한국심리학회지: 건강
The Korean Journal of Health Psychology
2007. Vol. 12, No. 1, 241 - 267

Two measurement models of testing measurement equivalence: an example of Child Behavior Checklist

Eun-Jeong Kang† 1)　E. Michael Foster2)　Michael J. Rovine3)

1) Research Fellow, Korea Institute for Health and Social Affairs

2) Professor, University of North Carolina at Chapel Hill

3) Associate Professor, The Pennsylvania State University

In children's mental health, many psychological assessments are critical to measure end outcomes. And equally valid measures are essential for health services outcome research in diverse populations. Examination of equal validity in tests measuring the emotional, motivational, attitudinal, and interpersonal characteristics of children can be achieved with factor analytic tools. Factor analysis assumes a particular measurement model, effect indicator model. However, some psychological measures such as Child Behavior Checklist are not clear whether effect indicator model is assumed. In this paper, an alternative measurement model, causal indicator model, was introduced. Measurement equivalence of CBCL was tested using multiple group analysis based on effect indicator model and using principal component analysis based on causal indicator model. The results showed that measurement equivalence tests based on the two measurement models in CBCL led to similar conclusions. We conclude that the two measurement models could be complementary for explaining differential validity of psychological measures across sub-populations.

Keywords : effect indicator, causal indicator, measurement equivalence, CBCL

† Corresponding author : Eun-Jeong Kang, Korea Institute for Health and Social Affairs, San 42-14 Bulkwang-dong, Eunpyong-gu, Seoul; E-mail: marchej@kihasa.re.kr

In children's mental health, many psychological assessments are critical to measure end outcomes. And equally valid measures are essential for health services outcome research in diverse populations. When comparing the outcome of interest among heterogeneous populations, researchers often administer the same measure without considering conceptual and psychometric adequacy of the measure for heterogeneous populations. However, since most psychological measures in US were developed for the majority culture-middle class whites, these measures may not reflect the construct that researchers intend to measure for racial and ethnic minorities (Stewart & Napoles-Springer, 2000).

Examination of bias in tests measuring the emotional, motivational, attitudinal, and interpersonal characteristics of children has been limited (Cohen & Kasen, 1999). "Bias" refers to differential validity of a given scale score across different sub-populations (Cole & Moss, 1989). Validity refers to the degree to which a measure reflects what we intend to measure, i.e., the construct. If a measure is biased, we cannot unequivocally determine whether observed differences in the measure reflect real differences in the underlying construct. As a result, we might take policy initiatives based on misleading differences (or similarities) in the measurement scores, which are originally the same (or different). In addition, when we predict a secondary outcome with a biased measure, we may obtain differential predictions because of the differential relationship between a biased measure and other variable.

Since measurement bias means differential validity, measurement bias can be detected by validity tests. In most cases, however, it is not easy to define a construct, and neither is to test validity. A single study cannot prove or disapprove the validity of a measure. Instead, we need accumulated evidence from various studies on the internal structure and the relationship with other variables in order to support the validity of the measure.

Testing equal construct validity across groups identifies the equivalence of the internal structure of a measure, while testing equal criterion validity determines whether the relationships of a measure with other variables are the same across groups. Factor analysis has recently been popular in testing construct validity by looking at the internal structure of a construct and its items (Ben-Porath, 1990 Bird, 1996; Hartman, Hox, Auerbach, Erol et al., 1999; Muthen, Hasin, & Wisnicki, 1993; Reynolds & Paget, 1981). According to the classical test theory, the observed variables are the indicators of a latent construct (Lord & Novick, 1968). If we intend to measure the same construct across racial/ethnic groups, differential patterns of correlations among the observed variables may imply that the observed variables might not represent the latent construct in the same way across groups (Vanderberg & Lance, 2000). In this case, the measure is biased for the groups. On the other hand, if the measure allows culture-variant aspects of a construct, differential patterns of correlations among the observed variables may not necessarily indicate the existence of bias, but the true differences in the differential relationship between the construct and the indicators.

We can test equivalence of construct validity among heterogeneous groups using both exploratory factor analysis and confirmatory factor analysis. Exploratory factor analysis does not test a priori hypotheses. Instead, exploratory factor analysis is useful (1) to determine the number of factors that best represent the items and (2) to interpret the factors (Spector, 1992). Measurement equivalence is found when the same number of factors consisting of the same items is found across groups and therefore we can interpret the factors in the same way across groups.

Confirmatory factor analysis test a priori hypothesized relationships between a construct and the items. We can test if each factor consists of the same items that we found from exploratory factor analysis or theories. With confirmatory factor analysis, we can also test the equivalence in various parameters in a hypothesized factor model such as factor loadings, factor intercepts, inter-factor correlations, and factor variance. Multiple group analysis is a useful tool to test measurement equivalence across heterogeneous groups (Meredith, 1964). In this analysis, we can test if a specific parameter is the same across groups.

Two distinct measurement models are defined depending upon the direction of the causal relationship between the latent construct of interest and the measures or items that measure the construct. In a measurement model with "effect indicators," (Blalock, 1964; Bollen & Lennox, 1991) also known as reflective indicators (Fornell & Bookstein, 1982), the latent construct causes the indicators. On the other hand, "causal indicators" (Blalock, 1964; Bollen & Lennox, 1991) or formative or composite indicators (Fornell & Bookstein, 1982) cause the latent construct.

Whether we assume the items to be effect indicators or causal indicators has implications on measurement equivalence studies. Tests of measurement equivalence from the factor structure are suitable for effect indicators, but not for causal indicators, because factor analysis assumes a construct, even if it is abstract and unobservable, which is real. Since no pre-existing constructs are assumed in causal indicator models, it is illogical to do factor analysis with causal indicators.

However, we can use principal component analysis to test measurement equivalence with causal indicator model. Even though principal component analysis is often used as exploratory "factor" analysis, its computation for extracting "components" is different from the way usual factor analysis does. Principal component analysis finds "components" in the items by maximizing all variance of the items (Nunnally & Bernstein, 1994).

The Child Behavior Checklist syndromes are empirically driven without a priori theories using principal component analysis. In this sense, the conventional measurement model where a pre-existing construct is assumed does not seem to fit for the CBCL. However, many researchers have used several models based on the conventional measurement model such as factor analysis (Dedrick et al., 1997; DeGroot et al., 1994; Lambert et al., 2003) and latent class analysis (Hudziak, Wadsworth, Heath, & Achenbach, 1999; Wadsworth, Hudziak, Heath, & Achenbach, 2001). In this study, measurement equivalence is tested in both models. Different results and inferences would suggest more caution in selecting the model.

