Accident Analysis and Prevention 121 (2018) 321–334

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

The measurement equivalence of a safety climate measure across five faultlines

Xiaohong Xu[a,*], Stephanie C. Payne[b], Mindy E. Bergman[b]

[a] Department of Psychology, Old Dominion University, Norfolk, VA, 23529, United States
[b] Department of Psychological and Brain Sciences, Texas A&M University, College Station, TX, 77843, United States

ARTICLE INFO

Keywords:
Safety climate
Multilevel measurement equivalence
Faultlines
National culture
Language
Organizational hierarchy
Employment arrangements
Work environments

ABSTRACT

This study examines the appropriateness of comparing safety climate survey responses across multiple faultlines—hypothetical dividing lines that split a group into subgroups based on one or more attributes. Using survey data from 8790 employees of a multinational chemical processing and manufacturing company from 76 work sites nested within 19 different countries, we examined the multilevel measurement equivalence of a safety climate measure across cultural dimensions, survey languages, organizational hierarchy, employment arrangements, and work environments. As simulation studies support the faultline at the individual-level requires measurement equivalence tests that are different from the faultline at the country-level, we used multi-group multilevel confirmatory factor analyses for the Level-3 faultline, and multilevel factor mixture models for known classes for the Level-1 faultlines. The results demonstrated that faultlines can prevent safety climate measurement equivalence, which prohibits the aggregation of individual-level scores to higher levels and making comparisons across faultlines. This first study on multilevel safety climate measurement equivalence serves as both a warning to safety climate researchers and practitioners regarding the importance of faultlines and reminds us to consider the level of the faultlines when testing measurement equivalence with multilevel data.

1. Introduction

Workplace safety is an important issue for organizations and their employees. In 2013, over 1 million employees in the US reported nonfatal injuries and illnesses that resulted in lost work days; in 2014, 4679 workers were killed on the job in the US (U.S. Bureau of Labor Statistics, 2013, 2014). Worldwide, more than 337 million workplace accidents occur each year (International Labour Organization, 2015). One of the strongest predictors of workplace safety behavior and workplace safety-related outcomes is safety climate, or 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 et al., 2010; Christian et al., 2009; Nahrgang et al., 2011). Thus, having a valid measure of safety climate, as well as a strong safety climate assessment program, are essential components to a workplace safety program.

One challenge in implementing a strong safety climate assessment program is that (a) the manifestation of safety climate varies across contexts and/or (b) employees from different contexts might not interpret a safety climate measure in the same way (Zohar, 2010). This raises a very important practical question: is it appropriate to compare safety climate scores across organizational groups? Within an organization, there are a multitude of meaningful groups. For example, worksites within the same organization are embedded in a variety of national cultures, have different processes and products associated with them, and use different operational languages, among other factors. Within worksites, there are employee differences in hierarchical position, work arrangements, jobs, language spoken, demographic variables, and other factors. Practically speaking, it is unknown whether comparisons of the observed safety climate scores across these groups can be made with any confidence.

These factors—both between and within worksites—have been conceptualized as faultlines, the hypothetical dividing lines that split a group (organization) into subgroups based on one or more attributes (Lau & Murnighan, 1998). Depending on the similarity and salience of individuals' attributes, many potential faultlines exist within groups, and each of these faultlines could activate important subgroupings (Lau & Murnighan, 1998). In reference to the practical question, if individuals from different subgroups created by faultlines interpret the safety climate scale differently, then combining or comparing subgroup safety climate scores is inappropriate (Vandenberg & Lance, 2000).

* Corresponding author.
E-mail address: x3xu@odu.edu (X. Xu).

https://doi.org/10.1016/j.aap.2018.05.011
Received 11 December 2017; Received in revised form 26 March 2018; Accepted 9 May 2018
Available online 21 May 2018

Complicating the matter further is the multilevel nature of organizations and corresponding data. Although the measurement equivalence of safety climate measures has been tested across a number of faultlines (e.g., hierarchical position (Beus et al., 2012; Cheyne et al., 2003; Huang et al., 2014); organizational heritage following mergers and acquisitions (Beus et al., 2012); countries (Barbaranelli et al., 2015; Reader et al., 2015); the combination of language and race (Cigularov et al., 2013)), the multilevel nature of the organizational data was not modeled in any of these studies. It is unclear if the conclusions drawn from these studies would be different if the multilevel nature of the data would have been taken into account. Failure to account for the interdependencies within the data results in an increased likelihood of finding measurement non-equivalence when there is equivalence (Kim et al., 2012; Kim et al., 2015). Further, a single-level approach to measurement invariance may result in researchers overlooking similarities and differences across levels of analysis (Sirotnik, 1980). Simultaneous examination of the factor loadings, means, and intercepts of safety climate scores across levels of analysis has the potential to reveal cross-group differences.

The purpose of this paper is to examine the measurement equivalence of a safety climate measure across theoretically-meaningful hierarchically-arranged faultlines using a sample of 8790 employees from a large multinational chemical processing and manufacturing company. We use an abbreviated variant of Zohar and Luria's (2005) safety climate measure. Although there is no consensus measure of safety climate, variants of Zohar's (1980; 2000; Zohar & Luria, 2005) measures have been used the most. Consistent with the common practices in safety climate assessment (Flin et al., 2000), we adapted the survey items to the organization we were working with and their needs, internal terminology, and included a process safety[1] (rather than occupational or personal safety) item, in response to Zohar's (2003, 2010) call for the inclusion of industry-specific safety climate items. The results will reveal if there are meaningful faultlines that limit comparisons of observed safety climate scores across groups. Testing the measurement equivalence of this safety climate measure provides some initial evidence of the appropriateness of comparing safety climate scores across the tested faultlines.

There are three levels within the current study data (Fig. 1). Level 1 is the employee-level characteristics and faultlines. Four different Level 1 faultlines were tested here: language chosen by employees to respond to the survey (n = 7), hierarchical position in the organization (n = 3), employment arrangement (i.e., core vs. contingent employee; n = 3), and work environment (n = 2). Language and work environment are theoretically and practically important faultlines that have not been previously tested and, as noted above, none of these four Level-1 faultlines have been tested in a multilevel model. Level 2 is the worksites[2] (n = 76) that the employees were embedded in. Level 3 is the cultural-level characteristics and faultlines. Six different Level 3 faultlines were tested here, operationalized as whether the country in which the employees were located is either high or low on each of Hofstede's six cultural dimensions (Hofstede, 1980, 1992; Hofstede et al., 2010).

The paper proceeds as follows. First, we briefly review the concept of measurement equivalence and describe the forms of measurement equivalence that we will test in this paper. Then we review the faultlines tested in this study and why they are expected to result in measurement non-equivalence. Then, we test the measurement equivalence of a shortened and slightly adapted version of a well-known safety climate measure (Zohar & Luria, 2005) across these faultlines while accounting for the hierarchical structure of the data. In the Discussion, we reflect on what these results mean to both theory and practice involving safety climate.

1.1. Measurement equivalence

When analyzing organizational survey data, it is common to generate scores for various subgroups based on meaningful faultlines within the organization (e.g., men vs. women; managers vs. line workers). When groups differ significantly on observed means, these mean differences reflect either or both (a) "true" construct differences in the populations they represent or (b) differences driven by measurement error or non-equivalence (Lord et al., 1968; Spearman, 1904). Measurement equivalence tests are designed to test the assumption that observed differences across a faultline are indeed true differences. It is essential to demonstrate that measurement is equivalent across groups before comparisons between groups can be made; otherwise, the comparisons can obscure true differences or show observed differences when there are no true differences. However, the use of composite scores collapsed across subgroups does not require the establishment of measure equivalence.

Consistent with the requirements for comparing latent mean scores, three common measurement invariance operationalizations will be tested in this study. From least restrictive to most restrictive, they are configural, metric, and scalar invariance[3] (Vandenberg & Lance, 2000). Configural invariance is established when groups conceptualize the dimensionality of a latent construct the same way. Metric invariance is established when the relative importance of survey items to the latent construct is the same across groups. Metric invariance is demonstrated by determining that the magnitude of the item regression slopes on the latent construct (i.e., factor loadings) are the same across groups (Jöreskog, 1969; Schmitt, 1982; Vandenberg & Self, 1993). Scalar invariance is established when individuals from different groups with equal standing on the latent construct interpret the scale anchors the same way. Scalar invariance is demonstrated by determining that the intercepts of the items, as well as the magnitude of the item regression slopes (i.e., metric invariance), on the latent construct are the same across groups.

These forms of measurement equivalence are hierarchically arranged, so failure to support the less restrictive forms of measurement equivalence automatically precludes the more restrictive forms of measurement equivalence (Schmitt & Kuljanin, 2008; Vandenberg & Lance, 2000); that is, configural invariance is necessary (but not sufficient) to demonstrate metric and scalar invariance whereas metric invariance is necessary (but not sufficient) to demonstrate scalar invariance (Fig. 2). When the more restrictive measurement invariance model (e.g. metric invariance) has equal or better model fit than the less restrictive measurement invariance model (e.g., configural invariance), the more restrictive measurement invariance (e.g., metric invariance) is established. Observed mean differences between groups (e.g., difference on scale scores or d values) are only interpretable if all three forms of measurement equivalence are found (Schmitt & Kuljanin, 2008; Vandenberg & Lance, 2000).

[1] Process safety refers to the safe operations of a process (e.g., chemical processing in an oil refinery), rather than the safety of an individual person in the workplace (e.g., wearing personal protective gear) or creating safety programs that influence individual behavior (e.g., safety training). See Mannan et al. (2016) for a review of process safety.

[2] Whereas there are likely to be practical and meaningful differences between worksites (e.g., variations in management, safety practices, and monitoring equipment), we did not have any Level 2 data that could be modeled to reflect these differences. Further, treating 76 worksites as 76 groups results in an unidentified model. As such, we do not address Level 2 faultlines in this paper. However, when testing the measurement equivalence of the Level-1 and Level 3 faultlines, we used the Type = COMPLEX TWOLEVEL routine of Mplus 7.0 (Muthén & Muthén, 1998-2012), to deal with the data dependency caused by the nested sampling (i.e., employees nested within 76 work sites, all members of a given worksite were assigned the same worksite number).

[3] Tests of residual invariance and factor variance (for a single-factor) measurement model are more restrictive, but not necessary for comparisons between latent mean scores (Meredith, 1993; Landenberg and Lance, 2000).

