한국심리학회지 : 일반
Korean Journal of Psychology
2001. Vol. 21, No. 1, 91-116

Equivalence of Constructs Measured by Two Different Language Versions of 16PF

Wonsook Sohn†

Ewha Womans University

Psychological tests such as attitudinal or personality tests are often adapted or translated for use in many languages and cultures. In this situation, the construct equivalence is a fundamental issue because having equivalence of measures is a prerequisite for obtaining valid comparisons across cultural groups. To completely accept results from quantitative comparisons across groups, evidence for construct equivalence must be established. The purpose of this study is to describe a general approach for empirically investigating the construct equivalence of personality tests. The Sixteen Personality Factor (16PF) Questionnaire was administered in English to 844 American college students and in Korean to 538 Korean college students. Two statistical methods were utilized in a complementary way: (a) Principal Component Analysis, and (b) Multi-group Confirmatory Factor analysis. The results showed that the extraversion scale of the 16PF had the same factor structure across the two groups only in that it had the same number of factors and the same items on each factor. However, the results indicated that the factor loadings and error variances were not equivalent across the two groups. Practical and theoretical implications are discussed.

Keywords : construct equivalence, cross-cultural research, personality tests, multi-group confirmatory factor analysis (CFA)

이 논문은 저자의 박사논문 중 일부분을 발췌하여, 수정보완 한 것입니다.

† 교신저자 : 손 원 숙, E-mail julie69@hanmail.net/wsohn@ewha.ac.kr

In educational and psychological testing, we have commonly found numerous examples of translating or adapting tests for use in many languages and cultures. These trends are expected to continue in the future due to an increase in international testing, more demand for licensure exams in multiple languages, and a growing interest in cross-cultural research(Hambleton, 1993; 1994). There are many good reasons why researchers would be interested in adapting the tests for use in cross-cultural research. First of all, cross-cultural researchers are basically interested in evaluating the universality or specificity of a psychological concept or in explaining similarities and differences in the behaviors of people from different cultures(Poortinga & Malpass, 1986). One of the earliest examples is the translation of the Binet-Simon Intelligence Scale for children from French into English in 1911. In addition, many personality tests, such as the Minnesota Multiphasic Personality Inventory (MMPI) and the Sixteen Personality Factor(16PF) Questionnaire, have been translated into various languages for cross-cultural use (Butcher, 1996; Philip, 1972). Another reason to adapt the tests is that some researchers in educational testing are involved with comparing the educational achievement of students in different countries. As a recent example, the International Association for the Evaluation of Educational Achievement(IEA) conducted the third and fourth International Mathematics and Science Study(TIMSS) in 1995 and 1999(Hambleton, 1994). The practice of test adaptation can also happen for economic reasons. Sometimes researchers want to assess a particular construct, but the local measures of the construct do not exist in their culture. In this case, it might be less expensive and faster to translate a test than to construct a new instrument.

Even though the practice of test adaptation or translation for use in cross-cultural research has been growing, there are several complex and challenging methodological issues that researchers should deal with(Church & Lonner, 1998; Hambleton & Kanjee, 1995): (a) methods and procedures for adapting tests while focusing on establishing the equivalence of scores(Hambleton, 1993; van de Vijver & Poortinga, 1991), (b) ways of interpreting and using cross-cultural and cross-national data(Hambleton & Kanjee, 1995; Poortinga & Malpass, 1986), and (c) the development and use of guidelines for adapting tests(Geisinger, 1992, 1994; Hambleton, 1994). The first issue is related to a fundamental concern in cross-cultural research, that of establishing the equivalence of the measures across cultures. Equivalence refers to the measurement level at which scores can be compared across different groups(Cunningham, 1991). There are various levels or types of equivalence proposed by many cross-cultural researchers. The researchers also argue that the cross-cultural equivalence should be checked in multiple aspects to obtain quantitative comparisons between groups(Drasgow, 1984; Hui & Triandis, 1985). One of the levels of equivalence is construct equivalence, which means that a test measures same constructs in two or more cultural groups or that the same latent trait underlies test performance across groups. In any cross-cultural study to compare groups from two or more language groups, the constructs measured by the tests must be equivalent. Thus, evidence for construct equivalence must be established before results from quantitative comparisons across different groups can be completely accepted. Additionally, Gierl(2001) observed that construct equivalence has implications for the valid use of measurement procedures such as item response theory(IRT). For example, differential item functioning techniques are based on the assumption that a single latent trait underlies test performance.

Until recently, many cross-cultural researchers relied only on judgmental procedures(e.g., decentering or back-translation method) to ensure the equivalence of the tests across cultures. The back-translation method is the most popular of the judgmental methods (Brislin, 1970; Hambleton, 1993), and it was successfully used as an initial check of translation quality in several empirical studies(Hulin & Mayer, 1986; Hulin, Drasgow, & Komocar, 1982). However, it has been found to be insufficient to establish cross-cultural equivalence, and there is a need for statistical methods to complement the judgmental methods(Ellis, Minsel, & Becker, 1989; Hulin, et al., 1982). As also seen in the International Test Commissions(ITC) Guidelines for adapting educational and psychological tests(summarized by Hambleton, 1994), the instrument developers are recommended to use appropriate statistical techniques, in addition to the systematic judgmental methods, in establishing the equivalence of the different versions of the test.

Of interest in the present study is the extent to which constructs are measured equivalently across different cultural groups. Confirmatory factor analysis(CFA) is adopted as an approach to assessing the construct equivalence of test scores. Equivalent constructs are hypothesized to underlie scores on two different language versions of tests. Several models of construct equivalence, differing in their assumptions about factor patterns, factor loadings, and variable uniquenesses, were assessed with multi-group CFA. An examination of the equivalence of constructs measured by the tests should increase scientific understanding of the tests and allow a determination of the extent to which the measures can be used for cross-cultural research with no loss of information about the underlying construct.

The main purpose of this study is to describe an approach to empirically assessing the cross-cultural equivalence of personality tests across cultures and languages. Specifically, the cross-cultural equivalence of the 16PF test was investigated in terms of comparability of constructs, or construct equivalence. The Korean translation of the 16PF test was compared with its original English version. Of the five second-order factors of the 16PF, only the extraversion factor, which has the five sub-scales was examined in this study. To measure the construct equivalence, the present study used both an exploratory method and a confirmatory method: principal components analysis(PCA) and multiple-group confirmatory factor analysis (CFA).

Equivalence

Researchers have discussed the concept of cross-cultural equivalence, which is a prerequisite for score comparisons across cultures. The cross-cultural researchers have not agreed on the levels of equivalence. Rather, they have multiple and ambiguous definitions and utilize multiple and overlapping measurement methods to examine the levels of equivalence(Butcher, 1996; Hui & Triandis, 1985; van de Vijver & Leung, 1997; van de Vijver & Poortinga, 1982). For example, van de Vijver and Poortinga(1982) summarized the differences between the levels of equivalence as follows:

> In sum, conceptual universals refer to molar, theoretical concepts without any reference to measurement scales; functionally equivalent universals are concepts for which empirical referents have been specified and that are measured in quantitatively the same way in each culture; metrically equivalent universals are concepts that have the same metric but not the same scale origin across cultures, and strictly equivalent universals have the same scale with the same origin in each culture(p. 391).