Literature review

The Child Behavior Checklist

While the Diagnostic and Statistical Manual approach attempts to classify people into two categories, i.e., pathologic or normal, an empirically based assessment like the CBCL tries to understand the features of the syndrome manifested by a child (Achenbach & McConaughy, 1997). After discussing the items with parents and mental health professionals, 118 items were developed. Respondents rate each behavior on a 3-level scale: 0=not true, 1=somewhat or sometime true, and 2=very true or often true. The three-point scale was chosen for the untrained raters which usually applied to parents (Achenbach, 1991). The CBCL is obtained by summing up the item scores. For example, the total score ranges from 0 to 118×2=236. The CBCL can be administered by parent themselves or by an interviewer. The CBCL 4/18, which is to be rated by parents, can be assessed from 4 through 18 years old (Achenbach, 1991).

A major aim of the 1991 CBCL was to provide cross-informant syndromes that are common across parents, teachers, and children themselves (Achenbach, 1991). Before obtaining the cross-informant syndromes, Achenbach identified core syndromes underlying across sex/age groups. He chose items that were found in the syndrome for a majority of the sex/age groups. He found 85 items that were common in core syndromes across the CBCL, Teacher's Report Form(TRF), and Youth Self-Report(YSR). Items that were found on core syndromes for at least two of the three instruments were included in cross-informant syndromes. The eight syndrome scales were as follows: withdrawn, somatic complaints, anxious/depressed, social problems, thought problems, attention problems, delinquent behavior, aggressive behavior (Achenbach, 1991). The original wordings of the CBCL items are listed in Appendix 1. It is important to note that each syndrome score is obtained from the sum of the item scores instead of factor scores.

These scales were further factor analyzed and allotted to three factors: internalizing problems, externalizing problems, and mixed problems. Achenbach (1991) performed principal factor analysis of the correlations among the syndrome scores separately for each sex/age group on the CBCL, YSR, and TRF. The two largest factors in terms of eigenvalues in each group were orthognally rotated. Then they obtained syndromes that had the largest average factor loadings across all groups on all three instruments. Internalizing problems include withdrawn, somatic complaints, and anxious/depressed. Externalizing problems include delinquent behavior and aggressive behavior. The rest of the syndromes (social problems, thought problems, attention problems) were classified as mixed problems.

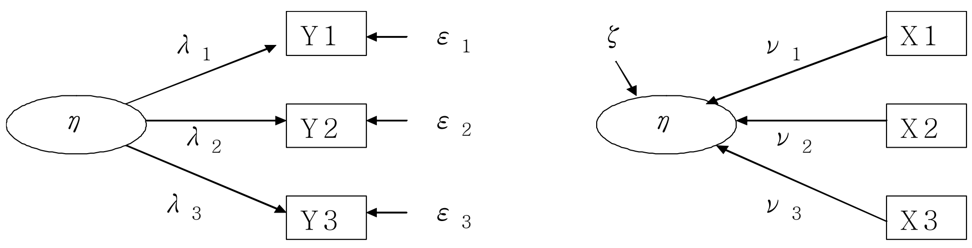
Two ways of modeling the Child Behavior Checklist

The effect indicator model is the one that classical test theory, reliability estimation, and factor analysis view (Edwards & Bagozzi, 2000). The algebraic expression of the effect indicators model is as follows:

$$y_i = \lambda_i \eta + \varepsilon_i \,, \qquad (1)$$

where y_i is the ith indicator, η is the latent construct, ϵ_i is the measurement error, and λ_i is the coefficient giving the effect of η on yi. It is assumed that the error is uncorrelated with η, covariance among the error terms is zero, and the expectation of the each error term is also zero.

In classical test theory, the indicators are



(a) Effect indicators (b) Causal indicators

Figure 1. Two measurement models

considered as a sample from the same domain. The indicators should be highly correlated, because all of them are positively associated with the same concept. For the same reason, these indicators should also constitute unidimensionality. If the indicators are representative of one particular concept, then the correlations among the indicators within the construct should be higher than the correlations with the indicators of a different construct. The above assumptions are the "conventional wisdom" on measurement (Bollen & Lennox, 1991).

However, an alternative model, the causal indicators model, presumes that the construct is the composite of several indicators.

$$\eta = \sum \gamma_i x_i + \zeta \quad (2)$$

where γ_i is the effect of x_i on η and ζ is the error term. Unlike the effect indicators, causal indicators are assumed to be error-free. As illustrated in Figure 1, all the indicators are correlated each other. Examples of the causal indicators model include socioeconomic status as measured by income, education, and occupation (Bollen & Lennox, 1991), quality of life of cancer patients as measured by pain, nausea & vomiting, and treatment symptoms (Fayers & Hand, 1997), and social support as measured by the number of incidents of positive and negative support in the past month (MacCallum & Browne, 1993).

The "conventional wisdom" mentioned above is not applied for causal indicators. First, causal indicators need not be highly correlated each other. In fact, it is more desirable that the indicators are measuring different aspects of the construct of interest. As a consequence, the more exhaustive the list of indicators is, the better these indicators explain the construct (Fayers & Hand, 1997). High correlations among the indicators could only cause a multicollinearity problem. Second, since high correlations among indicators are not expected for causal indicators, reliability measurement like internal consistency and factor analysis are not appropriate.

The discussion on which indicator model a measure implies is rather an ontological issue (Borsboom, Mellenbergh, & Heerden, 2003). In other words, if we accept that there exist mental health problems such as depression, anxiety, delinquent behaviors, and aggressive behaviors and these preexisting constructs causes the CBCL items accordingly, we are so-called realistsand the CBCL items are effect indicators. On the contrary, if we believe that mental health problems are nothing but a combination of phenomena that we can explain with what we observe, we are constructivists and the CBCL items are causal indicators.

A natural question, therefore, would be whether there exist the eight syndromes of the CBCL. We believe that we do not have a definitive answer. Until we gain more confidence toward either model, it seems to be fair to consider both types of models in mind when we analyze the CBCL. In this study, measurement equivalence of the CBCL was examined in both contexts of effect indicators and causal indicators accordingly.

Methods

Data and Sample

The data came from the evaluation of the Comprehensive Community Mental Health Services for Children and Their Families Program. The data used for this study come from the phase II of the program. The Center for Mental Health Services, which is a component of the Substance Abuse and Mental Health Services Administration, U.S. Department of Health and Human Services, administers the Federal grants.

Macro International Inc. has been conducting the outcome-based evaluation of the Comprehensive Community Mental Health Services for Children and Their Families Program since 1994. The evaluation components include child and family outcome data related to behavioral and functional status as well as demographic and descriptive information.

The evaluation study was longitudinally designed. Nineteen sites where the CBCL data were collected from 17 states across the country recruited children aged from less than 1 to 18 who had a serious emotional disturbance. The data were collected by interview from 2000 to 2002.

The study participants are limited to those who with serious emotional disturbance. Another selection criterion was involvement with multiple child-serving systems and greater risk for out-of-home placement (Macro International Inc, 1997). The children were enrolled in the evaluation study by self-referral or by external-referral through schools and mental health agencies.

