The concept of life satisfaction across cultures: An IRT analysis

Shigehiro Oishi

Department of Psychology, University of Virginia, PO Box 400400, Charlottesville, VA 22904-4400, USA

Available online 14 February 2006

Abstract

The present study examined measurement equivalence of the Satisfaction with Life Scale between American and Chinese samples using multigroup Structural Equation Modeling (SEM), Multiple indicator multiple cause model (MIMIC), and Item Response Theory (IRT). Whereas SEM and MIMIC identified only one biased item across cultures, the IRT analysis revealed that four of the five items had differential item functioning. According to IRT, Chinese whose latent life satisfaction scores were quite high did not endorse items such as "So far I have gotten the important things I want in life" and "If I could live my life over, I would change almost nothing." The IRT analysis also showed that even when the unbiased items were weighted more heavily than the biased items, the latent mean life satisfaction score of Chinese was substantially lower than that of Americans. The differences among SEM, MIMIC, and IRT are discussed.
© 2006 Elsevier Inc. All rights reserved.

Keywords: Life satisfaction; Culture; Measurement; Well-being judgments

1. Introduction

Kitayama and Markus (2000) presented a theoretical analysis of cultural differences in well-being, and argued that (a) well-being comes from cultural participation, and (b) to the extent that cultural participation requires different forms across cultures, well-being feels different and means something different across cultures. For instance, the Item Response Theory (IRT) analysis of the positive affect (PA) subscale of the Positive and Negative

E-mail address: soishi@virginia.edu

Affect Schedule (PANAS: Watson, Clark, & Tellegen, 1988) showed that "pride" was not endorsed by Chinese who endorsed other "positive" emotions, whereas it was endorsed by Americans who endorsed other "positive" emotions (Oishi, in press). This measurement discrepancy indicates that "pride" is not conceived as "positive" among Chinese and reveals the conceptual difference of "positive" emotions between Chinese and Americans (see Huang, Church, & Katigbak, 1997 on anxiety between Philippines and Americans). A main implication of Kitayama and Markus' theoretical analysis for culture and personality research is that it is crucial to examine not only mean-level differences in a construct (e.g., self-esteem) and the nomological net of this construct across cultures, but also the deeper structure of the construct because the traditional questions of mean-level difference across cultures presuppose conceptual equivalence.

The Satisfaction with Life Scale (SWLS; Diener, Emmons, Larsen, & Griffin, 1985) has been one of the most widely used scales for the measurement of global life satisfaction. Life satisfaction is one of the central constructs of well-being (Diener, 1984) and has been of great interest to both cultural and personality psychologists (Diener, Oishi, & Lucas, 2003; Diener, Suh, Lucas, & Smith, 1999 for review). Its psychometric properties have been well-established in the United States (Pavot & Diener, 1993). In contrast, the psychometric properties of the SWLS in non-American samples have not been extensively examined (see Vittersø, Røysamb, & Diener, 2002, however, for an initial effort in this direction). Therefore, although previous research found large international differences in the mean levels of life satisfaction (e.g., Diener, Suh, Smith, & Shao, 1995), it is unclear exactly how these mean differences can be interpreted because of the lack of information concerning measurement equivalence. The present study examines measurement equivalence of the SWLS between Chinese and American college student samples, using the structural equation modeling (SEM), multiple indicator multiple cause (MIMIC) modeling, and Differential Item Functioning (DIF) analysis.

1.1. DIF analysis

To examine measurement equivalence of the SWLS among Chinese and American college students, I employed the IRT analysis with a model-testing approach (Thissen, Stenberg, & Gerrard, 1986) using the Multilog 7.03 program. IRT is different from classical test theory (CTT) in several important ways (see Embretson & Reise, 2000; Hambleton & Swaminathan, 1985 for details). The most significant difference between CTT and IRT in the present context is concerned with the standard error of measurement. Whereas the standard error of measurement is assumed to apply to the whole sample in CTT, the standard error of measurement in IRT varies depending on the latent trait score (typically, there is less reliability for those with extreme latent scores). In other words, whereas the source of errors in CTT is either occasion (in the case of test–retest reliability) or item sampling (in the case of internal consistency), additional sources of error can be considered in IRT (as in Generalizability theory by Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson, Webb, & Rowley, 1989), such as a person's latent score and person-by-item interaction. Traditional reliability indices such as Cronbach's α and test–retest reliability coefficient do not provide information about person-by-item interaction, namely, whether some items measured some individuals better than others. In IRT, this interaction is considered. In addition, classical item parameters (e.g., item–total correlation) are sample-specific, whereas IRT parameters are not sample dependent. The score computed in IRT,