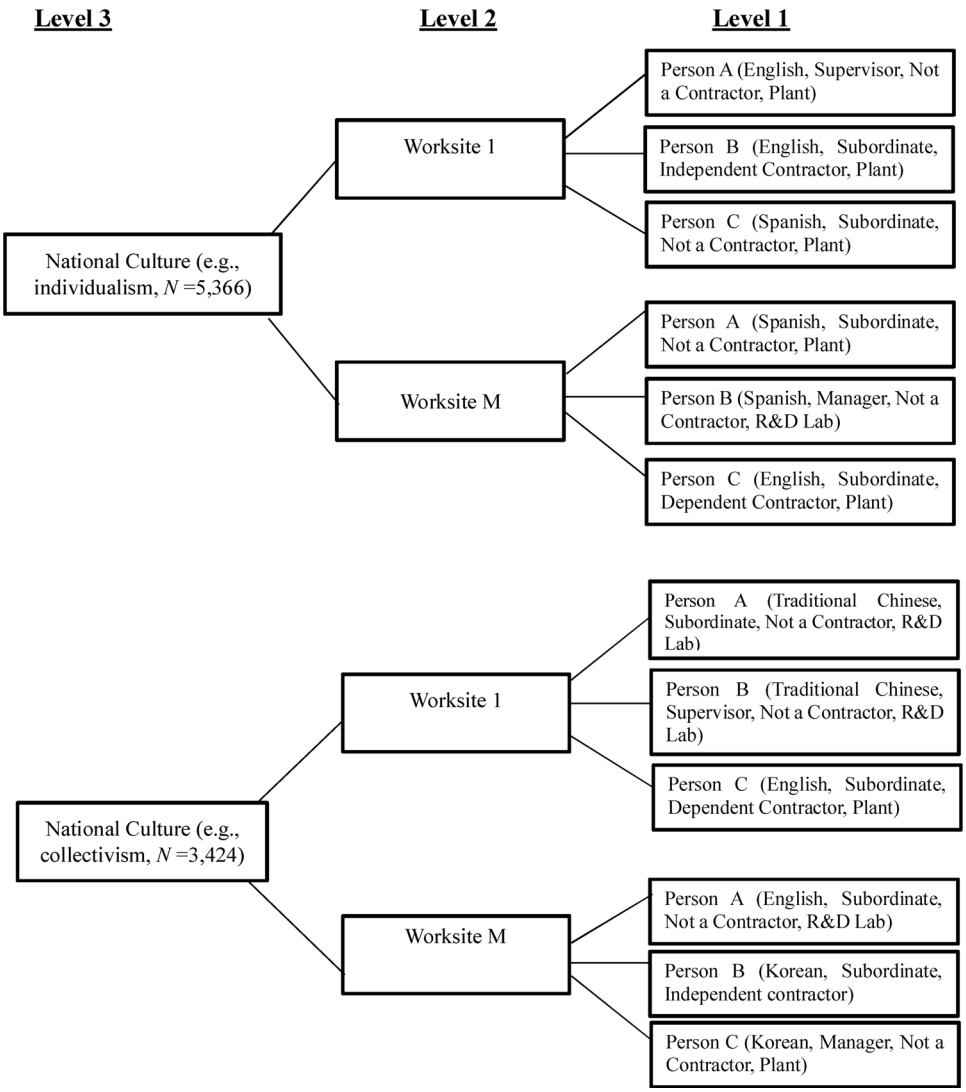

Fig. 1. Multilevel Structure of the Safety Climate Data.

1.2. Safety climate coherence and configural invariance

Safety climate—like all organizational climates—is a climate "for" a coherent part of organizational life rather than a general description of organizational life (Schneider & Reichers, 1983). Safety has numerous components, such as sufficient training, proper gear, managerial support and attention, and generally good housekeeping. Safety climate, then, focuses on the extent to which these safety components are rewarded, supported, and promoted in the organization. In this study, our measure of safety climate is short, touching upon several safety-related topics and couching all of them in terms of safety and, in particular, how this safety is supported by organizational leadership (e.g., supervisors, site management; see Appendix A).

We anticipate that configural invariance will be achieved across all faultlines in this study. This is because (a) the measure used herein was designed to reflect a single dimension of safety climate and (b) the dominant concept in safety climate is management commitment to safety, which is the focus of the included measure (Beus et al., 2010; Zohar, 2010). Thus, we anticipate that across all faultlines, people will perceive that all of the items reflect safety in some way, resulting in a single factor.

1.3. Faultlines and how they threaten measurement equivalence

In this section, we review the faultlines investigated in this study and describe how they theoretically threaten measurement equivalence. This list of faultlines is not meant to be exhaustive, nor are the underlying threats directly tested. Instead, they are illustrative of how measurement non-equivalence can arise because of the effects that faultlines have on individuals' experiences and perspectives, interpretations of items, and responses to items, etc. This list of faultlines was generated prior to administering the safety climate survey and the focal organization was amenable to including questions that would allow for us to measure them.

1.3.1. National culture (level 3)

Given globalization trends across industries, one of the most important faultlines to examine is national culture. Culture can have a significant impact on environment, health, and safety, and safety climate (Mearns & Yule, 2009). Nineteen countries were included in this study. Whereas it is possible that geographical boundaries between each country create meaningful faultlines, it is also likely that faultlines based on national culture characteristics will contribute to differences in safety climate scores. Six cultural dimensions (Hofstede, 1980, 1992; Hofstede et al., 2010) were operationalized as the scores Hofstede assigned to the 19 countries within which the employees were working:

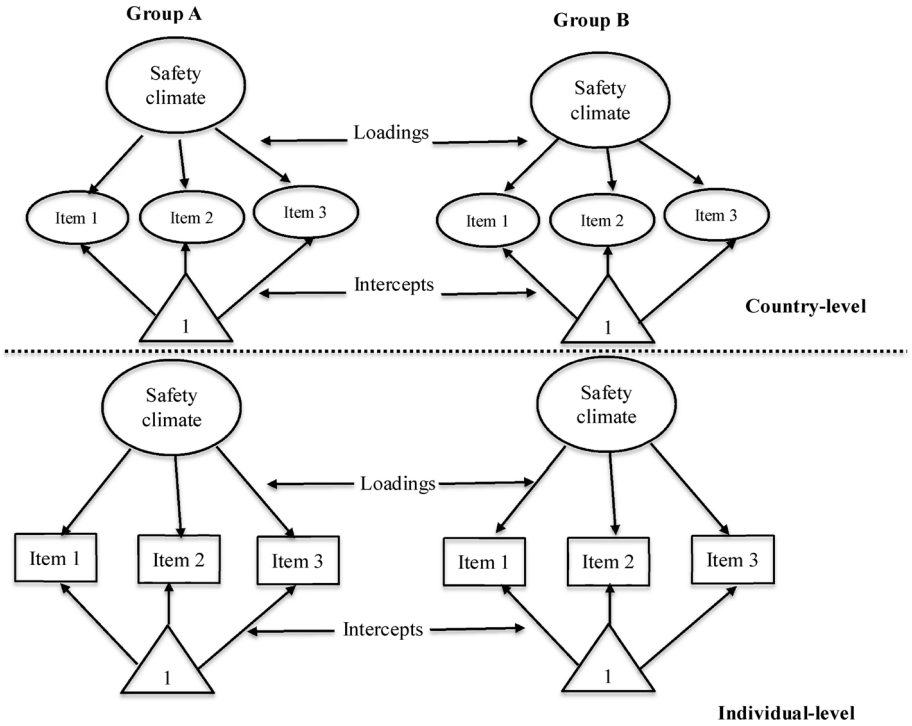

Fig. 2. Multilevel Measurement Invariance.

individualism-collectivism, power distance, uncertainty avoidance, masculinity-femininity, indulgence vs. restraint, and short-term vs. long-term orientation.

Individualism refers to a focus on rights above duties, concern for oneself and immediate family members, emphasis on personal autonomy and self-fulfillment, and basing one's identity on one's personal accomplishments; collectivism is its complement and refers to individuals being interdependent within their in-groups (family, tribe, nation, etc.), prioritizing the goals of their in-groups, shaping their behavior primarily on the basis of in-group norms, and behaving in a communal way (Hofstede, 1980). Uncertainty avoidance reflects the societal-level comfort with unpredictability and uncertainty; societies with low tolerance for uncertainty tend to have rigid social structures, be rule-oriented, and do not accept eccentricity (Hofstede, 1980, 1992). Power distance indexes the extent to which power inequalities are accepted and expected in a society, especially by people with lower power (Hofstede, 1980, 1992). Masculinity-femininity refers to the societal-level value of stereotypically masculine values and traits (e.g., achievement, competition, assertiveness) or stereotypically feminine values and traits (e.g., community, cooperation, consensus, caring, and modesty; Hofstede, 1980, 1992). Indulgent cultures permit relatively free gratification of basic and natural human drives or needs (e.g., having fun, enjoying life), whereas restrained cultures suppress gratification of human drives and needs using strict social norms (Hofstede et al., 2010). Finally, short-term orientation cultures tend to rely on tradition and have little future planning whereas long-term orientation cultures are more adaptive and focus on future goals and plans (Hofstede et al., 2010).

Most measures of safety climate in the literature were developed in a specific national culture, such as in individualistic culture (Flin et al., 2000). As a result, how measures operate within other national cultures is not well known (the etic approach, Berry, 1969; Church, 2001; van de Vijver & Leung, 1997). There are at least two reasons why national culture faultlines could prevent metric equivalence. First is item relevance (also known as item concreteness), which 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; mathematically, this results in different item slopes across groups (Chan, 2000). There is some evidence that national culture dimensions influence the relevance of item content (cf. Robert et al., 2006; van de Vijver & Leung, 1997; van de Vijver & Tanzer, 1997) and, therefore, the strength of the relationship between the safety climate items and the latent construct (i.e., metric equivalence, for a more technical explanation see Cheung and Rensvold, 2002). For example, Lin et al. (2008) found that respondents in China (collectivist, high power distance and long-term orientation, low indulgence/high restraint) and the US (individualist, low power distance and long-term orientation, high indulgence/low restraint) differed on which subdimensions accounted for the most variance in safety climate.

A second reason why safety climate scores may not be equal across national culture faultlines is response styles. Response styles refers to the tendency to use different parts of the response scale, (Cronbach, 1950) with differences in response styles resulting in different factor loadings across groups (Cheung and Rensvold, 2002). Empirical studies have supported associations between Hofstede's cultural dimensions and response styles (e.g., Bachman & O'Malley, 1984; Johnson et al., 2005; Shulruf et al., 2011; Triandis, 1994). For instance, individualists are more likely to have extreme response styles (i.e., the tendency to use outermost ends of the rating scale; Cronbach, 1950) than collectivists (Bachman & O'Malley, 1984; Harzing, 2006; Shulruf et al., 2011; Triandis, 1994); similar effects have been found for high levels of power distance, masculinity, and uncertainty avoidance (Harzing, 2006; Johnson et al., 2005). Thus, due to differences in item relevance and extreme responding, Hofstede's cultural dimensions are expected to be meaningful faultlines for metric equivalence.