The evaluation study requires the sites to choose at least 400 children. For those sites serving between 400 and 4,000 children, 400 children were drawn and included in the evaluation sample. For those sites serving over 4,000 children, a 10 percent sample was drawn (Macro International Inc, 1997). The data for this study include 1,188 children aged 4 to 18 at the first wave. Among these 1,188 children, 269 were further excluded from the analysis, because their respondents of the CBCL questionnaire were not biological parents, grandparents, or uncle/aunt who might share the same cultural backgrounds as the children. The final sample included 488 whites and 251 blacks. The final sample contained African-Americans and whites whose distributions of income and education of the parents match each other. Table 1 summarizes the descriptive statistics of socioeconomic characteristics of each sample.

Tests of the equivalence of validity

The equivalence of construct validity of the syndromes was tested across groups. Two different methods of testing the equivalence of construct validity were applied according to the two measurement models. Multiple group confirmatory factor analysis tested the equivalence of construct validity across groups for the effect indicator model, while principal

Table 1. Socioeconomic and Demographic Characteristics of the Cross-Racial/ethnic Study Sample

		White (n=195)	Black (n=198)
Child Gender	Boy	138 (70.8%)	139 (70.2%)
	Girl	57 (29.2%)	59 (29.8%)
Child Age	Mean	12.0	11.6
	4-11	71 (36.4%)	92 (46.5%)
	12-18	124 (63.6%)	106 (53.5%)
Respondent's Age	Mean	40.3	39.4
Relationship to Child			
	Biological Parent	172 (88.2%)	165 (83.3%)
	Grandparent	17 (8.7%)	21 (10.6%)
	Aunt or Uncle	6 (3.1%)	12 (6.1%)
Annual Income			
	Less than $5,000	21 (11.2%)	41 (22.2%)
	$5,000-$9,999	32 (17.0%)	33 (17.8%)
	$10,000-$14,999	31 (16.5%)	45 (24.3%)
	$15,000-$19,999	21 (11.2%)	28 (15.1%)
	$20,000-$24,999	21 (11.2%)	15 (8.1%)
	$25,000 and over	62 (32.9%)	23 (12.5%)
Respondent's Education			
	Less than High School	86 (47.3%)	26 (48.5%)
	High School Diploma	44 (23.7%)	54 (27.8%)
	Associate Degree	10 (5.4%)	15 (7.7%)
	Some College	33 (17.7%)	18 (9.3%)
	Bachelors Degree	3 (1.6%)	9 (4.6%)
	Graduate Degree	10 (4.3%)	4 (2.1%)
Eligible for Medicaid		113 (59.2%)	154 (78.6%)
S-CHIP		40 (21.5%)	33 (16.8%)
TANF		25 (13.2%)	43 (21.8%)

note) S-CHIP: State-Children's Health Insurance Program, TANF: Temporary Aid for Needy Family component analysis tested the same thing for the causal indicator model.

Tests of the equivalence of construct validity under effect indicator model

In multiple group analysis, we can test if a specific parameter is the same across groups. One of the assumptions made in Joreskog's multiple group analysis model is that the observed variables follow a multivariate normal distribution in each subpopulation. However, a lot of measures for psychological testing do not meet this assumption. Most of the psychological measures are rating scales with polychotomous items. The CBCL, for example, includes three categories in each item; 0 (not true), 1 (somewhat or sometimes true), and 2 (very true or often true). Rating scales like Likert variables may not have equidistant scale steps or more importantly, the strong skewness of some variables will distort ordinary Pearson product moment correlations (or covariance). When the normality assumption is violated, we cannot use the log-likelihood function proposed in Joreskog's model and therefore, chi-square goodness-of-fit statistic is not correct.

Estimator

Maximum Likelihood estimator is not appropriate for categorical variables because of the non-normality of categorical variables. Weighted Least Square (WLS) is, instead, often used for categorical variables. The fitting function, which is minimized to obtain consistent estimators of parameters, of WLS includes a weight matrix as well as the sample covariance matrix and the implied covariance matrix of structural parameters.

WLS is known to be inferior to Weighted Least Square Mean and Variance-corrected (WLSMV) when the sample size is small. WLS and WLSMV use the same weight matrix, but WLSMV uses only the diagonals in the weight matrix, whereas WLS uses the entire matrix. In small samples, additional information in WLS is noise, which may cause non-convergence. Since the sample size was not substantial, WLSMV was used in this study.

Testing Hypotheses

Multiple-group analysis is very flexible method to test numerous hypotheses of invariance across different populations. The process, however, can be centered on two kinds of invariance: measurement invariance and structural invariance (Byrne, Shavelson, & Muthen, 1989). The measurement invariance concerns the invariance of regression intercepts, factor loadings (regression slopes), and error/uniqueness variances (Byrne et al., 1989). The structural invariance addresses the invariance of factor mean and factor variance-covariance structures (Byrne et al., 1989).

The following hypotheses in each of measurement invariance and structure invariance are hierarchically ordered in that if the current hypothesis is rejected, the subsequent hypothesis does not need to be tested.

The first hypothesis is for testing *configural invariance*. If configural invariance exists, then each group has the same number of factors with the same items associated with each factor (Cheung & Rensvold, 2002; Meredith, 1993). One assumption underlying here is that the factor pattern matrix is a reasonable empirical representation of our concept. The model for the test of configural invariance is the baseline model, which will serve as a benchmark against which more restricted models are compared. If the data do not satisfy configural invariance, no further invariance tests are necessary.

The second hypothesis is related to metric invariance (or weak factorial invariance). Metric invariance exists when all factor loading parameters are equal across groups, whereas all the other parameters are free to vary (i.e., $\Lambda_1 = ... = \Lambda_g$, where g=gth group) (Cheung & Rensvold, 2002). The constrained model ($\Lambda_1 = ... = \Lambda_g$) across groups is tested against the baseline model. The premise underlying metric invariance is that factor loadings are regression slopes which represent the change in the response in the observed scores to their latent variable. For example, even if different racial/ethnic groups have the same items for Delinquency, specific behaviors of Delinquency might not be present to the same degree, possibly because the linkage between these behaviors and the underlying construct is weaker for one group than the others. As a result, if alcohol use and drug use are less frequently observed among blacks than among whites, the factor loadings of these two items of blacks may be significantly smaller than those of whites.

The third hypothesis proposes that the vectors of item intercepts are also invariant. The item intercepts are the values of each item corresponding to the zero value of the underlying construct. The invariance of item intercepts as well as factor loading parameters is also called *strong factorial invariance* (Meredith, 1993). Strong factorial invariance is the prerequisite for the comparison of latent means (Cheung & Rensvold, 2002). Without strong factorial invariance, the comparison of latent means is ambiguous, because the effects of a between-group difference in the latent means are confounded with differences in the scale and origin of the latent variable.

