한국심리학회지 : 일반
Korean Journal of Psychology
2003. Vol. 22, No. 1, 89-113

Effects of Different Types of Practice in Cross-Cultural Test Adaptation of Affective Measures

Ahyoung Kim† Eun-young Lim

Department of Psychology, Ewha Womans University

The purpose of the present study is to compare the effectiveness of three types of practice in enhancing the validity and equivalency of test instruments when cross-cultural adaptation of attitude measures is necessary. The three types of practice are: (1) translation and review; (2) translation, back translation, and review; (3) translation, back translation, review, and empirical validation study. Seven hundred and thirty four 5th graders from three public elementary schools in Seoul, Korea participated in this study. Reponses on the three test versions and two other motivation scales were collected within a 3-week period with approximately one-week intervals. Overall results show that the validation version is most superior in all aspects. Furthermore, back translation version is superior to the translation version in terms of its similarity to the validation version and construct-related evidence. Results from IRT analysis reveal that the item qualities of the validation version is superior to the other two versions. Discussions are provided in terms of the nature of the adapted attitude scales.

Key Words : cross-cultural research, test adaptation, back-translation, test equivalence, IRT

† 교신저자 : 김 아 영, (120-750) 서울시 서대문구 대현동 11-1, 이화여자대학교 심리학과
E-mail : aykim@ewha.ac.kr

When a researcher investigates certain human characteristics by adopting a theory that has been developed and tested in a foreign language and culture, replication of the findings and confirmation of applicability of the theory to his or her own culture are due procedures. These procedures also provide an expansion of the universality and generalizability of the theory. Therefore, researchers investigating cultural similarities and/or differences in human psychological traits, especially in the affective domain, need to have equivalent research materials including psychological testing instruments for measuring the traits in all involved cultures. Consequently, researchers should adapt the instrument written in the original researcher's language. An appropriate adaptation procedure is required to secure psychological equivalency between the original (source) and target language versions of the instrument. Test adaptation does not mean a simple translation of an original measure. Rather it involves a series of procedures that includes translation, back translation, group discussion for review, and empirical validation procedures.

The validity of psychological test adaptation has long been an issue for cross-cultural researchers (e.g., Cattell, 1970; Eysenck & Eysenck, 1983; Geisinger, 1994; Hambleton, 1993). To the extent that the adaptation is valid, acceptance of the research findings in the culture in which the test was originally developed is judged valid. Because of this reason, numerous attempts have been made all around the world to improve the equivalency and validity of cross-cultural test adaptation [e.g., Cheung (1985) in Hong Kong; Manos (1985) in Greece; Savasir & Erol (1990) in Turkey]. But unfortunately, to the present authors' knowledge, insufficient effort has been made to improve the validity of foreign instruments in Korean cross-cultural test adaptation practice.

Theoretical and Methodological Considerations in Cross-Cultural Test Adaptation

Psychological Equivalence

Berry and Dasen (1974) have pointed out that there are three aspects of psychological equivalence that should be taken into consideration when cross-cultural test adaptation is in need: These are functional, conceptual, and metric equivalencies. Some researchers such as Butcher and Garcia (1978), and Butcher and Han (1996) have proposed scalar equivalence in addition to the three aspects.

Functional Equivalence

Functional equivalence exists when certain behaviors that the instrument attempts to represent function identically in all involved cultures. For example, "when personality characteristics measured by one scale are highly related to those measured by another scale in a different culture, it can be said that these two scales, though manifestly different, are functionally equivalent across cultures (Butcher & Han, 1996, p. 45)." On the other hand, if a boy pushes a girl, the manifested pushing behavior can mean either a favorable gesture or a hostile gesture toward the girl. In this case, the manifested pushing behavior of the two situation are not functionally equivalent. Statistical analysis techniques, such as factor analysis and intercorrelation pattern analysis are applied to assess functional equivalence between scales(Butcher & Han, 1996). When the functional equivalence can be considered to be present, then securing conceptual equivalence is the next concern.

Conceptual Equivalence

When there are semantic similarities between the words in the source and target language versions, conceptual or linguistic equivalence is considered to be present. Translation, back translation, and small group discussion for review have been adopted to ensure conceptual or linguistic equivalence (Brislin, 1971; Hulin, 1987). Back translation in particular has been identified as an effective procedure to secure conceptual equivalence (Butcher, 1985).

Metric Equivalence

Metric or psychometric equivalence can be acquired when the instrument is validly adapted. Various statistical analyses have been proposed to ensure metric equivalence, such as: computation of intercorrelation among subcomponents, examination of point-biserial correlation between item responses, and the total scale score between the different language versions of the scales (Butcher & Han, 1996). Differences in item-total correlations are assumed to reflect psychometric differences introduced by the translation from the source to the target language.

Scalar Equivalence

Along with the above mentioned three types of equivalence, scalar equivalence has been proposed by some researchers (e.g., Butcher & Garcia, 1978; Butcher & Han, 1996). Scalar equivalence is said to be established when the two instruments measure certain characteristics with the same degree, intensity, or magnitude. Thus, mean score similarity is not sufficient to demonstrate scalar equivalence of two instruments. Scalar equivalence is more specific to the psychometric nature of a test.

Butcher and Han (1996) illustrate that the scalar equivalence has been established when two persons who have MMPI T scores of 75 on the social subscale are socially introverted to approximately the same degree (p. 48). However, scalar equivalence is the most difficult one to establish among the four types, and only indirect approaches have been provided.

Statistical Methods

To achieve the psychological equivalence between the source and target language versions, various statistical methods have been applied. Examination of response tendency of the source and target population, reliability indices, and the patterns of correlations between the response scores of the scale and the scores of the external variables are the recommended (APA, AERA, & NCME, 1999) and routinely used statistical methods to test the psychological equivalence.

Factor Analysis

One of the most rigorous and commonly applied statistical analyses to confirm the underlying factor structures of the source and target language versions of a scale is factor analysis. If two scales are representing the same traits, the factor structure obtained from the analyses of two response sets will be similar. Commonly used methods of factor structure comparison are examination of factor congruence coefficients, factor score correlation, and maximum likelihood confirmatory factor analysis is also recommended by some researchers [see Butcher & Han (1996) for details].

