Learning and Individual Differences 27 (2013) 16–25

Contents lists available at ScienceDirect

Learning and Individual Differences

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

Comparisons of latent factor region means of spatial ability based on measurement invariance☆

Hi-Lian Jeng [a,*], Yu-Fang Chen [b,1,2]

[a] Graduate Institute of Digital Learning and Education, National Taiwan University of Science and Technology, Taiwan
[b] Shuang-Xi High School, New Taipei City, Taiwan

ARTICLE INFO

Article history:
Received 24 April 2012
Received in revised form 5 March 2013
Accepted 24 June 2013

Keywords:
Measurement invariance
Region invariance
Latent factor mean structure
Multi-group confirmatory factor analysis
Spatial ability

ABSTRACT

Empirical investigation was conducted to examine the measurement invariance of the Standardized Spatial Ability Test in three regions. A full line of invariance testing (for configure, factor loadings, intercepts, residuals, latent factor variances, latent factor covariances, and latent factor mean structure) was employed. Special treatment was applied for the non-normal and dichotomous data in the study. As the region invariance of the measurement model was sustained, the results from the two mean comparisons (the latent factor region means and the observed score region means) were the same accordingly. The establishment of measurement invariance of the measure can facilitate meaningful group comparisons on the same test structure and augment the measure's validity in generalization in large scale application and test development.

1. Introduction

Spatial ability is an important component of intellectual abilities (Gardner, 1983; Guilford, 1988; Thurstone, 1938), and is correlated with academic, vocational, and everyday performances such as general chemistry (Bodner & Guay, 1997), gender and educational differences in an air traffic control training program (Contreras, Colom, Shih, Alava, & Santacreu, 2001), high school geometry and creativity (Guzel & Sener, 2009), dental education (Hegarty, Keehner, Khooshabeh, & Montello, 2009), in using e-map, Google Earth, and GPS (Lei, Kao, Lin, & Sun, 2009), introductory physics (Pallrand & Seeber, 1984), map learning (Sanchez & Branaghan, 2009), and wayfinding (Silverman et al., 2000). Spatial ability is usually measured by tests of paper-and-pencil or performance type, and these tests are usually designed with different foci of spatial factors underlying the spatial ability. In the repertoire of spatial factors reported in literature, Visualization (Vz), Spatial Orientation (SO), Speeded Relation (SR) and Mental Rotation (MR) are most commonly studied (Linn & Petersen, 1985; Lohman, 1979, 1988; McGee, 1979). However, it is rather inconclusive in synthesizing the results and conclusions about these spatial factors, partly because of the overlaps in the definitions of the spatial factors, and partly because of the other aspects such as speed, level, and complexity intertwined in the test tasks (Lohman, 1988). Therefore, the resultant factor loadings of certain spatial factors were sometimes found switched from the presumed spatial factors (Chien et al., 2008; Jeng & Chen, 2007; Lohman, 1988), which would make the spatial factors indistinguishable and the test structure distracted from the original test design.

In developing a large scale standardized test, a stable and common test structure to its designated population and samples is particularly important in that the test development takes a lot of effort, the test influences its stakeholders, and therefore the validity in the comparisons of test scores is crucial in making decisions about the reflection of individual differences. An important objective of psychological measurement is to evaluate the subject's latent traits with minimum measurement error. Most measures with good test reliability can reduce measurement error, but it is more complicated when referring to the extent that the latent traits can be assessed. Among subjects, there are individual differences as well as possible structural or systematic variations due to demographics, genders, cultures, languages, countries, or time. If these variations do influence, without being formally recognized and extracted from the observed test scores, it is often the consequence that they are treated as part of the subject's latent traits, as would be derived from the true score model (Lord & Novick, 1968), and the interpretation, generalization, and comparisons of individual differences based on the observed test scores would then be biased (Cole & Moss, 1989).

It is usual practice to assume that measurement is invariant and generalizable to groups for those who can be subsumed in the population.

☆ Parts of the research work and their abstracts have been reported in
1. The 7th Conference of the International Test Commission, held at the Chinese University of Hong Kong in July, 2010.
2. The 9th Conference of the Psychological and Educational Measurement and Testing in Taiwan and China, held at the National Taiwan Normal University in October, 2010, for receiving award to Yu-Fang Chen's master's thesis.

* Corresponding author. Tel.: +886 2 27376500; fax: +886 2 27376433.
E-mail addresses: JENGHL@mail.ntust.edu.tw (H.-L. Jeng), gtf@ntpc.edu.tw (Y.-F. Chen).
[1] Tel.: +886 920221418.
[2] Yu-Fang Chen was Jeng's research assistant and he helped with the work of this research.

1041-6080/$ – see front matter
http://dx.doi.org/10.1016/j.lindif.2013.06.012

This practice "has worked well in homogeneous and stable communities where people spoke the same language, had the same social background and shared the same culture, that is, in well bounded regions with low social and geographical mobility (Roe, 2010)." However, in crossing the passage to globalization, many societies have undergone enormous and fast mobility with different cultures and languages emerging, therefore the assumption of measurement invariance (MI) is getting more challenged and attended.

