Assessing Measurement Invariance in the Attitude to Marriage Scale across East Asian Societies

Xiaowen Zhu

Xi'an Jiaotong University

Yanjie Bian

Xi'an Jiaotong University

Paper to be presented at the annual meeting of the World Association for Public Opinion Research (WAPOR) in Hongkong, June 14-16, 2012

Abstract

In comparative social science research, it is essential to establish measurement invariance for relevant constructs across different cultures and political economies. If this "measurement invariance (MI)" assumption is violated, any comparisons and interpretations will be invalid and misleading. In this paper we reviewed the main approaches for assessing measurement invariance – confirmatory factor analysis (CFA) and item response theory (IRT). CFA is the most popular and traditional method for evaluating measurement invariance. IRT is another technique that researchers have recently used to test MI. Based on the 2006 EASS module on family, we applied these two methods to the assessment of the measurement invariance in the attitude to marriage scale across four East Asian societies (China, Japan, Korea, and Taiwan). In the CFA framework, a hierarchical multiple-group CFA models were estimated and compared using the LISREL8.8 program in order to test different types of invariance – configural equivalence, weak equivalence, and strong equivalence. In the IRT framework, the logistic regression procedure was used to detect the items with differential item functioning (DIF). The methods provided different information regarding the invariance of the EASS marriage scale.

Assessing Measurement Invariance in the Attitude to Marriage Scale across East Asian Societies

Introduction

In social research, substantive studies focus on multi-group comparisons which investigate group differences on constructs of interest. The groups can be many kinds such as different gender or ethnic groups, different societies, different cultures, or different countries. A critical assumption for cross-group comparative research is "measurement invariance (MI)". It requires that the instrument of measurement (e.g., ability tests, surveys, questionnaires) measures the same underlying construct in all groups. In other words, the construct should have the same theoretical structure and meaning across the groups of interest. When the items on a scale display the same relationship to a latent variable across two or more groups, this measurement scale is said to be "invariant" across those groups. If the MI assumption is violated, any comparisons and interpretations will be invalid and misleading. Mean differences between groups in observed scores may not be attributed solely to construct differences, and they may be due to some construct-irrelevant variables (e.g., translation, socio-cultural factors). Unless evidence is demonstrated, construct comparability should never be naively assumed. Therefore, testing for measurement invariance is a prerequisite step for meaningful comparisons across groups (Meredith, 1993).

Different methods have been proposed to evaluate measurement invariance. Multiple-group confirmatory factor analysis (MG-CFA) in the structural equation modeling (SEM) framework is the most popular technique. Differential item functioning (DIF) in the item response theory (IRT) framework is another useful approach.

MG-CFA

In a CFA model, observed variables (e.g., survey items) are linked to latent variable(s) or factor(s) through a linear function. Specifically, the observed response on an item is a linear combination of latent variables, factor loadings, intercept, and residual/error score for that item. The MG-CFA approach compares a hierarchical set of measurement models to evaluate different levels of invariance: (1) configural invariance, (2) weak invariance, (3) strong invariance (Meredith, 1993).

The lower level of invariance is configural invariance, which requires that the model specification (i.e., number of factors and their loading pattern) be same across groups. It investigates whether respondents from different groups employ the same conceptual framework to answer the test or survey items (Vandenberg & Lance, 2000). To test configural invariance using MG-CFA, different groups are examined simultaneously. The same items are forced to load on the same factors across groups, but parameter estimates themselves are allowed to vary in each group. That is, we only need constrain the factor structure (i.e., number of factor(s) and their loading pattern) to be the same across groups. The fit of this configural model provides the baseline value against which all subsequently specified invariance models are compared. It should be noted that if this model exhibits some evidence of misfit, model modifications can be done to improve the fit. A well-fitting baseline model should be established for all groups in order to test the weak and strong invariance.

Weak invariance requires equal factor loadings across groups. A test of the weak invariance using MG-CFA is conducted by examining a model identical to the baseline model except that factor loadings are constrained to be equal across groups and then comparing this more restricted model with the baseline model.