Item Response Theory

Item response theory (IRT) based techniques have been acknowledged as a very important method in the cross-cultural test adaptation. While factor analysis techniques do not allow individual item comparisons, IRT method provides assessment of the similarity of invariant individual item characteristics across samples (Butcher & Han, 1996; Bontempo, 1993). Differences in the item characteristic curve (ICC) indicate that the two items are not equivalent. Thus, such items will produce nonequivalent scales. IRT, although indirect, can be used to ensure translation adequacy. Securing high-fidelity translations from source to target language is essential to ensuring metric equivalence in the two versions. IRT based method also provides item quality and equivalency information.

As Hulin (1987) noted, metric equivalence is determined by the equivalence of responses to two different versions. If two versions of an item elicit equal probabilities of a specified response from individuals at the same level of the trait assessed by the item, metric equivalence of the two items is supported (Hulin, 1987). On this ground, cross-cultural test adaptation researchers have acknowledged the effectiveness of IRT-based techniques in ensuring the quality and equivalence of test items between the source and target language versions (e.g., Candell & Hulin, 1986; Ellis, Becker, & Kimmel, 1993; Drasgow, 1984; Hulin, Drasgow, & Komoar, 1982). These researchers claim that the classical test theory-based item analysis techniques can not achieve psychometric equivalence between the target and source language versions because of the sample-specific nature of item difficulties and discriminations.

Since earlier version of IRT method presumes dichotomous response items, other response scales such as rating scale measures have often been treated as dichotomous ones, which brings about serious limitations in the adoption of the IRT method to affective scales. But this problem has been solved with the development of a graded response model which can handle polytomous responses obtained from multiple choice or Likert-type items (Samejima, 1969; Tissen, 1992).

As Butcher and Han (1996) have noted, it is difficult to distinguish and establish the four types of equivalence separately. Thus, it is proposed that cross-cultural test adaptation researchers should first improve an instrument through proper translation techniques, and then establish conceptual and functional equivalence by constructing nomological network or by factor analysis, followed by application of IRT or regression methods to test item/metric equivalence and scalar equivalence (Hui & Triandis, 1985).

Back Translation

Back translation involves, first, the process of translating the translated target language version back to the source language by a bilingual person. The back translated version is then compared with the original version in terms of general meaning of the sentences, complexity levels, forms, semantic similarity of words, and grammatical structures. Items that don't match the original version are retranslated, back translated, and compared again. Multiple iterations are recommended to produce equivalence between the two language versions. A small group of bilinguals are involved in the translation, back translation, and review discussion process for item revision. Functional and conceptual equivalence are tested and secured via psychometric procedures. In this sense, rigorous procedure of translation of the original into target language version is fundamental prior condition for achieving the validity and equivalence of the two.

Korean Adaptation Practice

For valid test adaptation, it is proposed to follow all of the above mentioned procedures through empirical research (Butcher & Han, 1996; Geisinger, 1994). Nevertheless, few Korean cross-cultural test adaptation researchers have applied the recommended procedures adequately. In Korea, it is observed that four different practices have been attempted in cross-cultural test adaptation. These practices are based on either a partial procedure or the whole procedure that has been proposed by the researchers, such as Bracken and Barona (1991), Butcher (1985), Geisinger (1994), and Hambleton and Kanjee (1993) and others. The four types of practice applied in Korea will be described below.

The adaptation procedure starts with the translation of the original scale into a new language version. Thus, the first and simplest way of adapting the original scale is to translate the original version into Korean and use it without any further endeavor for evaluating equivalence. The second and the most commonly used practice in Korea is to translate the scale, then set up a small review committee which edits or revises the translated items to ensure correct understanding and content validity of the instrument. In some instances, if certain items are not appropriate in Korean culture, those are eliminated. The third practice is that, after first translation, back translation procedure is adopted. Items for which the original version and back translated version do not match are subjected to another translation by the first translator (this procedure is called double back translation), or sent to a review committee to be edited or revised as was mentioned in the second type of practice above, i.e., without any double back translation. The fourth and the most desirable practice is that, after both second and third practice procedures are completed, empirical validation study is conducted. That is, after back or double back translation and editing and revising items, a test is assembled and administered to a sample from the target population. Item analysis and factor analysis are conducted to select good items, and the factor structure and other validity evidences are examined to ensure equivalency to the original instrument. However, there have been very limited reports on the use of back translation, let alone with validation studies. The most popular practice in Korea is the second one mentioned above. Therefore, in this study, we attempt to test the relative effectiveness of these different practices and provide empirical evidence to alert Korean researchers and maybe researchers in different cultures the importance of the valid adaptation procedure.

Purpose of the Present Study

In the present study, we are concerned with the relative effectiveness of the second (translated and reviewed version), third (translated, back translated, and reviewed version), and fourth types of practice for the following reasons: (1) The second practice is the most commonly used in Korea and some researchers (e.g., Hambleton, 1993) claimed that back translation did not significantly improve the validity of the translated version in many empirical studies; (2) nevertheless, some researchers (e.g., American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999; Butcher, 1985) contend that back translation enhances the validity of cross-cultural test adaptation; (3) the simplest first practice is least recommended. We are going to use the fourth type (validation version) as the criterion in examining the relative effectiveness of the second and third types.

We will judge the differential effectiveness in enhancing equivalence and validity of the two types of procedure by comparing the similarity of the translated and the back translated versions to the validated version in the following aspects: (1) a general tendency of subjects' response, (2) the total and subscales reliability coefficients, (3) patterns of item-total correlations, (4) factor structures, (5) patterns of intercorrelation among factors, (6) patterns of relationships with external variables, such as other motivation variables like general self-efficacy and locus of control that have been included in the previous studies, and (7) item parameters estimated via IRT method.

Methods

Participants

The participants in the present study were 734 5th graders attending three typical public elementary schools in a middle class residential area of metropolitan Seoul, Korea. Intact classrooms were the unit of sampling. Data from 711 students' (357 males, 354 females) were used in the final analysis. Data from 10 students were excluded due to the incompleteness of the responses in three repeated administrations of three versions of the scales used in this study.

Materials

Academic Failure Tolerance Scale

To examine the effects of test adaptation practices, this study used Margaret M. Clifford's Academic Failure Tolerance Scale (Clifford, 1988, 1991, hereafter AFT) as the original test instrument (Appendix 1). The AFT was developed as an academic motivation measure that assesses students' reactions following failure experience. The AFT consists of 27 6-point (1: strongly disagree to 6: strongly agree) Likert-type scale items with three 9-item subscales, each measuring preferred task difficulty, feelings following failure, and behavior following failure. High scores represent positive attitude following failure. Technical properties, such as validity and reliability, of the original instrument were already reported from US samples (Clifford, 1988, 1991) and the original AFT has been adapted into Korean version. The Korean version of AFT (K-AFT) scale is one of few available instruments for measuring attitude, which has applied a valid adaptation procedure which includes translation, double back translation, review, and empirical validation studies (Kim, 1993, 1994, 1997).

