Educational and Psychological Measurement
Volume XX Number X
Month XXXX xx-xx
© 2007 Sage Publications
10.1177/0013164407310127
http://epm.sagepub.com
hosted at
http://online.sagepub.com

Internalizing and Externalizing Behavior Problem Scores

Cross-Ethnic and Longitudinal Measurement Invariance of the Behavior Problem Index

Katarina Guttmannova
Jason M. Szanyi
Northwestern University
Philip W. Cali
University of Illinois at Chicago

Accurate measurement of behavioral functioning is a cornerstone of research on disparities in child development. This study used the National Longitudinal Survey of Youth 1979 (NLSY79) data to test measurement invariance of the Behavior Problem Index (BPI) during middle childhood across three ethnic groups. Using the internalizing and externalizing behavior problem division derived by Parcel and Menaghan (1988) and suggested for use with NLSY79 data, the configural invariance hypothesis was not supported. The BPI factor structure model was revised based on theoretical considerations using the division of items from the Child Behavior Checklist. This model demonstrated configural invariance across ethnic groups and over time. Moreover, measurement invariance of factor loadings and thresholds across ethnic groups at each time point and within each ethnic group over time was also supported. The implications of these findings for educational and cross-cultural research are outlined.

Keywords: *behavior problems; internalizing; externalizing; measurement invariance; ethnicity; longitudinal*

Behavioral functioning represents a key developmental outcome and serves as a strong predictor of future adjustment. A majority of child development researchers agree on a classification of behavior problems that distinguishes between internalizing and externalizing manifestations of dysfunction (Cicchetti & Toth, 1991). Externalizing behavior problems—behaviors characterized by an undercontrol of emotions—include difficulties with interpersonal relationships and rule breaking as well as displays of irritability and belligerence (Achenbach & Edelbrock, 1978;

Authors' Note: Please address correspondence to Katarina Guttmannova, 2040 Sheridan Road, Evanston, IL 60208; e-mail: k-guttmannova@northwestern.edu.

Hinshaw, 1992). Conversely, internalizing behavior problems—defined as an over-control of emotions—include social withdrawal, demand for attention, feelings of worthlessness or inferiority, and dependency (Achenbach & Edelbrock, 1978; McCulloch, Wiggins, Joshi, & Sachdev, 2000).

Research consistently links both types of behavior problems to broad issues during middle childhood, including a lack of effortful control (Murray & Kochanska, 2002), peer rejection (Wood, Cowan, & Baker, 2002), and scholastic difficulties (Hinshaw, 1992). Behavior problems are also an important predictor of maladjustment in later life. For example, long-term associations exist between childhood externalizing behavior problems and substance abuse (King, Iacono, & McGue, 2004), smoking (Helstrom, Bryan, Hutchison, Riggs, & Blechman, 2004), antisocial outcomes (Lynam, 1996; Moffitt, 1993), underachievement (Hinshaw, 1992), and lower graduation rates from high school (for review, see McLeod & Kaiser, 2004). Childhood internalizing behavior problems are linked to major depression in adolescence (Reinherz et al., 1993), initiation of illegal substance use in early adolescence (King et al., 2004), and elevated risks of high school dropout (McLeod & Kaiser, 2004).

Behavior problems affect a substantial number of children. In a recent assessment of behavior problems based on a nationally representative sample of children, Achenbach, Dumenci, and Rescorla (2003) found that more than one in five children exhibited general behavior problems in the borderline or clinical range based on parental reports. However, there is significant variability over time in the persistence and consequences of behavior problems for individual children (for review, see Ackerman, Brown, & Izard, 2004). For example, the manifestation of behavior problems can change over time as children move from one developmental phase to another and broaden their behavioral repertoire. Additionally, developmentalists emphasize the importance of viewing the appropriateness of individual behavior within the proper context (i.e., the goodness of fit between a person's behavior and his or her social environment). As children get older, the environment changes, as do external demands and behavioral expectations. Consequently, the meaning and composition of the behavior problem constructs might change over time. Behaviors considered normative during one developmental phase might be considered inappropriate or even pathological at other stages of development (e.g., Barr, 2000). For example, crying would be considered an appropriate developmental cue for food or sleep in infancy but would not be a normative means to an end in middle childhood. Or, to highlight the importance of the goodness-of-fit concept, a behavior that would be labeled lively and spirited in a 2-year-old child might be considered hyperactive in a 4-year-old preschooler.