The last hypothesis for measurement invariance is for testing *strict factorial invariance* (Meredith, 1993). Additional constraint on the unique variances is imposed in addition to unique means and factor loadings. This model forces the unique variances, i.e., the combined specific and random error components of each variable, to be equivalent across groups such that differences in variance across groups are permitted only at the latent variable level. In other words, mean differences in the items between groups will all be conveyed through mean differences in the common factor among groups.

There are two hypotheses for testing structural invariance. First, one can test if the variance/covariance of constructs (i.e., latent variables) is invariant across groups. This hypothesis must be supported before comparing the correlations of constructs across groups (Byrne et al., 1989). Equal variance of a construct is important as a prerequisite of comparing item reliability (Vanderberg & Lance, 2000).

Second, one can hypothesize that the latent means are invariant across groups. As stated above, strong factorial invariance is the prerequisite of this hypothesis test. The model constraining on factor loadings, item intercepts, and construct means is tested against the strong factorial invariance model.

For the purpose of testing construct validity across groups, only configural invariance and metric invariance were tested. These two hypotheses were considered to be more pertinent to our cognitive process of conceptualizing a construct (Vandenberg, 2002), whereas the other hypotheses were more necessary for mean comparisons (intercepts) or reliability comparisons (e.g., invariance of covariance/correlation and errors). In other words, configural invariance and metric invariance can answer the questions, first, if various racial/ethnic groups have the same indicators representing the construct of interest and, second, if the strength of the representativeness of each indicator is the same across racial/ethnic groups.

We cannot use chi-square estimate for the model when WLSMV is applied because it uses mean- and covariance-adjusted robust chi-square (Muthen & Muthen, 1998-2001). Difference in CFI (>.01) was the criterion instead.

Tests of the equivalence of construct validity under causal indicator model

Principal component analysis is useful to transform the original variables into uncorrelated variables or to reduce the number of variables. Suppose we have p variables. The maximum number of components that we can obtain from these p variables is p. Components are the linear combinations of the original variables. Coefficients of each component (Ci) are decided in such a way that satisfies the following three requirements (Afifi & Clark, 1996):

A. Variance (C1) >Variance (C2)>…> Variance (Cp)

B. All the pairs of two components are uncorrelated.

C. The sum of the squares of the coefficients for any component is one.

Once identifying the components, we can construct the scores of the components by summing up the scores of the original variables that belong to each component. In this way, we can reduce the number of variables. In addition, we can be sure that these transformed variables are not correlated each other.

There are several criteria that determine the number of components. Some people choose the number of components from scree plots where the variance of the component, i.e., eigenvalue, is plotted on the vertical axis and the component number is plotted on the horizontal axis. They choose the number of components at the joining point where its left side is steep and its right side is flat. Other people discard components that have variance less than 5% of the total variance (Afifi & Clark, 1996). In this study, eight components were chosen, because Achenbach (1991) retained eight components from the CBCL. In this way, we can examine whether the current samples replicate the original components.

Principal component analysis is an appropriate analytic method for testing the equivalence of construct validity when the items are causal indicators. First, like the causal indicator model, principal component analysis does not assume any a priori underlying constructs that cause the variables. The sum of variance of components is the same as the sum of variance of the original variables, which means that principal component analysis is merely a linear transformation of the original variables.

Second, causal indicators do not have to be correlated each other. High correlations among causal indicators only cause problems such as multicollinearity. If causal indicators are not highly correlated each other, the scores of some summated scales out of these indicators should not be highly correlated, either. Therefore, principal component analysis that provides uncorrelated components is appropriate for the analysis of causal indicators.

Results

Tests of the equivalence of construct validity using multiple group confirmatory factor analysis

Configural invariance was tested when the CBCL items were treated as categorical variables. Before testing measurement invariance using multiple group analysis, it is important to find a reasonably good model for all the groups (Byrne et al., 1989). Even though one-factor model for each subscale of the CBCL has been assumed in previous studies, unidimensionality was examined before any conclusions regarding measurement invariance. Song et al. (1994) suggested that the root mean square residual might be a useful tool for determining unidimensionality of a scale. More than one factor is suspected, when (1) the RMSR of one-factor model is large (>.07), (2) the ratio of the RMSR's for the one-factor model and the two-factor model is less than 3, and (3) the difference of the two RMSR's is less than .02. Song et al.(1994) also suggested that the ratio of eigenvalues that is less than 2 may imply the existence of more than one factor. According to these three criteria, a two-factor model was better for Withdrawn, Delinquent Behaviors, Aggressive Behaviors, Social Problems, and Attention Problems.

Since chi-squares and degrees of freedom cannot be used when WLSMV estimator was applied, only CFI, TLI, and RMSEA were included in Table 2. Tucker Lewis Index (TLI), also known as non-normed fit index (Bentler & Bonett, 1980), was proposed originally by (Tucker & Lewis, 1973). Hu and Bentler (1999) recommended a cutoff value of TLI close to 0.95. Bentler (1990) proposed comparative fit index (CFI). Hu and Bentler (1999) recommended that CFI be close to 0.95. It has been recommended that root mean square error of approximation (RMSEA: Steiger,1990) be smaller than 0.06 (Hu & Bentler, 1999) or in the range of 0.05 and 0.08 (Browne & Cudeck, 1992). Therefore, in this study, cutoff values of 0.95, 0.95, and 0.06 were adopted as the bound for TLI, CFI, and RMSEA, respectively.

Table 2 shows that only Somatic Complaints and the two factor model of Social Problems met configural invariance. Therefore, further metric invariance testing equal factor loadings was tested on these two scales.

Table 2. Model Fit Indices for Configural Invariance Models

	one-factor			two-factor		
	CFI	TLI	RMSEA	CFI	TLI	RMSEA
Withdrawn	.88	.90	.09	.93	.93	.07
Somatic	.98	.99	.04			
Anxious/depressed	.93	.96	.09			
Delinquent	.90	.92	.09	.94	.95	.07
Aggressive	.89	.95	.12	.92	.97	.09
Social	.92	.93	.11	.96	.96	.08
Thought	.95	.94	.15			
Attention	.88	.91	.11	.92	.94	.09

Test of metric invariance

Since a test of chi-square differences is not allowed with WLSMV estimator, the difference in CFI was used as the criterion of a better model. In other words, if the CFI when constraining the factor loadings to be equal between whites and blacks increased significantly, namely ΔCFI>-0.01, compared to the configural invariance model, then the factor loadings of the scale were considered to be equal. However, if ΔCFI was less than -0.01, then the metric invariance model was considered to significantly become worse than configural invariance, hence metric invariance model was rejected.

Since only Somatic Complaints and the two factor model of Social Problems met configural invariance, only these two were tested for metric invariance(Table 3). Both scales met metric invariance(ΔCFI>-0.01).

Tests of the equivalence of construct validity using principal component analysis

Principal component analysis was used to find the underlying constructs from the causal indicators. Differences in the patterns of component coefficients across races implied differences in the underlying constructs. Since 5 items in the CBCL 4/18 were included in two syndromes(items 1, 45, 62, 80, 103), 80 different items were included in the analysis. The 80 items were analyzed to see if the sample replicated the eight syndromes in the two racial groups.

