Accident Analysis and Prevention 103 (2017) 96–104

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

An item-response theory approach to safety climate measurement: The Liberty Mutual Safety Climate Short Scales

Yueng-hsiang Huang[a,*], Jin Lee[a,b], Zhuo Chen[a,c], MacKenna Perry[a,d], Janelle H. Cheung[a,e], Mo Wang[f]

[a] Liberty Mutual Research Institute for Safety, Hopkinton, MA, USA
[b] Kansas State University, Manhattan, KS, USA
[c] University of Connecticut, Storrs, CT, USA
[d] Portland State University, Portland, OR, USA
[e] Oregon Health & Science University, Portland, OR, USA
[f] University of Florida, Gainesville, FL, USA

ARTICLE INFO

Keywords:
Safety climate
Item response theory
Shortened scales

ABSTRACT

Zohar and Luria's (2005) safety climate (SC) scale, measuring organization- and group- level SC each with 16 items, is widely used in research and practice. To improve the utility of the SC scale, we shortened the original full-length SC scales. Item response theory (IRT) analysis was conducted using a sample of 29,179 frontline workers from various industries. Based on graded response models, we shortened the original scales in two ways: (1) selecting items with above-average discriminating ability (i.e. offering more than 6.25% of the original total scale information), resulting in 8-item organization-level and 11-item group-level SC scales; and (2) selecting the most informative items that together retain at least 30% of original scale information, resulting in 4-item organization-level and 4-item group-level SC scales. All four shortened scales had acceptable reliability (≥0.89) and high correlations (≥0.95) with the original scale scores. The shortened scales will be valuable for academic research and practical survey implementation in improving occupational safety.

1. Introduction

1.1. Safety climate

Safety climate research has been ongoing for more than 35 years, since Zohar published his seminal work in 1980 defining this construct as workers' shared perceptions regarding their organization's policies, procedures, and practices in relation to the value and importance of safety within that organization (Zohar, 1980; Griffin and Neal, 2000; Zohar, 2000, 2002, 2003). The study of safety climate is based on perceptions of workers, with the major factors relating to (a) management commitment to safety and (b) communication pertaining to safety as a true priority from top management and direct supervisors (Dejoy et al., 2004). Prior research has stated that safety climate is a multilevel construct encompassing two managerial levels: (1) organization-level safety climate, which refers to employees' perceptions of the company's or top management's commitment to and prioritization of safety, and (2) group-level safety climate, meaning employees' perceptions of their direct supervisors' commitment to and prioritization of safety (e.g., Zohar and Luria, 2005; Huang et al., 2013a,b). Several meta-analyses have provided robust evidence that safety climate is one of the best leading indicators of organizational safety outcomes, such as frequency or severity of injury incidents (Christian et al., 2009; Beus et al., 2010; Nahrgang et al., 2011). Overall, safety climate influences employees' motivation and knowledge to act in a safe manner, which in turn lead to safer behaviors and fewer accidents and injuries (Griffin and Neal, 2000; Christian et al., 2009).

Since the inception of safety climate research, many safety climate scales have been developed and validated in the scientific literature. One of the most widely used safety climate scales published in the field, which has robust evidence of reliability and validity, is a generic safety climate scale developed by Zohar and Luria (2005). Their scale includes 32 total items: 16 items to measure organization-level safety climate and 16 items to measure group-level safety climate. In Zohar and Luria's (2005) study, the Cronbach's alpha of the scale was 0.92 for organizational-level safety climate (OSC) and 0.95 for group-level safety climate (GSC). In terms of criterion-related validity, OSC was correlated with safety audit/observation scores at 0.46, and GSC was correlated with safety behavior observations at 0.38. According to Google Scholar (retrieved January, 2017), their paper has been cited by

* Corresponding author at: Center for Behavioral Sciences, Liberty Mutual Research Institute for Safety, 71 Frankland Road, Hopkinton, MA 01748, USA. (Y.H. Huang)
E-mail address: Yueng-hsiang.Huang@Libertymutual.com (Y.-h. Huang).

http://dx.doi.org/10.1016/j.aap.2017.03.015
Received 8 December 2016; Received in revised form 15 February 2017; Accepted 19 March 2017
Available online 07 April 2017

nearly 800 publications, many of which use their measure. For example, one of the heavily cited papers (Johnson, 2007) found that GSC was significantly correlated with injury frequency at −0.50 and safety behaviors at 0.78. Examining OSC, Martínez-Córcoles et al. (2011) found a correlation with safety behaviors at 0.43, while Brondino et al. (2012) found correlations with safety compliance and safety participation ranging from 0.27 to 0.36. Due to its increasingly high usage in research and practice, the current study focuses on increasing the utility of this scale by shortening the number of items required while maximizing information provided.

1.2. Length of safety climate scales

Safety researchers are frequently faced with a dilemma in field research: whether to use brief measures or longer, more exhaustive and thorough measures. A longer measure can capture a fuller range of construct content and variance of interest, whereas a brief measure can boost both participant engagement and the efficiency of data collection. There are times when a longer scale is preferable, but shorter scales may be more effective in other cases.