Although the clinical criteria for problem behavior can change over time, individual perceptions of what constitutes problem behavior can also vary. This concern is particularly important because most research on childhood behavior problems utilizes parental reports, and parents' perceptions of the appropriateness, severity, and quality of their child's behaviors are influenced by many factors. It is possible that the measurement and assessment of behavior problems could be influenced by cultural factors that vary by ethnicity (Spencer, Fitch, Grogan-Kaylor, & McBeath, 2005). For example, the experience and manifestation of depression has been found to vary across cultural groups, with non-European populations being more likely to experience somatic and quasi-somatic symptoms as opposed to the "expected" feelings of worthlessness, inferiority, and guilt (for review of this and other relevant ethnocultural examples, see Harkness & Super, 2000).

The assessment of behavior problems might be further complicated by the interaction between the aforementioned cultural factors and the differential developmental appropriateness of some behaviors. Specifically, different behaviors can be considered problematic during different developmental phases, but this developmental appropriateness might be viewed differently in various cultural groups. Considering again the example of the energetic or hyperactive 4-year-old, if his or her behavior is labeled as problematic primarily because of its lack of fit with the new social environment (e.g., the child's entry into preschool or daycare and the associated new behavioral expectations), then this would be a problem only in cultural groups where one would expect a 4-year-old to enter institutionalized care. Taken together, these potential developmental and cultural sources of variability could influence the validity and reliability of research that utilizes parental reports of behavior problems.

Assessment of Behavior Problems and Methodological Issues

Researchers interested in examining group differences (such as gender, race/ethnicity, or treatment versus control group differences in the dependent variable) must address issues related to measurement bias with respect to the groups under consideration (Chan, 1998; Tyson, 2004; Vandenberg & Lance, 2000). Measurement unbias, henceforth termed measurement invariance (Vandenberg & Lance, 2000), across groups under consideration can be expressed as a conditional independence that holds if, and only if, the scores on a measured or manifest variable (Y) that reflect the underlying latent construct (η) are independent of group membership (ν), as expressed in the following equation (Mellenbergh, 1989):

$$F(Y|\eta,\nu) = F(Y|\eta) \quad \text{for all } Y, \eta, \nu.$$

In other words, the scores on a behavior problem measure should reflect only the underlying behavior problem construct and should not be affected by group membership such as ethnicity or culture. In cross-ethnic research on group differences, there are several sources of potential measurement bias. The three major categories of measurement bias involve construct, item, and method bias (van de Vijver & Poortinga, 1997; van de Vijver & Tanzer, 2004). Construct bias, and the related cultural bias (for review, see Tyson, 2004), suggest that the construct under consideration has different content across different cultural groups or that individuals from different groups attach a different meaning to the construct. Item bias, related to the issue of differential item functioning, refers to group differences on an item level and can stem from poor item translation or the inappropriateness of item content. Finally, method bias results from problems related to test administration including differential response patterns across cultural groups, such as possible ethnic differences in the use of ordinal rating scales as well as yes/no categories. For example, some studies indicate that adult research participants of Hispanic origin are more likely to adopt an acquiescent response style (Marin, Gamba, & Marin, 1992) and exhibit extreme checking on Likert-type scales (Hui & Triandis, 1989; Marin et al., 1992), as compared to non-Hispanic White counterparts. Similarly, adult African American research participants have been found to be more likely than Whites to use the extreme response categories on Likert-type scales (Bachman & O'Malley, 1984).

A number of well-known instruments exist to assess and categorize problem behaviors in children. One of the most respected instruments in this area is the Child Behavior Checklist (CBCL; Achenbach, 1991). The CBCL is a paper-based questionnaire that presents caregivers with a series of 113 statements that relate to emotional and behavioral problems and competencies, using a 3-point response format to establish the frequency of problem behaviors. The CBCL produces a total problems score, internalizing and externalizing problem scores, and six narrow-band subscores that can be compared to norms and clinical cutoffs for groups based on age and sex. Follow-up studies by the author of the CBCL (Achenbach, 1995; Achenbach & Howell, 1993) and the hundreds of studies that rely on the instrument find a strong support of adequate reliability and validity of the CBCL scores in various populations, although there have been some challenges (e.g., Raadal, Milgrom, Cauce, & Mancl, 1994).