The results from the three validation studies for K-AFT were relatively satisfactory to conclude the equivalency to American AFT (Kim, 1994; 1997). Reliability of the subscales, factor structures and factor coefficients, patterns of intercorrelation among subscales, the predictability in academic achievement, developmental trend, and gender differences among subscales were all quite similar to the original version (Kim, 1994). In addition to these two studies, Item analysis via polytomous IRT technique also shows that K-AFT is a fairly good test for measuring academic failure tolerance (Seong, 1998). Upon completion of the full adaptation procedure and validation studies, the K-AFT resulted in 24 items while the original AFT had 27 items.

Instrument

Materials used in the present study were based on Clifford's 1991 AFT scale. Excluding three items that were eliminated in K-AFT, the remaining 24 corresponding AFT items were translated and reviewed, composing the first set (translation plus review version: T, hereafter). This first set items was back translated (Appendix 2). Back translated items were compared with the original English items and 10 out of 24 items didn't sufficiently converge with the original meanings. These items were then revised, back translated (Appendix 3), and revised again. These 10 items were merged with the remaining items, which resulted in the second set (back translation plus review version: BT, hereafter). The third set items were from K-AFT scale (validation version: V, hereafter).

Since comparisons among the three procedural types were our purpose, repeated responses to all three sets from all participants were required. Items from each version were scrambled with items of two other scales (Korean General Self-efficacy Scale: K-GS; Korean Locus of Control Scale: K-LC). The Korean General Self-efficacy Scale (24 Likert-type items) was developed and modified by Kim and Cha (Kim & Cha, 1996; Kim, 1997), and Korean Locus of Control Scale (16 Likert-type items), developed by Clifford (1988), has been adapted by Kim (1996, 1997). These two scales were used as external criterion variables to test concurrent and construct validity as was done in Kim's validation study (Kim, 1997).

Translators and Reviewers

Translation was conducted by a Korean male who lived for seven years and received B.A. degree in business in the US. His English proficiency level was within top 5% among Korean college students who are studying in the US. His TOEFLE score was over 600. Back translation was conducted by a Korean bilingual female who lived for 15 years and received B.A. degree in English in the US. Her English proficiency level was similar to average English major in an American college. She also has experiences in translating psychological test items both in Korean and English, before. The review group consisted of four female psychology majors in a Korean graduate school whose English reading comprehension and writing proficiency levels were very high. Also they all have item construction experience in the field of psychological testing.

Procedure

Subjects received three forms of test booklets, each of them consisting of 48, 40, and 24 items, respectively. To eliminate order effects of the administration sequence of the three adaptation versions, Latin-square design was employed by counterbalancing three administration sequences to each of the three groups (A, B, C). Each administration sequence consisted of three alternative forms that contained three versions (T, BT, V). For effective use of test administration, items of K-GS and K-LC were included in two of the three administrations. As a result, nine different booklets were prepared. Table 1 shows the content and order of the administered test booklets.

Test administrations were repeated three times to intact classrooms by homeroom teachers in a manner similar to standardized testing situations. There was at least a one-week separation between the three sessions for all repeated administrations. Instructions were read aloud and explained by the teachers and sample items were answered together following teachers' request for sincere response. Average testing time was 15 to 20 minutes depending on the test booklets. As is shown in Table 1, to eliminate school effect, all three forms of the test booklets were distributed to the classes of all three schools.

Table 1. Counterbalanced Content and Order of Test Administration Sequences

Order	Group		
	Group A (7 classes)	Group B (6 classes)	Group C (6 classes)
1st administ.	Booklet A1 (48 items) T-version (24) + K-GS scale (24)	Booklet B1 (24 items) BT-version	Booklet C1 (40 items) V-version (24) + K-LC scale (16)
2nd administ.	Booklet A2 (40 items) BT-version (24) + K-LC scale (16)	Booklet B2 (48 items) V-version (24) + K-GS scale (24)	Booklet C2 (24 items) T-version
3rd administ.	Booklet A3 (24 items) V-version	Booklet B3 (40 items) T-version (24) + K-LC scale (16)	Booklet C3 (48 items) BT-version (24) + K-GS scale (24)

Note. All three groups included three different schools. To avoid confusion, we marked each envelop to indicate which class should go on which day.

Analyses

The scrambled items were sorted to restore the original scale sets, representing T, BT, V, K-GS, and K-LC. Since V can be assumed to be valid and equivalent to the original AFT, comparisons were to be made between the T and V, and BT and V sets.

Differences were examined as follows: Basic descriptive statistics, item-total correlations, and reliability indices were compared. Factor analysis was conducted and factor structures and factor coefficients were examined and compared. Item qualities were examined using item parameters estimated from graded response model (Samejima, 1969; Thissen, 1991). For the comparison of the pertinent construct-related validity evidence, correlational analysis was conducted and the patterns of interrelationship among subscale scores, general self-efficacy scale scores, locus of control scale scores were compared. Statistical Analyses System (SAS Institute Inc., 1996) and *MULTILOG* 6.0 (Thissen, 1991) programs were used for statistical analyses.

Results and Discussion

Response Tendency

Preliminary analyses of the subjects' responses to individual items showed that the responses for each item were normally distributed and that the means and the score variability of the total scale and the feeling subscale (Feel), preferred task difficulty subscale (PD), and behavior subscale (Beh) of the three versions (T; BT; V) were similar. The score variabilities of all the scales were similar to the results of antecedent studies (Kim, 1994; 1996). However, while the means of Feel in the three versions were somewhat higher in the present study than those of the 5th graders in the Kim's 1996 data, the means of the Beh subscales were somewhat lower in the present study. Since the subjects of Kim's 1996 study were from six representative regional strata in Korea and the subjects of the present study were from one of such strata, this discrepancy can be interpreted as group differences.

Since sex differences were not our primary concern, the data were not analyzed separately. Table 2 shows basic descriptive statistics of the total and subscales of the three versions and those from Kim's 1996 data.

