Personality and Individual Differences 90 (2016) 190–198

Contents lists available at ScienceDirect

Personality and Individual Differences

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

Measurement equivalence of the Wong and Law Emotional Intelligence Scale across cultures: An item response theory approach

Matthew L. LaPalme [a,*], Wei Wang [a], Dana L. Joseph [a], Donald H. Saklofske [b], Gonggu Yan [c]

[a] University of Central Florida, 4000 Central Florida Blvd, Orlando, FL, USA
[b] University of Western Ontario, 1151 Richmond St, London, Ontario, Canada
[c] Beijing Normal University, 19 Xinjiekou Outer St, Haidian, Beijing, China

ARTICLE INFO

Article history:
Received 26 August 2015
Received in revised form 23 October 2015
Accepted 24 October 2015
Available online 12 November 2015

Keywords:
Emotional intelligence
WLEIS
Item response theory
Ideal-point models
Differential item functioning
Cross-cultural measurement equivalence

ABSTRACT

The global popularity of emotional intelligence (EI) makes understanding its measurement equivalence across cultures an important issue. Although previous research examining the measurement equivalence of self-reported EI has failed to detect cultural differences, these results may be due to the use of measurement equivalence models that do not adequately specify item level differences between cultures or quantify the magnitude of differences. In this study, we adopted an item response theory (IRT) approach to examine differential item functioning (DIF) in the Wong and Law Emotional Intelligence Scale (WLEIS; Wong & Law, 2002) across American and Chinese cultures using both a dominance and ideal point IRT model. Our results revealed: 1) The dominance IRT model had good fit with the WLEIS compared to the ideal point model; 2) the WLEIS items varied in their DIF, which ranged from negligible to moderately large across American and Chinese cultures; and 3) the largest DIF was found for the Other Emotional Appraisal (OEA) dimension of the WLEIS, which indicated that Chinese respondents found these items substantially more difficult to endorse. Implications and future research directions are discussed.

1. Introduction

In recent years, emotional intelligence (EI) has garnered the attention of researchers and practitioners alike because of its relationship with job performance (Joseph, Jin, Newman, & O'Boyle, 2015; Joseph & Newman, 2010a; O'Boyle, Humphrey, Pollack, Hawver, & Story, 2011), leadership (Harms & Crede, 2010), social competence in adults and children, academic achievement (see Mayer, Roberts, & Barsade, 2008 for a review) and both psychological and physiological wellbeing (Martins, Ramalho, & Morin, 2010; Stough, Saklofske, & Parker, 2009). As such, EI measures are often used for a variety of purposes, including personnel training and selection (Cherniss, 2000), leadership training and selection (Caruso, Mayer, & Salovey, 2001; Wong & Law, 2002), distinguishing between healthy and bipolar/depressed individuals (Hertel, Schutz, & Lammers, 2009), improving learning process and group processes for students and teachers (Vesely, Saklofske, & Nordstokke, 2014), and improving the quality and therapeutic value of nursing care (Freshwater & Stickley, 2004).

With the popularization of EI, it is important to examine measurement equivalence of EI scales across demographic groups (e.g., age, gender, culture). Measurement equivalence studies examine the extent to which a scale measures the same construct in the same way across groups (Vandenberg & Lance, 2000) to ensure that differences in responses are interpreted in a meaningful and unbiased manner. In the current study, we use the global popularity of EI scales as the impetus for an examination of measurement equivalence in self-report EI across cultures. Previous research has shown promising results indicating measurement invariance of self-report EI measures across cultural groups (e.g., Li, Saklofske, Bowden, Yan, & Fung, 2012; Whitman, Van Rooy, Viswesvaran, & Kraus, 2009). These results are surprising, however, given that cultural perceptions about emotions vary widely. For example, in western-European cultures, the public display of anger, contempt and disgust is more permissible than in Asian cultures which tend to emphasize the public display of happiness (Safdar et al., 2009). Subsequently, this type of cross-cultural difference in how much or how little an emotion should be displayed may translate to cross-cultural differences in how individuals interpret items about emotional displays. Given these known cultural differences, we argue that the conclusions drawn from prior measurement equivalence investigations require further exploration because of the methods used. First, most previous studies have only examined scale- or factor-level measurement equivalence rather than examining the item-level psychometric properties of EI scales, which may hide item-level nonequivalence (i.e., although previous research found measurement equivalence, item-level nonequivalence may still exist). Moreover, item-level nonequivalence is arguably the most meaningful if one wishes to identify the type of item content that is causing any