Although well established, the length and cost of the CBCL make it difficult to administer. In response to these limitations, Peterson and Zill (1986) developed the 28-item Behavior Problem Index (BPI), modeled after the CBCL, to more conveniently measure the incidence and severity of child behavior problems in a survey setting. Like the CBCL, the BPI produces a total problems score and six narrow-band scores developed by Zill (1985). Some researchers use one or more of the BPI's six subscales in their analyses, but many others rely on the scores derived from the internalizing and externalizing division of items originally generated by Parcel and Menaghan (1988). This broadband division of the BPI has been widely used as the primary indicator of behavior problems (for a list of studies using this division and National Longitudinal Survey of Youth [NLSY] data, see Bureau of Labor Statistics, 2005).

Many studies use the BPI items and their corresponding internalizing and externalizing division, but several issues pertaining to the psychometric properties of scores derived from this measure remain largely unaddressed. Presently, most peer-reviewed empirical studies provide evidence of basic psychometric properties of scores derived from their assessment tools using statements about reliability and validity; however, when group comparisons are involved, further considerations must be made (Vandenberg & Lance, 2000). Measurement invariance, as an important property of scores, should be assessed prior to any multigroup and/or longitudinal analyses in order to ensure not only that the same constructs are measured, but also that they are measured with equivalent accuracy across different groups and/or time points; otherwise, the interpretation of test scores, prevalence rates, and developmental trajectories could be compromised (Chan, 1998; Vandenberg & Lance, 2000).

Only one study to date has assessed measurement invariance of BPI scores (Spencer et al., 2005). This study used NLSY data to evaluate the measurement invariance of the BPI in three ethnic groups. The analyses revealed that the behavior problem scores were not measured with equivalent accuracy among these three groups using a one-factor, two-factor, and six-factor model of the BPI. However, the sample of children used in this study was of a wide age range (i.e., 4-14 years of age, $M = 10.45$ years, $SD = 2.83$), a choice that may have obfuscated the findings for two reasons. First, the age range of 4 to 14 years includes children from several qualitatively different developmental periods. For example, preschool children and their caregivers often face different challenges and concerns than those in middle childhood or puberty. Thus, the measurement bias found by the authors might be an artifact of grouping together developmentally distinct groups of children into a single sample.

A second and equally important reason for using a more narrow age group pertains to the fact that, by design, some items are not age appropriate for certain children (i.e., 2 items out of the total 28 are to be administered only to children older than 5 years of age and another 2 items only to children younger than 12 years of age). It is unclear how this omission influenced the earlier findings of measurement bias in BPI scores. Furthermore, the factor structure and associated remedies offered by Spencer and colleagues (2005) included several controversial analytic practices including the use of double loadings (where several indicators are allowed to load on both factors) and correlated error terms among several items measured concurrently (for review of issues related to multidimensional measurement, see Kline, 2005). Thus, the results of the first and only published study on measurement invariance of scores derived from the BPI in U.S. ethnic groups are inconclusive. Equally important, despite the wide use of the BPI in developmental research, no known studies have been published on the longitudinal invariance of the internalizing and externalizing scores derived from the BPI.

Purpose

This study addressed these outstanding measurement issues by investigating the measurement invariance of the BPI scores of children from different ethnic groups during their middle childhood. Middle childhood is a distinct developmental phase that has many important markers and milestones; many key abilities, capacities, and relations do not emerge or become consolidated until this developmental period (e.g., Collins, 1984). This developmental period begins about the time children begin formal schooling, an event that has tremendous effect on all areas of development, and ends prior to the biological changes that demarcate the onset of adolescence. This corresponds approximately to the age range of 6 to 12 years (e.g., Santrock, 2000). To evaluate the measurement invariance of the BPI scores over time and across different ethnic groups, this study assessed the measurement bias within and across three ethnic groups of children during the course of middle childhood using a series of cross-sectional and longitudinal models.

Method

Participants

The NLSY originated in 1979 as a successor of earlier efforts to generate nationally representative data sets, focusing on the factors predictive of labor market experience (Chase-Lansdale, Mott, Brooks-Gunn, & Phillips, 1991). The original sample included 6,283 females representing noninstitutionalized young women between 14 and 21 years of age as of 1978. This study focused on children born to these women, in particular a cohort of children aged 5 to 7 years (61 to 94 months) in 1990 and followed biannually until 1994 when they were 9 to 11 years (108 to 143 months) of age. Initially, there were 1,251 children in this cohort, but this number also included 146 sibling dyads and 3 triads. In the present study, only 1 child from each household was selected at random. The resulting sample of 1,099 children included 233 Hispanic; 352 Black; and 514 non-Hispanic, non-Black (henceforth termed White) children.