Correlations among the Three Versions in All Scales

Table 3 shows the Pearson product moment correlations among the three versions in the total scale and the subscales. As can be seen in Table 3, the patterns of correlations among three versions are quite similar in all total and in

Table 2. Means and Standard Deviations of Total and Subscale Scores of the Three Versions(N=711)

	Version			
	T-version	BT-version	V-version	Kim data*
Mean	3.44	3.50	3.47	3.44
SD	0.68	0.66	0.69	0.73
Mean	3.23	3.12	3.35	2.96
SD	1.15	1.03	1.11	1.00
		PD		
Mean	3.31	3.53	3.31	3.31
SD	0.94	0.99	1.01	1.14
Mean	3.77	3.87	3.76	4.06
SD	0.74	0.78	0.84	0.97

Table 3. Intercorrelations Among 3 Versions in Total Scale and Subscales(N=711)

		1	2	3	4	5	6	7	8	9	10	11	12
TOT	1. T-vers	1.00											
	2. BT-vers	.76**	1.00										
	3. V-vers	.81**	.80**	1.00									
FEEL	4. T-vers	.71**	.50**	.54**	1.00								
	5. BT-vers	.49**	.62**	.48**	.74**	1.00							
	6. V-vers	.54**	.48**	.63**	.78**	.76**	1.00						
PD	7. T-vers	.74**	.63**	.65**	.16**	.10*	.11*	1.00					
	8. BT-vers	.58**	.77**	.64**	.12*	.10	.09	.76**	1.00				
	9. V-vers	.63**	.65**	.77**	.14**	.08	.09	.78**	.77**	1.00			
BEH	10. T-vers	.70**	.50**	.56**	.18**	.06	.10*	.50**	.44**	.50**	1.00		
	11. BT-vers	.48**	.66**	.53**	.07	.05	.03	.46**	.50**	.51**	.60**	1.00	
	12. V-vers	.47**	.51**	.64**	.02	.00	.01	.48**	.50**	.52**	.65**	.65**	1.00

*p 〈.01 **p 〈.001

subscales. To be specific, the correlations between V and any of the other two versions are virtually the same for each scale. However, the correlations between T and BT are consistently lower than the correlation between V and any of the other versions. This reveals that the

Table 4. Item-Total Correlations of 3 Versions of 3 Subscales

		T-version		BT-version		V-version	
	Item No.	Item-total Correlations	α Changed*	Item-total Correlations	α Changed	Item-total Correlations	α Changed
		α= .84		α= .80		α= .82	
	1	.663	.810	.592	.772	.685	.781
F	2	.426	.839	.271	.815	.424	.817
E	3	.625	.814	.563	.776	.572	.797
E	4	.552	.824	.531	.781	.472	.811
L	5	.660	.809	.597	.770	.584	.796
	6	.643	.812	.694	.755	.621	.790
	7	.357	.846	.312	.812	.396	.821
	8	.647	.812	.588	.772	.588	.796
		α= .78		α= .84		α= .84	
	9	.576	.738	.693	.808	.697	.803
	10	.158	.803	.549	.826	.472	.831
P	11	.594	.732	.609	.818	.602	.815
D	12	.282	.782	.377	.846	.487	.829
	13	.591	.733	.668	.810	.622	.812
	14	.607	.731	.608	.818	.569	.819
	15	.569	.737	.509	.830	.541	.823
	16	.469	.755	.578	.822	.564	.819
		α= .64		α= .69		α= .73	
	17	.367	.605	.240	.690	.415	.700
	18	.471	.580	.402	.655	.526	.678
B	19	.305	.622	.289	.682	.329	.719
E	20	-.086	.724	.255	.694	.126	.758
H	21	.545	.556	.576	.611	.620	.657
	22	.476	.578	.532	.626	.549	.675
	23	.258	.634	.313	.675	.367	.711
	24	.500	.572	.497	.635	.486	.687

relationship between T and BT is the least among the possible correlations between any pair of the three versions. However, we can say that the three correlations between any pair of the three versions are large enough to support or extract one super-ordinate method factor. And the correlation coefficients were close enough to treat the three versions as alternative measures for each other.

Reliability and Item-total Correlations

The Cronbach's α coefficients for internal consistency were obtained to assess the reliability of the total and subscales in the three versions. Although α coefficients of Beh in T and BT are .64 and .69 which are not very high, α coefficients of all other scales are satisfactory for attitude measures, ranging from .73 to .84. In PD and Beh, V and BT show reliability better than T. However, T shows the highest reliability in the Feel subscale.

The similarity in the patterns of item-total correlation among the three versions was examined. Table 4 shows the item-total correlations and changes of α when the given item is removed from the scale for each subscale in the three versions. For the Feel subscale, only 1 item of BT has item-total correlation lower than .30. For the PD subscale, 2 items of T have item-total correlation lower than .30. For the Beh subscale, 2 of T, 3 of BT, and 1 of V have this pattern. In summary, V has less poor items than the other two versions, but BT turned out to be no better than T in regard to the quality of items.

Table 5. Factor Analysis Result for T-version

Item No.	Factor 1	Factor 2	Factor 3
	Feel	Beh	PD
T1	***.755***	.032	.017
T8	***.736***	.065	.068
T5	***.723***	.021	.043
T6	***.715***	.032	-.014
T3	***.686***	-.144	.107
T4	***.598***	-.109	.120
T2	***.464***	.113	.028
T7	***.394***	.227	.389
* **T20**	.167	-.156	-.084
T21	-.064	***.733***	.140
T24	-.087	***.691***	.119
T22	-.067	***.624***	.170
T18	.058	***.608***	.159
T17	.267	***.351***	.256
T19	.140	***.337***	.134
* **T23**	.110	.267	.147
* **T12**	-.128	.296	.239
T14	-.025	.318	***.664***
T11	-.057	.349	***.611***
T9	.041	.316	***.591***
T13	-.021	.398	***.572***
T16	.175	.164	***.557***
T15	.050	.356	***.557***
* **T10**	.079	-.093	.264
Eigen Value	3.556	2.942	2.623
% of Variance	39	32	29

Note. * less interpretable items that show loading values lower than .30

Factor Structures

Factor analysis was performed to compare the underlying factor structures of the three versions. As was done in the previous studies (Clifford, 1988; Kim, 1994), the common factor model (method=prinit, priors=SMC, nfactor=3 in SAS PROC FACTOR) with varimax rotation was