On the other hand, Hui and Triandis(1985) proposed an equivalence model for the cross-cultural comparison of test performance. Their model relates various cross-cultural equivalence assumptions to a specified domain of cross-cultural measurement strategies. The notions of equivalence are explained within the abstraction-concreteness and the universality-cultural specificity continua. Their model accounts for both measurement and relational equivalence(as discussed by Drasgow, 1984) and proposes four types of equivalence: (a) conceptual / functional, (b) construct operationalization, (c) item, and (d) scalar equivalence. They also argue that one should check whether the measures have cross-cultural equivalence in all four aspects in order to obtain informative and quantitative comparisons between the cultures. As mentioned earlier, it is not easy to draw clear boundaries between the levels of equivalence. However, there are two things that most researchers agree about regarding cross-cultural equivalence. One is that establishing more abstract types of equivalence is a prerequisite for considering more specific types (Hui & Triandis, 1985), even though each equivalence is important and complementary to the others. The other is that it is more important to use multiple methods to evaluate equivalence than to determine which particular method measures which kind of equivalence and which method should be used before another method (Malpass & Poortinga, 1986). This study adopts the levels of equivalence proposed by van de Vijver and Leung(1997), where a distinction is made between hierarchically-linked types of equivalence. There are three types of equivalence: (a) construct equivalence, (b) measurement unit equivalence, and (c) full scale, or scalar, equivalence. This study deals with only construct equivalence.

Construct Equivalence

Construct equivalence is a very general term stating that the same psychological construct is measured across all studied groups (Singh, 1994; Sireci, Bastari, & Allalouf, 1998; Sireci, Xing, & Fitzgerald, 1999). This equivalence is also called conceptual equivalence within the framework of the Triandis and Hui model. Some researchers have attempted to evaluate this level of equivalence within the context of the construct validation process proposed by Cronbach and Meehl (1955). These two researchers describe construct validity as the process of forming a nomological network, which is accomplished by correlating variables of interest with measures of other psychological constructs. Thus, this level of equivalence can be established by showing similar patterns of correlations among variables and constructs in both cultures(e.g., convergent or discriminant validity). However, establishing construct equivalence by a nomological network is not simple in practice because the cross-cultural equivalence of other psychological instruments is also required. As pointed out by Drasgow (1984), equivalence of relations with external variables (called relational equivalence) should be studied after the measurement equivalence is examined. On the other hand, the interpretation of the construct equivalence can be related to the "etic" or "emic" position, which are frequently used in cross-cultural psychology. Construct equivalence means the universal or culture-independent validity of the underlying psychological construct, so it can be associated with an "etic" position. In contrast, construct inequivalence implies that an instrument measures different constructs in two different cultural groups or that the concepts of the construct overlap only partially across cultures. This might happen because the constructs are associated with different characteristics across cultural groups. This culture-specific feature can be associated with an "emic" position.

The present study explores only one specific aspect of construct equivalence that is particularly important in cross language versions of a test. If the dimensional structure of a test is found to be consistent across cultures and languages, then evidence that the test is measuring the same construct across cultures and languages is provided (Ben Porath, 1990; van de Vijver & Leung, 1997).

Statistical Assessment of Construct Equivalence

Three main statistical methods are identified in the cross-cultural studies evaluating the construct equivalence of different language versions of a test: (a) exploratory factor analysis(EFA), (b) confirmatory factor analysis(CFA), and (c) multidimensional scaling(MDS). Basically, these methods are used to determine whether the internal dimensional structure of a test is consistent across its different language versions or whether the test is measuring the same construct in all groups. The EFA procedures such as common factor analysis and principal components analysis(PCA) have some limitations in that they test each language group separately and, thus, make it difficult to compare factor loadings across groups. Also, there are no statistical tests or numerical indices to help determine the degree to which structural or construct equivalence holds across groups. On the other hand, the CFA and MDS allow for the evaluation of the test structure simultaneously across multiple groups and the discovery of departures from a common structure across groups. They are somewhat different in that the CFA is confirmatory, while the MDS is exploratory. In practice, Reise, Widaman, & Pugh(1993) and Gierl(2001) support the use of the CFA for testing construct equivalence of a test across groups. The CFA is attractive in this situation because it can handle multiple groups simultaneously and it provides both statistical tests of model fit and descriptive indices of model fit(Jöreskog, 1971). The MDS has an advantage that it does not assume a linear relationship among test items and it is relatively easy to use and interpret (Sireci, Foster, Robin, & Olsen, 1997). Recently, Sireci, et al. (1998) compared several methods of evaluating structural equivalence with real and simulation data and found that the CFA and the MDS were useful procedures.

There are other useful methods utilized for testing construct equivalence such as DIMTEST (Li & Stout, 1995) and Agglomerative hierarchical cluster(HAC) analysis(Roussos, Stout, & Marden, 1997), which can deal with both the polytomous and dichotomous items. These methods also have some limitations in that they are unable to deal with all groups simultaneously, and the DIMTEST needs at least 20 items for each scale or dimension to obtain accurate results. Nevertheless, one empirical study successfully tested the construct equivalence of an achievement test in combination with these two methods and the multi-group CFA (Ryan, Wardrop, Pyo, & Sohn, 2001). This study also supports the findings of the previous studies (Sireci et al., 1998) that a multiple method approach involving exploratory and confirmatory procedures would be helpful to evaluate construct equivalence of tests.

Method

Samples

This study examined data from two groups that differ in culture and language: American and Korean. The Korean sample consisted of 538 college students obtained from two large private universities in Korea. The American sample consisted of 844 college students from two sources: 431 subjects from an experiment subject pool at the University of Illinois at Urbana-Champaign and 413 subjects from a sample collected by the Institute for Personality and Ability Testing(IPAT). In cross-cultural research, sample comparability is a critical issue because non-comparable samples can lead to alternative explanations for any differences in results across the two cultures. Sample comparability was

Table 1. Sample Demographics

		U.S. (*n*=844)			Korea (*n*=538)	
		n	percents		*n*	percents
Gender	Male	224	26.5		207	38.5
	Female	575	68.1		319	59.3
	Missing	45	5.3		12	2.2
Year	Freshman	52	6.2	(12.1)[a]	198	36.8
	Sophomore	109	12.9	(25.3)	121	22.5
	Junior	56	6.6	(13.0)	160	29.7
	Senior	92	10.9	(21.3)	22	4.1
	Graduate	41	4.9	(9.5)		
	Missing	494	58.5		37	6.9
Major	Social Science, Liberal Arts, Education	311	36.9	(76.2)	314	58.4
	Engineering, Sciences	39	4.6	(9.1)	106	19.7
	Business	3	0.4	(0.7)	50	9.3
	Arts, Architecture	40	4.7	(9.3)	50	9.3
	Missing	451	53.4	(4.7)	18	3.3

Note. Some demographic information for the 413 subjects obtained from IPAT was unavailable.