Principal component analysis with the white sample

The first component almost replicated Aggressive Behaviors. Among the 20 items, two items, *no guilt* and *lie/cheat*, came from Delinquent Behaviors and two each from Social Problems (*not get along*) and Attention Problems (*impulsive*). At the same time, only three items of original Aggressive Behaviors (*destroy own things*, *show off*, and *talk much*) did not belong to this component. For these reasons, this component was named "aggressive behaviors."

The second component consisted of 12 items was named "anxiety/depression." Among these

Table 3. Results of the Tests of Metric Invariance

	CFI		
	configural	metric	Δ
Somatic	.983	.991	+0.008
Social(two-factor)	.960	.967	+0.007

Note: Δ <-0.01 rejected metric invariance

12 items, 3 items came from Withdrawn(*sulks, sad, withdrawn*), and 2 from Social Problems(*teased, not liked*). The third component almost replicated Delinquent Behaviors. Only two items, *secretive* and *clings*, originally belonged to Withdrawn and Social Problems, respectively. In addition, only two original items of Delinquent Behaviors were missing (*no guilt*, *lie/cheat*), which were found in "aggressive behaviors." As a result, the third component was named "delinquency."

The white sample exactly replicated Somatic Complaints in the forth component. The fifth component, which was named "psychosis,"

Table 4. Principal Component Analysis of the White Sample (n=159)

1 aggressive behaviors		2 anxiety/depression	3 delinquency
26 no guilt[d] (.389)	57 attacks[e] (.743)	88 sulks[a] (.460)	69 secretive[a] (.385)
43 lie/cheat[d] (.464)	68 screams[e] (.682)	103 sad[a] (.761)	39 bad company[d] (.451)
90 swears[d] (.609)	86 stubborn[e] (.459)	111 withdrawn[a] (.407)	63 prefers older[d] (.334)
3 argues[e] (.629)	87 mood change[e] (.466)	12 lonely[b] (.464)	67 run away[d] (.559)
7 brags[e] (.453)	94 teases[e] (.604)	33 unloved[b] (.606)	72 set fires[d] (.325)
16 mean[e] (.799)	95 temper tantrum[e] (.617)	34 out to get[b] (.495)	81 steal home[d] (.573)
19 demand attention[e] (.499)	97 threaten[e] (.812)	35 worthless[b] (.727)	82 steal out[d] (.452)
21 destroy others[e] (.609)	104 loud[e] (.589)	71 self-conscious[b] (.556)	101 truant[d] (.431)
23 disobey school[e] (.575)	25 not get along[f] (.523)	89 suspicious[b] (.324)	105 alcohol/drug[d] (.695)
27 jealous[e] (.300)	41 impulsive[h] (.566)	112 worries[b] (.524)	11 clings[f] (-.410)
37 fights[e] (.683)		38 teased[f] (.442)	
		48 not liked[f] (.555)	
4 somatic complaints	**5 psychosis**	**6 social problems**	**7.**
51 dizzy[c] (.550)	80 stares[a] (.540)	32 perfect[b] (-.542)	42 rather alone[a] (.408)
54 tired[c] (.486)	14 cries[b] (.374)	52 guilty[b] (-.506)	75 shy[a] (.306)
56a aches[c] (.643)	31 fear do bad[b] (.357)	1 acts young[f] (.381)	45 nervous[b] (.651)
56b headaches[c] (.666)	50 fearful[b] (.457)	62 clumsy[f] (.413)	20 destroy own[e] (.432)
56c nausea[c] (.716)	9 mind off[g] (.559)	61 poor school[h] (.305)	10 sit still[h] (.460)
56d eye[c] (.261)	40 hears things[g] (.680)		**8.**
56e skin[c] (.369)	66 repeats acts[g] (.450)		65 won't talk[a] (-.470)
56f stomach[c] (.681)	70 sees things[g] (.651)		102 underactive[a] (-.342)
56g vomit[c] (.511)	84 strange behavior[g] (.652)		74 show off[e] (.435)
	85 strange ideas[g] (.574)		93 talk much[e] (.503)
	8 concentrate[h] (.340)		64 prefers young[f] (.303)
	13 confused[h] (.500)		
	17daydream[h] (.529)		

note: Item numbers indicate the CBCL item numbers. Items in each component are listed by the order of the CBCL 4/18 syndromes (see Appendix). The superscripts, a, b, c, d, e, f, g, h represent withdrawn, anxious/depressed, somatic complaints, delinquent behaviors, aggressive behaviors, social problems, thought problems, and attention problems, respectively, in the CBCL 4/18.

included all the original six items of Thought Problems. This component also involved Anxious/depressed(*cries, fear do bad, fearful*), Withdrawn(*stares*), and Attention Problems (*concentrate, confuse, daydream*).

The sixth component, "social problems," consisted of five items; two each from Anxious/depressed(*perfect, guilty*) and Social Problems(*acts young, clumsy*) and one from Attention Problems(*poor school*). The last two components seemed to represent some kinds of covert aggressive problems mixed with items from Aggressive Behaviors and Withdrawn.

In sum, only four CBCL syndromes, Anxiety/depression, Somatic Complaints, Aggressive Behaviors, and Delinquent Behaviors were replicated in the white sample, though the consisting items were not exactly the same as those in the CBCL syndromes. Thought Problems syndrome hung together with internalizing problems and was named "psychosis." Social problems was not the same as the CBCL Social problems, either. Withdrawn and Attention Problems did not appear in the white sample. Instead, two components representing covert aggressive behaviors emerged.

Principal component analysis with the black sample

The eight components for the black sample were presented in Table 5. The first component, which was named "conduct problems," consisted of items dominantly from Aggressive Behaviors and some from Delinquent Behaviors(*no guilt, bad company, lie/cheat, sets fire, swears*), Attention Problems(*concentrate, sit still, impulsive, poor school*), Social Problems(*not get along, not liked*), and Thought Problems (*strange behavior*).

The second component was named "withdrawn." In addition to four items of the original Withdrawn scale, two from Anxious/depressed(*nervous, fearful*), two from Social Problems (*acts young, clumsy*), one each from Thought Problems(*repeats acts*) and Attention Problems(*day dream*) were also included in this component.

The third component was dominantly consisted of items from Anxious/depressed. So it was called "anxiety/depression." It was different from the black sample that *hears things* and *sees things* were included in this construct rather than "psychosis." Two Withdrawn items(*rather alone, sad*), one Social Problems item(*teased*), and one Attention Problems item(*confuse*) also hung together in this component.

The forth component retained all the items of Somatic Complaints. In addition, *underactive* was also in this component. The fifth component included 3 items from Delinquent Behaviors(*run away, steal home, steal out*), and so it was named "delinquency." We note that one withdrawn item, *withdrawn* and one Thought Problem item, *strange ideas*, hung together with these 3 delinquency items. Loeber et al.(1998) found that shy/withdrawn behavior were related to delinquent behavior such as stealing.