Measures

The BPI (Peterson & Zill, 1986) was designed as a parent report of child behavior to measure the frequency, type, and scope of behavior problems in children aged 4 to 17. The 28 items in the BPI target specific behaviors potentially exhibited by children during the previous 3 months. These items were derived from the CBCL and maintain the same 3-point ordinal scale (*often*, *sometimes*, and *not true*). The majority of these items have exact or near exact wording of the CBCL items, yet other items differ and some are not present in the CBCL at all. In the BPI, 27 of the 28 items are relevant to the internalizing and externalizing division of scores, with 7 items referring to internalizing problems, 17 items referring to externalizing problems, and 3 items referring to both types of problems (Center for Human Resource Research, 2000; Parcel & Menaghan, 1988). Table 1 describes the individual items and the related scores for both the original division of items as well as

Table 1
Original and CBCL-Based Internalizing and Externalizing Classification of BPI Items

Item Description	BPI-Based (Original)	CBCL-Based
Cheats or tells lies	E	E
Bullies or is cruel/mean to others	E	E
Does not feel sorry for misbehaving	O	O
Breaks things deliberately <12 years	E	E
Disobedient at school >5 years	E	E
Trouble getting along with teachers >5 years	E	O
Sudden changes in mood/feeling	E	E
Feels/complains no one loves him/her	I	I
Too fearful or anxious	E/I	I
Feels worthless or inferior	I	I
Unhappy, sad, or depressed	E/I	I
Clings to adults <12 years	I	O
Cries too much <12 years	I	I
Demands a lot of attention <12 years	I	E
Too dependent on others <12 years	I	O
High-strung, tense, nervous	E	I
Argues too much	E	E
Disobedient at home	E	E
Stubborn, sullen, or irritable	E	E
Strong temper, loses easily	E	E
Difficulty concentrating/paying attention	E	O
Easily confused/in a fog	E/I	O
Impulsive, acts without thinking	E	O
Trouble with obsessions, etc.	E	O
Restless, overly active, etc.	E	O
Trouble getting along with others	E	O
Not liked by other children	E	O
Withdrawn, not involved with others	I	I

	Scale *M* (*SD*)	Cronbach's α	95% CI
BPI externalizing 1990	7.664 (5.774)	.868	(.856, .880)
BPI internalizing 1990	3.094 (2.852)	.744	(.720, .766)
BPI externalizing 1992	8.161 (6.167)	.884	(.873, .894)
BPI internalizing 1992	3.219 (2.937)	.754	(.731, .776)
BPI externalizing 1994	8.008 (6.270)	.890	(.880, .900)
BPI internalizing 1994	3.048 (3.038)	.781	(.761, .800)
CBCL externalizing 1990	4.831 (3.416)	.799	(.780, .817)
CBCL internalizing 1990	1.772 (2.009)	.686	(.657, .714)
CBCL externalizing 1992	5.023 (3.527)	.812	(.794, .829)
CBCL internalizing 1992	2.009 (2.156)	.714	(.687, .740)

(continued)

Table 1 (continued)

	Scale *M* (*SD*)	Cronbach's α	95% CI
CBCL externalizing 1994	4.900 (3.617)	.823	(.806, .838)
CBCL internalizing 1994	2.017 (2.227)	.746	(.722, .769)

Note: BPI = Behavior Problem Index; CBCL = Child Behavior Checklist; E = externalizing; I = internalizing; E/I = both externalizing and internalizing; O = neither externalizing nor internalizing; *SD* = standard deviation, 95% CI = 95% confidence interval for α.

the revised CBCL-based division of items. Scores were computed by summing the items reflecting the following coding: *not true* = 0, *sometimes* = 1, and *often* = 2 (Center for Human Resource Research, 2000). Cronbach's alpha and the associated confidence intervals were computed for each of the factors at each of the three waves (Fan & Thompson, 2001) and indicate a satisfactory internal consistency (Henson, 2001).