Strong invariance requires not only equal factor loadings but also equal intercepts across groups. Item intercepts can be interpreted as systematic biases in the response of a group to an item. As a result, the observed group mean response can be systematically higher or lower than one would expect based on the group's latent trait and factor loadings. If strong invariance does not hold, the difference on the observed scores between two groups may not reflect their difference on the measured construct. To test the strong invariance, in a MG-CFA model, both the factor loadings and intercepts are constrained to be same for different groups, and this model is compared to the previous model for weak invariance.

In sum, in order to conduct a comparison of different groups on a construct and interpret it meaningfully, the above-mentioned three levels of invariance are required. Only if all three types of invariance are supported can we confidently carry out comparisons based on group means. It also should be noted that researchers are left to choose additional tests to best suit their needs. Possible follow-up invariance tests could include tests of error variances and covariances, and tests of factor variances and covariances.

Differential Item Functioning (DIF)

Different from the SEM framework, the IRT framework uses a log-linear, rather than a linear, model to describe the relationship between observed item responses and the underlying latent variable. The probability of a person choosing a particular response category on an item is determined by this person's latent trait and the property of this item. The item property is typically described using two parameters – item difficulty and item discrimination (Embertson & Reise, 2000). The non-linear function is called item characteristic curve (ICC).

Under the IRT framework, if an item does not have the same relationship to a latent variable across different groups, it is said that this item shows differential item functioning

(DIF). More specifically, an item displays DIF if respondents from two groups who are equal in level on the latent trait do not have the same probability of endorsing a response. If the ICCs are identical for each group, or very close to identical, it can be said that the item does not display DIF. If, however, the ICCs are significantly different from one another across groups, then the item is said to show DIF. When many items on a scale display DIF, that scale will not be invariant across groups.

It is worthy to note that DIF has two types: uniform and non-uniform DIF. An item with uniform DIF indicates that this item favor one group consistently across all trait levels. It occurs when the ICCs from two groups do not cross but have very large area. An item with non-uniform DIF indicates that this item favor one group for some levels of trait, but favor another group for the remaining levels of trait. It occurs when the ICCs from two groups cross with each other.

Many DIF-detection procedures have been developed. Zumbo's (1999) ordinal logistic regression (OLR) approach is a very flexible one. In this OLS model, the item response is the dependent variable, and the grouping variable, total scale score, and a group-by-total interaction are independent variables. The total score is entered into the model first. Next, the grouping variable is entered to detect the uniform DIF. Finally, the interaction term is entered to examine the non-uniform DIF.

The East Asian Social Survey (EASS) is a collaborative social survey framework among four East Asian societies (China, Japan, Korea, Taiwai). Since 2006, the four societies have been working together to develop a set of topical modules, construct the questionnaires, conduct the module surveys every two years in each society. Thus far, EASS has the modules of family

(2006), globalization and culture (2008), and social health (2010). The 2012 module of network social capital is under construction.

One of the main purposes of EASS is to provide a major source of data for comparing these four societies. In order to make the surveys comparable, the developers should pay serious attention to many aspects, such as exact translations of the items into different languages, comparable sampling procedures, and similar data collection techniques. Although these steps are necessary for establishing measurement invariance across different societies, they cannot guarantee invariance. Therefore, once we have data collected, measurement invariance of constructs should be assessed or tested based on the data. The purpose of this study is to apply the MG-CFA and DIF approaches to assess the measurement invariance of the marriage scale on the 2006 EASS family module and also compare the results.

Methods

Measure

The marriage scale on the 2006 EASS family module includes 7 items (Table 1). Each item states one aspect related to marriage and asks respondents "to what extent do you agree or disagree with this statement". The response scale is likert-type with 7 categories: strongly agree, fairly agree, somewhat degree, neither agree nor disagree, somewhat disagree, fairly disagree, and strongly disagree. Since the overall attitude to marriage can be measured by summing up the item scores, these seven items should be in the same direction. As seen from Table 1, items 2, 5, and 7 have the different direction from other items. So we recoded these three items before any analysis was conducted.