Hypothesis 1. The slopes of the safety climate items (i.e., factor loadings) are not equivalent between countries that vary on (a) individualism, (b) power distance, (c) uncertainty avoidance, (d) masculinity, (e) indulgence, and (f) long-term orientation.

National culture is also expected to threaten scalar equivalence due to group differences in acquiescence response styles, socially desirable responding, and reference groups (Bernardi, 2006; Johnson et al., 2005; van Herk et al., 2004). First, the acquiescence response style refers to

the tendency to agree with an item regardless of the content (Billiet & McClendon, 2000). Nonequivalence due to the acquiescence response style occurs when one group systematically agrees with more items than another group regardless of the item content, resulting in scale displacement (Mullen, 1995). Empirical studies have supported that individualism, uncertainty avoidance, power distance, and masculinity are negatively related to acquiescence response styles (Harzing, 2006; Johnson et al., 2005; van Herk et al., 2004)[4]. Second, social desirability bias refers to the tendency to both consciously and unconsciously respond in a way that is socially acceptable based on cultural norms (Paulhus & Reid, 1991). Empirical studies have supported the associations between Hofstede's cultural dimensions and socially desirable responding (e.g., Bernardi, 2006; Triandis, 1994). For example, cross-cultural studies demonstrate that uncertainty avoidance, individualism, and power distance are significantly related to socially desirable responding (Bernardi, 2006; Triandis, 1994; Triandis et al., 2001; Triandis & Suh, 2002).

Finally, the frame-of-reference effect (Robert et al., 2006), also known as the reference-group effect (Heine et al., 2002), states that individuals understand themselves and evaluate their perceptions, attitudes, values, and beliefs by comparing themselves with similar others (Festinger, 1954) and these evaluations and understandings depend in part on the reference group an individual uses. The reference group effect has been recognized as a potential threat to the validity of comparisons of observed scale scores between different cultural referent groups due to the confounding effect of national culture (cf. Heine et al., 2002; Robert et al., 2006). That said, national culture influences the reference group that an individual uses when evaluating one's perceptions (Heine et al., 2002), such as safety climate perception. Thus, individuals from different cultural backgrounds in terms of Hofstede's six cultural dimensions are likely to use different reference groups (i.e., the implicit comparison with those around them), because they are more likely to use reference groups they are familiar with rather than a global comparison (Heine et al., 2002). This will lead to different degrees of item endorsement, regardless of the item content, to the extent that two groups differ on safety climate or differ in the standards/norms by which members of those groups are evaluated.

Hypothesis 2. The intercepts of safety climate items are not equivalent between countries that vary on (a) individualism, (b) power distance, (c) uncertainty avoidance, (d) masculinity, (e) indulgence, and (f) long-term orientation.

1.3.2. Language ofsurvey administration (level 1)

The second faultline predicted to influence safety climate scores is language of survey administration. Language is related to culture. Hence, the mechanisms by which language threatens measurement equivalence of a safety climate measure are likely to be the same mechanisms by which national culture dimensions threaten measurement equivalence (e.g., item relevance and response style). Non-equivalence across languages 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). Therefore, the slopes and intercepts of safety climate items are not expected to be equivalent across linguistic groups.

Hypothesis 3. The (a) slopes and (b) intercepts of safety climate items are not equivalent across the language faultline.

[4] We do not intend to take the position of or adopt the view of one culture and any statement that seems to do so is merely an unintentional reflection of our own cultural biases while describing cultures and the behaviors of people within them. The interpretation of behaviors lies in the eye of the beholder. For instance, collectivists may view social desirability bias as an effort to be "harmonious" rather than socially acceptable and acquiescence as "humility" rather than being agreeable.

1.3.3. Hierarchical position (level 1)

Safety climate is likely to differ across 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). In the focal organization, managers are at the top of the hierarchy, followed by supervisors, and then subordinates. Managers develop safety policies and procedures, supervisors 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 experienced by subordinates. Further, managers and supervisors are likely to differ based on span of control.

Hierarchical position is expected to be a meaningful faultline resulting in metric non-equivalence. Individuals in different positions have different responsibilities, views, and insights about organizational processes and phenomena, affecting the relevance of the information in some items to their position in the organizational hierarchy. For example, for the item "Site management considers health and safety when setting production rates and schedules," managers and supervisors likely have more information about the process that sets production rates and schedules than would front line workers. In measurement invariance terms, such differences in information will lead to larger factor loadings for these items for managers relative to supervisors and employees because they will better represent the latent construct of safety climate.

Hypothesis 4a. The slopes of safety climate items are not equivalent across the hierarchical position faultline.

Hierarchical position is also expected to influence scalar invariance because of differences in socially desirable responding and reference groups, as well as item evocativeness. First, 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 like safety. As a result, they are more likely to provide responses that are socially acceptable (cf. Huang et al., 2014). Additionally, safety climate consists of employee inferences regarding management commitment to safety (Hofmann & Stetzer, 1996; 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). In sum, social desirability bias moves the responses of managers and supervisors up the scale of the safety climate measure as they are the foci of some safety climate items.

Second, employees at different levels within the organization could use different frames of reference. Respondents are most likely to use employees at the same level for their reference group (Heine et al., 2002; Robert et al., 2006). In addition, to the extent that employees at different levels differ in the standards/norms by which they evaluate safety climate, the frame-of-reference effect occurs. 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. Managers tend to spend more of their time in locations physically separated from where the front-line employees work (Cole & Bruch, 2006). Further, managers and supervisors will be (by definition) less aware of unreported workplace unsafe incidents (Arthur et al., 2005; Probst et al., 2008) and less likely to witness and be aware of close calls (Crowl & Louvar, 2002). Thus, managers and supervisors are likely to have different frames-of-reference, making hierarchical level a meaningful faultline.

Third, hierarchical position may also influence item evocativeness (Oort, 1998). Item evocativeness can be interpreted as the location on the latent construct continuum that determines the mean response; the more evocative the item is, the more the respondent is likely to endorse the item resulting in higher response levels on average (i.e., higher

intercept; Lanning, 1991). Whereas item attractiveness influences the association of the items with the latent construct (i.e., factor loadings), item evocativeness influences item intercepts. To the extent that safety climate items focus on these responsibilities, they may be more or less evocative to certain employees. An individuals' position might make some item content more evocative 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 evocative for supervisors and managers and not very evocative for front-line workers because audits and the like are relevant to supervisors' and managers' conceptualization of how safety is accomplished. This would result in higher scores for supervisors and managers than for subordinates even when they have the same standing on the latent construct, because the priority or focus of safety is determined more by the leaders than the employees.

Hypothesis 4b. The intercepts of safety climate items are not equivalent across the hierarchical position faultline.

1.3.4. Employment arrangement (level 1)

Employment arrangement refers to the formal relationship between the worker and the organization (Feldman, 1995; Hulin & Glomb, 1999). In this study, we focus on faultlines arising among core employees and two types of contractors employed in the focal organization. Core employees are employed directly and managed by the focal organization. Independent contractors were hired temporarily to provide services to the focal organization (client, for the contractors) on a fixed-term or a project basis. They are not under the direct day-to-day supervision of the focal/host company and perform a specific scope of work. 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 (Rebitzer, 1995). Dependent contractors work daily alongside core employees but they are officially employees of another company contracted to the client organization. They are under the direct day-to-day supervision of the focal/host company and have specific roles that do not have a defined termination point. Some example job titles include security guards, cafeteria workers, and tubers. Outsourcing or utilizing contractors is quite common in the oil and gas industry, as well as chemical processing industries, but it is not limited to those industries. The general phenomenon of outsourcing has increased considerably over the past two decades (Kakabadse & Kakabadse, 2002). The potentially unique aspect of outsourcing in the focal company and like industries is that contractors often work side-by-side sometimes on a daily basis with the core employees.

We propose that employment arrangement is a meaningful faultline, because contract workers and employees are likely to use different frames of reference (Festinger, 1954; Heine et al., 2002; Robert et al., 2006), resulting in different safety climate item factor loadings and intercepts. Generally, contractors are likely to have lower standards for safety than the host company. First, contractors tend to be less experienced, receive less safety training, have lower levels of familiarity with the host company's practices and procedures, receive less communication from the organization and have higher levels of injuries and incidents (Clarke, 2003; Feldman, 1995; Hulin & Glomb, 1999; Rousseau & Libuser, 1997). As such, contractors are likely to have lower safety climate perceptions. Second, McDonald and Ryan (1992) argued that the development of safety climate is constrained by control over the work process/tasks and contractors have less control over their work (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 et al. (1998) found that contractors had significantly more negative safety attitudes concerning management commitment to safety and incident and accident reporting.

Hypothesis 5. The (a) slopes and (b) intercepts of safety climate items are not equivalent across the employment arrangement faultlines.

1.3.5. Work environment (level 1)

Employees work in a wide variety of work environments, even within the same organization. Adverse job characteristics and conditions—such as noise, heat, chemical exposure, high demands, and overcrowding—are critical factors that influence work-related injuries (Frone, 1998; Nahrgang et al., 2011; Picard et al., 2008; Rabinowitz, 2000; Ramsey et al., 1983). Work environments differ in terms of the types of hazards and risks that are present. These factors are likely to influence employees' perception of safety climate. Employees working in a highly hazardous environment may have different expectations and standards related to workplace safety, which may influence their interpretation of and responses to safety climate items (cf. Cigularov et al., 2013).

In the focal organization, there are three primary work environments: manufacturing plant, research and development laboratory (R&D lab), or office. Safety personnel from the participating organization deemed some of the safety climate items irrelevant to office personnel, especially considering that some of the offices were not sited near manufacturing plants or R&D labs; thus, only a few safety climate items were administered to office personnel. As a result, measurement equivalence across only the plant and R&D work environments can be evaluated.

Some safety climate items may be more effective at differentiating a good safety climate for employees working in plants than in R&D labs. For example, the item "Site management provides all necessary safety equipment for workers" is likely to be more relevant to employees working in plants than those working in R&D labs, as plant employees use more safety equipment than R&D personnel.

Hypothesis 6a. The slopes of safety climate items are not equivalent across the work environment faultline.

The intercepts of the safety climate items may not be equivalent across different work environments because of item evocativeness and the reference group effect. First, the environment that people work in makes certain item content more evocative as a marker of the underlying construct of safety climate. For instance, the item "my supervisor insists we wear our protective equipment even if it is uncomfortable" is likely to be more evocative to a plant employee, because plant employees work in more adverse conditions where protective equipment is more likely to be required and in many cases is more extensive and therefore less comfortable. Second, once again employees are likely to think of people more like themselves and thus working in the same environment than a different work environment (Heine et al., 2002). For example, plant employees are more likely to think of other plant employees than office or R&D employees when responding to "Site management considers health and safety when setting production rates and schedules." In summary, item evocativeness and the reference group effect are expected to lead to nonequivalence of the item intercepts across groups.