The sixth component represented "covert aggression." This component included six Aggressive Behaviors items in the CBCL(*brags, demand attention, jealous, screams, talk much, loud*). Other items were from Anxious/depressed (*perfect*), Delinquent Behaviors(*prefers older*), Social Problems(*clings*), and Thought Problems (*mind off*).

Table 5. Principal Component Analysis of the Black Sample (n=163)

1 conduct problems		2 withdrawn	3 anxiety/depression
26 no guilt[d] (.584) 39 bad company[d] (.461) 43 lie/cheat[d] (.477) 72 sets fire[d] (.319) 90 swears[d] (.458) 3 argues[e] (.537) 16 mean[e] (.693) 20 destroy own[e] (.554) 21 destroy other[e] (.663) 23 disobey school[e] (.689) 37 fights[e] (.674) 57 attacks[e] (.733) 74 show off[e] (.479) 86 stubborn[e] (.546)	87 mood change[e] (.471) 94 teases[e] (.395) 95 temper tantrum[e] (.532) 97 threaten[e] (.610) 25 not get along[f] (.621) 48 not liked[f] (.376) 84 strange behavior[g] (.384) 8 concentrate[h] (.543) 10 sit still[h] (.475) 41 impulsive[h] (.401) 61 poor school[h] (.475)	65 won't talk[a] (.333) 69 secretive[a] (.449) 75 shy[a] (.537) 80 stares[a] (.639) 45 nervous[b] (.637) 50 fearful[b] (.417) 1 acts young[f] (.654) 62 clumsy[f] (.350) 66 repeats acts[g] (.360) 17 day dream[h] (.478)	42 rather alone[a] (.226) 103 sad[a] (.762) 12 lonely[b] (.494) 14 cries[b] (.456) 31 fear do bad[b] (.450) 33 unloved[b] (.593) 34 out to get[b] (.530) 35 worthless[b] (.738) 52 guilty[b] (.432) 71 self-conscious[b] (.423) 89 suspicious[b] (.406) 112 worries[b] (.390) 38 teased[f] (.451) 40 hears things[g] (.542) 70 sees things[g] (.519) 13 confused[h] (.538)
4 somatic complaint	5 delinquency	6 covert aggression	7.
102 underactive[a] (.454) 51 dizzy[c] (.601) 54 tired[c] (.514) 56a aches[c] (.643) 56b headaches[c] (.666) 56c nausea[c] (.716) 56d eye[c] (.455) 56e skin[c] (.470) 56f stomach[c] (.681) 56g vomit[c] (.511)	111 withdrawn[a] (.384) 67 run away[d] (.377) 81 steal home[d] (.758) 82 steal out[d] (.729) 85 strange ideas[g] (.349)	32 perfect[b] (.471) 63 prefers older[d] (.380) 7 brags[e] (.517) 19 demand attention[e] (.439) 27 jealous[e] (.500) 68 screams[e] (.447) 93 talk much[e] (.658) 104 loud[e] (.605) 11 clings[f] (.512) 9 mind off[g] (.477)	88 sulks[a] (-.404) 64 prefers young[f] (-.447) **8.** 101 truant[d] (.697) 105 alcohol/drug[d] (.757)

note: Item numbers indicate the CBCL item numbers. Items in each component are listed by the order of the CBCL 4/18 syndromes (see Appendix). The superscripts, a, b, c, d, e, f, g, h represent withdrawn, anxious/depressed, somatic complaints, delinquent behaviors, aggressive behaviors, social problems, thought problems, and attention problems, respectively, in the CBCL 4/18.

In sum, like the white sample, the black sample did not replicate the original CBCL syndromes. Especially, mixed problems syndromes such as Social Problems, Thought Problems, and Attention Problems did not emerge. Among externalizing problems syndromes, Aggressive Behaviors and Delinquent Behaviors were reorganized and emerged as three new syndromes: conduct problems, delinquency, and covert aggression. On the other hand, three Internalizing syndromes, Withdrawn, Anxiety/depression, and Somatic Complaints, were relatively well replicated.

Discussion

Based on the two measurement models, it was tested that construct validity of the cross-informant CBCL syndromes was equivalent between whites and blacks. Under the assumption that the CBCL items were effect indicators and causal indicators, multiple group analysis and principal component analysis were used, respectively.

Multiple group analysis indicated that only Somatic Complaints and two-factor model of Social Problems had identical items and the identical strength of relationship with the items for the two racial groups. One explanation of this result might be that the other six syndromes had different concept between whites and blacks and therefore consisted of different items for each syndrome. The lack of configural invariance might also be due to non-clarity of a construct to some of or all the study groups.

The results of principal component analyses did not quite replicate the original eight syndromes in both groups. There were at least two reasons of these results. First, Achenbach (1991) developed the eight syndromes based on four separate principal component analyses for the four groups based on age and gender, while this study used a whole range of ages and both genders. Achenbach found the common components and items across age and gender groups and therefore there was no room for interactions of age and gender with the items. Since this study used the whole range of ages and both genders, age and gender might have interacted with items and thus affected the correlations among the items and the components found.

In addition, the relatively small sample size might undermine the reliability of the principal component analysis results. It is conventionally believed that at least 5 to 10 times larger observations than the number of items are needed for principal component analysis. Future studies with a larger sample size should replicate this study to see if we can obtain the same components as in this study.

Nevertheless, the eight syndromes that each of the white sample and the black sample retained had some similarities. First, Somatic Complaints, though there was one item difference, emerged in both samples. This result coincided with the result from the multiple group analysis. Second, Attention Problems did not emerge in either group.

However, there were more differences in the components found between the two groups than similarities. Except Somatic Complaints, basically all the other components did not match in the two groups. For example, while Aggressive Behaviors was almost exactly replicated as the first component among the whites, Aggressive Behaviors was manifested in two components (covert and overt conduct problems) for the blacks. The component for "delinquency" in the black sample was different from the other group in that it did not include *truant* and *alcohol/drug use* in it. The black sample also had different patterns with *hears things* and *sees things* from the white group. These two items belonged to a component dominated by internalizing problems in the black sample. However, they belonged to a component that could be named "psychosis,"in the other group. There were no components in the black sample that might represent mixed problems such as Social Problems, Thought Problems, or Withdrawn.

In conclusion, both confirmatory factor analysis and principal component analysis are useful to find that several CBCL syndromes have differential construct validity. It can also be concluded that the two measurement models can be complementary in that principal component analysis can give insights into how the items correlate each other across the syndromes, when multiple group analysis rejected configural invariance. This study results showed that except Somatic Complaints all the other syndromes may not be appropriate to compare white and black children in their emotional and behavioral problems. Social Problems may be appropriate to compare the two groups, when it is divided to two factors. One way of avoiding the problems of differential construct validity of the CBCL syndromes may be using these syndromes with other closely related constructs, for example, Withdrawn with Anxious/depressed, Delinquent with Aggressive, and Attention with Social or Thought. In other words, broadband scale scores are preferred to syndrome scores in racial comparisons. Alternatively, we should consider constructing "cross-race syndromes" like the cross-informant syndromes that Achenbach(1991) suggested. Cross-race syndromes should be comprised of common items across different racial groups for a given children's emotional and behavioral problem.