[a] Percents were calculated based on n = 431, after eliminating missing data.

addressed in the present study by selecting respondents from the same occupational group and from large regional universities within each of the countries. In addition to controlling for sample comparability, college students were chosen for this study because they tend to be readily available and are accustomed to taking tests. Also, they usually produce relatively normal-range personality profiles when compared with the American norms (Butcher, 1996). Demographic information for both samples is summarized in Table 1.

Instrument

The Sixteen Personality Factor(16PF) Questionnaire (Cattell & Cattell, 1995) is one of the most widely used personality tests in cross-cultural settings, and it has a long history of testing normal adult personality(Cattell, Eber, & Tatsuoka, 1970). There are many translations available, some with standardization, and there are numerous examples of previous cross-cultural applications such as Abrahams and Mauer(1999), Cattell and Krug(1986), Ellis and Mead(1998), Golden(1978), Tsujioka and Cattell(1965). These researchers who extended the 16PF to other countries, have the identical goal of determining if the same personality structure exists elsewhere. The 16PF was developed by Raymond E. Cattell and his associate at the Institute for Personality and Ability Testing(IPAT) and was first published in 1949. The 16PF Fifth Edition contains 185 items grouped into sixteen primary scales and one validity scale called Impression Management (IM). The sixteen primary scales can be further grouped into five secondary-order dimensions or global factors. The test uses a three-level forced choice response format and can be administered individually or in groups. The administration time is approximately 45 minutes. The primary factors include warmth(Factor A), reasoning(Factor B), emotional stability(Factor C), dominance(Factor E), liveliness(Factor F), rule-consciousness(Factor G), social boldness(Factor H), sensitivity(Factor I), vigilance(Factor L), abstractness(Factor M), privateness (Factor N), apprehension(Factor O), openness to change(Factor Q1), self-reliance(Factor Q2), perfectionism(Factor Q3), and tension(Factor Q4) (See Appendix D). Additionally, the IM scale helps gauge the level of impression management and was designed to measure both intentional positive distortion and self-deception(Cattell & Cattell, 1995). The five global factors are extraversion(A+, F+, H+, N-, Q2-: *Note.* "+" refers high scores while "–" refers low scores), anxiety(C-, L+, O+, Q4+), tough-mindedness (A-, I-, M-, Q1-), independence(E+, H+, L+, Q1+), and self-control(G+, F-, M-, Q3+). Previous research shows that the narrower primary scales are far more useful for prediction than broad global scales, so this study focused on these more unidimensional primary factor scales, which contribute to the extraversion scale(See Table 2).

Table 2. Extraversion Scale (51 items)

Primary Factor	# of Items	Low Sten Score Description	High Sten Score Description
Warmth (A)	11	Reserved, Detached, Critical, Aloof	Warmhearted, Easygoing
Liveliness (F)	10	Sober, Taciturn, Serious	Enthusiastic, Happy-go-lucky
Social Boldness (H)	10	Shy, Timid, Threat-Sensitive	"Thick-skinned", Socially Bold
Privateness (N)	10	Forthright, Unpretentious	Astute, Worldly
Self-Reliance (Q2)	10	Sociably Group Dependent	Self-sufficient, Resourceful

Note. Raw scores are converted into standardized (sten) scores by using the norm table:
Stens are based on a 10-point scale with a mean of 5.5 and a standard deviation of 2.

Procedures

Translating the 16PF

The Korean translation of the 16PF for this study was developed based on a Korean translation that was used in a comparative study of Korean and U.S. gifted children(Shaughnessy & Kang, 1998). Two bilinguals meeting the same criteria as the test translators carefully reviewed the quality of the translation. For this editorial review, the two bilinguals were selected from the graduate students in the areas of Education and Psychology. These two bilingual translators were fluent in both languages, knowledgeable of both cultures, and familiar with both the characteristics and the content measured by the instrument (Geisinger, 1994). This editorial review was accomplished in a group meeting, with individual reviews given by the two translators. To ensure that the translation was conducted appropriately, the translators (a) reviewed the items and reacted in writing, (b) shared their comments with one another, and (c) met to consider the points made by one another and to reconcile any differences of opinion.

Several translation problems arose during the translation process, even though the sentences for the 16PF were relatively simple and easy to translate. The first and most important decision to be made was whether to literally translate the items with little or no content change or to modify them through item substitution to fit local behavior norms. This study tried to maintain the meaning of the items as closely as possible. Therefore, only items considered irrelevant or meaningless in the Korean culture were altered substantially. No items needed drastic cultural adaptation, but a few items required minor content changes. The second problem was idiomatic expressions. These items were translated either nonidiomatically or idiomatically with the most appropriate Korean idiom in an effort to convey the meaning as well as possible. The third problem was passive sentences. It is unusual for Koreans to use the passive voice, so most of the passive sentences were translated as active sentences. The fourth problem was associated with personal and possessive pronouns. Koreans like to use our instead of my, so depending on the context, “I” and “my” were replaced by “we” and “our.”

Recoding of Items

The 16PF test uses a three-level forced choice response format. Except for the items on the reasoning scale(Factor B), the items on the remaining scales have a middle response, “?,” that is used only when neither of the two opposing statements is acceptable. Also, the items have opposing response options with varying anchors, such as true and false, hardly ever and often, or ooperative and assertive. In scoring the items, the keyed response receives two points, the “?” response one point, and the remaining response no points.

A Principal Components Analysis(PCA)

The Big Five construct extraversion is measured using five facets or subscales: (a) warmth (Factor A), (b) liveliness (Factor F), (c) social boldness (Factor H), (d) privateness (Factor N), and (e) self-reliance (Factor Q2). Each of the five facets was examined by sample to determine if it met the assumption of unidimensionality. First, separate inter-item polychoric correlation matrices for the polytomous items were derived, and then a principal components analysis (PCA) was performed on each scale using PRELIS 2.30 (Jöreskog & Sörbom, 1986). A unidimensional test would contain a dominant factor if the eigenvalue for the first factor was appreciably larger than the eigenvalues for the remaining factors.

A Multiple-Group Confirmatory Factor Analysis (CFA)

CFA was used to indicate the extent to which tests measure equivalence constructs across groups. LISREL provides the simultaneous assessment of relationship in a hypothesized model. In CFA, LISREL allows the examination of models that are based on various assumptions about factor loadings, factor intercorrelations, and uniquenessess that are fixed to be equal to specific values or to other parameter estimates, or are allowed to be free to take on any value.

Item parceling.