Table 1:

Item	Content
1	Husband should be older than wife
2	It is not necessary to have children in marriage

3	Married men are generally happier than unmarried men
4	Married women are generally happier than unmarried women
5	It is all right for a couple to live together without intending to get married
6	People who want to divorce must wait until children are grown up
7	Divorce is usually the best solution when a couple can't seem to work out their marriage

Sample

The total sample included 9045 respondents, 3208 from China, 2130 from Japan, 1605 from Korea, and 2102 from Taiwan. The MG-CFA and DIF analyses require the data without missing, so the respondents who had missing on any of the 7 items were excluded from the analysis. The valid sample size is 3208 for China, 2080 for Japan, 1585 for Korea, and 2089 from Taiwan.

Analysis

The MG-CFA analyses were conducted in the LISREL/SIMPLIS 8.8 program (Jöreskog & Sörbom, 1996) using the maximum likelihood estimation. This EASS marriage scale was designed to measure one construct "attitude to marriage", so the hypothesized model is one-factor model with all 7 items loading on it. This initial one-factor seven-indicator CFA model was first fit to each group data separately using the covariance structure and the goodness-of-fit was evaluated. A modification was done to improve the fit in order to find a well-fitting baseline model. There are many kinds of fit statistics for assessing the model fit. Research had shown the commonly-used chi-squared fit statistic is a function of the sample size. It rejects the null hypothesis with two much power if the sample size is large. It may reject trivial model-data differences and tends to lose practical usefulness when used as the sole decision rule. Therefore, a variety of fit indices have been proposed to accommodate the problems with sample size and model complexity. Two powerful indices are comparative fit index (CFI, Bentler, 1990) and

Root Mean Square Error of Approximation (RMSEA, Steiger, 1989). For a CFA model, CFA≥0.95 indicate a good fit. RMSEA≤ 0.06 suggests a good fit and RMSEA≤ 0.08 indicating reasonable fit (Hu & Bentler, 1999).

Next, the welling-fitting model was fit simultaneously to test configural invariance for each of six pairs of groups: China vs. Japan, China vs. Korea, China vs. Taiwan, Japan vs. Korea, Japan vs. Taiwan, and Korea vs. Taiwan. The mean and covariance structure was used. For this MG-CFA model, RMSEA≤0.06 and CFI≥0.95 were used as the criteria to indicate a good fit.

If the configural invariance holds for the data, the next-level weak invariance is tested. The same model was fit simultaneously to each pair of groups and the factor loadings were constrained to be the same across the two groups. This constrained model was compared to the configural baseline model to assess the weak invariance. If the data meets the weak invariance, testing for strong invariance is followed. In this multiple CFA model, both the intercepts and factor loadings were constrained to be the same across two groups. This more constrained model was compared with the weak invariance model.

As discussed above, to test weak or strong invariance, the more restricted models are compared with the previous less restricted models. It is common practice to use change in chi-squared test statistics $\Delta\chi^2$ between two models to evaluate if the more restricted model fit the data better than the less restricted. However, Cheung and Rensvold (2002) showed that, like the general χ^2 test, $\Delta\chi^2$ test is also susceptible to sample size and/or model complexity and has less value in making practical decisions about measurement invariance. They also showed that most of the fit indices (e.g., CFI) were susceptible to model complexity for the MG-CFA nested models and should not be trusted as the sole criterion in making decisions about MI. Thus, there

has been a trend toward using difference between the CFI values (ΔCFI) as a more practical way to determine the MI. The ΔCFI is calculated as the difference between the CFI of the MI model being tested with that of a one-level less constrained MI model. Cheung and Rensvold (2002) suggested that an absolute value of ΔCFI not exceeding 0.01($|\Delta CFI| \leq 0.01$) indicates the MI.