The study results also shed some lights on the adoption of psychological measures developed in other countries. Since the study results showed that a psychological measure may not have equal validity across different races which have different culture and norm, it is possible that a measure may not be equally valid across samples of different nationalities. We should be cautious in using health-related psychological measures adopted from foreign countries, questioning whether the measure is culturally valid in our country.

Finally, it needs to be addressed that a single study cannot prove or disapprove the validity of a measure. Instead, we need accumulated evidence from various studies on the internal structure and the relationship with other variables in order to support the validity of the measure. There are many other analyses of testing measurement equivalence than multiple group analysis and principal component analysis. For example, regression analysis can be a good tool to test the equivalent criterion validity across groups and multiple test and multiple methods(MTMM) can also be another method to test equivalent construct validity. Further studies using these methods will help us better understand the measurement equivalence of a measure.

Reference

Achenbach, T. M. (1991). *Manual for the Child Behavior Checklist/4-18 and 1991 Profile*: Department of Psychiatry, University of Vermont.

Achenbach, T. M., & McConaughy, S. H. (1997). Empirically Based Assessment of Child and Adolescent Psychopathology: Practical Applications (2 ed.) (Vol. 13). Thousand Oaks, CA: SAGE publications.

Afifi, A. A., & Clark, V. (1996). Computer-aided multivariate analysis (3 ed.). New York, NY: Chapman & Hall.

Ben-Porath, Y. S. (1990). Cross-cultural assessment of personality: The case for replicatory factor analysis. In J. N. Butcher, Spielberger, C.D. (Ed.), Advances in personlity assessment (Vol. 8, pp. 27-85). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107(238-246).

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, 88(3), 588-606.

Bird, H. R. (1996). Epidemiology of childhood disorders in a cross-cultural context. *Journal of Child Psychology and Psychiatry*, 37(1), 35-49.

Blalock, H. M. (1964). Causal inferences in nonexperimental research. Chapel Hill: University of North Carolina Press.

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: a structural equation perspective. *Psychological Bulletin*, 110(2), 305-314.

Borsboom, D., Mellenbergh, G. J., & Heerden, J. v. (2003). The theoretical status of latent variables. *Psychological Review*, 110(2), 203-219

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, 21(2), 230-258.

Byrne, B. M., Shavelson, R. J., & Muthen, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, 105(3), 456-466.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9(2), 233-255.

Cohen, P., & Kasen, S. (1999). The context of assessment: Culture, race, and socioeconomic status as influences on the assessment of children. In D. Shaffer, Lucas, C.P., Richters, J.E. (Ed.), Diagnostic assessment in child and adolescent psychopathology (pp. 299-318). New York, NY: The Guilford Press.

Cole, N. S., & Moss, P. A. (1989). Bias in Test Use. In L. L. Robert (Ed.), Educational Measurement (3 ed., pp. 201-219). New York, NY: Macmillan.

Dedrick, R. F., Greenbaum, P. E., Friedman, R. M., Wetherington, C. M., & Knoff, H. M. (1997). Testing the structure of the Child Behavior Checklist/4-18 using confirmatory factor analysis. *Educational and Psychological Measurement*, 57(2), 306-313.

DeGroot, A., Koot, H. M., & Verhulst, F. C. (1994). Cross-cultural generalizability of the Child Behavior Checklist cross-informant syndromes.

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods,* 5(2), 155-174.

Fayers, P. M., & Hand, D. J. (1997). Factor analysis, causal indicators and quality of life. *Quality of Life Research*, 6, 139-150.

Fornell, C., & Bookstein, F. L. (1982). Two structural equation models: LISREL and PLS applied to consumer exit-voice theory. *Journal of Marketing Research*, 19, 440-452

Hartman, C. A., Hox, J., Auerbach, J., Erol, N., Fonseca, A. C., Mellenbergh, G. J., Novik, T. S., Oosterlaan, J., Roussos, A. C., Shalev, R. S., & Zilber, N. (1999). Syndrome dimensions of the Child Behavior Checklist and the Teacher Report Form: A critical empirical evaluation. *Journal of Child Psychology and Psychiatry*, 40(7), 1095-1116.

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

Hudziak, J. J., Wadsworth, M. E., Heath, A. C., & Achenbach, T. M. (1999). Latent class analysis of Child Behavior Checklist attention problems. *Journal of American Academy of Child and Adolescent Psychiatry*, 38(8), 985-991.

Lambert, M. C., Schmitt, N., Samms-Vaughan, M. E., An, J. S., Fairclough, M., & Nutter, C. A. (2003). Is it prudent to administer all items for each Child Behavior Checklist Cross-Informant syndrome? Evaluating the psychometric properties of the Youth Self-Report dimensions with confirmatory factor analysis and item response theory. *Psychological Assessment*, 15(4), 550-568.

Loeber, R., Farrington D. P., Stouthamer-Loeber M., & Van Kammen, W. B. (1998) Antisocial behavior and mental health problems. Explanatory factors in childhood and adolescence. Lawrence Erlbaum Associates, Inc., Mahwah, NJ.

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

MacCallum, R. C., & Browne, M. W. (1993). The use of causal indicators in covariance structure models: some practical issues. *Psychological Bulletin*, 114(3), 533-541.

Macro International Inc. (1997). Annual Report To Congress on the Evaluation of the Comprehensive Community Mental Health Services For Children and Their Families Program . Atlanta, GA: Macro International Inc.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4), 525-543.

Meredith, W. (1964). Notes on factorial invariance. *Psychometrika*, 29(2), 177-185.

Muthen, B. O., Hasin, D., & Wisnicki, K. S. (1993). Factor analysis of ICD-10 symptom items in the 1988 National Health Interview Survey on alcohol

Muthen, L. K., & Muthen, B. O. (1998-2001). Mplus User's Guide (2 ed.). Los Angeles, CA: Muthen & Muthen.

Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric theory (3 ed.). New York: McGroaw-Hill, Inc.

Reynolds, C. R., & Paget, K. D. (1981). Factor analysis of the Revised Children's Manifest Anxiety Scale for Blacks, Whites, Males, and Females with a national normative sample. *Journal of Counseling and Clinical Psychology*, 49(3), 352-359.

Song, L., Sing, J., & Singer, M. (1994). The Youth Self-report Inventory: A study of its measurement fidelity. *Psychological Assessment*, 6(3), 236-245.

Spector, P. E. (1992). Summated rating scale construction (Vol. 82). Newbury Park: SAGE.