Item parcels were used as the unit of analysis in the CFA. In general, item parcels are preferred to be factor analyzed because they are more reliable than individual items and produce more stable, easily interpretable factors (Bernstein

& Teng, 1989; Cattell & Cattell, 1995). For example, item parceling tends to produce indicators that are normally distributed, which is a key assumption for maximum likelihood(ML) parameter estimation. It also results in stronger indicators with increased reliability and decreased error variances. It was found that item parceling was useful to detect the correct number of dimensions especially when the data were unidimensional (Egan, Sireci, Swaminathan, & Sweeny, 1998). However, parceling results in a loss of information about individual items, and the resulting parameter estimates are dependent on the items assigned to each parcel(Gierl, 2001). For this study, the use of item parceling was partly supported by the results of PCA that the five subscales were unidimensional for both groups. A total of fifteen parcels were created by summing three to four items on the basis of their descriptive statistics and inter-item correlations. Namely, items that correlated most highly were assigned to the same parcel(Cattell & Cattell, 1995). Therefore each factor or subscale had three indicator variables. Table 3 shows both the items comprising each parcel and the means, the standard deviations, and the values for skewness and kurtosis for the item parcels. As shown in Table 3, item parcels were, in general, skewed or (and) kurtotic in both groups. The tests of multivariate normality were conducted by PRELIS 2.30, which gives tests of zero multivariate skewness and zero multivariate kurtosis (Bollen, 1989; pp.420-425). The results showed that χ^2 values for U.S. and Korea were 238.01(p < .01) and 12.04 (p < .01), respectively, and thus the assumption of multivariate normality cannot be assumed. Because of the nonnormal distribution of item parcels, the asymptotic distribution-free (ADF) or weighted least square(WLS) estimation procedure (Bollen, 1989; Jöreskog & Sörbom, 1993) may be more appropriate for model evaluation than the ML estimation procedure. However, the ML estimation procedure has been found to be relatively robust to the multivariate normal distribution assumption(e.g., Chou, Bentler, & Satorra, 1991). For this study, the results from the WLS estimation procedure were relatively similar to those from the ML estimation in terms of parameter estimates and overall fit indexes (*Note.* The results reported for this study were based on the ML estimation).

Baseline model

As a prerequisite to testing for factorial equivalence, a baseline model was estimated separately for each group. The five-factor model based on each item-parcel covariance matrix was freely estimated using LISREL 8.30(Jöreskog & Sörbom, 1993), while the variances of the factors were set to one. Analyses were performed on within-group covariance matrices rather than within-group correlation matrices(Drasgow & Kanfer, 1985; Jöreskog, 1971). Standardizing the data separately for each of the groups would lead

Table 3. Item Parcels

Parcels	Items	U. S. (n=844)				Korean (n=522)			
		M	SD	Skewness	Kurtosis	M	SD	Skewness	Kurtosis
A1	96,98,159	4.70	1.52	-1.00*	.27*	4.87	1.32	-.98*	.22
A2	31,161,65,129	5.23	2.18	-.52*	-.51*	4.10	2.19	-.04	-.72*
A3	33,63,127,1	4.65	1.59	-1.00*	.51*	3.88	1.53	.001	-.72*
F1	103,39,68	4.15	1.70	-.63*	-.45*	3.96	1.52	-.47*	-.21
F2	6,134,164,37	5.89	2.10	-.81*	-.16	5.14	2.11	-.48*	-.48*
F3	100,70,4	4.23	1.60	-.67*	-.21	2.66	1.66	.11	-.73*
H1	9,73,41,135	4.43	2.89	-.17	-1.40*	4.41	2.92	-.18	-1.40*
H2	137,105,169	3.32	2.24	-.21	-1.40*	3.31	2.07	-.20	-1.20*
H3	71,107,167	3.01	2.16	-.01	-1.31*	2.72	2.01	.23	-1.10*
N1	47,50,143,117	3.73	2.51	.16	-1.00*	4.22	2.20	-.11	-.80*
N2	80,15,84	3.20	2.01	-.13	-1.10*	3.61	1.72	-.40*	-.63*
N3	113,148,18	3.64	1.73	-.24*	-.75*	3.56	1.60	-.29	-.47*
Q21	27,59,89	3.31	1.99	-.20*	-1.10*	3.15	1.88	-.10	-.98*
Q22	121,25,152,56	2.70	2.36	.58*	-.63*	3.42	2.29	.23	-.89*
Q23	92,123,156	1.56	1.75	.92*	-.04	1.04	1.45	1.35*	1.16*

Note. Standard error for skewness and for kurtosis is .08 and .17 for the U.S. group; .11 and .21 for the Korean group.
* $p < .01$

to different rescaling of measured variables within each group and thus it would remove information about variability that is essential for a correct analysis.

Multi-group CFA

Then, a multiple-group CFA was conducted to evaluate the equivalence of the factor structure, factor loadings, and error variances across the U.S. and Korean groups (Jöreskog, 1971). Three different multiple group models were tested using the ML estimation. All models specified the same oblique factor structures but differed in the equality constraints used across the two groups. The first, least restrictive model specified a common five-factor structure underlying the data for both groups. This model is termed the congeneric or essentially tau-equivalence(Lord & Novick, 1968), model of equivalence. The congeneric model implies that equivalent tests load on a shared construct, although the loadings may differ in magnitude. The second model, termed tau equivalence, constrained the factor loadings to be equal across the U.S. and Korean groups. Tau equivalent tests measure the shared constructs to the same degree but are not equally reliable. The third model. termed parallelism, added the constraint that the errors associated with the factor loadings also be equivalent across the two groups. Hence, parallel tests load on the same construct to the same degree and have equal error. Thus observed scores are interchangeable only when the parallel model holds.

Measures of goodness-of-fit

The differences among the three models were examined using five measures of goodness-of-fit: (a) chi-square ratio test (b) χ^2/df ratio, (c) goodness of fit index(GFI), (d) root mean square error of approximation(RMSEA), and (e) root-mean square residual(RMR). The first index is chi-square statistic. The chi-square likelihood ratio test is used as a criterion for assessing the extent to which a proposed model fits the observed data. However, problems associated with the chi-square statistic as a criterion of fit have been noted(see, e.g., Bentler & Bonett, 1980) and in fact have led Jöreskog(1979) to conclude that the decision to accept or reject a model cannot be made on a purely statistical basis. Rather, interpretations of the data should be "based on substantive, theoretical and conceptual considerations (Bollen, 1989)." Thus, so-called practical indices of fit as well as the chi-square statistics are recommended when judging the fit of a model to data(Jöreskog, 1971). The second index is the χ^2/df ratio. A ratio ranging from 1.00 to 5.00 indicates a reasonable fit to the data(Wheaton, Muthen, Alwin, & Summers, 1977). The third index is GFI, which does not depend on sample size explicitly and measures how much better the model fits as compared to no model at all. In general, the larger the values of GFI (i.e., values above .90), the better fitting the model (Bollen, 1989). The fourth index is RMSEA. RMSEA provides a measure of parsimony by assessing the discrepancy per degree of freedom in the model. It takes into account the number of free parameters required in order to achieve a given level of fit. Browne and Cudek(1993) suggested that a value of 0.08 or less for the RMSEA would indicate a reasonable error of approximation and a model with a RMSEA greater than 0.1 would not be acceptable. The fifth index is RMR, which is an average of the squared residuals between the observed and the hypothesized covariances. Generally, a value of 0.05 of the standardized RMR indicates a close fit, and values above 0.08 represent a fairly sizable difference between the covariance reproduced from the factor models and the actual data(Bollen, 1989).