In the NLSY79, mothers filled out the BPI inventory for 4- to 14-year-olds every 2 years from 1986. The overall completion rate for the BPI was very high, averaging about 93% (Center for Human Resource Research, 2000). Although the NLSY survey maintained the BPI's original 28 items, response format, and division of internalizing and externalizing items, there were some changes to the BPI's original administration and format. For example, whereas parents completed all items in the original version of the BPI, certain questions were administered only to specific age groups in the NLSY: 5 items were administered only to children younger than 12, and 2 items were administered only to children older than 5. Additionally, 4 items were added to the BPI in the NLSY to address issues relevant to older children, although the new items were not included in the computation of internalizing and externalizing scores. In total, data from the 27 original BPI items that comprised the internalizing and externalizing subscales were used in this study. This classification of items is termed the "BPI-based division" of items.

Analysis

Structural equation modeling was used to test the invariance hypotheses across the three ethnic groups and over time in a series of multigroup confirmatory factor analyses (CFAs). Two types of invariance hypotheses were tested: configural and measurement invariance. The configural invariance analyses were designed to examine whether the patterns of zero and nonzero factor pattern coefficients (henceforth called loadings) were equivalent across groups (with no equality constraints imposed across groups) and to establish baseline models with adequate fit for the subsequent measurement invariance testing (Kline, 2005). The measurement invariance hypotheses involved testing whether the magnitude of factor loadings was equivalent across groups (i.e., construct-level metric invariance; Bollen, 1989; Vandenberg & Lance, 2000). Although the equality of factor loadings is not the sole standard that can be used to test measurement invariance, it is the most important one (Raffalovich & Bohrnstedt, 1987) and it is generally agreed that holding real-life data to more stringent standards is not only too demanding, but also unrealistic (Chan, 1998). However, the analyses involved skewed outcomes measured on a 3-point Likert-type scale, and studies indicate that measures of fit based on the assumption of continuous, normally distributed data can produce biased estimates with categorical data (Lubke & Muthén, 2004). Consequently, models and estimation methods for ordered categorical outcomes were utilized (i.e., the robust mean-adjusted weighted least square estimation described below). This allowed for inclusion of an additional parameter in testing the invariance hypotheses: the threshold. In summary, the measurement invariance hypothesis was tested by comparing the fit of two nested models: one with factor loadings and thresholds constrained to be equal and the other with the parameters free to vary.

Maximum likelihood (ML) has been shown to produce biased estimates when the distributional requirements are seriously violated, particularly with nonnormal ordinal data (e.g., Lubke & Muthén, 2004; Muthén & Kaplan, 1985). A robust weighted least squares estimation method has been proposed by Muthén, du Toit, and Spisic (in press) and is recommended for CFAs that use ordinal data (e.g., Flora & Curran, 2004). Consequently, we used robust mean-adjusted weighted least square (WLSM), available in Mplus 4.1 (Muthén & Muthén, 2005).

A chi-square difference test was used to evaluate an incremental fit in nested model comparisons using a scaling correction, which is analogous to the Satorra–Bentler robust chi-square (Muthén & Muthén, 2005). A nonsignificant chi-square difference test is taken as an indication that the nested model should not be rejected. In these analyses, the nonsignificant chi-square test indicated that the constrained model fit equally well as the model in which the specified parameters were free to vary, suggesting that there are no differences in thresholds and factor loadings across groups under comparison. However, the chi-square test is known to be unduly influenced by sample size, and it is common for the chi-square difference to be large even when the constrained model fits the data well. Cheung and Rensvold (2002) proposed the difference in comparative fit index (ΔCFI) as a useful and robust complementary statistic to the likelihood ratio test ($\Delta\chi^2$) to assess the difference in fit between two nested models. However, their simulation study that investigated the properties of ΔCFI was based on the ML estimation procedure. Thus, when CFI is computed via WLSM estimation, using ΔCFI is not recommended because its distributional properties are unknown.

Consequently, three additional measures of fit statistics were used including the root mean square error of approximation (RMSEA; Steiger, 1990); the CFI (Bentler & Wu, 1995); and the Tucker–Lewis index (TLI; Tucker & Lewis, 1973), also known as the nonnormed fit index (Bentler & Bonnett, 1980). The RMSEA, which measures the discrepancy in the covariance matrices, equals zero if the model provides an exact fit; according to arbitrary but practical guidelines, a value of approximately .05 or less suggests a close fit of the model, and a value of .1 or more indicates a poor fit (e.g., Browne & Cudeck, 1993). The RMSEA is a recommended measure of fit in the analyses with ordered categorical data as it has been shown to be influenced by neither the size of the model nor the sample size (Hutchinson & Olmos, 1998). Similarly, the CFI compares the fit of a model with a null model and is both uncorrelated with the overall fit measures and independent of sample size and model complexity (Cheung & Rensvold, 2002). The TLI, another incremental fit index, has also been found to be relatively independent of sample size (e.g., Marsh, Balla, & McDonald, 1988). TLI and CFI values range between 0 and 1, with values above .95 considered to indicate a good fit of the model to the data (e.g., Hu & Bentler, 1999).