Table 6. Factor Analysis Result for BT-version

Item No.	Factor 1 Feel	Factor 2 Beh	Factor3 PD
B9	*.743*	-.001	.253
B13	*.726*	-.022	.256
B16	*.640*	.117	.193
B14	*.630*	.018	.205
B10	*.587*	.117	.178
B11	*.551*	-.031	.411
B15	*.442*	-.014	.345
B6	.053	*.777*	.072
B5	.025	*.669*	.049
B1	-.007	*.660*	.076
B8	.060	*.651*	.110
B3	.099	*.639*	-.149
B4	.008	*.582*	-.039
B7	.305	*.343*	.291
B2	-.050	*.321*	-.076
B21	.236	-.064	*.699*
B22	.247	-.068	*.651*
B24	.244	-.106	*.611*
B18	.185	-.101	*.488*
#B12	.295	-.072	.368
B23	.190	.104	*.350*
B19	.181	.058	*.317*
B20	.029	.264	*.311*
B17	.098	.156	*.304*
Eigen Value	3.218	3.053	2.740
% of Variance	36	34	30

Note. # items that seem to be an indicator of other factors than originally expected.

Table 7. Factor Analysis Result for V-version

Item No.	Factor 1 Feel	Factor 2 Beh	Factor 3 PD
V9	*.702*	.011	.322
V16	*.662*	.104	.163
V13	*.661*	-.007	.251
V14	*.625*	-.040	.268
V11	*.596*	-.028	.283
V10	*.547*	.124	.088
V15	*.506*	.014	.322
V12	*.456*	.033	.277
V1	.013	*.781*	-.009
V6	-.042	*.714*	.035
V3	-.019	*.661*	-.154
V8	.055	*.656*	.034
V5	-.044	*.645*	-.080
V4	.086	*.498*	.007
V2	-.014	*.475*	.026
V7	.257	*.428*	.193
* **V20**	.141	.183	.049
V21	.228	-.043	*.756*
V22	.181	-.042	*.663*
V18	.277	.006	*.611*
V24	.265	-.118	*.564*
V23	.173	.038	*.403*
V17	.321	.178	*.389*
V19	.206	.068	*.304*
Eigen Value	3.383	3.179	2.716
% of Variance	36	34	29

Note. * uninterpretable item

estimated. Results are given in Tables 5, 6, and 7.

In terms of the size of explained common variance, V and BT are virtually the same, ordered as PD(36%), Feel(34%), and Beh(30%, 29%). However, T shows quite a different pattern from the other two versions: Feel factortakes the largest portion(39%) of explained common variance, PD factor the least(29%), and Beh factor the medium(32%). It seems that BT is closer to V than T is.

For T, 4 items are less interpretable. For BT, 1 item originally from PD seems to be a better indicator of the Beh factor. Other than that all the other items are consistent with V. With respect to the quality of items indicating the values lower than .30. factors, T is the worst, while BT and V perform similarly and are better than T.

Factor loadings of items on the three factors in the three versions were compared. Items are rearranged by the size of factor loadings in the validation version. Factor loadings and their ranks of corresponding items of the other two versions are also presented (Table 8). If the three versions are equivalent, the ranks of the factor loadings of the three versions should coincide. Spearman's rank-ordcorrelation coefficients between each pair of versions for each subscale were computed. Rank-order correlation coefficients between T and V, and BT and V are .81 and .76 in the Feel factor, respectively; these coefficients are .55 and .95 in the PD factors and .86 and .92 in the Beh factor, respectively. According to these results, BT is more similar to V in their factor loading pattern than the T in the PD and Beh factors, but not in the Feel factor.

FEEL				PD				BEH			
Item #	V-vers (rank)	BT-vers (rank)	T-vers (rank)	Item #	V-vers (rank)	BT-vers (rank)	T-vers (rank)	Item #	V-vers (rank)	BT-vers (rank)	T-vers (rank)
1	.781(1)	.660(3)	.755(1)	9	.702(1)	.743(1)	.591(3)	21	.756(1)	.699(1)	.733(1)
6	.724(2)	.777(1)	.715(4)	16	.662(2)	.640(3)	.557(5)	22	.663(2)	.651(2)	.624(3)
3	.661(3)	.639(5)	.686(5)	13	.661(3)	.726(2)	.572(4)	18	.611(3)		.608(4)
8	.656(4)	.651(4)	.736(2)	14	.625(4)	.630(4)	.664(1)	24	.564(4)	.611(3)	691(2)
5	.645(5)	.669(2)	.732(3)	11	.596(5)	.551(6)	.611(2)	23	.404(5)	.350(5)	.267(7)
4	.498(6)	.582(6)	.598(6)	10	.547(6)	.587(5)	.264(7)	17	.389(6)	.304(7)	.351(5)
2	.475(7)	.321(8)	.464(7)	15	.506(7)	.442(7)	.557(5)	19	.305(7)	.317(6)	.337(6)
7	.428(8)	.343(7)	.394(8)	12	.456(8)	.295(8)	.239(8)	20	.049(8)	.304(7)	-.156(8)

Intercorrelations between Three Versions and External Variables

It is recommended to examine the relationship between focal variables and external criterion variables in assessing the validity of the focal variables. In the present study we use K-GS and K-LC as the external variables which are expected to have a certain degree of correlation with the three subscales. The relations in each subscale and both K-GS and K-LC have been studied earlier by Kim (1996; 1997). The correlations are given in Table 9.

In Table 9, we present the result from Kim's data as evidence of convergent validity for the validation version. The results from Kim's data and V are virtually the same. We then compared the similarity of T and BT to V. Regarding the Feel subscale, no version shows a significant correlation with K-LC and all the versions show significant correlation with K-GS. Judging from the size of correlation between both T and BT, and V, BT is more similar to V than T is. Regarding the PD subscale, all the versions have significant correlations with the two external variables. BT is less similar to V than T is in its correlation with K-LC. However, BT is more similar to V than T is in its correlation with K-GS. Regarding the Beh subscale, BT is more similar to V than T is in its correlation with both K-LC and K-GS. All in all, the BT shows more similarity to V than T does, yielding additional evidence favoring for BT over T.

Table 9. Correlations with External Variables (N=711)

		K-LC	K-GS
Feel	T-version	-.05**	.34**
	BT-version	-.08**	.27**
	V-version	-.07**	.29**
	Kim data#	-.06**	.20**
PD	T-version	.41**	.61**
	BT-version	.38**	.58**
	V-version	.44**	.59**
	Kim data	.44**	.61**
Beh	T-version	.49**	.59**
	BT-version	.43**	.54**
	V-version	.46**	.49**
	Kim data	.45**	.53**

Item Response Theory

Since the factor analysis shows three distinct subscale factors as expected, we applied IRT to analyze each subscale. Items of each subscales were analyzed with *MULTILOG* program. For each subscale, items from the three versions were entered simultaneously in the model to estimate the item parameters and test information function.