Results

Descriptive Statistics

The descriptive statistics for the Extraversion scale raw scores are given in Table 4. The results of the *t*-test showed that there were significant mean differences between the U.S. and Korean groups in three of the Extraversion scales ($p < .05$). The U.S. sample showed significantly higher mean scores than the Korean sample in scales A and F. On the other hand, the mean score of scale N was significantly higher in the Korean sample. The *t*-test was also performed to examine gender differences within each group on the 16PF scales($p < .05$). Data for the U.S. group showed that in three of the Extraversion scales there were significant gender-related differences (Factors A, N, and Q2). For the Korean data, significant gender-related differences emerged on four scales(Factors A, F, H, and Q2).

Cronbachs alpha coefficients for the Extraversion scales were calculated by gender for both the U.S. and the Korean samples. As shown in Table 5, values ranged from .69 to .87, with a median of .78, for the U.S. sample and from .54 to .87, with a median of .62, for the Korean sample. These values for the U.S. sample were extremely similar to the estimates of internal consistency reliabilities presented by the 16PF Fifth Edition Administrators Manual(ranging from .66 to .86 with a median of .75) (Russell & Karol, 1994). On the other hand, the reliabilities of the Extraversion scales for the Korean sample were generally lower than those for the U.S. sample. In addition, data for both the males and the females showed a similar pattern of reliabilities for the U.S. sample, whereas the males tended to have lower reliabilities than the females for the Korean sample.

Table 4. Extraversion Scale Raw Score Means and Standard Deviations for the U.S. and Korean Samples

	U. S.						Korean					
	Male (*n*=224)		Female (*n*=575)		Total (*n*=844)		Male (*n*=200)		Female (*n*=321)		Total (*n*=522)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
A	14.14	4.33	17.25	3.90	16.31	4.26	12.70	3.49	14.42	3.86	13.77	3.81
F	14.33	4.27	14.28	4.33	14.27	4.30	12.39	3.43	11.35	4.27	11.76	4.00
H	10.38	6.16	10.84	6.53	10.76	6.40	9.75	5.85	10.86	6.25	10.45	6.13
N	11.93	5.01	10.02	5.27	10.56	5.26	11.39	3.75	11.40	4.33	11.37	4.14
Q2	8.50	4.90	7.14	5.14	7.59	5.11	7.11	4.17	7.97	4.73	7.62	4.55

Table 5. Internal Reliabilities[a] of the Extraversions Scales for the U.S. and Korean Samples

Primary Factor (# of items)	U.S.			Korean		
	Male (*n*=224)	Female (*n*=575)	Total (*n*=844)	Male (*n*=200)	Female (*n*=321)	Total (*n*=521)
A (11)	.63	.67	.69	.46	.55	.54
F (10)	.69	.71	.71	.49	.67	.62
H (10)	.85	.88	.87	.85	.88	.87
N (10)	.75	.79	.78	.48	.67	.61
Q2 (10)	.75	.80	.79	.71	.78	.75

Note. [a]Cronbach's Alpha.

Principal Components Analysis (PCA)

Previous research(Ellis & Mead, 1998; Flanagan, Raju, & Haygood, 1998) has supported a high degree of unidimensionality for each of the 16PF scales. Each of the Extraversion scales for the U.S. and Korean groups was separately subjected to PCA using PRELIS 2.30.

Table 6 shows the results of PCA for the Extraversion items. Looking at each of the five scales, the first factors in scales H and Q2 accounted for over 40% of the total variances in both samples. On the other hand, the explained variances by the first factors in scales A, F, and N were relatively lower for the Korean sample when compared to those of the U.S. group. Namely, the first factor in scale A accounted for 34.6% of the total variance for the U.S. group, but 23.4% for the Korean group. The first factor in scale F accounted for 38.8% and 30% of the

Table 6. Principal Components Analysis (PCA) Results for Extraversion Scales by Sample

	U.S. (*n*=844)			Korean (*n*=521)		
	PC_1	PC_2	PC_3	PC_1	PC_2	PC_3
A	3.81[a](34.6)[b]	1.62(14.7)	0.97(8.8)	2.57(23.4)	1.36(12.4)	1.28(11.7)
F	3.88(38.8)	1.15(11.5)	0.92(9.2)	3.00(30.0)	1.37(13.7)	1.04(10.5)
H	6.08(60.8)	0.96(9.6)	0.59(5.9)	5.96(59.6)	1.00(10.0)	0.64(6.4)
N	4.69(46.9)	1.20(12.0)	0.92(9.2)	2.93(29.3)	1.34(13.4)	1.11(11.2)
Q2	4.76(47.6)	1.02(10.2)	0.98(9.9)	4.29(42.9)	1.06(10.6)	0.85(8.5)

Note. [a] Eigenvalues; [b] % Variance; PC=Principal Components.

total variance for the U.S. and Korean groups, respectively. The first factor in scale N accounted for 46.9% of the total variance for the U.S. group and 29.3% for the Korean group. However, these explained variances accounted for by the first factors were generally well above the 20 percent level suggested by Reckase (1979), and thus, the scales appeared to satisfy the unidimensionality assumption across the U.S. and Korean samples for each of the 16PF scales. In addition, the CFA results for the baseline models, which are discussed in the next section, supported this unidimensional assumption for each of the five subscales.

Table 7. Estimated Parameter Matrices and Fit Statistics for the Baseline Models by Sample

Parameter	U.S. (*n*=844)	Korean(*n*=522)
Factor A		
A1	0.63	0.33
A2	0.57	0.48
A3	0.63	0.53
Factor F		
F1	0.66	0.44
F2	0.82	0.81
F3	0.58	0.51
Factor H		
H1	0.92	0.89
H2	0.90	0.89
H3	0.68	0.72
Factor N		
N1	0.91	0.86
N2	0.69	0.42
N3	0.71	0.46
Factor Q2		
Q21	0.69	0.66
Q22	0.85	0.79
Q23	0.81	0.76
Chi-Square	333.05	195.47
Df	80	80
Chi-Square/Df	4.16	2.44
GFI	0.95	0.95
RMSEA	0.061	0.053
RMR	0.041	0.042

Note. GFI = goodness-of-fit index;
RMSEA = root mean square error of approximation;
RMR = (Standardized) root mean square residual.