Hypothesis 6b. The intercepts of safety climate items are not equivalent across the work environment faultline.

1.4. Summary

This study examines the configural, metric, and scalar measurement equivalence of a measure of safety climate across faultlines at multiple levels. Although safety climate is conceptualized by some researchers as a multidimensional construct, we focused on the most common and central component of safety climate (Flin et al., 2000) that is the strongest predictor of safety outcomes (management commitment to safety; Beus et al., 2010) that were extracted from a measure developed by one of the most prominent and prolific safety climate researchers (Zohar & Luria, 2005). To summarize our hypotheses, configural invariance is expected to hold across all faultlines, as all safety climate items reflect the broad dimension of management commitment to

Table 1
Responses by Countries.

Country	N	Individualism	Power Distance	Uncertainty Avoidance	Masculinity	Indulgence	Long-term Orientation	Safety Climate Mean (SD)
United States	3361	high	low	low	high	high	low	4.00 (0.75)
Mexico	1306	low	high	high	high	high	low	4.11 (0.62)
China	777	low	high	low	high	low	high	4.11 (0.54)
Brazil	644	low	high	high	low	high	low	4.06 (0.63)
Canada	597	high	low	low	high	high	low	3.85 (0.73)
Germany	572	high	low	high	high	low	high	4.09 (0.66)
United Kingdom	488	high	low	low	high	high	high	4.07 (0.70)
Singapore	312	low	high	low	low	low	high	4.15 (0.59)
Netherlands	295	high	low	high	low	high	high	3.96 (0.64)
Taiwan	209	low	high	high	low	low	high	4.12 (0.62)
Argentina	83	low	low	high	high	high	low	4.23 (0.53)
Switzerland	36	high	low	high	high	high	high	3.45 (0.71)
Colombia	32	low	high	high	high	high	low	3.73 (0.81)
Japan	30	low	high	high	high	low	high	3.88 (0.80)
Korea	13	low	high	high	low	low	high	4.04 (0.78)
Italy	13	low	low	high	high	low	high	3.95 (0.77)
Australia	9	high	low	high	high	high	low	3.94 (0.56)
France	8	high	high	high	low	low	high	4.04 (1.33)
Thailand	5	low	high	high	low	low	low	3.63 (1.80)

Note. Classification into high and low based on Hofstede (1980) and Hofstede et al.'s (2010) 100-point data using 50 as a cut-score.

safety, but metric invariance and scalar invariance are not expected to hold for any of the faultlines.

2. Materials and methods

2.1. Participants and procedure

A health and safety survey was conducted at all sites of an international chemical processing and manufacturing organization. The online questionnaire was made available to 20,260 employees and contractors, of which 8790 individuals (77% male) participated, providing a response rate of 43%. Respondents were from 76 work sites (ranging from 3 to 1063 employees, M = 219, SD = 248) in 19 countries. Based on Hofstede's data and a cut-score of 50 (Hofstede, 1980; Hofstede et al. 2010), countries were dichotomized as: high versus low individualism, power distance, uncertainty avoidance, masculinity, indulgence, and long-term orientation. Table 1 lists the countries, their categorization on each of these dimensions, and the number of respondents per country.

Respondents were given the option to complete the survey in one of nine languages (Table 2). However, only seven language-based groups were included in the relevant analyses, as the French and Japanese groups were too small to model and draw reliable conclusions. Although most respondents from a given country completed the survey in the country's primary spoken language (e.g., English in the United States, 3% of the respondents completed the survey in a language that was not the primary language of that country, making language a faultline distinct from the cultural dimension faultlines. A majority of respondents were subordinates (n = 6238), followed by managers (n = 1058), supervisors (n = 902), dependent contractors (n = 362), and independent contractors (n = 230). A majority of respondents worked in the plant (n = 5517), with fewer employees in office (n = 2922) and R&D lab (n = 351) environments.

Table 2
Responses by Languages.

Language	Frequency	Percent	Safety Climate M (SD)
Simplified Chinese	757	8.6	4.10 (0.52)
Traditional Chinese	215	2.4	4.10 (0.63)
Dutch	270	3.1	3.95 (0.63)
English	4962	56.5	4.00 (0.74)
French[a]	10	0.1	4.00 (1.09)
German	534	6.1	4.09 (0.67)
Japanese[a]	24	0.3	3.94 (0.84)
Portuguese	628	7.1	4.06 (0.62)
Spanish	1390	15.8	4.11 (0.62)

Note. [a]When examining measurement equivalence across the language faultline, Japanese and French were excluded because these samples were too small for reliable results.

2.2. Measures

2.2.1. Safety climate

Safety climate was assessed with seven items slightly adapted from Zohar and Luria (2005; see Appendix A for English-language version). Professional translators, contracted by the survey delivery company, translated the survey items from English into eight additional languages. All items were administered on a 5-point agreement scale (1 = strongly disagree, 5 = strongly agree) plus a "not applicable" (NA) option. The percentage of NA responses ranged from 0.8%–1%; these were treated as missing data (listwise deletion). The Cronbach's alpha coefficient was 0.91. Prior to the administration of the survey, one item was deemed by organizational management to be irrelevant to the R&D employees (i.e., "Site management focuses on process safety in audits, self-assessments, and inspections."), so skip logic removed this item for employees who identified themselves as R&D; analyses comparing work environments thus used this reduced set of items for both groups (R&D, plant).

2.2.2. Faultlines

By choosing a language, respondents entered into the corresponding translated survey. They also indicated the country in which they worked, their hierarchical position, their employment arrangement, and their work environment from multiple choice lists.

2.3. Data analysis

All models were estimated with Mplus 7.0 (Muthén & Muthén, 1998-2012). The nature of the data examined is multilevel, with national culture dimensions as Level-3 variables, worksite (n = 76) as a Level-2 variable, and survey language, hierarchical position, employment arrangement, and work environment as a Level-1 variables. Multigroup multilevel confirmatory factor analyses (Kim et al., 2012) were conducted for the Level-3 faultlines, whereas multilevel factor mixture

models for known classes (Kim et al., 2015) were conducted for the Level-1 faultlines.[5]

When testing measurement invariance for the Level-3 faultlines, the fit indices were examined across different levels of analyses. For Level-1 (i.e., individual-level), the standardized root mean square residual for the within-level model (SRMR-W), the root mean square error of approximation (RMSEA; Steiger & Lind, 1980), and the comparative fit index (CFI) were examined. Good fit is indicated when CFI is greater than 0.90, RMSEA is less than .06, and SRMR-W is less than .08 (Hu & Bentler, 1999). For Level-2 (i.e., work site-level), good fit is indicated when the standardized root mean square residual for the between-level model (SRMR-B) is less than 0.14 (Hsu et al., 2015). Finally, the Satorra-Bentler scaled chi-square difference test (SBχ^2; Satorra & Bentler, 1994) is considered, with a significant p value indicating that the relaxed model is preferred over the constrained model.[6]

When testing measurement invariance for the Level-1 faultlines, there are no absolute fit indices (e.g., CFI and RMSEA) for the multilevel factor mixture models. Instead, the Satorra-Bentler scaled likelihood ratio (SBLR: Satorra & Bentler, 1994) is used for model comparison. Akaike information criterion (AIC; Akaike, 1987), Bayesian information criterion (BIC; Schwartz, 1978), and sample-size adjusted BIC (SBIC; Sclove, 1987) were also considered. Models associated with smaller AIC, BIC, and SBIC values are considered to be better (Kim et al., 2015).[7]

3. Results

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 worksites is not small (n = 76), the model-based approach was used to take into account the clustering within worksites by specifying a model for the worksite level and for the individual level, respectively, whereas the design-based approach was used to adjust the overall model chi-square value and the standard errors of parameter estimates for the clustering within countries (n = 19). The results of the multilevel CFA for the 7-item safety climate measure indicated the one-factor measurement model had good model fit (RMSEA = .02; SRMR-W = 0.03; CFI = .96; SRMR-B = .12), supporting the single-factor model (i.e., management commitment to safety).

3.1. National culture

The results of multi-group multilevel CFAs indicated for all six cultural dimensions that the one-factor configural equivalence model did not have acceptable fit (SRMR-B was larger than 0.14). That is, for each of the six cultural dimensions, the safety climate items did not evoke the same conceptual framework in defining the latent construct across the individual-level and the worksite-level (Table 3). In other words, the proposed single dimension for safety climate was not representative of the data across the various culture faultlines.

Table 3
Results of Measurement Equivalence Tests for Level 3 National Culture Dimensions.

	χ^2 (df)	RMSEA	CFI	SRMR-W	SRMR-B
Individualism-Collectivism					
Configural Equivalence	476.58(56)*	.04	.95	.03	.16
Metric Equivalence	532.93(62)*	.04	.94	.03	.18
Scalar Equivalence	696.11(68)*	.05	.92	.03	.24
Power Distance					
Configural Equivalence	515.08(56)*	.04	.95	.03	.16
Metric Equivalence	567.62(62)*	.04	.94	.03	.18
Scalar Equivalence	737.45(68)*	.05	.92	.03	.24
Masculinity					
Configural Equivalence	382.90(56)*	.04	92	.03	.17
Metric Equivalence	421.09(62)*	.05	.91	.03	.31
Scalar Equivalence	440.87(68)*	.04	.90	.03	.32
Uncertainty Avoidance					
Configural Equivalence	601.75(56)*	.05	.95	.03	.19
Metric Equivalence	644.88(62)*	.05	.94	.03	.22
Scalar Equivalence	667.79(68)*	.05	.94	.03	.21
Indulgence					
Configural Equivalence	262.59(56)*	.03	.93	.03	.34
Metric Equivalence	303.63(62)*	.03	.92	.03	.29
Scalar Equivalence	309.36(68)*	.03	.92	.03	.30
Long-Term Orientation					
Configural Equivalence	577.59(56)*	.05	.92	.03	.23
Metric Equivalence	542.27(62)*	.04	.92	.03	.26
Scalar Equivalence	585.66(68)*	.05	.92	.03	.27

Note. RMSEA = root mean square error of approximation; CFI = comparative fit index; SRMR-W = standardized root mean square residual for the within-level model; SRMR-B = standardized root mean square residual for the between-level model. The CFI, SRMR-W, RMSEA were for Level-1 (i.e., individual-level), SRMR-B was for Level-2 (i.e., site-level), and all estimates are adjusted based on the Level-3 clustering (i.e., the data dependence due to 19 countries).
* $p < .05$, dropping were also supported.