The DIF analysis was conducted in SPSS using Zumbo's OLR program. For each pair of groups, three steps were performed. In Step 1, a logistic regression model with the total scale score was estimated. In Step 2, the group indicator variable was entered into the model to detect the uniform DIF. In Step 3, the interaction term was added in order to detect the non-uniform DIF. In each step, the chi-squared test statistic and effect size R^2 were calculated for each item. The chi-square change and R^2-change from Step 1 to Step 2 were used to flag items with uniform DIF, and the chi-square change and R^2-change from Step 2 to Step 3 were used to flag items with non-uniform DIF.Based on recommendations made by Zumbo (1999) and Jodoin and Gierl (2001), when the two degree of freedom chi-squared test for DIF must have a *p* value less than 0.01, and the effect size R^2-change (i.e., R^2 due to DIF) must surpass 0.035, an item is said to exhibit DIF.

Results

MG-CFA

The theoretical one-factor seven-indicator FA model for the marriage scale did not fit each group data well. The values of CFI and RMSEA were 0.80 and 0.12 for China, 0.88 and 0.11 for Japan, 0.82 and 0.12 for Korea, 0.75 and 0.10 for Taiwan. They all did not meet the good-fit criteria (CFI>0.95 and RMSEA<0.06). A review of the modification indices revealed two error covariances (tem3 and Item4, Item 2 and Item5) to be markedly misspecified for all four groups. Given the obvious overlap of content between Item3 and Item4, and possible content overlap between Item2 and Item5, the initial model was respecified and reestimated with

two error convariances included. This modification resulted in a big improvement for model fit. The CFI and RMSEA indices both showed an excellent fit of the modified model to each group data. Specifically, the CFI value was 0.96 for China, 0.97 for Japan, 0.98 for Korea, and 0.98 for Taiwan, and the RMSEA value was 0.057 for China, 0.056 for Japan, 0.037 for Korea, and 0.035 for Taiwan. Since the modification indices for this modified model did not provide further clear evidence of poorly specified parameters, this model was treated as the well-fitting baseline model for each group.

As discussed earlier, the first multigroup test for invariance is to test configural invariance. The configural model simply incorporates the baseline models for both groups and allows for their simultaneous analyses. Table 2 presents the sample sizes and the results for each-level invariance test. As can be seen, the goodness-of-fit statistics related to the testing of the configural model yielded a very good fit for each pair of comparisons. The RMSEA ranged from 0.036 to 0.056 which marginally met the criteria (RMSEA<0.06). The CFI index was 0.97 for four pairs of groups, and 0.98 for the other two pairs of groups, and they were all greater than the cut-off value of 0.95. These results indicated that the configural model represented the data fairly well. Thus, all six comparisons passed the weak invariance test. In other words, it can be assumed that those four groups had similar factor structure (one factor and the similar pattern of loadings).

Table 2: Fit Indices for Three Invariance Models

	CN vs. JP	CN vs. KR	CN vs. TW	JP vs. KR	JP vs. TW	KR vs. TW
N	5288	4793	5297	3665	4169	3674
Configural Invariance Model						
RMSEA	0.056	0.051	0.049	0.049	0.047	0.036
CFI	0.97	0.97	0.97	0.98	0.97	0.98
Weak Invariance Model						
RMSEA	0.066	0.055	0.057	0.049	0.047	0.044
CFI	0.95	0.96	0.95	0.97	0.97	0.96

ΔCFI	**-0.02**	-0.01	**-0.02**	-0.01	-0.00	**-0.02**
Strong Invariance Model						
RMSEA	0.12	0.16	0.10	0.16	0.11	0.11
CFI	0.68	0.39	0.77	0.50	0.76	0.66
ΔCFI	**-0.27**	**-0.60**	**-0.20**	**-0.47**	**-0.21**	**-0.30**

The next test was for the weak invariance. A MG-CFA model was fit to each pair of groups with the factor loadings were estimated only for the first group and then constrained equal for the second group. For example, for China vs. Japan, the factor loadings values for Japan were fixed to be same as the values estimated for China. The fit statistics results are shown in Table 2. Compared with the configural model, this constrained model for weak invariance revealed very negligible decrement in overall fit. All CFI values were greater than 0.95, and all RMSEA values except China vs. Japan were less than 0.06. The MG-CFA model for China and Japan had a RMSEA value of 0.66, which was only slightly higher than the cut-off value. Overall, the weak invariance model fit each pair of groups very well.