Overall, a survey instrument should not overwhelm respondents with too many questions. Previous research has demonstrated that survey length can negatively impact response rates (e.g., Crawford et al., 2001). By shortening the length of a survey, individuals may be more likely to perceive that they have time to participate in survey research, even when they do not feel participation will directly benefit themselves (Woods and Hampson, 2005). Furthermore, in cases where measures contain many items focused on a very similar topic, many participants may interpret items as redundant and may have negative reactions toward the overall survey assessment (Wanous et al., 1997).

An additional issue with longer measures is that their use can limit the nature of models that can be tested to explore relations among various constructs (Fisher et al., 2016). Zohar and Luria's (2005) generic safety climate scale includes 32 items, which is a fairly long measurement scale. Despite the existence of this psychometrically solid and widely accepted scale, Zohar (2010) stated that more work is needed to explore how safety climate emerges and how safety climate is influenced or changed (i.e., which factors contribute to the development of safety climate perceptions). In order to fill this gap, researchers need to collect additional data on many other variables simultaneously with safety climate. With the current length of the safety climate scale, it is challenging to achieve this goal within realistic limitations that researchers face. In order to further explore potential factors influencing safety climate, a shorter and valid generic safety climate scale is needed.

1.3. Item response theory (IRT)

We propose an Item Response Theory (IRT) approach because it assesses multiple psychometric features of individual scale items. In comparison, Classical Test Theory (CTT) places more emphasis on the scale's composite score. IRT is a probabilistic non-linear modeling technique for developing and evaluating psychological measurement scales. For example, it can be posited that items of a scale are designed to assess a certain psychological attribute (e.g., safety perception) such that endorsing higher values on the items suggests a stronger underlying psychological attribute (e.g., stronger safety perception). If respondents give undiscriminating endorsements to an item when they indeed differ in terms of the underlying psychological attribute, the item should be deemed improper as a measure of the psychological attribute. To this end, IRT calculates the respondents' probability of endorsing particular response options of each scale item and estimates each item's ability to differentiate respondents, which can be used for strategic tailoring of lengthy psychological scales.

It needs to be noted that even though IRT has been frequently used with educational and psychological tests which have correct or wrong answers, it can be applied to Likert scale-based measures (i.e., item with ordered categorical – polytonomous – response options) of psychological trait/attribute such as perceived job security (Probst, 2003) and personality (e.g., Reise and Henson, 2000). Likewise, higher levels of underlying trait/attribute are assumed to lead to higher probabilities of stronger endorsement (e.g., choosing the category 'strongly agree' on a 5-point Likert scale). IRT is free from limitations faced by conventional linear regression-based development and validation techniques such as circular sample dependency of item/person statistics (Fan, 1998). Furthermore, IRT considers the differentiating/discrimination ability and difficulty of each item as information to be incorporated in the scale. It allows researchers to more efficiently assemble the items that offer the most information for measuring the targeted underlying trait/attribute.

The unique parameters offered by IRT, such as slope and difficulty parameters, can be derived based on the probability of responses, which is illustrated by the item option response functions (ORFs). For a five-point Likert scale, each item has five response options. In the polytomous IRT model, ORFs are used to describe participants' response patterns. Each option has an ORF curve, with the x-axis representing the trait being measured (θ) and the y-axis representing the probability of endorsing this particular option; an ORF thus depicts the relationship between the participants' trait and their responses to an item.

The slope, discrimination, or differentiation parameter determines the slope of the option response functions (ORF) for each item. Every item will have one slope parameter. If all other difficulty parameters are equal, items with high slope parameters will have smaller overlap of θ values between the option response functions, representing better differentiation. In the current study, the slope parameter represents each item's sensitivity to the overall level of safety climate.

The difficulty parameter determines the location of the ORF along the θ axis and indicates on which part of the range of θ the item is most informative, or the θ value at which people have a 50% chance of selecting specified responses (i.e., the cutoff points that separate the response option categories). In the current study, each item was rated on a 5-point Likert scale. Therefore, each item has four ORFs and four difficulty parameters (i.e., the cutoff points that separate response 1 from responses 2–5, responses 1–2 from 3 to 5, responses 1–3 from 4 to 5, and, finally, responses 1–4 from 5). These four difficulty parameters jointly indicate the overall difficulty of an item. In the current study, the item's difficulty represents whether an item is more informative (i.e., sensitive in differentiating the level/strength of estimated target trait) at lower or higher ranges of safety climate scores.

The item information curve (IIC) for each item is a function of both the slope and difficulty parameters. The amount of information that a particular item provides depends on both the size of the slope parameter and the spread of the category thresholds. An IIC represents the amount of information provided by a specific item across the entire continuum of the latent construct of interest. The area of the IIC above the x-axis (θ) equals the item information. If an item has a larger amount of item information, the item has higher discriminating ability to differentiate respondents along the θ axis. Depending on the slope and difficulty parameters, the amounts of information offered by items will differ. By aggregating the IICs of items in a measure, the test information function (TIF) for a scale can be generated. Similar to IICs, the area of the TIF above the θ axis equals the total test information. If a scale has a larger amount of total test information, the scale score has higher discriminating ability along the latent θ value.

Overall, the current study aims to utilize IRT to shorten Zohar and Luria's (2005) 32-item safety climate scale. Both slope and difficulty parameters for each item in the existing scale were calculated, and all information available was carefully considered to decide on the best items to include in the final shortened scales and the ideal number of items to include. The new, shortened scales are expected to benefit future safety climate research and practice by allowing for more diverse data collection opportunities and addressing concerns that organiza-