations based on the analysis of 121 instruments for quantifying culture. *Journal of International Management, 15,* 357-373.

Taris, T. W., Bok, I. A., & Meijer, Z. Y. (1998). Assessing stability and change of psychometric properties of multi-item concepts across different situations: A general approach. *Journal of Psychology, 132,* 301-316.

Triandis, H. C. (1989). The self and social behavior in differing cultural contexts. *Psychological Review, 96,* 506-520.

Triandis, H. C. (2004). The many dimensions of culture. *The Academy of Management Executive, 18,* 88-93.

Triandis, H. C., Carnevale, P., Gelfand, M., Robert, C., Wasti, A., Probst, T., ... & Schmidt, P. (2001). Culture, personality and deception: A multilevel approach. *International Journal of Cross-Cultural Management, 1,* 73-90.

Triandis, H. C., Leung, K., Villareal, M. J., & Clack, F. I. (1985). Allocentric versus idiocentric tendencies: Convergent and discriminant validation. *Journal of Research in Personality, 19,* 395-415.

Triandis, H. C., & Marin, G. (1983). Etic plus emic versus pseudoetic: A test of a basic assumption of contemporary cross-cultural psychology. *Journal of Cross-Cultural Psychology, 14,* 489-500.

Triandis, H. C., & Suh, E. M. (2002). Cultural influences on personality. *Annual review of psychology, 53,* 133-160.

Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods, 5,* 139-158.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3,* 4-70.

Vandenberg, R. J., & Self, R. M. (1993). Assessing newcomers' changing commitment to the organization during the first 6 months of work. *Journal of Applied Psychology, 78,* 557-568.

Van Hemert, D. A., Van de Vijver, F. J., Poortinga, Y. H., & Georgas, J. (2002). Structural and functional equivalence of the Eysenck Personality Questionnaire within and between countries. *Personality and Individual Differences, 33,* 1229-1249.

Van Herk, H., Poortinga, Y. H., & Verhallen, T. M. (2004). Response styles in rating scales evidence of method bias in data from six EU countries. *Journal of Cross-Cultural Psychology, 35,* 346-360.

Van de Vijver, F. J., & Leung, K. (1997). *Methods and data analysis for cross-cultural research* (Vol. 1). Thousand Oaks, CA: Sage.

Van de Vijver, F., & Tanzer, N. K. (1997). Bias and equivalence in cross-cultural assessment. *European Review of Applied Psychology, 47,* 263-279.

Volkmann, J. (1951). Scales of judgment and their implications for social psychology. In J. H. Rohrer & M. Sherif (Eds.), *Social psychology at the crossroads*. New York: Harper.

Voronov, M., & Singer, J. A. (2002). The myth of individualism-collectivism: A critical review. *The Journal of Social Psychology, 142,* 461-480.

Wallace, J. C., Popp, E., & Mondore, S. (2006). Safety climate as a mediator between foundation climates and occupational accidents: A group-level investigation. *Journal of Applied Psychology, 91,* 681.

Winkler, J. D., Kanouse, D. E., & Ware, J. E. (1982). Controlling for acquiescence response set in scale development. *Journal of Applied Psychology, 67,* 555-561.

Woehr, D. J., Sheehan, M. K., & Bennett Jr, W. (2005). Assessing measurement equivalence across rating sources: A multitrait-multirater approach. *Journal of Applied Psychology, 90,* 592-600.

Wu, J., & Kwok, O. (2012). Using SEM to analyze complex survey data: A comparison between design-based single-level and model-based multilevel approaches. *Structural Equation Modeling, 19,* 16–35.

Yoon, M., & Millsap, R. E. (2007). Detecting violations of factorial invariance using data-based specification searches: A Monte Carlo study. *Structural Equation Modeling, 14,* 435–463.

Zax, M., & Takahashi, S. (1967). Cultural influences on response style: Comparisons of Japanese and American college students. *The Journal of Social Psychology, 71,* 3-10.

Zhou, Q., Fang, D., & Wang, X. (2008). A method to identify strategies for the improvement of human safety behavior by considering safety climate and personal experience. *Safety Science, 46,* 1406-1419.

Zohar, D., (1980). Safety climate in industrial organizations: Theoretical and applied implications. *Journal of Applied Psychology 65,* 96–102.

Zohar, D. (2000). A group-level model of safety climate: Testing the effect of group climate on micro-accidents in manufacturing jobs. *Journal of Applied Psychology, 85,* 587-596.

Zohar, D., (2002). The effects of leadership dimensions, safety climate, and assigned priorities on minor injuries in work groups. *Journal of Organizational Behavior 23,* 75–92.

Zohar, D. (2003). The influence of leadership and climate on occupational health and safety. In: Hofmann, D.A., Tetrick, L.E. (Eds.), *Health and safety in organizations: A multilevel perspective* (pp. 201-230). San Francisco, CA: Jossey-Bass.

Zohar, D. (2010). Thirty years of safety climate research: Reflections and future directions. *Accident Analysis & Prevention, 42,* 1517-1522.

Zohar, D., & Luria, G., (2004). Climate as a social-cognitive construction of supervisory safety practices: Scripts as proxy of behavior patterns. *Journal of Applied Psychology, 89,* 322–333.

Zohar, D., & Luria, G. (2005). A multilevel model of safety climate: Cross-level relationships between organization and group-level climates. *Journal of Applied Psychology, 90,* 616-628.

Zyphur, M. J., Kaplan, S. A., & Christian, M. S. (2008). Assumptions of cross-level measurement and structural invariance in the analysis of multilevel data: Problems and solutions. *Group Dynamics Theory Research and Practice, 12,* 127–140.

APPENDIX A

Safety Climate

Please indicate your level of agreement with each of the following statements. 5- point agreement scale (1=strongly disagree, 5 = strongly agree, NA).

1. Site management focuses on process safety in audits, self-assessments, and inspections.

2. Site management considers health and safety when setting production rates and schedules.

3. Site management provides all necessary safety equipment for workers.

4. Site management focuses on safety in audits, self-assessments, and inspections.

5. My supervisor is strict about working safely at all times even when we are tired or stressed.

6. Site management is strict about working safely at all times even when work falls behind schedule.

7. My supervisor frequently discusses health and safety issues throughout the work week.[a]

8. My supervisor insists we wear our protective equipment even if it is uncomfortable.

Note. [a] This item was dropped from the measurement equivalence tests as the multilevel confirmatory factor analysis indicated that this item correlates strongly with items 6 and 8.

APPENDIX B

Results of Measurement Equivalence Tests for Individualistic Countries

	χ^2 (*df*)	RMSEA	CFI	SRMR
Configural Equivalence	642.28(70)*	.09	.93	.05
Metric Equivalence	731.04(94)*	.08	.82	.08
Scalar Equivalence	1236.65(118)*	.10	.87	.11

Note. * p <.05.

Congfigural equivalence held between individualistic countries.

APPENDIX C

Results of Measurement Equivalence Tests for Collectivistic Countries

	χ^2 (*df*)	RMSEA	CFI	SRMR
Configural Equivalence	281.10(70)*	.07	.95	.04
Metric Equivalence	398.04(94)*	.07	.92	.10
Scalar Equivalence	887.69(118)*	.10	.80	.13

Note. * p <.05.

Configural equivalence held between Collectivistic Countries.