Steiger, J. H. (1990). Structural model evaluation and modification: an interval estimation approach. *Multivariate Behavioral Research*, 25, 173-180.

Stewart, A. L., & Napoles-Springer, A. (2000). Health-related quality-of-life assessments in diverse population groups in the United States. *Medical Care*, 38(9 Supplement II), II102-II124

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38(1), 1-10.

Vandenberg, R. J. (2002). Toward a Further Understanding of and Improvement in Measurement Invariance Methods and Procedures. *Organizational Research Methods*, 5(2), 139-158.

Vanderberg, R. J., & Lance, C. E. (2000). A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. *Organizational Research Methods,* 3(1), 4-69.

Wadsworth, M. E., Hudziak, J. J., Heath, A. C., & Achenbach, T. M. (2001). Latent class analysis of Child Behavior Checklist anxiety/depression in children and adolescents. *Journal of American Academy of Child and Adolescent Psychiatry*, 40(1), 106-114.

논문접수일: 2007년 2월 12일
게재결정일: 2007년 3월 4일

한국심리학회지: 건강
The Korean Journal of Health Psychology
2007. Vol. 12, No. 1, 241 - 267

측정 동등성을 평가하는 두 가지 모형: Child Behavior Checklist의 예

강 은 정[1] E. Michael Foster[2] Michael J. Rovine[3]
1) 한국보건사회연구원 부연구위원
2) University of North Carolina at Chapel Hill, Professor
3) The Pennsylvania State University, Associate Professor

아동 및 청소년의 정신건강에서, 결과변수를 측정하는데 많은 심리학적 도구들이 사용된다. 이때 다양한 인구집단에서 동등하게 타당한 도구를 사용하는 것은 보건의료서비스의 결과 평가에 있어서 결정적으로 중요하다. 아동 및 청소년의 정서, 동기부여, 태도, 대인 특성 등의 측정 도구에 있어서 동등한 타당성의 검토는 요인분석을 통해 이루어질 수 있다. 요인분석은 특정한 측정 모형인 '효과 지표 모형'을 가정한다. 하지만 Child Behavior Checklist와 같은 어떤 도구들은 이러한 모형에 기반 하지 않은 것들이 있다. 이 연구에서는 대안적인 모형인 '인과 지표 모형'을 소개하고 있다. '효과 지표 모형'을 근거로 한 다중 집단 분석과 '인과 지표 모형'을 근거로 한 주요인분석을 각각 사용하여 CBCL의 측정 동등성을 분석하였다. 분석 결과 CBCL의 경우 두 모형에 근거로 한 측정 동등성 결과는 비슷하였다. 결론적으로 이 두 측정 모형은 서로 다른 하위집단 사이의 타당성의 차이를 설명하는데 보완적인 자료로 사용될 수 있다고 보인다.

주요어 : 효과 지표 모형, 인과 지표 모형, 측정 동등성, CBCL

Appendix.

Original wordings of the Cross-informant CBCL items

(Source: Achenbach(1991). Manual for the Child Behavior Checklist 4/18 and 1991 Profile. pp.11-12)

Withdrawn
CBCL42 Would rather be alone than with others
CBCL65 Refuses to talk
CBCL69 Secretive, keeps things to self
CBCL75 Shy or timid
CBCL80 Stares blankly
CBCL88 Sulks a lot
CBCL102 Underactive, slow moving, or lacks energy
CBCL103 Unhappy, sad, or depressed
CBCL111 Withdrawn, doesn't get involved with others

Anxious/Depressed
CBCL12 Complains of loneliness
CBCL14 Cries a lot
CBCL31 Fears he/she might think or do something bad
CBCL32 Feels he/she has to be perfect
CBCL33 Feels or complains that no one loves him/her
CBCL34 Feels others are out to get him/her
CBCL35 Feels worthless or inferior
CBCL45 Nervous, highstrung, or tense
CBCL50 Too fearful or anxious
CBCL52 Feels too guilty
CBCL71 Self-conscious or easily embarrassed
CBCL89 Suspicious
CBCL103 Unhappy, sad, or depressed
CBCL112 Worries

Somatic Complaints
CBCL51 Feels dizzy
CBCL54 Overtired

CBCL56a Aches or pains (not headaches)
CBCL56b Headaches
CBCL56c Nausea, feels sick
CBCL56d Problems with eyes (describe) _____
CBCL56e Rashes or other skin problems
CBCL56f Stomachaches or cramps
CBCL56g Vomiting, throwing up

Delinquent Behaviors
CBCL26 Doesn't seem to feel guilty after misbehaving
CBCL39 Hangs around with others who get in trouble
CBCL43 Lying or cheating
CBCL63 Prefers being with older kids
CBCL67 Runs away from home
CBCL72 Sets fires
CBCL81 Steals at home
CBCL82 Steals outside home
CBCL90 Swearing or obscene language
CBCL101 Truancy, skips school
CBCL105 Uses alcohol or drugs for nonmedical purposes (describe) _____

Aggressive Behaviors
CBCL3 Argues a lot
CBCL7 Bragging, boasting
CBCL16 Cruelty, bullying, or meanness to others
CBCL19 Demands a lot of attention
CBCL20 Destroys his/her own things
CBCL21 Destroys things belonging to his/her family or others
CBCL23 Disobedient at school
CBCL27 Easily jealous
CBCL37 Gets in many fights
CBCL57 Physically attacks people
CBCL68 Screams a lot
CBCL74 Showing off or clowning
CBCL86 Stubborn, sullen, or irritable
CBCL87 Sudden changes in mood or feelings

CBCL93 Talks too much
CBCL94 Teases a lot
CBCL95 Temper tantrums or hot temper
CBCL97 Threatens people
CBCL104 Unusually loud

Social Problems
CBCL1 Acts too young for his/her age
CBCL11 Clings to adults or too dependent
CBCL25 Doesn't get along with other kids
CBCL38 Gets teased a lot
CBCL48 Not liked by other kids
CBCL62 Poorly coordinated or clumsy
CBCL64 Prefers being with younger kids

Thought Problems
CBCL9 Can't get his/her mind off certain thoughts; obsessions (describe) ______
CBCL40 Hears sounds or voices that aren't there (describe) ______
CBCL66 Repeats certain acts over and over; compulsions (describe) ______
CBCL70 Sees things that aren't there (describe) ______
CBCL84 Strange behavior (describe) ______
CBCL85 Strange ideas (describe) ______

Attention Problems
CBCL1 Acts too young for his/her age
CBCL8 Can't concentrate, can't pay attention for long
CBCL10 Can't sit still, restless, or hyperactive
CBCL13 Confused or seems to be in a fog
CBCL17 Daydreams or gets lost in his/her thoughts
CBCL41 Impulsive or acts without thinking
CBCL45 Nervous, highstrung, or tense
CBCL61 Poor school work
CBCL62 Poorly coordinated or clumsy
CBCL80 Stares blankly