To test if the weak invariance was met or not, the values of ΔCFI were calculated and also reported in the table. Using the $|\Delta CFI| \leq -0.01$ as the decision rule, three comparisons (China vs. Korea, Japan vs. Korea, Japan vs. Taiwan) passed the weak invariance test with the ΔCFI values were all equal to -0.01, but the remaining comparisons (China vs. Japan, China vs. Taiwan, Korea vs. Taiwan) failed. The modification indices results for these three comparisons further showed the loadings of Item 7 were probably different for China and Japan, the loadings of Items1, 6 and 7 were different between China and Taiwan, and the loadings of Items 6 and 7 were different between Korea and Taiwan. It may be worthy to note that given that the CFI and RMSEA values for this weak invariance model were all good and the ΔCFI value of -0.02 was just slightly exceeding the criterion, this finding probably represents statistical significance, but

not necessarily practical significance (i.e., a non-negligible effect size). Therefore, it may be appropriate to say that the marriage scale essentially meet the weak invariance.

For testing the strong invariance, a MG-CFA model was fit to each pair of groups with the factor loadings and intercepts were estimated only for the first group and then constrained equal for the second group. Table 2 also shows the results for the strong invariance test. As can be seen, the absolute values of ΔCFI were all much higher than 0.01, indicating that all six comparisons failed the strong invariance test. An examination of the modification indicates showed some misfit. For example, Item1 had different intercepts for China and Japan, also between Japan and Taiwan; Item 2 had different intercepts for China and Korea, also for Japan and Korea.

In summary, the results from the MG-CFA analyses indicated that the EASS marriage scale met the configural invariance, and essentially met the weak invariance, but did not meet the strong invariance. The configural invariance provided the evidence that the one-factor model structure was invariant for all groups. The factor loading invariance indicated that the strength of the relationship between the observed and latent variables were essentially the same. Thus, the construct "attitude to marriage" had similar meaning for different groups and the marriage scores can be interpreted to have similar meaning for different groups. The lack of the intercept invariance implied that the bias in item response was systematically different across groups. Therefore, the cross-group differences in the means of the observed items did not reflect their differences in the construct (or factor) only. So it will be inappropriate to compare their mean scores.

DIF

The DIF results based on the ordinal logistic regression method were presented in Tables 3.1 to 3.6 for all group comparisons. They include the chi-square test statistics and R2 for each step and the final DIF chi-squared test and DIF R2. DIF chi-squared test statistics were the chi-square difference between Step 3 and Step 1. DIF R2 was the R2 difference between Step 3 and Step 1. Adopting the criteria of DIF R2>0.035, Item1 performed differently between China and Japan (DIF R2=0.082), between China and Korea (DIF R2=0.038), between Japan and Taiwan (DIF R2=0.105), and also between Korea and Taiwan (DIF R2=0.061). In addition, Item 6 appeared to perform not similarly between Japan and Taiwan (DIF R2=0.037). Moreover, these two items all showed uniform DIF.

Table 3.1: DIF results for China vs. Japan:

	Step1	Step2	Step3	DIF $\chi^2(2)$	DIF R^2
Item1	Chi=1081.45 R2=0.209	Chi=1599.00 R2=0.291	Chi=1599.00 R2=0.291	517.55 P=0.0000	Uniform DIF 0.082>0.035
Item2	Chi=1891.45 R2=0.309	Chi=1891.46 R2=0.309	Chi=1893.62 R2=0.310	2.17 P=0.3379	0.001
Item3	Chi=1953.52 R2=0.341	Chi=1970.76 R2=0.343	Chi=1971.05 R2=0.343	17.53 P=0.0002	0.002
Item4	Chi=1850.12 R2=0.328	Chi=1861.06 R2=0.330	Chi=1862.54 R2=0.329	12.42 P=0.0020	0.001
Item5	Chi=1637.41 R2=0.275	Chi=1638.81 R2=0.275	Chi=1639.08 R2=0.275	1.67 P=0.4339	0.000
Item6	Chi=1187.40 R2=0.212	Chi=1224.70 R2=0.218	Chi=1229.51 R2=0.219	42.11 P=0.0000	0.007
Item7	Chi=548.11 R2=0.105	Chi=621.54 R2=0.116	Chi=621.68 R2=0.116	73.57 P=0.0000	0.011