* Corresponding author at: University of Central Florida, 4111 Pictor Lane, Psychology Building 99 Ste. 320, Orlando, FL 32816, USA.
E-mail addresses: m.lapalme@knights.ucf.edu (M.L. LaPalme), wei.wang@ucf.edu (W. Wang).

http://dx.doi.org/10.1016/j.paid.2015.10.045
0191-8869/

measurement nonequivalence (see Gignac & Ekermans, 2010; Nye, Newman, & Joseph, 2010). Second, previous studies of cross-cultural measurement equivalence in self-report EI have only used null hypothesis significance testing (NHST) to evaluate measurement equivalence. NHST only provides information about whether measurement equivalence exists and fails to provide an estimate of the magnitude of measurement non-equivalence (i.e., if measurement nonequivalence is found, it is unclear precisely how big the difference is). Finally, prior investigations of cultural measurement equivalence in self-report EI used dominance models which assume that as a person's latent standing on the trait increases, the probability of endorsing the item also increases (Coombs, 1964; Wang, Tay, & Drasgow, 2013). Although dominance models may be appropriate for ability measures of EI (e.g., measures that are scored correct/incorrect), research shows that an ideal point model may be more appropriate for measures that require introspection (i.e., self-report trait measures; Tay, Ali, Drasgow, & Williams, 2011; Tay & Drasgow, 2012). Therefore, we use both a dominance model and an ideal point model when examining cultural measurement equivalence in the current paper.

To summarize, this study was aimed at examining the measurement equivalence of self-reported EI across American and Chinese cultures by adopting an item response theory (IRT) differential item functioning (DIF) approach to examine responses on the widely-used Wong and Law Emotional Intelligence Scale (WLEIS; Wong & Law, 2002). Specifically, we applied two distinct IRT models under two different psychometric frameworks (i.e., Samejima's Graded Response [SGR] model under the dominance framework and Generalized Graded Unfolding Model [GGUM] under the ideal-point framework) and we assessed the model fit for each model with the WLEIS data. We then estimated whether differential item functioning (DIF) exists (via NHST) and the magnitude of DIF (with DIF effect sizes) across cultures. Finally, we examined the DIF manifested in item difficulty and discrimination using item characteristics (ICC) to evaluate measurement equivalence at the most precise level of specificity (i.e., the response option).

1.1. Emotional intelligence

Previous literature has drawn a distinction between ability measures of EI and trait measures of EI (Brackett & Mayer, 2003; Côté, 2014; Mayer, Caruso, & Salovey, 1999; Petrides, Pita, & Kokkinaki, 2007b; Siegling, Saklofske, & Petrides, 2015). Ability EI, as espoused by Mayer and Salovey (1997), is viewed as an intelligence and has four facets that reflect the ability to: 1) accurately perceive emotion, 2) use emotion to facilitate thought, 3) understand emotion, and 4) manage emotion. Ability EI measures therefore attempt to capture the latent ability to perceive, use, understand, and regulate emotion. Trait EI is conceptualized as "a constellation of emotion-related self-perceptions and dispositions at the lower levels of personality hierarchies" (p. 26, Petrides, Perez-Gonzalez, & Furnham, 2007a). That is, trait EI is conceptualized as the shared affective variance within the personality domain that is sampled by lower order facets (Siegling et al., 2015) such as adaptability, emotion regulation, and trait empathy (to name a few).