A Multiple-Group CFA

As a preliminary step to the multiple-group CFA, the five-factor baseline model for each group was first tested based on each item parcel covariance matrix. Although the chi-square values were statistically significant for the U.S. and Korean samples, GFI, RMSEA and RMR indicated an acceptable model fit. Also, the χ^2/df ratio for each group, which is 4.16 for U.S. and 2.44 for Korea, seemed to indicate a reasonable fit to the data(Wheaton, et al., 1977). Looking at Table 7, the value of chi-square with 80 degrees of freedom was 333.05 with the GFI at 0.95, the RMSEA at 0.06, and the RMR at 0.04 for the U.S. sample. On the other hand, the value of chi-square with 80 degrees of freedom was 195.47, with the GFI at 0.95, the RMSEA at 0.05, and the RMR at 0.04 for the Korean

Table 8. Factor Correlations for the Baseline Models by Sample

	Factor A	Factor F	Factor H	Factor N	Factor Q2
Factor A		0.48	0.37	-0.65	-0.65
Factor F	0.74		0.56	-0.41	-0.75
Factor H	0.74	0.47		-0.47	-0.44
Factor N	-0.71	-0.44	-0.62		0.50
Factor Q2	-0.67	-0.71	-0.29	0.51	

Note. Coefficients for the U.S. group are presented above the diagonal; Coefficients for the Korean group are presented below the diagonal.

sample(See also Figures 1 and 2). Considering the χ^2/df ratios and the practical indices of fit such as the GFI, RMSEA and RMR, there was evidence to suggest that the five-factor model for the extraversion scale provided acceptable fit to both groups. As shown in Table 7, there was a fair degree of consistency across groups with respect to high and low factor loadings, except for item parcels A1, A2, F1, N2, and N3, which showed fairly lower factor loadings for the Korean sample. This might be partly due to the lower reliabilities of scales A, F, and N for the Korean data. Examinations of the factor inter- correlations(See Table 8) showed that their patterns were consistent across groups. However, the correlations between Factors A and F(r = 0.48) and between Factors A and H(r = 0.37) were relatively low in the U.S. sample. For the Korean sample, these correlations were larger: r = 0.74 for Factors A and F and r = 0.74 for Factors A and H.

Using a multiple group CFA (Jöreskog, 1971), factor analysis models were compared both with and without constraints across the two groups to test the equality of the factor structures, the factor loadings, and the error variances for the U.S. and Korean samples. Namely, three nested models were sequentially tested by equating the number of factors(model 1), factor loadings (model 2), and error variances(model 3). To decide which models would be best, chi-squared differences were used, and the results of these tests for invariant models are summarized in Table 9. The differences between models 1 and 2 ($\Delta\chi2$= 56.04, Δdf = 15, $p < .01$) and models 2 and 3 ($\Delta\chi2$= 143.1, Δdf = 15, $p < .01$) were statistically significant, indicating that the factor loadings and the error variances were not invariant across groups. Given practical considerations, the congeneric equivalent model (Model 1) provided the closest fit to the data. As can be seen in Table 7, estimated factor loadings especially for factors A, F, and N in the Korean group were substantially smaller than the

Table 9. Tests for Invariant Models Between the U.S. and Korean Samples

Models	Chi-square	DF	GFI	RMSEA	RMR
Model #1 Equated Number of Factors	528.51	160	0.95	0.058	0.042
Model #2 Equated Number of Factors Equated Factor Loadings	584.55	175	0.94	0.059	0.091
Model #3 Equated Number of Factors Equated Number of Loadings Equated Errors	727.65	190	0.93	0.064	0.077

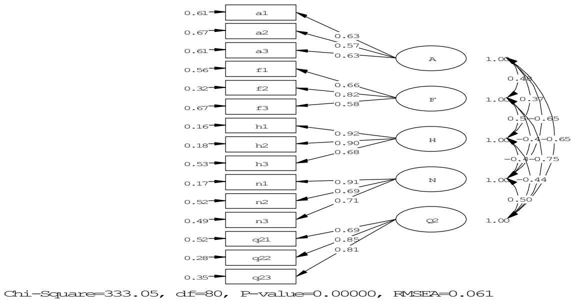
Note. GFI = Goodness-of-fit index; RMSEA = Root mean square error of approximation; RMR=(Standardized) root mean square residual; ΔX^2(Model#1-Model#2) = 56.04, Δdf = 15, $p < .01$; ΔX^2(Model#2-Model#3) = 143.1, Δdf = 15, $p < .01$

corresponding loadings in the U. S. group.

As a result, LISREL multi-group CFA support the conclusion that tests loaded on their hypothesized shared constructs. Namely, only the number of factors was invariant across the U.S. and Korean groups. Invariance of loadings across the groups also provided adequate practical fit, though not superior to the less restrictive congeneric model. Again, observed scores on tests would be expected to differ, and tests may contain differential specific variance or reliabilities across the two groups.

Conclusion and Discussion

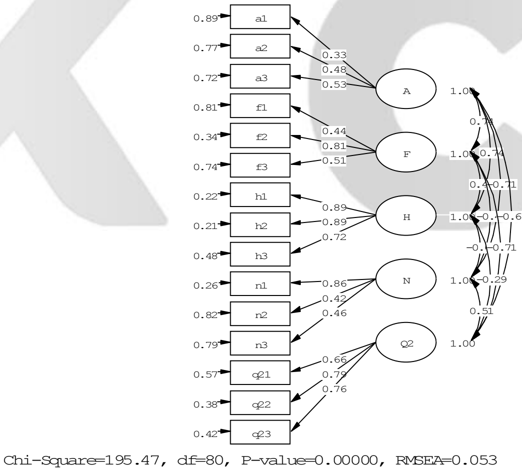
This study explored the equivalence of constructs measured by the English version of 16PF test with the constructs underlying scores on the Korean 16PF. To test the construct equivalence of 16PF across two different cultures and languages, two statistical methods were utilized: (a) a PCA, and (b) a multi-group CFA. Results from principal components analysis(PCA) partly supported the assumption of unidimensionality for each scale of extraversion separately for each group. A comparison of the results between the groups showed that the eigenvalues for the first factors in scales A, F, and N tended to be relatively lower for the Korean group than for the U.S. group. Nevertheless, the eigenvalues for the first factors across the groups were appreciably larger than the eigenvalues for the second factors. Thus, it can be concluded that a single latent trait underlay test performance on each of the five scales across groups. Results from confirmatory factor analysis(CFA) also suported the unidimensional



Chi-Square=333.05, df=80, P-value=0.00000, RMSEA=0.061

Figure 1. The Baseline Model for the U.S. Sample.

Figure 1. The Baseline Model for the U.S Sample



Chi-Square=195.47, df=80, P-value=0.00000, RMSEA=0.053

Figure 2. The Baseline model for Korean Sample.

Figure 2. The Baseline Model for the Korea Sample

assumption across groups. The CFA model with the five factors(A, F, H, N, and Q2) measuring the construct Extraversion provided an adequate fit to both groups. However, the results from the multiple-group CFA suggested that the parameters in the five-factor model were not equivalent across the groups. That is, the U.S. and Korean groups had a comparable number of factors but not comparable factor loadings or error variances. The locus of inequivalence was found to be Factors A, F, and N, where some of the item parcels had lower factor loadings for the Korean group.

The Korean 16PF test provided congeneric measures of the Extraversion construct assessed with the English 16PF test. Hence, the extraversion scale of the 16PF had the same factor structure across the U.S. and Korean groups only in that it had the same number of factors and the same items on each factor. To put it another way, the same construct might be interpreted, but understood in different ways by members in each group. Therefore, group comparisons must be made with caution until the nature of this difference is evaluated. There are some possible explanations for this difference.