Table 3.2: DIF results for China vs. Korea:

	Step1	Step2	Step3	DIF $\chi^2(2)$	DIF R^2
Item1	Chi=884.45 R2=0.182	Chi=1106.15 R2=0.219	Chi=1106.66 R2=0.220	222.21 P=0.0000	Uniform DIF 0.038>0.035
Item2	Chi=2053.60 R2=0.364	Chi=2192.68 R2=0.383	Chi=2199.01 R2=0.380	145.41 P=0.0000	0.016
Item3	Chi=2205.05 R2=0.391	Chi=2226.40 R2=0.394	Chi=2238.25 R2=0.399	33.20 P=0.0000	0.008
Item4	Chi=1799.571 R2=0.3350	Chi=1799.802 R2=0.3352	Chi=1805.829 R2=0.3375	6.26 P=0.0437	0.003

Item5	Chi=1639.651 R2=0.3046	Chi=1643.569 R2=0.3053	Chi=1643.965 R2=0.3046	4.32 P=0.1153	0.000
Item6	Chi=1473.58 R2=0.279	Chi=1478.61 R2=0.280	Chi=1500.15 R2=0.288	26.57 P=0.0000	0.009
Item7	Chi=811.08 R2=0.169	Chi=813.59 R2=0.170	Chi=847.84 R2=0.180	36.76 P=0.0000	0.011

Table 3.3: DIF results for China vs. Taiwan:

	Step1	Step2	Step3	DIF $\chi^2(2)$	DIF R^2
Item1	Chi=970.59 R2=0.179	Chi=1030.51 R2=0.188	Chi=1030.55 R2=0.188	59.96 P=0.0000	0.009
Item2	Chi=1792.38 R2=0.292	Chi=1803.29 R2=0.294	Chi=1804.81 R2=0.295	12.43 P=0.0020	0.003
Item3	Chi=1683.92 R2=0.288	Chi=1697.02 R2=0.290	Chi=1702.79 R2=0.292	18.87 P=0.0001	0.004
Item4	Chi=1621.59 R2=0.281	Chi=1745.68 R2=0.297	Chi=1753.37 R2=0.300	131.78 P=0.0000	0.019
Item5	Chi=1799.60 R2=0.295	Chi=1817.75 R2=0.298	Chi=1838.82 R2=0.304	39.22 P=0.0000	0.009
Item6	Chi=839.35 R2=0.155	Chi=949.63 R2=0.170	Chi=956.19 R2=0.169	116.84 P=0.0000	0.014
Item7	Chi=546.13 R2=0.105	Chi=597.05 R2=0.115	Chi=616.68 R2=0.122	70.55 P=0.0000	0.017

Table 3.4: DIF results for Japan vs. Korea:

	Step1	Step2	Step3	DIF $\chi^2(2)$	DIF R^2
Item1	Chi=1390.07 R2=0.359	Chi=1392.86 R2=0.360	Chi=1393.18 R2=0.360	3.11 P=0.2111	0.001
Item2	Chi=2140.09 R2=0.467	Chi=2298.06 R2=0.489	Chi=2301.64 R2=0.487	161.55 P=0.0000	0.020
Item3	Chi=2331.78 R2=0.506	Chi=2332.45 R2=0.506	Chi=2333.49 R2=0.507	1.71 P=0.4253	0.001
Item4	Chi=1996.57 R2=0.454	Chi=2010.08 R2=0.457	Chi=2010.10 R2=0.457	13.53 P=0.0012	0.003
Item5	Chi=1648.69 R2=0.380	Chi=1660.33 R2=0.382	Chi=1660.47 R2=0.382	11.78 P=0.0028	0.002
Item6	Chi=1258.40 R2=0.303	Chi=1293.24 R2=0.310	Chi=1330.81 R2=0.322	72.41 P=0.0000	0.019
Item7	Chi=813.92 R2=0.213	Chi=824.09 R2=0.215	Chi=864.91 R2=0.231	50.99 P=0.0000	0.018