Parameter Estimation

Item parameters for the three versions of Feel, Beh, PD are shown in Tables 10, 11, and 12,

Table 10. Estimated Item Parameters of the 3 Versions of FEEL

Item	trait levels (*b*k)					
	a	1	2	3	4	5
T1	1.46	-1.14	-.25	.70	1.25	2.22
BT1	1.43	-1.08	-.01	1.05	1.63	2.51
V1	1.99	-1.00	-.21	.56	.96	1.62
T2	.86	-2.49	-1.10	.26	1.02	2.35
***BT2**	**.54**	**-4.26**	**-2.23**	**-.10**	**1.59**	**4.29**
V2	.87	-2.45	-.97	.14	.96	2.44
T3	1.67	-1.06	-.18	.63	1.06	1.77
BT3	1.34	-.71	.30	1.33	1.80	2.69
V3	1.42	-1.23	-.19	.66	1.10	2.15
T4	1.33	-1.26	-.13	.80	1.28	2.16
BT4	1.17	-1.22	.02	1.00	1.63	2.71
V4	1.03	-1.37	-.17	.96	1.52	2.65
T5	1.63	-1.23	-.42	.28	.70	1.34
BT5	1.66	-1.25	-.46	.31	.74	1.35
V5	1.62	-1.52	-.69	.05	.51	1.28
T6	1.79	-.95	-.06	.62	.97	1.68
BT6	1.95	-.81	.06	.81	1.31	2.11
V6	1.65	-1.15	-.29	.50	.90	1.82
T7	.65	-2.27	-.67	.95	1.97	3.87
BT7	.67	-3.04	-1.22	.43	1.33	3.09
V7	.82	-2.47	-1.21	.06	.90	2.43
T8	1.61	-1.59	-.59	.28	.79	1.70
BT8	1.40	-1.52	-.61	.37	1.01	1.97
V8	1.37	-1.82	-.83	.26	.92	1.96

Note. * poor quality item

respectively.

Items were judged by the discrimination parameter (*a*) and location parameters of boundary characteristics curve (*bk*). Tables show these parameters for the 8 items in the three versions of the three subscales.

Table 11. Estimated Item Parameters of the 3 versions of PD

Item	trait levels (*b*k)					
	a	1	2	3	4	5
T9	1.45	-1.82	-.80	.28	1.26	2.22
BT9	1.89	-1.64	-.69	.20	1.27	2.20
V9	1.91	-1.44	-.48	.49	1.46	2.47
*** T10**	**.30**	**-1.82**	**2.34**	**4.99**	**6.46**	**9.03**
BT10	1.18	-1.91	-.59	.52	1.44	2.66
V10	.97	-1.55	-.18	1.03	1.94	3.45
T11	1.56	-1.85	-.97	-.16	.62	1.53
BT11	1.42	-2.16	-1.29	-.43	.70	1.78
V11	1.50	-1.95	-.88	.13	1.08	2.15
*** T12**	**.51**	**-6.59**	**-4.80**	**-2.99**	**-.59**	**1.93**
BT12	.70	-4.83	-3.34	-1.75	-.10	1.82
V12	.97	-3.08	-1.68	-.49	.82	2.19
T13	1.47	-1.71	-.76	-.07	.82	1.78
BT13	1.84	-1.77	-.84	-.01	.98	2.11
V13	1.73	-1.58	-.81	-.03	1.02	2.08
T14	1.72	-1.67	-.79	.12	1.02	1.88
BT14	1.53	-1.95	-1.02	.03	1.15	2.28
V14	1.59	-1.70	-.76	.21	1.35	2.24
T15	1.32	-2.07	-.92	-.01	.90	1.85
BT15	.99	-3.72	-2.25	-1.03	.36	1.68
V15	1.24	-2.58	-1.38	-.57	.43	1.62
T16	1.11	-1.17	.06	1.42	2.37	3.27
BT16	1.30	-1.35	-.39	.66	1.40	2.72
V16	1.23	-1.67	-.84	.31	1.33	2.55

Note. * poor quality item

Items with high discrimination power and equally spread range of category boundary span are judged to be good (Baker, 1992). Baker suggested that the item discrimination parameter estimates could be judged according to the following criteria: *a* below 0.65 is low; from

Table 12. Estimated Item Parameters of the 3 versions of Beh

Item	trait levels (bk)					
	a	1	2	3	4	5
T17	.78	-4.57	-2.87	-1.30	.08	1.96
*** BT17**	**.49**	**-7.20**	**-4.99**	**-2.66**	**-.62**	**2.27**
V17	.93	-3.72	-2.38	-.98	.02	1.98
T18	1.15	-3.31	-2.27	-1.21	.24	1.85
BT18	1.01	-3.42	-2.27	-.97	.77	2.51
V18	1.34	-2.33	-1.47	-.56	.61	1.94
T19	.78	-2.79	-.90	.34	1.80	3.35
BT19	.80	-2.60	-.94	.29	1..71	3.31
V19	.75	-2.60	-1.10	.25	1.77	3.53
*** T20**	**.21**	**-6.87**	**-.53**	**3.54**	**6.97**	**11.41**
*** BT20**	**.44**	**-5.07**	**-2.12**	**-.27**	**1.35**	**3.77**
*** V20**	**.39**	**-6.47**	**-2.18**	**.38**	**2.41**	**5.27**
T21	1.88	-2.21	-1.30	-.54	.57	1.64
BT21	1.75	-2.13	-1.32	-.57	.50	1.65
V21	2.00	-1.82	-1.08	-.29	.61	1.81
T22	1.33	-2.59	-1.31	-.29	1.15	2.49
BT22	1.51	-2.47	-1.27	-.30	1.00	2.36
V22	1.64	-2.26	-1.26	-.36	.88	2.23
*** T23**	**.62**	**-4.65**	**-2.80**	**-.71**	**1.04**	**2.93**
BT23	.69	-4.63	-2.97	-1.10	.48	2.73
V23	.79	-3.00	-1.18	.10	1.18	2.99
T24	1.62	-2.64	-1.68	-.78	.43	1.73
BT24	1.53	-2.67	-1.63	-.60	.68	2.03
V24	1.36	-2.90	-1.87	-.99	.05	1.30

0.65 to 1.34 is appropriate; from 1.35 to 1.69 is high; above 1.70 is very high. The attribute (attitude trait) of the person being measured by the test (Θ) is usually arbitrarily placed on a z-score scale, thus in practice, ranges roughly from -3.0 to +3.0.