Because measurement equivalence is established hierarchically, metric invariance cannot be found when configural equivalence is not found. Therefore, the lack of configural invariance also indicates a lack of metric invariance, supporting Hypotheses 1a–f. Further, scalar invariance cannot be found when metric equivalence is not found. Therefore, Hypotheses 2a–f

3.2. Language of survey administration

Multilevel factor mixture models were conducted to test for measurement equivalence across the Level-1 language faultline (Table 4). All model fit statistics supported configural invariance across the language faultline, indicating that a single factor (as planned) was found in all language groups. Next, restrictions were placed in the model to test for metric invariance. The results indicated that two of the three information criteria (i.e., BIC and SBIC) supported the fit of the metric equivalence model over the configural equivalence model, whereas the opposite was true for AIC. However, the SBLR (0.2) = 3610.98, $p < .05$ provided further support that metric invariance cannot be achieved for language of survey administration. Because of the hierarchical nature of measurement equivalence tests, the lack of metric invariance also means a lack of scalar invariance. Thus, the slopes and the intercepts of safety climate items across the seven linguistic groups were not equivalent, supporting Hypotheses 3a and b.

3.3. Hierarchical position[8]

Multilevel factor mixture models were conducted to test for

[5] Mplus codes are available from Kim et al. (2015) and at http://scholarcommons.usf.edu/edq_facpub/3/.

[6] The Level-3 faultline model (i.e., national culture dimensions) 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. Therefore, a design-based approach was used for the Level-3 model, which does not generate fit indices.

[7] All the subgroup sizes are adequate for the analyses performed ($n > 200$, Kline, 2011), but subgroup sizes differed. In multi-group analyses, model estimates are largely driven by the subgroup with the largest sample size (Kline, 2011). For instance, for the employment arrangement faultline, there were many fewer independent contractors (n = 230) and dependent contractors ($n = 362$) than core employees ($n = 8{,}189$). 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, the unbalanced sample sizes of the subgroups for each faultline did not appear to affect the results.

[8] An alternative hierarchical position faultine that incorporates the contractors

Table 4
Results of Measurement Equivalence Tests for Level 1 Language, Hierarchical Position, Employment Arrangement, and Work Environment.

	Loglikelihood	AIC	BIC	SBIC
Language				
Configural Equivalence	−43883.99	88076	89123	88634
Metric Equivalence	−43924.11	88084	88887	88512
Scalar Equivalence	−44088.73	88341	88899	88638
Hierarchical Position				
Configural Equivalence	−55614.53	111369	111859	111636
Metric Equivalence	−55627.63	111371	111777	111593
Scalar Equivalence	−55823.24	111738	112060	111914
Employment Arrangement				
Configural Equivalence	−56082.53	112306	112800	112577
Metric Equivalence	−56094.72	112305	112715	112531
Scalar Equivalence	−56119.07	112330	112655	112508
Work Environment				
Configural Equivalence	−53083.57	106287	106711	106520
Metric Equivalence	−53093.83	106288	106641	106482
Scalar Equivalence	−53185.22	106450	106733	106606

Note. AIC = Akaike information criterion; BIC = Bayesian information criterion; SBIC = sample-size adjusted BIC.

measurement equivalence across the Level-1 hierarchical position faultline (Table 4). Results supported configural invariance, with a single planned factor evident across the faultline. The results indicated that two of the three information criteria (i.e., BIC and SBIC) supported the fit of the metric equivalence model over the configural equivalence model, whereas the opposite was true for AIC. However, the SBLR (1.8) = 14.54, $p < .05$ provided further support that metric invariance cannot be achieved for the hierarchical position. Because of the hierarchical nature of measurement equivalence tests, the lack of metric invariance also means a lack of scalar invariance. Thus, the slopes and the intercepts of safety climate items across managers, supervisors and employees are not equivalent, supporting Hypotheses 4a and b.

3.4. Employment arrangements

Multilevel factor mixture models were conducted to test for measurement equivalence across the Level-1 employment arrangements faultline (Table 4). Configural invariance was supported across employment arrangements, with the single planned factor found for independent contractors, dependent contractors, and core employees. Constraints were then added to the model to test for metric invariance. Contrary to predictions, all three information criteria indicated that the metric equivalence model provided at least as good if not better fit than the configural equivalence model for the employment arrangement faultline. Thus, contrary to Hypothesis 5a, metric invariance was achieved across independent contractors, dependent contractors, and core employees for the employment arrangements faultline.

Additional constraints were placed in the model to test for scalar invariance. The results were mixed. As Table 4 shows, all three information criteria supported the fit of the scalar equivalence model over the metric equivalence model, suggesting that the scalar equivalence model provided better model fit than the metric invariance model. Thus, contrary to Hypothesis 5b, the conclusion is that scalar invariance is achieved for the employment arrangement faultline.

(*footnote continued*)
subgroups from the employment arrangement faultline would classify all respondents into one of five subgroups in the following order: managers, supervisors, subordinates, dependent contractors, and independent contractors. An examination of measurement equivalence of the safety climate measure for this alternative faultline supported metric equivalence but not scalar equivalence.

3.5. Work environment

Multilevel factor mixture models were conducted to test for measurement equivalence across the Level-1 work environment faultline (Table 4). 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 were not administered one of the seven items. The configural invariance model provided a satisfactory fit (Table 4). The fit statistics for the metric equivalence model relative to the configural invariance model were mixed, with BIC and SBIC favoring the metric equivalence model over the configural equivalence model but AIC favoring the configural equivalence model over the metric equivalence model. However, the SBLR (0.7) = 28.44, $p < .05$ also indicated that the metric equivalence model fit the data worse than the configural equivalence model. Thus, as a group, the fit statistics suggested that the metric invariance model was not supported and, therefore, the scalar model cannot be supported when comparing plant vs. R&D lab employees. Therefore, Hypothesis 6a and b were supported.

3.6. Does modeling multilevel data matter?

For comparison purposes, measurement equivalence analyses were repeated without modeling the multilevel nature of the data. Instead, all faultline variables were treated as Level-1 properties of the individual and single-level multigroup confirmatory factor analyses (Jöreskog, 1971; Little, 1997; Sörbom, 1974) were used to assess measurement equivalence across groups. These types of analyses are common in research, where the multilevel dependencies are not accounted for. The factor structure (configural invariance) was equivalent across all faultlines. However, consistent with previous claims (Kim et al., 2012; Kim et al., 2015), metric equivalence did not hold across all the faultlines except the hierarchical position faultline when the nested nature of the data was ignored, whereas multilevel measurement equivalence tests indicated that metric equivalence and scalar equivalence were achieved for employment arrangement.[9] Given the overall lack of equivalence between groups across faultlines, participant sex was also examined as an exploratory Level-1 faultline. Analyses revealed configural and metric but not scalar equivalence between male and female participants.

4. Discussion

Within a multilevel organizational dataset, potentially important faultlines to the measurement of safety climate were examined including: six dimensions of national culture (i.e., individualism, power distance, uncertainty avoidance, masculinity, indulgence, and long term orientation), language, hierarchical position, employment arrangement, and work environment. At the individual level, regardless of which faultline, configural invariance was supported; however, for all of the national cultures (Level 3), configural invariance was not supported. Metric invariance analyses revealed that the factor loadings of safety climate items and the intercepts of safety climate measure were equivalent across employment arrangements (employees, independent contractors, and dependent contractors). The factor loadings of safety climate items and the intercepts of safety climate measure were not equivalent across language, work environment, hierarchical position, or any of the national culture faultlines. In summary, any observed mean differences (i.e., differences on scale scores or d values) between groups defined by these faultlines except the employment arrangement cannot be directly interpreted because the differences in the group means occur, at least in part, because of measurement-related issues.

[9] These results are available from the first author.

There are a variety of causes for metric variance including psychological interpretation differences as well as technical within-study cross-group differences (Cheung and Rensvold, 2002). However, we acknowledge that we cannot differentiate the language faultline from translation errors. We used a professional translation service and ensured that the survey administration was as similar as possible across groups, so the possible problems that could arise from these mechanistic issues were likely reduced. Therefore, the lack of measurement equivalence is interpreted to be more a function of psychological differences than mechanistic differences.

4.1. Measurement invariance with multilevel data

Our study provides the first empirical evidence beyond simulation studies that demonstrate that neglecting the multilevel structures of data sets leads to understating the equivalence of measures across Level 1 faultlines (Kim et al., 2015). Although accounting for the multilevel nature of the data in this study did not result in many faultlines exhibiting metric or scalar equivalence, one faultline (i.e., employment arrangement) exhibited scalar equivalence when the multilevel structure within the data was modeled. This leads to two questions: to what extent have organizational scientists underestimated the amount of metric equivalence in our data sets? And, to what extent have organizational scientists overlooked important faultlines? It is important to recognize that many populations have a multilevel structure in the real world, even if our data sets and samples do not allow us to include the multilevel structure in our analyses. As an example, consider a two site data collection from a 50 site organization. Organizational site might be a faultline, yet it might be difficult to test the site faultline with only two representatives. Organizational scientists should make a stronger case to organizational partners that multilevel components should be part of sampling plans. Further, organizational scientists need to account for multilevel structures in analyses to the extent possible. When they cannot, then they should address these issues further when considering limitations of their work.

4.2. Faultlines

We proposed that the faultlines in the present data were likely to disrupt the ability to make comparisons of a safety climate measure across groups. The reasons for anticipated differences were the relevance of survey items to the latent construct (van de Vijver & Leung, 1997), response styles (e.g., extreme response style, acquiescence; Cheung and Rensvold, 2002), social desirability bias (Heine & Lehman, 1997; Lalwani et al., 2006), and frames of reference (Heine et al., 2002). Generally, our perspective was supported: measurement equivalence was not supported across faultlines with the exception of the employment arrangement faultline. This lack of measurement equivalence underscores the importance of examining measurement equivalence regardless of what constructs are under investigation (Schmitt & Kuljanin, 2008) and the importance of emic, rather than etic, approaches to theory and measure development (e.g., van de Vijver & Leung, 1997). Although it is always important, from a research and practice standpoint, to have good measurement, it is essential to have good measurement when the construct is critical to the safety, well-being, and life-sustainment of workers, as is the case for safety climate. Safety climate is literally a life-or-death issue.

Importantly, the included faultlines and the theoretical mechanisms through which these faultlines may threaten the measurement invariance of the safety climate measure do not represent an exhaustive list. Rather, they represent a common set of cross-sample differences (i.e., faultlines) that are likely to threaten the measurement invariance of all psychological measures in organizational research, particularly cross-cultural research. Future research studies should have a theory of faultlines when examining relationships; that is, researchers should take the time to evaluate possible faultlines before proceeding with their analyses. These concerns are accounted for in many studies where control variables such as sex, age, or race are included, but control variables do not account for non-equivalent measurement. It is essential that measurement equivalence be demonstrated in order to make reliable judgments of differences between groups.