First, it might be due to translation error that can change the validity of the Korean translation and adversely influence the comparability of the construct. The importance of the translation was pronounced in many previous studies that the translation process plays an important role in establishing cross-cultural equivalence of measures even though it is only one of the first steps in test adaption (Kim & Lim, 1999; Sohn, 2001). Thus, an effective translation technique such as back translation or decentering techniques should be developed for avoiding translation difficulty.

Second, this difference may occur if the construct measured is not identical across cultural groups; that is, it differs to a substantial degree across cultural groups. Specifically, the definitions of the concept under study do not fully overlap. For example, the Korean concept of filial piety refers to taking care of ones parents, conforming to their requests, and treating them well. This Korean concept is much broader than the Western concept of being a good son or daughter. Or it is also related to the term "construct underrepresentation," which is proposed by Embretson(1983) and Messick(1989) as a source of invalidity. Broad constructs are often represented by a relatively small number of items in a test, and, as a result, the test fails to include important dimensions of the construct.

Also, the two samples used for this study might not be comparable. Even though the present study used college students for both groups to control for sample comparability, some demographic information(e.g., year and major) was not available for 431 U. S. subjects(See Table 1). Thus this might adversely affect the comparability of constructs. Therefore, future research would replicate this study using more comparable samples.

This study intends to provide an example for how researchers can assess construct equivalence of tests across languages and cultures. However, it should be noted that the problem of cross-cultural equivalence cannot be solved by any single method; instead, several methods that deal with different kinds of equivalence should be used. For example, statistical evaluation of a test structure is only one aspect of evaluating construct equivalence across different language versions of a test. Examination of the internal structure of the test might not be sufficient, to establish cross-cultural comparability or generalizability of the construct and the test. After the equivalence of an internal factor structure is assumed across the groups, a higher level of construct equivalence could be established by forming either a nomological network(Cronbach & Meehl, 1955) or equivalence of relations with external variables (Drasgow, 1984). Namely, successive efforts to improve equivalence on all levels are needed.

Although statistical issues were the main focus of this study, the results have implications for substantive issues related to why construct inequivalence occurs. In this sense, the qualitative review of items will be very informative that it can provide some possible explanations, which cannot be provided by the statistical procedures, as to why some of items are interpreted differently by different groups. These findings would be useful for subsequent test development or test adaptations in cross-cultural settings.

Finally, the construct equivalence techniques described in this study may provide a useful approach to exploring the constructs underlying scores on various other attitudinal tests and survey type questionnaires in cross-cultural settings. The results of such investigations may serve to broaden the scientific understanding of test scores and the extent to which tests measure constructs equivalently across cultural groups.

References

Abrahams, F., & Mauer, K. F. (1999). Qualitative and statistical impacts of home language on responses to the items of the Sixteen Personality Factor Questionnaire (16PF) in South Africa. *South African Journal of Psychology, 29*(2), 76-86.

Ben-Porah, Y. S. (1990). Cross-cultural assessment of personality: The case for replicatory factor analysis. In J. N. Butcher & C. D. Spielberger (Eds.), *Advances in personality assessment* (Vol. 8, pp. 1-26). Hillsdale, NJ: Erlbaum

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness-of-fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588-606.

Bernstein, I. H., & Teng, G. (1989). Factoring items and factoring scales are different: Spurious evidence for multidimensionality due to item categorization. *Psychological Bulletin, 10*(3), 467-477.

Bollen, K. A. (1989). *Structural equations with latent variables.* NY: Wiley.

Brislin, R. W. (1970). Back translation for cross-cultural research. *Journal of Cross-Cultural Psychology, 1,* 185-216.

Browne, M. W., & Cudek, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long(Eds.), *Testing structural equation models.* Newbury Park, CA: Sage.

Butcher, J. N. (1996). *International Adaptations of the MMPI-2: Research and Clinical Applications.* Minneapolis, MN: University of Minnesota Press.

Cattell, R. B., & Cattell, H. E. (1995). Personality structure and the new fifth edition of the 16 PF. *Educational and Psychological Measurement, 55*(6), 926-937.

Cattell, R. B., Eber, H. W., & Tatsuoka, M. M. (1970). *Handbook for the 16PF.* Champaign, IL: Institute for personality and ability testing, INC.

Cattell, R. B., & Krug, S. E. (1986). The number of factors in the 16PF: A review of the evidence with special emphasis on methodological problems. *Educational and Psychological Measurement, 46,* 509-522

Chou, C., Bentler, P. M., & Satora, A. (1991). Scaled test statistics and robust standard errors for nonnormal data in covariance structure analysis: A Monte Carlo study. *British Journal of Mathematical and Statistical Psychology,* 44, 347-357.

Church, A.T., & Lonner, W. T. (1998). The cross-cultural perspective in the study of personality: Rationale and current research. *Journal of Cross-Cultural Psychology, 29*(1), 32-62.

Cronbach, L. J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 32*(4), 281-302.

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are the central issues. *Psychological Bulletin, 95*(1), 134-145.

Drasgow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogenous populations. *Journal of Applied Psychology, 70,* 662-680.

Egan, K. L., Sireci, S. G., Swaminathan, H., & Sweeney, K. P. (1998, April). *Effect of item bundling on the assessment of test dimensionality.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Ellis, B. B., & Mead, A. D. (1998, August). *An application of the DFIT framework to assess the measurement equivalence of a Spanish translation of the 16PF questionnaire.* Paper presented at the annual meeting of the International Congress of Applied Psychology, San Francisco, CA.

Ellis, B. B., Minsel, B., & Becker, P. (1989). Evaluation of attitude survey translations: An investigation using item response theory. *International Journal of Psychology, 24,* 665-684.

Embretson, S. (1983). Construct validity: construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179-197.

Flanagan, W., Raju, N. S., & Haygood, J. M. (1998, August). *Impression management, measurement equivalence, and personality factors: Can IRT be* *used to determine the impact of faking.* Paper presented at the 13th annual conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Geisinger, K. F. (1992). The metamorphosis of test validation. *Educational Psychologists, 27,* 197-222.

Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment,* 6(4), 204-312.

Gierl, M. J. (2001). *Construct equivalence on translated achievement tests.* Unpublished Manuscript.

Golden, C. J. (1979). Cross-Cultural second order factor structures of the 16PF. *Journal of Personality Assessment, 42*(2), 167-170.

Hambleton, R. K. (1993). Translating achievement tests for use in cross-national studies. *European Journal of Psychological Assessment, 9*(1), 57-68.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A Progress Report. *European Journal of Psychology Assessment, 10*(3), 229-244.

Hambleton, R. K., & Kanjee, A. (1995). Increasing the validity of cross-cultural assessments: Use of improved methods for test adaptations. *European Journal of Psychological Assessment, 11(*3), 147-157.

Hui, C. H., & Triandis, H. C. (1985). Measurement in cross-cultural psychology: A review and comparison of strategies. *Journal of Cross-Cultural Psychology, 16*(2), 131-152.

Hulin, C. L., Drasgow, F., & Komocar, J. (1982). Applications of item response theory to analysis of scale translations. *Journal of Applied Psychology, 67,* 818-825.