Results

Baseline Models: Configural Invariance of the BPI-Based Division

As described in the analysis section, in the two-factor models the factors were allowed to correlate and the estimated correlations between the two factors across these models ranged between .672 and .841. Table 2 details results of the configural invariance analyses and tests of the same number of factors across groups separately at each time point for the original BPI-based classification. The results suggest that although the two-factor models fit the data significantly better than the one-factor models, the two-factor models still fit the data inadequately (i.e., although the RMSEA was below .10 in all models, the CFI was above .95 in only two out of nine models). Furthermore, in most of these models there was at least one indicator with a factor pattern coefficient that was not significantly different from zero, and these zero-loading indicators varied across groups. Specifically, these indicators pertain to the items that were expected to load on both the internalizing and externalizing factors according to the BPI-based division (Parcel & Menaghan, 1988). For example, the pattern coefficients between the items “too fearful or anxious” and “unhappy, sad, or depressed” and the externalizing construct were significantly different from zero only for Black participants in 1990. The pattern coefficient between the item “easily confused/in a fog” and the externalizing construct was significantly different from zero only for Black and White participants in 1990, while the pattern coefficient between this item and the internalizing behavior problem construct was statically significant only for Hispanic participants in this assessment wave. In addition, these patterns did not remain stable over time and varied within the ethnic groups from assessment to assessment (see Table 3). Because the baseline models in which the patterns of zero and nonzero loadings were held equal across ethnic groups for each assessment year did not accurately represent the data and the configural invariance for the original classification of items was not

Table 2
Baseline Models: Configural Invariance of the Original Division of Items by Race/Ethnicity

	RMSEA	CFI	TLI	$\Delta\chi^2$	ΔCorr	Δ*df*	*p*
White							
1990							
Two-factor	.068	.945	.940	151.361	1.339	4	<.001
One-factor	.076	.931	.925				
1992							
Two-factor	.086	.925	.918	233.531	0.927	4	<.001
One-factor	.092	.912	.904				
1994							
Two-factor	.092	.933	.927	438.490	0.967	4	<.001
One-factor	.104	.913	.905				
Black							
1990							
Two-factor	.077	.946	.940	214.419	1.533	4	<.001
One-factor	.093	.919	.913				
1992							
Two-factor	.087	.923	.916	157.765	0.740	4	<.001
One-factor	.093	.912	.905				
1994							
Two-factor	.074	.955	.951	109.082	0.863	4	<.001
One-factor	.079	.948	.944				
Hispanic							
1990							
Two-factor	.071	.925	.918	72.782	0.951	4	<.001
One-factor	.077	.912	.904				
1992							
Two-factor	.078	.957	.953	149.868	0.847	4	<.001
One-factor	.087	.945	.941				
1994							
Two-factor	.084	.939	.933	72.297	1.167	4	<.001
One-factor	.090	.929	.924				

Note: RMSEA = root mean square error of approximation; CFI = comparative fit index; TLI = Tucker–Lewis index; $\Delta\chi^2$ = chi-square of the difference test; ΔCorr = scaling correction factor difference; Δ*df* = degrees of freedom of the difference; *p* = probability level associated with the χ^2 of the difference test.

supported, no further measurement invariance analyses were conducted using this factor structure. To avoid inferring a new model from the observed data, we used a theory-based approach; we turned to the CBCL, a measure on which the BPI was originally based, and followed the internalizing and externalizing classification of items suggested by Achenbach (1991).