4.3. Theoretical implications

This study has several theoretical implications. The primary theoretical lens in this paper is faultline theory, which argues that faultlines influence employees' sensemaking (Lau & Murnighan, 1998) and sensemaking is the key process of developing climate perceptions (Lau & Murnighan, 1998; Ostroff et al., 2012; Weick, 1995). The specific faultlines we examine are theoretically relevant based on theories about each of them including cultural values (e.g., Hofstede, 1983) and organizational structure (Mintzberg, 2008), as well as variables and phenomena related to them (e.g., span of control, communication, etc.). Our results further demonstrate the importance of faultlines as they influence employees' interpretations of a safety climate measure. Incorporating faultline theory into organizational climate theory may provide new insights into the mechanisms through which the group-level climate emerge from individual-level climate (Ostroff et al., 2012). Such research could also lead to the identification of faultline triggers that make the faultlines more salient, which would cause weaker overall group-level climate. Such research could be used to develop interventions to reduce either the activation or the effect of these triggers, ultimately facilitating the emergence of group-level safety climate (Jehn & Bezrukova, 2010).

Faultline theory suggests that faultlines can be dormant for years without any changes in the group process (Lau & Murnighan, 1998). That is, while some characteristics (e.g., hierarchical position, demographics) may provide the potential for particular subgroupings (Lau & Murnighan, 1998), individuals may actually feel like one group as the faultline is inactive. Faultline activation refers to the process of triggering social categorization (i.e., subgroupings) based on the salience of characteristic differences (Lau & Murnighan, 1998) or features of the situation that stimulate recognition of these differences. It is important to conduct empirical studies to understand when, why, and how faultlines are activated, drawing upon existing relevant theories. Based on the results of this study, it appears that national culture, language, hierarchical position, and work environments are all salient and activated faultlines for employees' shared perceptions of the extent to which safety policies, procedures, and practices are enforced.

Our study also has important theoretical implications for safety climate. Mearns and Yule (2009) identified a number of ways in which national culture is likely to influence workplace safety including safety management practices, involvement, and communication. We proposed that national culture dimensions were meaningful faultlines due to differences in item relevance, extreme responding, acquiescence, and frame-of-reference, extending the reasons why national culture is important to our understanding of workplace safety. Future research could probe national culture faultlines further to determine the underlying mechanisms for the group differences that emerged. It will also be important to test the replicability of these findings with a multidimensional measure of safety climate.

As is the case with all studies of metric equivalence, it is unknown at this stage whether the low number of faultlines achieving metric equivalence is an idiosyncrasy of the items (and/or, for some faultlines, their translations), a feature of the faultlines in this organization, or a feature of the faultlines relative to the safety climate construct broadly. It is clear that additional research is needed on the metric equivalence of safety climate measures across faultlines, especially while accounting for the multi-level nature of data. Although appropriately representing the nature of the data in analyses is important regardless of the construct of interest, it is especially important when there are a priori factors that interact with the construct of interest. In the case of safety

climate, for example, it is clear that there are industry, national or world-regional (e.g., European Union), and company differences in safety regulation and oversight. Where possible, these should be tested ahead of other analyses, regardless of whether it is relevant to the particular research question in the study; this is because measurement non-equivalence that is undetected within a group (e.g., differences between men and women in a worksite) can result in unreliable and incorrect results.

Further, to the best of our knowledge, this is the only nonsimulation-based empirical study on measurement equivalence that takes into account the multilevel nature of the data. Without modeling the nested nature of organizational data, it is unclear how much the multilevel data impacts conclusions about measurement equivalence. Testing the same faultlines with measures of different constructs while accounting for the multilevel nature of the data would begin to answer this question. Further, although it is often ideal to have a measure that is impervious to faultlines (i.e., allows for benchmarking across operational sites, ease of use, ease of communicating results), there may be some faultlines that are so deeply ingrained that no measure can bridge the gap because the different groups define, perceive, and experience the construct itself differently. Multiple measures of the same construct across faultlines would address this. Such research could shed light on the impermeability of faultlines.

4.4. Practical implications

This study has several practical implications. First, we used a slightly modified and abbreviated version of Zohar and Luriaös (2005) safety climate measure. The present findings indicated that for the data examined in this study, safety climate scores based on this measure were not equivalent across national culture, language, hierarchical position, and work environment. These findings are consistent with Zohar's (2003, 2010) argument that safety climate perceptions are context-dependent, as well as his call for the development of industry-specific safety climate scales. Zohar (2010) implied that no safety climate measures are able to adequately assess safety climate across different contexts, industries, and levels of analyses. In other words, the influence of faultlines on the measurement equivalence of safety climate measures is expected and appropriate given the theories of organizational climate in general and safety climate in particular. For instance, Zohar (2010) argued that top managers and supervisors might have inconsistent perceptions or interpretation of safety climate within the organization (i.e., the misalignment between enacted and espoused safety priority), which is consistent with the present finding that the safety climate measure was not equivalent between managers and supervisors. The existence of faultlines in the organization may prevent meaningful comparisons of observed scores between groups. Therefore, researchers and practitioners need to establish the equivalence of their safety climate measures before comparing groups. Alternatively, it may be that unique safety climate measures need to be developed within each group identified by safety climate-relevant faultlines. Organizations would then need to set standards that each group should achieve, rather than comparing groups in order to determine which groups are doing well and which are doing poorly.

Second, multilevel researchers advocate confirming there is sufficient agreement across individual-level ratings before aggregating to a higher level (e.g., group or organizational level; Bliese, 2000). This study provides empirical evidence that researchers should not only rely on traditional statistical indices for agreement (e.g., rwg and ICC). In addition, researchers should consider additional faultlines before aggregating as well, as measurement equivalence tests may suggest that the same measure assesses different things across groups, which makes aggregating individual-level scores inappropriate.

Third, the use of different referents/standards in responding to scale items by different groups (i.e., the frame-of-reference effect) may be why intercepts were 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 across groups. For instance, consistent with the hierarchical position 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 reporting of incidents and close calls) 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 groups.

Research should interpret the results of measurement equivalence tests with caution, as measurement equivalence is measure-specific, sample-specific, and faultine-specific (Vandenberg & Lance, 2000). The safety climate measure used in this study may be equivalent for other faultlines and for other samples, whereas other safety climate measures might be equivalent for the faultlines examined in the present study. For instance, Reader et al. (2015) found that the safety culture measure they used is equivalent across eight subgroups (combinations of one occupation and one region). We encourage researchers to conduct qualitative research, such as interviews, theoretical analysis, reviewing organizational records and documents (i.e., accidents reports), to identify concepts that have equivalent meaning/interpretation across different subgroups to increase the probability of measurement equivalence in the future.

There are a few additional analytical options to pursue when measurement equivalence does not hold (Davidov et al., 2014; Beuckelaer & Swinnen, 2011). One option is to examine a subset of groups to determine which ones are comparable. This can be done based on conceptual similarity or through cluster analysis (Byrne & van de Vijver, 2010; Welkenhuysen-Gybels et al., 2007). As an exploration of this idea, we reexamined our data, comparing Simplified Chinese to Traditional Chinese responses; we found that scalar equivalence was achieved for these two subgroups, revealing more refined measurement equivalence results.

Another option is to demonstrate partial measurement equivalence by identifying a subset of survey items that are equivalent across groups before conducting cross-group comparisons (Byrne et al., 1989; Cheung & Rensvold, 1998). Cheung and Rensvold (1998) provide detailed procedures for identifying non-invariant items and the item characteristics most conducive to this approach. However, there are a number of impediments to this approach, including unclear criteria for establishing partial equivalence and insufficiency of partial equivalence for meaningful cross-group comparisons (Beuckelaer & Swinnen, 2011).[10]

[10] In response to a reviewer request, we conducted some item-level measurement invariance tests for four of the five faultlines (culture values was excluded). Consistent with the scale-level measurement invariance tests, item-level measurement invariance tests suggest that all items are equivalent for the employment arrangement faultline. For hierarchical position, language, and work environment faultlines, the items that are not equivalent between subgroups are not the same ones. For instance, the item "Site Management is strict about working safely at all times even when work falls behind schedule" is not equivalent for the hierarchical position and work environment faultlines but is equivalent for the language faultline. Unfortunately, we do not have data to test the multiple possible explanations for measurement non-equivalence including response

Finally, researchers can retain non-equivalent items to identify sources of measurement non-equivalence (e.g., Cheung & Rensvold, 1998; Poortinga, 1989). For single-level data, researchers could use a multiple indicators-multiple causes model to identify sources of scalar non-equivalence (e.g., Jak et al., 2013) and latent interaction modeling to detect sources of metric non-equivalence (e.g., Little et al., 2006). Multilevel structural equation modeling could be used to explain sources of multilevel measurement non-equivalence by accounting for cross-group differences in the estimated parameters (e.g., intercepts and factor loadings) by including individual and or/contextual predictors (e.g., cultural values) into the model (Hox, 2010). Investigation of the sources of measurement non-equivalence could provide useful information as to how scales can be improved.

4.5. Limitation and future directions

Despite the numerous strengths to this study including a large, multinational field sample with multiple theoretically-meaningful faultlines, there are some limitations to acknowledge. First, national culture dimensions were operationalized using the country in which the respondent worked. This assumes culture is homogenous within a country and that the respondent is from that country (rather than just working there). 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 predictive than national culture of a given individual's responses to the safety climate measure. Future studies should investigate how individual-level culture influences employees' interpretation and perceptions of safety climate.

Second, worksite (Level-2) is a potentially practically meaningful faultline. Conceptually, worksites differ in a number of ways that could affect safety climate, including: the amount of communication amongst managers, supervisors, and employees; safety policies and laws; and subcultures that vary across different worksites. Future studies are needed to explore how these characteristics of the worksite influence the measurement equivalence of safety climate measures. However, many of these factors can also be examined at Level 3, as groups of worksites that share characteristics. This is what we did here, grouping individual worksites into categories of national culture types and examining those differences. Although these analyses would be more fine-grained at Level 2, the Level 3 analyses herein have provided information about differences across worksites that differ on these Level 3 national culture characteristics.

It was impossible to identify the exact source(s) that led to measurement non-equivalence between groups in this study. For example, it is impossible to know whether the lack of equivalence between respondents from individualist and collectivist countries is due to 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 all the other faultlines as well. Additional research is needed to differentiate all these potential sources of nonequivalence (e.g., Davidov et al., 2014).

Finally, the present study focused on single faultlines rather than the combinations of faultlines (e.g., individualistic English-speaking employees versus collectivistic Chinese-speaking employees). However, because measurement equivalence did not hold across the faultlines under investigation, the measurement equivalence of groups based on different combinations of these faultlines is extremely unlikely.

(*footnote continued*)
styles, socially desirable responding, and frame-of-reference. We look forward to future research that does.