Hulin, C. L., & Mayer, L. J. (1986). Psychometric Equivalence of a translation of the Job Descriptive Index into Hebrew. *Journal of Applied Psychology, 71*(1), 83-94.

Jöreskog, K. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36,* 409-426.

Jöreskog, K., & Sörbom, D. (1986). *PRELIS 2: Users reference guide.* Chicago: Scientific software international.

Jöreskog, K., & Sörbom, D. (1993). *LISREL 8.3: Structural equation modeling with the SIMPLIS command language.* Chicago: Scientific software international.

Kim, A. & Lim, E-Y. (1999, April). *Comparison of effectiveness among different types of practices in cross-cultural test adaptation of attitude measures.* Paper presented at Annual Meeting of American Educational Research Association, Montreal, Canada.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Li, H., & Stout, W. (1995). *A version of DIMTEST to assess latent trait unidimensionality for mixed polytomous and dichotomous item response data.* Unpublished Manuscript.

Malpass, R. S., & Poortinga, Y. H. (1986). Strategies for design and analysis. In W. J. Lonner & J. W. Berry (Eds.), *Field methods in cross-cultural research* (pp. 47-84), Newbury Park, CA: Sage.

Messick, S. (1989). Validity. In R. L. Linn (Eds.), *Educational measurement* (pp. 13-103). New York: Macmillan.

Philip, A. E. (1972). Cross-cultural stability of second-order factors in the 16 PF. *British*

Journal of Social and Clinical Psychology, 2, 276-283.

Poortinga, Y. H., & Malpass, R. S. (1986). Making inferences from cross-cultural data. In. W. J. Lonner & J. W. Berry (Eds.), *Field methods in cross-cultural psychology* (pp. 17-46). Newbury Park, CA: Sage.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*(3), 207-230.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*(3), 552-566.

Roussos, L. A., Stout, W. F., & Marden, J. I. (1997). *Using new proximity measures with hierarchical cluster analysis to detect multidimensionality.* Unpublished Manuscript.

Russell, M. & Karol, D. (1994) *16PF Fifth Edition: Administrators manual.* Champaign, IL: The Institute for Personality and Ability Testing, Inc.

Ryan, K., Wardrop, J., Pyo, K., & Sohn, W. J. (2001, April). *Investigating the construct equivalence for constructed response and selected resonse items within a standards-based framework.* Paper presented at annual meeting of American Education Research Association, Seattle, Washington.

Shaughenssy, M. F., & Kang, M. H. (1998). *Personality profile of gifted children: The 16PF Fifth Edition- A Comparative study of Korean and US Children.* Unpublished manuscript.

Singh, J. (1994). Measurement issues in cross-national research, *Journal of International Business Studies, Third Quarter,* 597-619.

Sireci, S. G., Bastari, B., & Allalouf, A. (1998, August). *Evaluating construct equivalence across adapted tests.* Paper presented at the annual meeting of the American Psychological Association, San Franscisco, CA.

Sireci, S. G., Xing, D., & Fitzgerald, (1999, April). *Evaluating adapted tests across multiple groups: Lessons learned from the information technology industry.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.

Sireci, S. G., Foster, D. F., Robin, F., & Olsen, J. (1997, April). *Comparing dual-language versions of an Internationoal Computerized-Adaptive Certification Exam.* Paper presented at the annual meeting of the National Council on Measurement in Education.

Sohn, W. J. (2001). *Using differential item functioining techniques for investigating the cross-cultural equivalence of personsality tests.* Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *Psychometrica, 27,* 229-239.

Tsujioka, B., & Cattell, R. B. (1965). Constancy and difference in personality structure and mean profile, in the questionnaire medium, from applying the 16 P. F. test in America and Japan. *British Journal of Social and Clinical Psychology, 4,* 287-297.

Van de Vijer, F., & Leung, K. (1997). Methods and data analysis for comparative research. In J. W. Berry, Y. H. Poortinga, & J. Pandey (Eds.).

Handbook of Cross-Cultural Psychology (vol. 1), (pp. 257-300).

Van de Vijver, F., & Poortinga, Y. H. (1982). Cross-cultural generalization and universality. *Journal of Cross-Cultural Psychology, 13,* 387-408.

Van de Vijver, F., & Poortinga, Y. H. (1991). Testing across cultures. In R. K. Hambleton & J. N. Zaal (Eds.), *Advances in educational and psychological testing* (pp. 277-309). Dodrecht, the Netherlands: Kluwer.

Wheaton, B., Muthen, B., Alwin, D. F., & Summers, G. F. (1977). Assessing reliability and stability in panel models. In D. R. Heise (Ed.), *Sociological methodology* (pp. 84-136). San Francisco: Jossey-Bass.

1차원고접수: 2002. 4. 8
수정원고접수: 2002. 6. 6
최종게재결정: 2002. 6. 10

한국심리학회지 : 일반
Korean Journal of Psychology
2001. Vol. 20, No. 1, 91-116

한국어판 16PF와 영어판 16PF 검사간의 구인 평형성

손 원 숙

이화여자대학교 심리학과

우리는 종종 외국의 심리검사를 번역하여서 그것을 비교문화 연구에 사용하는 경향이 있다. 이런 상황에서 타당한 그룹간의 비교를 위해서는 그 검사들이 재고 있는 구인(construct)이 서로 다른 언어 버젼에서 평형적(equivalent)임을 항상 경험적으로 검증해 보아야 한다. 이 연구에서는 한국어로 번역된 성격요인검사(Sixteen Personality Factor Questionnaire: 16PF)와 영어로 된 16PF의 구인 평형성(Construct Equivalence)을 통계적으로 검토하였다. 이를 위하여 두 개의 통계방법들이 상호보완적으로 사용되었는데 즉, 탐색적인 목적으로 주성분분석을, 확인적인 목적으로 중다그룹 확인적 요인분석(multi-group confirmatory factor analysis)이 사용되었다. 주성분 분석과 확인적 요인분석의 결과에 따르면 외향성을 재고 있는 하위 다섯 개의 척도는 두 그룹 모두에서 일차원성(a unidimensional construct)을 가지고 있음이 밝혀졌다. 반면, 중다 그룹 확인적 요인분석의 결과에 따르면, 오직 동일한 요인의 수와 각 요인에 같은 문항들이 속해져 있다는 면에서 이 두 개의 검사가 동일한 요인구조를 있다고 말 할 수 있었다. 그러나, 요인 부하값(factor loadings)과 오차분산(error variance)은 이 두 그룹에서 평형적이지 않았다. 마지막으로, 서로 다른 언어로 된 검사들이 사용되는 연구에서는 하나의 선행조건으로 반드시 구인평형성의 문제는 신중히 검토되어야 하며, 이를 위해서 한 개 이상의 통계적인 방법들을 사용하는 것이 바람직 할 것이라는 점이 이 논문에서 토의되고 있다.

주요어 : 구인 평형성, 비교문화 연구, 성격검사, 중다 그룹 확인적 요인분석