1.2. The Wong and Law Emotional Intelligence Scale

Wong and Law (2002) developed the Wong and Law Emotional Intelligence Scale (WLEIS), a self-reported 16-item EI measure based on the ability model which has four dimensions: self emotional appraisal (SEA), other emotional appraisal (OEA), use of emotions (UOE), and regulation of emotions (ROE). The WLEIS' brief, easy-to-administer self-report format and strong theoretical grounding in ability EI theory make it a popular measure for research (e.g., a Google Scholar citation search indicated this measure has been cited over 1300 times by May 2015).

As evidence of the construct validity of the WLEIS, the WLEIS has shown criterion-related validity to job satisfaction ($r = .40$), job performance ($r = .21$), and peer-rated task performance ($r = .27$; Hui-Hua & Schutte, 2015; Wong & Law, 2002), discriminant validity from personality (Wong & Law, 2002), and equivalent measurement across peer and self-reports (Joseph & Newman, 2010b).

1.3. Previous measurement equivalence studies on the WLEIS

The purpose of measurement equivalence is to examine if a scale measures the same construct in the same way across groups with different characteristics (Vandenberg & Lance, 2000). Finding that tests are interpreted differently across populations threaten the validity of test score interpretation (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). According to the Standards for Educational and Psychological Testing, 2014, a test is considered fair if scores have the same meaning across target populations and if individuals are not disadvantaged because of characteristics that are irrelevant to the measured construct (for example, one's culture). Measurement non-equivalence, therefore, is a strong indicator that bias may exist and that scores on a scale should be interpreted with caution across subgroups.

There are two major approaches to examining measurement equivalence: Confirmatory Factor Analysis (CFA) and Differential Item Functioning (DIF). The CFA approach to measurement equivalence (also called the mean and covariance structures method, or MACS) involves testing three stages of invariance: configural invariance (i.e., a test of whether the same factor structure is found across groups), metric invariance (i.e., a test of whether the factor loadings are the same across groups), and scalar invariance (i.e., a test of equality of item intercepts across groups; Vandenberg & Lance, 2000). In contrast, DIF is distinguishable from the MACS measurement equivalence approach because it analyzes metric and scalar invariance in one step and is better able to account for the effects of mean differences of latent-trait distributions between subpopulations (i.e., impact; Stark, Chernyshenko, & Drasgow, 2006a). Moreover, the IRT approach performs better in large samples such as those found in the current study (Stark, Chernyshenko, Drasgow, & Williams, 2006b). Finally, dominance model assumptions are implicit to all CFA analyses (and therefore MACS) whereas IRT DIF approaches are able to examine DIF with either a dominance or ideal point models (Drasgow, Chernyshenko, & Stark, 2010), making the IRT approach more suitable for measures like self-reported EI wherein the most appropriate model is unclear.

Previous work has primarily used the MACS approach to demonstrate measurement equivalence of the WLEIS across self and other-reports (Joseph & Newman, 2010b), sex, race (Whitman et al., 2009), and language of the measure (Li et al., 2012). Related to the current study, Wang, Kim, and Ng (2011)) used the MACS approach and found weak measurement invariance (i.e., configural and metric invariance only) across East Asian, European, and Indian samples. CFA approaches have also been applied to other EI measures (such as the TEIQue) and have similarly found configural invariance (Mavroveli, 2012) and modest measurement differences (Gökçen, Furnham, Mavroveli, & Petrides, 2014) between European and Asian samples.

While the research noted above indicates measurement invariance of the WLEIS holds across a myriad of conditions and groups, because these prior studies have primarily used MACS to test measurement equivalence, there are several methodological limitations. First, the MACS approach detects measurement non-equivalence by testing for significant decrements in model fit (as factor structure, factor loadings, and intercepts are sequentially constrained). Thus, although prior studies examined whether measurement equivalence existed, these studies failed to quantify the magnitude of measurement non-equivalence on the WLEIS across groups. Second, prior work primarily used the MACS approach to examine scale-level measurement equivalence. As such, few studies have explained the underlying item-level measurement non-equivalence in the WLEIS. Finally, because the MACS approach is inherently tied to a dominance model, no previous studies have