Therefore, Items that have location parameters within this range and have approximately equal intervals between *b*k's are judged to be good.

An examination of the quality of the items using item parameter estimates reveals that 9 items of T, 9 items of BT, and 5 items of V have unrealistic *b*k values (below -3.0 and over +3.0) and that 4 items of T, 3 items of BT, and 1 item of V have *a* lower than 0.65. Overall, 4 items (#10, #12, #20, & #23) of T, 3 items (#2, #17, & #20) of BT, and 1 item (#20) of V have both low *a* and unrealistic value of *b*k's. These results show that BT is slightly better than or similar to T in their item qualities, and V is better than the other two.

Test Information Function

Table 13 shows the test information function for the subscales of the three versions. The test information function values are generally similar across the attribute levels(Θ) of -1.0 to 1.5 in Feel, -1.5 to 1.5 in PD, -2.0 to 2.0 in Beh, showing that the Beh subscale provides similar information over the widest range. Regarding the Feel subscale, T shows the most information and BT the least. However, V shows the best information for the PD and Beh subscales. BT shows more information than T for the PD subscale, but the reverse is observed for the Beh subscale.

From the overall results based on the IRT analyses, we can conclude that item quality of V is definitely superior to the other two versions and BT is not particularly superior to T in its item quality.

Table 13. Test Information Functions of the Subscales

Scale	version	-2.0	-1.5	-1.0	-.5	Θ0	.5	1.0	1.5	2.0
FEEL	T	2.91	4.05	4.77	5.04	5.14	5.21	5.17	4.88	4.16
	BT	2.35	3.32	4.07	4.41	4.53	4.56	4.57	4.40	4.04
	V	3.01	4.02	4.65	4.87	4.95	4.97	4.93	4.64	3.95
PD	T	3.36	3.83	4.00	4.05	4.06	4.07	4.04	3.90	3.50
	BT	4.14	4.70	4.87	4.91	4.88	4.83	4.83	4.74	4.52
	V	4.02	4.76	5.01	5.06	5.05	5.02	4.99	4.94	4.70
BEH	T	3.29	3.38	3.37	3.33	3.26	3.26	3.22	3.15	2.82
	BT	3.07	3.17	3.18	3.14	3.08	3.07	3.03	2.98	2.78
	V	3.62	3.87	3.91	3.89	3.83	3.77	3.67	3.55	3.28

Conclusions

The purpose of the present study is to compare the relative effectiveness of three types of practice in enhancing the validity and equivalency of test instruments in the cross-cultural test adaptation, particularly in the measurement of affective characteristics. Prevalent practice of ignoring proper adaptation procedures in Korea would bring about adverse effects on the generalization of certain theories originated from different cultures. Although numerous international studies have provided accumulated evidences that back translation is an essential technique of ensuring psychological equivalence between source and target language versions (Brislin, 1970; Butcher, 1993; Thorndike, 1974), cross-culturally adapted Korean instruments rarely report such practices. In this respect, this paper attempts to emphasize the importance of valid adaptation procedure including back translation for securing psychological equivalence and provides empirical evidences which are supportive to its purpose.

The results of the present study confirm the importance of valid adaptation procedure. The results also show that the back translation version is more similar to the validation version in the pattern of intercorrelation among subscales, of factor structure, and of its relations with external variables. However, the similarity in the response tendency, item-total correlations, and the item quality are not particularly in favor of the effectiveness of back translation. This result can be understood from the fact that the complexity level of the meaning and sentences used in the AFT is very simple and clear. As Thorndike noted, "maintaining comparability under translation becomes a progressively more serious problem as the material to be translated becomes more difficult (Thorndike, 1974, p. 9)," which implies that the relative efficiency of back translation procedure may vary with the nature of the sentences used. The material used in the present study was not complex enough to reveal the problem of misunderstanding caused by inaccurate translation. The similarity of response tendency and of item quality supports this interpretation. The item quality assessed by IRT suggests that all three versions can be judged to be an acceptable measure of academic failure tolerance, evidencing the scalar equivalence.

However, an adoption of back translation procedure enhances construct-related validity which results in conceptual and metric equivalence. Especially, the factor similarity of BT to V is more salient than that of T to V. In addition, the more equivalent relations with the two external variables support this contention.

All in all, as was evidenced by Brislin's early work, back translation procedure can confirm the quality of translator and translation (Brislin, 1970), which leads to functional, conceptual, metric, and even scalar equivalence between the source and target language versions. With the results of the present study, we can strongly recommend the inclusion of back translation procedure in the cross-cultural test adaptation. It is suggested that in Korea future research should be conducted in the area of personality assessment and in clinical settings where abstract and complex psychological instruments are frequently used.

Since multiple alternatives of translation and back translation versions are possible, a further research with multitrait-multimethod design to examine the equivalency of different versions would be worth trying.

However, it should be emphasized that the consistent superiority of the validation version in terms of its reliability, factor structure clarity, and item quality confirms the importance of a proper validation procedure in cross-cultural test adaptation.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington D. C.: American Psychological Association.

Baker, F. B. (1992). *Item Response Theory: Parameter Estimation Techniques.* New York: Marcel Dekker, Inc.

Berry, J. W., & Dasen, P. (Eds.) (1974). *Culture and cognition.* London: Methuen.

Bontempo, R. (1993). Translation fidelity of psychological scales: An item response theory analysis of an individualism-collectivism scale. *Journal of Cross-Cultural_Psychology, 24.* 149-166.

Bracken, B. A., & Barona, A. (1991). State of the art procedures for translating, validating and using psychological tests in cross-cultural assessment. *School Psychology International, 12,* 119-132.

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

Butcher, J. N. (1985). Current developments in MMPI use: An international perspective. In J. N. Butcher & C. D. Spielberger(Eds.), *Advances in personality assessment,* (Vol. 4, 83-94). Hillsdale, NJ: Erlbaum.

Butcher, J. N., & Garcia, R. (1978). Cross-national application of psychological tests. *Personnel and Guidance, 56,* 472-475.