It is the notion of MI that direct evidence of invariance in measurement model (test structure) should be provided, for otherwise it would be difficult to determine whether the group differences are real or whether the measurement model means differently to the groups. For psychological measures, MI is considered as a critical property (Drasgow, 1984, 1987; Rock, Werts, & Flaugher, 1978), a necessary condition (Bowden, Lange, Weiss, & Saklofske, 2008), and a logical prerequisite (Vandenberg & Lance, 2000). The concept of MI and its analytic procedures (Byrne, Shavelson, & Muthén, 1989; Drasgow, 1984, 1987; Meredith, 1993; Rock et al., 1978; Stark, Chernyshenko, & Drasgow, 2006; Vandenberg & Lance, 2000) facilitate the examination of test structure in the groups of interest. The well-known Wechsler series is a prominent application of MI. The four-latent-variable measurement model of WAIS-III scores has been shown to be MI for the U. S. and Canadian standardization samples (Bowden et al., 2008). The factorial structure of WISC-IV is gender invariant in the U. S. standardization samples (Chen & Zhu, 2008). With the recent publication of WISC-IV in Asia, testing MI in Hong Kong, Macau, Taiwan, and mainland China is conducted using the structure of U. S. WISC-IV as the baseline model (Chen, Weiss, & Li, 2010). The structure of U. S. WISC-IV is also used as the baseline model to develop the Chinese Intelligence Scale for Young Children (CISYC), and MI of CISYC is tested between rural and urban children in China (Guo, Aveyard, & Dai, 2009). As shown in these applications, a standardized test with MI confirmed would support and provide more comparable results, especially in the case of larger scale comparisons. With the world becoming flat, the issue of MI would further challenge international and cross national test developers to concern about whether their tests evaluate the human resources on the same scale.

1.1. Tests of measurement invariance

A measurement model provides a description of the numerical and theoretical relationship between the observed variables (items) and their corresponding latent variables, factors, traits or constructs (Bowden et al., 2008). This model should be able to explain the relationship in a consistent manner across groups. The establishment of MI can facilitate meaningful cross-group comparisons of construct measurement of the test.

To examine MI across groups, the first test is to test the variance–covariance matrices relating items in the measure across the groups. If the first test indicates nonsignificant differences in the groups, the usual conclusion is that MI is established and no further tests are required. However this is a rather strict test and is seldom seen applied as shown in Table 1. Instead the sequential test analyses as in the following are introduced. Tests (1–4) are categorized as the measurement components defining the measurement of the latent variables in each group, and the reliability associated with the measurement of the observed variables, while the latter tests (5–7) are categorized as the structure components testing for the relationships of the latent factors (Bowden et al., 2008; Schmitt & Kuljanin, 2008).

(1) Test for configural invariance: This test examines whether the overall test structure is the same across the groups, which is equivalent to determining if the same latent factors connect to the same observed variables in all groups.
(2) Test for metric invariance: This test examines whether the factor loadings of the observed variables on the latent factors are the same in all groups.

Table 1
Parameter tests of invariance used between 2004 and 2009.

References	Total # of tests	Var.- cov. matrices	Configure	Metric	Scalar	Residuals	Latent factor var.	Latent factor cov.	Latent factor mean
1. Ang et al. (2009)	4		◎	◎			◎	◎	
2. Beuckelaer and Lievens (2009)	3		◎	◎	◎				
3. Bowden et al. (2008)	4		◎	◎	◎	◎			
4. Chen, Hsieh, and Ye (2007)	6		◎	◎	◎	◎	◎	◎	
5. Chen and Hwang (2006)	7		◎	◎	◎	◎	◎	◎	◎
6. Chen and Hwang (2008)	7		◎	◎	◎	◎	◎	◎	◎
7. Chen and Zhu (2008)	5		◎	◎		◎	◎	◎	
8. Chen and Zhu (2009)	3		◎	◎		◎			
9. Deng et al. (2008)	2		◎	◎					
10. Gomez (2006).	5		◎	◎			◎	◎	◎
11. Grouzet, Otis, and Pelletier (2006)	6		◎	◎		◎	◎	◎	◎
12. Guo et al. (2009)	6		◎	◎	◎		◎	◎	◎
13. Hansen, Deitz, Tokman, Marino, and Weaver (2009)	7		◎	◎	◎	◎	◎	◎	◎
14. Hilton, Schau, and Olsen (2004)	4		◎	◎			◎	◎	
15. Hu (2008)	6		◎	◎	◎	◎	◎	◎	
16. Kim, Nair, Knight, Roosa, and Updegraff (2009)	4		◎	◎	◎	◎			
17. Lievens, Anseel, Harris, and Eisenberg (2007)	4		◎	◎			◎	◎	
18. Limbers, Newman, and Varni (2008)	2		◎	◎					
19. Lindwall and Palmeira (2009)	5		◎	◎		◎	◎	◎	
20. Motl and Conroy (2001)	6		◎	◎		◎	◎	◎	◎
21. Revell, Caskie, Willis, and Schaie (2009)	6		◎	◎	◎		◎	◎	◎
22. Sawang et al. (2009)	2			◎	◎				
23. Shih and Wu (2008)	5		◎	◎		◎	◎	◎	
24. Vlachopoulos (2008)	4		◎	◎	◎	◎			
25. Wasti, Tan, Brower, and Önder (2007).	2			◎	◎				
26. Wu and Yao (2006)	3		◎	◎		◎			
27. Xu (2008)	5		◎	◎	◎	◎			◎
Total papers			25	27	14	16	16	16	9
Percentage (out of 27 papers)			93	100	52	59	59	59	33
Total papers			59	66	8	33	39	22	14
Percentage (out of 67 papers from Vandenberg and Lance (2000))			88	99	12	49	58	33	21