5. Conclusion

In summary, this study examined the measurement equivalence of a safety climate measure with a sample of 8790 employees across multiple faultlines at different measurement levels. 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 different national cultures operationalized in terms of Hofstede's cultural dimensions. Multilevel factor mixture models indicated that the dimensionality of the latent construct of safety climate was conceptualized in the same way across subgroups of all the examined faultlines, but the relative importance of survey items to the latent construct and the scale anchors were not equivalent across language, hierarchical position, or work environment subgroupings. Interestingly, the relative importance of the items to the latent construct and the scale anchors were equivalent across the employment arrangement faultline (employees vs. two kinds of contractors). Researchers and practitioners should confirm measurement equivalence before benchmarking safety climate scores across these faultlines.

Authors' notes

This paper is based on the first author's dissertation. We thank the other committee members Drs. Winfred Arthur, Jr. and Myeongsun Yoon for their comments on earlier versions of the paper and the leadership of the focal organization for their help and access for this assessment. This work was supported by the Mary Kay O'Connor Process Safety Center in the Dwight Look College of Engineering at Texas A&M University.

Appendix A

Safety Climate

1. 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 [a]My supervisor insists we wear our protective equipment even if it is uncomfortable.

Note. [a]This item was deemed irrelevant to the R&D employees, so that employees who identified themselves as R&D workers skipped this item.

References

Akaike, H., 1987. Factor analysis and AIC. Psychometrika 52, 317–332. http://dx.doi.org/10.1007/BF02294359.
Arthur Jr., W., 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. Hum. 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 Opin. Q. 48, 491–509. http://dx.doi.org/10.1086/268845.

Barbaranelli, C., Petitta, L., Probst, T.M., 2015. Does safety climate predict safety performance in Italy and the USA? Cross-cultural validation of a theoretical model of safety climate. Accid. Anal. Prev. 77, 35–44.
Bernardi, R.A., 2006. Associations between Hofstede's cultural constructs and social desirability response bias. J. Bus. Ethics 65, 43–53. http://dx.doi.org/10.1007/sl0551-005-5353-0.
Bernardin, H.J., Smith, P.C., 1981. A clarification of some issues regarding the development and use of behaviorally anchored ratings scales (BARS). J. Appl. Psychol. 66, 458–463.
Berry, J.W., 1969. On cross-cultural comparability. Int. J. Psychol. 4, 119–128. http://dx.doi.org/10.1080/00207596908247261.
Beuckelaer, A.D., Swinnen, G., 2011. Biased latent variable mean comparisons due to measurement noninvariance: a simulation study. Cross-Cultural Analysis. Methods and Applications. Taylor and Francis, NY/Hove, pp. 117–148.
Beus, J.M., Payne, S.C., Bergman, M.E., Arthur Jr, W., 2010. Safety climate and injuries: an examination of theoretical and empirical relationships. J. Appl. Psychol. 95, 713–727. http://dx.doi.org/10.1037/a0019164.
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. J. Occup. Organ. Psychol. 85, 454–471. http://dx.doi.org/10.1111/j.2044-8325.2011.02049.x.
Billiet, J.B., McClendon, M.J., 2000. Modeling acquiescence in measurement models for two balanced sets of items. Struct. Equ. Model. 7, 608–628. http://dx.doi.org/10.1207/S15328007SEM0704_5.
Bliese, P.D., 2000. Within-group agreement, non-independence, and reliability: implications for data aggregation and analysis. In: Klein, K.J., Kozlowski, S.W. (Eds.), Multilevel Theory, Research, and Methods in Organizations. Jossey-Bass, San Francisco pp. 349–381.
Bureau of Labor Statistics, 2013. Census of fatal occupational injuries (CFOI) – Current and revised data. http://www.bls.gov/iif/oshcfoi1.htm.
Bureau of Labor Statistics, 2014. National census of fatal occupational injuries in 2014. Retrieved from. https://www.bls.gov/news.release/archives/cfoi_09172015.pdf.
Byrne, B.M., van De Vijver, F.J., 2010. Testing for measurement and structural equivalence in large-scale cross-cultural studies: addressing the issue of nonequivalence. Int. J. Test. 10, 107–132. http://dx.doi.org/10.1080/15305051003637306.
Byrne, B.M., Shavelson, R.J., Muthén, B., 1989. Testing for the equivalence of factor covariance and mean structures: the issue of partial measurement invariance. Psychol. Bull. 105, 456–466. http://dx.doi.org/10.1037/0033-2909.105.3.456.
Candell, G.L., Hulin, C.L., 1986. Cross-language and cross-cultural comparisons in scale translations independent sources of information about item nonequivalence. J. Cross-Cult. Psychol. 17, 417–440.
Chao, G.T., Moon, H., 2005. The cultural mosaic: a metatheory for understanding the complexity of culture. J. Appl. Psychol. 90, 1128–1140. http://dx.doi.org/10.1037/0021-9010.90.6.1128.
Cheung, G.W., Rensvold, R.B., 1998. Cross-cultural comparisons using non-invariant measurement items. Appl. Behav. Sci. Rev. 6, 93–110. http://dx.doi.org/10.1016/S1068-8595(99)80006-3.
Cheung, G.W., Rensvold, R.B., 2002. Evaluating goodness-of-fit indexes for testing measurement invariance. Struct. Equ. Model. 9, 233–255. http://dx.doi.org/10.1207/S15328007SEM0902_5.
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. J. Appl. Psychol. 94, 1103–1127.
Church, A.T., 2001. Personality measurement in cross-cultural perspective. J. Pers. 69, 979–1006. http://dx.doi.org/10.1111/1467-6494.696172.
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. Saf. Sci. 54, 58–68. http://dx.doi.org/10.1016/j.ssci.2012.11.006.
Clarke, S., 2003. The contemporary workforce: implications for organisational safety culture. Pers. Rev. 32, 40–57.
Cole, M.S., Bruch, H., 2006. Organizational identity strength, identification, and commitment and their relationships to turnover intention: does organizational hierarchy matter? J. Organ. Behav. 27, 585–605.
Cox, S.J., Cheyne, A.J.T., 2000. Assessing safety culture in offshore environments. Saf. Sci. 34, 111–129.
Cronbach, L.J., 1950. Further evidence on response sets and test design. Educ. Psychol. Meas. 10, 3–31.
Crowl, D.A., Louvar, J.F., 2002. Chemical Process Safety: Fundamentals With Applications, 2nd ed. Prentice Hall, Upper Saddle River, NJ.
Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., Billiet, J., 2014. Measurement equivalence in cross-national research. Annu. Rev. Sociol. 40, 55–75. http://dx.doi.org/10.1146/annurev-soc-071913-043137.
Feldman, D.C., 1995. Managing part-time and temporary employment relationships: individual needs and organizational demands. In: London, M. (Ed.), Employees, Careers, and Job Creation: Developing Growth-Oriented Human Resource Strategies and Programs. Josey Bass, San Francisco pp. 121–141.
Festinger, L., 1954. A theory of social comparison processes. Hum. Relat. 7, 117–140. http://dx.doi.org/10.1177/001872675400700202.
Flin, R., Mearns, K., O'Connor, P., Bryden, R., 2000. Measuring safety climate: identifying the common features. Saf. Sci. 34, 177–192. http://dx.doi.org/10.1016/S0925-7535(00)00012-6.
Frone, M.R., 1998. Predictors of work injuries among employed adolescents. J. Appl. Psychol. 83, 565–576.
Glendon, A.I., Litherland, D.K., 2001. Safety climate factors, group differences and safety behaviour in road construction. Saf. Sci. 39, 157–188.
Harzing, A.W., 2006. Response styles in cross-national survey research a 26-country study. Int. J. Cross Cult. Manage. 6, 243–266. http://dx.doi.org/10.1177/1470595806066332.
Heine, S.J., Lehman, D.R., 1997. The cultural construction of self-enhancement: an examination of group-serving biases. J. Pers. Soc. Psychol. 72, 1268–1283. http://dx.doi.org/10.1037/0022-3514.72.6.1268.
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. J. Pers. Soc. Psychol. 82, 903–918. http://dx.doi.org/10.1037//0022-3514.82.6.903.903.
Hofmann, D.A., Stetzer, A., 1996. A cross-level investigation of factors influencing unsafe behaviors and accidents. Pers. Psychol. 49, 307–339.
Hofstede, G., Hofstede, G.J., Minkov, M., 2010. Cultures and Organizations: Software of the Mind. Revised and Expanded. McGraw-Hill, New York.
Hofstede, G.H., 1980. Culture's Consequences: International Differences in Work-Related Values. Sage, Beverly Hills, CA.
Hofstede, G., 1983. The cultural relativity of organizational practices and theories. J. Int. Bus. Stud. 14, 75–89. http://dx.doi.org/10.1057/palgrave.jibs.8490867.
Hofstede, G., 1992. Cultural Dimensions in People Management-The Socialization Perspective. Pucik, Vladimir.
Hox, J.J., 2010. Multilevel Analysis: Techniques and Applications. Lawrence Erlbaum, Mahwah, NJ.
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. Multivar. Behav. Res. 50, 197–215. http://dx.doi.org/10.1080/00273171.2014.977429.
Hu, L.T., Bentler, P.M., 1999. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct. Equ. Model.: A Multidisc. J. 6, 1–55. http://dx.doi.org/10.1080/10705519909540118.
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. Transp. Res. Part F: Traffic Psychol. Behav. 26, 348–360. http://dx.doi.org/10.1016/j.trf.2014.04.006.
Hulin, C.L., Glomb, T.M., 1999. Contingent employees: individual and organizational considerations. In: Ilgen, D.R., Pulakos, E.D. (Eds.), The Changing Nature of Performance: Implications for Staffing, Motivation, and Development. Jossey-Bass, San Francisco pp. 87–118.
Hulin, C.L., 1987. A psychometric theory of evaluations of item and scale translations fidelity across languages. J. Cross Cult. Psychol. 18, 115–142. http://dx.doi.org/10.1177/0022002187018002001.
International Labour Organization, 2015. Retrieved from http://www.ilo.org/global/topics/safety-and-health-at-work/lang–en/index.htm.
Jöreskog, K.G., 1971. Simultaneous factor analysis in several populations. Psychometrika 36, 409–426. http://dx.doi.org/10.1007/BF02291366.
Jak, S., Oort, F.J., Dolan, C.V., 2013. A test for cluster bias: detecting violations of measurement invariance across clusters in multilevel data. Struct. Equ. Model.: A Multidisc. J. 20, 265–282. http://dx.doi.org/10.1080/10705511.2013.769392.
Jehn, K.A., Bezrukova, K., 2010. The faultline activation process and the effects of activated faultlines on coalition formation, conflict, and group outcomes. Organ. Behav. Hum. Decis. Processes 112, 24–42. http://dx.doi.org/10.1016/j.obhdp.2009.11.008.
Jöreskog, K.G., 1969. A general approach to confirmatory maximum likelihood factor analysis. Psychometrika 34, 183–202.
Johnson, T., Kulesa, P., Cho, Y.I., Shavitt, S., 2005. The relation between culture and response styles evidence from 19 countries. J. Cross Cult. Psychol. 36, 264–277. http://dx.doi.org/10.1177/0022022104272905.
Kakabadse, A., Kakabadse, N., 2002. Trends in outsourcing: contrasting USA and Europe. Eur. Manage. J. 20, 189–198.
Kim, E.S., Kwok, O.M., Yoon, M., 2012. Testing factorial invariance in multilevel data: a Monte Carlo study. Struct. Equ. Model.: A Multidisc. J. 19, 250–267. http://dx.doi.org/10.1080/10705511.2012.659623.
Kim, E.S., Yoon, M., Wen, Y., Luo, W., Kwok, O.M., 2015. Within-level group factorial invariance with multilevel data: multilevel factor mixture and multilevel mimic models. Struct. Equ. Model.: A Multidisc. J. 22, 603–616. http://dx.doi.org/10.1080/10705511.2011.
Kline, R.B., 2011. Principles and Practice of Structural Equation Modeling, 3rd ed. Guilford Press, New York.
Lalwani, A.K., Shavitt, S., Johnson, T., 2006. What is the relation between cultural orientation and socially desirable responding? J. Pers. Soc. Psychol. 90, 165–178.
Lanning, K., 1991. Consistency, Scalability and Personality Measurement. Springer-Verlag, New York.
Lau, D.C., Murnighan, J.K., 1998. Demographic diversity and faultlines: the compositional dynamics of organizational groups. Acad. Manage. Rev. 23, 325–340. http://dx.doi.org/10.5465/AMR.1998.533229.
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. Saf. Sci. 46, 1037–1046. http://dx.doi.org/10.1016/j.ssci.2007.05.001.
Little, T.D., Bovaird, J.A., Widaman, K.F., 2006. On the merits of orthogonalizing powered and product terms: implications for modeling interactions among latent variables. Struct. Equ. Model. 13, 497–519. http://dx.doi.org/10.1207/s15328007sem1304_1.
Little, T.D., 1997. Mean and covariance structures (MACS) analyses of cross-cultural data: practical and theoretical issues. Multivar. Behav. Res. 32, 53–76.
Lonner, W.J., 1990. An overview of cross-cultural testing and assessment. In: Brislin, R.W. (Ed.), Cross-cultural Research and Methodology Series, Vol. 14. Applied Cross-cultural Psychology. Sage Publications, Inc, Thousand Oaks, CA, US, pp. 56–76.
Lord, F.M., Novick, M.R., Birnbaum, A., 1968. Statistical Theories of Mental Test Scores.