Table 3.5: DIF results for Japan vs. Taiwan:

	Step1	Step2	Step3	DIF $\chi^2(2)$	DIF R^2
Item1	Chi=917.73 R2=0.222	Chi=1478.11 R2=0.327	Chi=1479.25 R2=0.327	561.52 P=0.0000	0.105

Item2	Chi=1882.94 R2=0.373	Chi=1893.09 R2=0.375	Chi=1900.52 R2=0.378	17.58 P=0.0002	0.005
Item3	Chi=1843.27 R2=0.382	Chi=1902.26 R2=0.390	Chi=1904.59 R2=0.391	61.32 P=0.0000	0.009
Item4	Chi=1840.51 R2=0.384	Chi=2040.20 R2=0.413	Chi=2041.21 R2=0.414	200.70 P=0.0000	0.030
Item5	Chi=1760.53 R2=0.352	Chi=1791.03 R2=0.358	Chi=1820.61 R2=0.366	60.08 P=0.0000	0.014
Item6	Chi=746.91 R2=0.172	Chi=954.64 R2=0.210	Chi=955.57 R2=0.209	208.66 P=0.0000	0.037
Item7	Chi=612.56 R2=0.142	Chi=616.68 R2=0.143	Chi=638.30 R2=0.154	25.74 P=0.0000	0.012

Table 3.6: DIF results for Korea vs. Taiwan:

	Step1	Step2	Step3	DIF $\chi^2(2)$	DIF R^2
Item1	Chi=655.06 R2=0.169	Chi=940.32 R2=0.230	Chi=941.45 R2=0.230	286.39 P=0.0000	0.061
Item2	Chi=1776.23 R2=0.394	Chi=1822.70 R2=0.401	Chi=1830.50 R2=0.400	54.27 P=0.0000	0.006
Item3	Chi=1876.19 R2=0.408	Chi=1903.51 R2=0.412	Chi=1906.14 R2=0.413	29.95 P=0.0000	0.005
Item4	Chi=1784.84 R2=0.394	Chi=1826.26 R2=0.401	Chi=1826.32 R2=0.401	41.42 P=0.0000	0.007
Item5	Chi=1520.41 R2=0.349	Chi=1527.35 R2=0.350	Chi=1548.02 R2=0.351	27.61 P=0.0000	0.002
Item6	Chi=995.72 R2=0.239	Chi=1029.48 R2=0.245	Chi=1071.63 R2=0.256	75.91 P=0.0000	0.017
Item7	Chi=634.47 R2=0.161	Chi=660.21 R2=0.167	Chi=663.40 R2=0.168	28.93 P=0.0000	0.007

References

Bryne, B., Shavelson, R. & Muthen, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*,456-466.

Cheung, G. & Rensvold, R. (2002). Evaluating goodness of fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal,* 9, 233-245.

Embretson, S. E. & Reise, S. P. (2000). Item Response Theory for psychologists. Mahwah, NJ: Erlbaum.

Hu, L-T, & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1-55.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating power and Type I error rates using an effect size with the Logistic Regression procedure for DIF. *Applied Measurement in Education, 14,* 329-349.

Jöreskog, K.G., & Sörbom, D. (1996). *LISREL 8: User's reference guide*. Chicago, IL: Scientific Software International.

Meredith, W. (1993). Measurment invariance, factor analysis and factorial invariance. *Psychometrika, 58*, 525-543.

Steiger, J. H. (1989). *EzPATH: Causal modeling*. Evanston, IL: SYSTAT.

Vandenberg, R.J., & Lance, C.E. (2000). A review and synthesis of the measurement equivalence literature: Suggestions, practices and recommendations for organizational research. *Organizational Research Methods, 3*, 4-70.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores.* Ottawa ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.