Butcher, J. N., & Han, K. (1996). Methods of establishing cross-cultural equivalence. In J. N. Butcher (Eds.), *International adaptations of the MMPI-2* (pp. 44-63). Minneapolis, Mn.: University of Minnesota Press.

Candell, G. L., & Hulin, C. l. (1986). Cross-language and cross-cultural comparisons in scale translation: Independent sources of information about item nonequivalence, *Journal of Cross-Cultural Psychology, 17,* 417-440.

Cattell, R. B. (1970). The isopodic and equipotent principles for comparing factor scores across different populations. *The British Journal of*

Mathematical and Statistical Psychology, 23, 23-41.

Cheung, F. M. (1985). Cross-cultural considerations for the translation and adaptation of the Chinese MMPI in Hong Kong. In J. N. Butcher & C. D. Spielberger, (Eds.), *Advances in personality assessment,* (Vol 4, pp. 131-158) Hillsdale, NJ: Erlbaum.

Clifford, M. M. (1988). Failure Tolerance and risk-taking in ten-to twelve-year-old students. *British Journal of Educational Psychology, 58,* 15-27.

Clifford, M. M. (1991). Risk-taking: Theoretical, empirical, and educational considerations. *Educational Psychologist, 26,* 263-297.

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variable are the central issues. *Psychological Bulletin, 95.* 134-135.

Ellis, B. B., Becker, P., & Kimmel, H. D. (1993). An item response theory evaluation of an English version of the Trier personality inventory (TPI), *Journal of Cross-Cultural Psychology, 24,* 133-148.

Eysenck, H. J., & Eysenck, S. B. G. (1983). Recent advances in the cross-cultural study of personality. In J. N. Butcher & C. D. Spielberger (Eds.), *Advances in personality assessment,* (Vol. 2, pp. 41-69). Hillsdale, NJ: Erlbaum.

Geisinger, K F. (1994). Cross-cultural normative assessment: Translation and adaption issues influencing the normative interpretation of assessment instruments. *Psychological Assessment, 6,* 304-312.

Hambleton, R. K. (1993). Technical standards for translating tests and establishing test score equivalence. In Symposium conducted at the 1001st Annual Convention of the American Psychological Association, Toronto, Ontario, Canada.

Hambleton, R. K., & Kanjee, A. (1993). Enhancing the validity of cross-cultural studies: Improvements in instrument translation methods. Paper presented at the Annual Meeting of American Educational Research Association. Atlanta, Georgia.

Hulin, C. l. (1987) A psychometric theory of evaluations of item and scale translations: Fidelity across languages, *Journal of Cross-Cultural Psychology, 18.* 115-142.

Hulin, C. I., Drasgow, F., & Komoar, J. (1982). Applications of item response theory to analysis of attitude scale translations. *Journal of Applied Psychology, 17.* 417-440.

Kim, A. (1993). Failure tolerance in Korean grade school students: Development and sex differences. Paper presented at the Annual Meeting of American Educational Research Association. Atlanta, Georgia.

Kim, A. (1994). Development of Korean version of the academic failure tolerance scale. *Korean Journal of Educational Research. 32,* 59-75. (Korean).

Kim, A. (1996). Data base '96.

Kim, A. (1997). A study on the academic failure tolerance and its correlates. *Korean Journal of Educational Psychology, 11*, 1-19. (Korean).

Kim, A., & Cha, J. (1996). General self-efficacy and its measurement. Paper presented at the Annual Meeting of Korean Industrial and Organizational Psychological Association, Seoul, Korea. (Korean).

Manos, N. (1985). Adaption of the MMPI in Greece: Translation, standardization, and cross-cultural comparison. *Advances in personality assessment* (Vol, 4, pp.159-208). Hillsdale, NJ: Erlbaum.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement,* No. 17.

Savasir, I., & Erol, N. (1990). The Turkish MMPI: Translation, standardization and validation. *Advances in personality assessment* (Vol. 8, pp. 49-62). Hillsdale, NJ: Erlbaum.

Seong, T. J. (1998). Item parameter estimation of the Korean Academic Failure Tolerance Scale and person's trait estimation via the grade response model. *Korean Journal of Educational Psychology, 12.* 203-218. (Korean).

Thorndike, R. L. (1974). Methodological problems in developing instruments for cross-national studies. Paper presented at the Annual Meeting of the National Council on Measurement in Education. Chicago, Illinois, April 1975.

Thissen, D. (1991). *MULTILOG Version 6.0 user's guide.* Chicago, Illinois: Scientific Software, Inc.

1차원고접수 : 2002. 10. 20.
수정원고접수 : 2003. 4. 10.
최종게재결정 : 2003. 4. 17.

한국심리학회지 : 일반
Korean Journal of Psychology
2003. Vol. 22, No. 1, 89-113

타문화권 척도 번안과정에서 적용되는 절차들 간의 효과 비교

김 아 영 임 은 영

이화여자대학교

이 연구는 타문화권에서 만들어진 척도를 사용하기 위해 필요한 '타문화 검사 번안' 방법으로 국내에서 주로 적용하고 있는 세 가지 번안 절차, 즉, '번역 후 문항검토(번역본)', '번역, 역번역 후 문항검토(역번역본)', '번역, 역번역, 검사 타당화 검증(타당화본)'의 세 가지 절차들 간의 효과를 비교하는 것이 목적이다. 효과의 차이는 각 절차를 적용해 제작된 세 가지 검사지를 711명의 초등학교 남녀 학생들의 세 번에 걸친 반복 측정 결과를 분석함으로써 검증하였다. 구체적으로, 세 가지 척도에 대한 반응 경향성의 비교, 신뢰도, 문항-총점간 상관 패턴, 요인간 상호상관 패턴, 외적 준거변인과의 관련 경향성, IRT를 적용한 문항모수치의 특성 및 양호도 비교 등을 통해 원 척도와 번안된 척도의 동등성을 비교하였다. 전체적인 연구 결과는 타당화본이 모든 측면에서 가장 양호한 척도임을 보여주었고 역번역본이 번역본보다 구성요인에 있어서 원본과 더 유사함을 보여주었다. 본 연구의 결과는 타문화권 척도를 도입할 때 필수적으로 따라야 하는 역번역 절차와 타당화 절차의 중요성을 보여주었다. 결론적으로 국내에서 타문화권에서 제작된 검사를 도입할 때의 적법한 절차 도입의 필요성을 논의하였다.

주요어 : 타문화권 척도, 검사 번안, 역번역, IRT