Addison-Wesley, Oxford, England.
Mannan, M.S., Reyes-Valdes, O., Jain, P., Tamim, N., Ahammad, M., 2016. The evolution of process safety: current status and future direction. Ann. Rev. Chem. Biomol. Eng. 7, 135–162. http://dx.doi.org/10.1146/annurev-chembioeng-080615-033640.
McDonald, N., Ryan, F., 1992. Constraints on the development of safety culture: a preliminary analysis. Ir. J. Psychol. 13, 273–281. http://dx.doi.org/10.1080/03033910.1992.10557886.
Mearns, K., Yule, S., 2009. The role of national culture in determining safety performance: challenges for the global oil and gas industry. Saf. Sci. 47, 777–785. http://dx.doi.org/10.1016/j.ssci.2008.01.009.
Mearns, K., Flin, R., Gordon, R., Fleming, M., 1998. Measuring safety climate on offshore installations. Work Stress 12, 238–254. http://dx.doi.org/10.1080/02678379808256864.
Meredith, W., 1993. Measurement invariance, factor analysis and factorial invariance. Psychometrika 58, 525–543.
Mintzberg, H., 2008. Structure in Sevens. Prentice-Hall, Upper Saddle River, NJ.
Mullen, M.R., 1995. Diagnosing measurement equivalence in cross-national research. J. Int. Bus. Stud. 26, 573–596. http://dx.doi.org/10.1057/palgrave.jibs.8490187.
Muthén, L.K., Muthén, B.O., 1998. Mplus User'S Guide. 1998–2012. 7th ed. Muthén & Muthén, Los Angeles, CA.
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. J. Appl. Psychol. 96, 71–94.
Oort, F.J., 1998. Simulation study of item bias detection with restricted factor analysis. Struct. Equ. Model.: Multidiscip. J. 5, 107–124.
Ostroff, C., Kinicki, A.J., Muhammad, R.S., 2012. Organizational culture and climate. In: 2nd ed. In: Weiner, I.B., Schmitt, N.W., Highhouse, S. (Eds.), Handbook of Psychology: Industrial and Organizational Psychology, vol. 12 John Wiley, New York, NY p. 643–676.
Paulhus, D.L., Reid, D.B., 1991. Enhancement and denial in socially desirable responding. J. Pers. Soc. Psychol. 60, 307–317. http://dx.doi.org/10.1037/0022-3514.60.2.307.
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. Accid. Anal. Prev. 40, 1644–1652.
Poortinga, Y.H., 1989. Equivalence of cross-cultural data: an overview of basic issues. Int. J. Psychol. 24, 737–756. http://dx.doi.org/10.1080/00207598908247842.
Probst, T.M., Brubaker, T.L., Barsotti, A., 2008. Organizational injury rate underreporting: the moderating effect of organizational safety climate. J. Appl. Psychol. 93, 1147–1154.
Rabinowitz, P.M., 2000. Noise-induced hearing loss. Am. Fam. Physician 61, 2759–2760.
Ramsey, J.D., Burford, C.L., Beshir, M.Y., Jensen, R.C., 1983. Effects of workplace thermal conditions on safe work behavior. J. Saf. Res. 14, 105–114.
Reader, T.W., Noort, M.C., Shorrock, S., Kirwan, B., 2015. Safety sans frontières: an international safety culture model. Risk Anal. 35, 770–789.
Rebitzer, J.B., 1995. Job safety and contract workers in the petrochemical industry. Ind. Relat. 34, 40–57.
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. Pers. Psychol. 59, 65–99. http://dx.doi.org/10.1111/j.1744-6570.2006.00804.x.
Rousseau, D.M., Libuser, C., 1997. Contingent workers in high risk environments. Calif. Manage. Rev. 39, 103–123. http://dx.doi.org/10.2307/41165889.
Sörbom, D., 1974. A general method for studying differences in factor means and factor structures between groups. Br. J. Math. Stat. Psychol. 27, 229–239.
Satorra, A., Bentler, P.M., 1994. Corrections to test statistics and standard errors in covariance structure analysis. In: von Eye, A., Clogg, C.C. (Eds.), Latent Variables Analysis: Applications for Developmental Research. Sage, Thousand Oaks, CA pp. 399–419.
Schmitt, N., Kuljanin, G., 2008. Measurement invariance: review of practice and implications. Hum. Resour. Manage. Rev. 18, 210–222. http://dx.doi.org/10.1016/j.hrmr.2008.03.003.
Schmitt, N., 1982. The use of analysis of covariance structures to assess beta and gamma change. Multivar. Behav.Res. 17, 343–358.
Schneider, B., Reichers, A.E., 1983. On the etiology of climates. Pers. Psychol. 36, 19–39. http://dx.doi.org/10.1111/j.1744-6570.1983.tb00500.x.
Schwartz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6, 461–464. http://dx.doi.org/10.2307/2958889.
Sclove, L., 1987. Application of model-selection criteria to some problems in multivariate analysis. Psychometrika 52, 333–343. http://dx.doi.org/10.1007/BF02294360.
Shulruf, B., Hattie, J., Dixon, R., 2011. Intertwinement of individualist and collectivist attributes and response sets. J. Soc. Evol. Cult. Psychol. 5, 51–65. http://dx.doi.org/10.1037/h0099275.
Sirotnik, K.A., 1980. Psychometric implications of the unit-of-analysis problem (with examples from the measurement of organizational climate). J. Educ. Meas. 17, 245–282.
Spearman, C., 1904. General intelligence, objectively determined and measured. Am. J. Psychol. 15 (2), 201–292.
Steiger, J.H., Lind, J.C., 1980. Statistically based tests for the number of factors. Paper Presented at the Annual Spring Meeting of the Psychometric Society, Iowa City, IA May.
Triandis, H.C., Suh, E.M., 2002. Cultural influences on personality. Annu. Rev. Psychol. 53, 133–160. http://dx.doi.org/10.1146/annurev.psych.53.100901.135200.
Triandis, H.C., Carnevale, P., Gelfand, M., Robert, C., Wasti, A., Probst, T., et al., 2001. Culture, personality and deception: a multilevel approach. Int. J. Cross-Cult. Manage. 1, 73–90. http://dx.doi.org/10.1177/147059580111008.
Triandis, H.C., 1994. Culture and Social Behavior. McGraw-Hill Book Company.
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. J. Cross Cult. Psychol. 35, 346–360. http://dx.doi.org/10.1177/0022022104264126.
Van de Vijver, F.J., Leung, K., 1997. Methods and Data Analysis for Cross-Cultural Research, vol. 1 Sage.
Van de Vijver, F., Tanzer, N.K., 1997. Bias and equivalence in cross-cultural assessment. Eur. Rev. Appl. Psychol. 47, 263–279. http://dx.doi.org/10.1027/1015-5759.13.1.29.
Vandenberg, R.J., Lance, C.E., 2000. A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. Organ. Res. Methods 3, 4–70. http://dx.doi.org/10.1177/109442810031002.
Vandenberg, R.J., Self, R.M., 1993. Assessing newcomers' changing commitment to the organization during the first 6 months of work. J. Appl. Psychol. 78, 557–568.
Weick, K.E., 1995. Sensemaking in Organizations. Sage, London.
Welkenhuysen-Gybels, J., Van de Vijver, F.J.R., Cambré, B., 2007. A comparison of methods for the evaluation of construct equivalence in a multi-group setting. In: Loosveldt, G., Swyngedouw, M., Cambré, B. (Eds.), Measuring Meaningful Data in Social Research. Acco, Leuven, Belgium pp. 357–372.
Zohar, D., Luria, G., 2004. Climate as a social-cognitive construction of supervisory safety practices: scripts as proxy of behavior patterns. J. Appl. Psychol. 89, 322–333.
Zohar, D., Luria, G., 2005. A multilevel model of safety climate: cross-level relationships between organization and group-level climates. J. Appl. Psychol. 90, 616–628. http://dx.doi.org/10.1037/0021-9010.90.4.616.
Zohar, D., 1980. Safety climate in industrial organizations: theoretical and applied implications. J. Appl. Psychol. 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. J. Appl. Psychol. 85, 587–596.
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. Jossey-Bass, San Francisco, CA, pp. 201–230.
Zohar, D., 2010. Thirty years of safety climate research: reflections and future directions. Accid. Anal. Prev. 42, 1517–1522. http://dx.doi.org/10.1016/j.aap.2009.12.019.