Table 3
Pattern Coefficients for Problematic Items From the Original Division of Scores

	Hispanic		Black		White	
	Est.	*SE*	Est.	*SE*	Est.	*SE*
Externalizing						
Too fearful or anxious 1990	.092*	.169	.198	.094	.129*	.095
Unhappy/sad/depressed 1990	−.24*	.189	.307	.101	.129*	.109
Easily confused/in a fog 1990	.204*	.183	.517	.095	.617	.121
Internalizing						
Easily confused/in a fog 1990	.376	.186	.195*	.106	.043*	.135
Externalizing						
Too fearful or anxious 1992	.35	.119	.329	.121	.111*	.102
Unhappy/sad/depressed 1992	.309	.134	.221*	.113	–.012*	.119
Easily confused/in a fog 1992	.39	.135	.626	.107	.44	.123
Internalizing						
Easily confused/in a fog 1992	.276*	.148	.088*	.116	.261	.129
Externalizing						
Too fearful or anxious 1994	.228*	.146	.203*	.156	.164*	.087
Unhappy/sad/depressed 1994	.031*	.147	–.033*	.145	–.029*	.088
Easily confused/in a fog 1994	.297	.137	.558	.156	.545	.096
Internalizing						
Easily confused/in a fog 1994	.484	.143	.114*	.164	.134*	.106

Note: Est. = pattern coefficient (factor loading); *SE* = standard error.
*Nonsignificant factor loading, $p > .05$.

Baseline Models: Configural Invariance of the CBCL-Based Division

This second approach involved using a simplified factor structure based on the CBCL division of items (Achenbach, 1991) and was based on theoretical considerations from an established and extensively tested classification of items. Applying this CBCL-based division to BPI items yielded a good overall fit of the model, did not require double loadings of items on both factors, and included 17 items: 7 referring to internalizing problems and 10 referring to externalizing problems. See Table 1 for the description of items and the related scores. Cronbach's alpha and the associated confidence intervals (Fan & Thompson, 2001) were computed for each of the factors at each of the three waves and indicate a satisfactory internal consistency (Henson, 2001). As described in the analysis section, in the two-factor models the factors were allowed to correlate and the estimated correlations between the two factors across these models ranged between .711 and .880. Table 4 describes results of the configural invariance tests for the CBCL-based classification of items. The results suggest that the two-factor model fit the data significantly better than the one-factor

Table 4
Baseline Models: Configural Invariance of the CBCL-Based Division of Items by Race/Ethnicity

	RMSEA	CFI	TLI	$\Delta\chi^2$	ΔCorr	Δ*df*	*p*
White							
1990							
Two-factor	.065	.959	.953	68.933	0.928	1	<.001
One-factor	.072	.949	.941				
1992							
Two-factor	.070	.957	.951	593.198	0.187	1	<.001
One-factor	.082	.941	.933				
1994							
Two-factor	.078	.957	.950	464.414	0.391	1	<.001
One-factor	.095	.936	.927				
Black							
1990							
Two-factor	.070	.961	.955	80.292	1.361	1	<.001
One-factor	.086	.940	.931				
1992							
Two-factor	.071	.948	.940	178.333	0.775	1	<.001
One-factor	.092	.911	.899				
1994							
Two-factor	.056	.979	.976	58.402	0.596	1	<.001
One-factor	.063	.973	.970				
Hispanic							
1990							
Two-factor	.062	.960	.954	103.852	0.680	1	<.001
One-factor	.080	.934	.925				
1992							
Two-factor	.079	.966	.960	223.319	0.345	1	<.001
One-factor	.095	.950	.943				
1994							
Two-factor	.062	.974	.971	71.308	0.780	1	<.001
One-factor	.076	.961	.955				

Note: CBCL = Child Behavior Checklist; RMSEA = root mean square error of approximation; CFI = comparative fit index; TLI = Tucker–Lewis index; $\Delta\chi^2$ = chi-square of the difference test using scaling correction; ΔCorr = scaling correction factor difference; Δ*df* = degrees of freedom of the difference; *p* = probability level associated with the χ^2 of the difference test.

model and the fit of these models ranges between acceptable to good. Likewise, the pattern of zero and nonzero factor pattern coefficients was equivalent across groups and over time (i.e., all of the specified factor pattern coefficients were significantly different from zero). Consequently, after establishing well-fitting baseline models, the next analytic step involved testing measurement invariance within each ethnic group during the period of middle childhood.

Table 5
Cross-Ethnic Measurement Invariance Test of the CBCL-Based Division of Scores

	RMSEA	CFI	TLI	$\Delta\chi^2$	ΔCorr	Δ*df*	*p*
Cross-ethnic 1990							
Constrained	.065	.954	.955	113.591	1.306	64	<.001
Unconstrained	.066	.960	.954				
Cross-ethnic 1992							
Constrained	.071	.952	.953	108.979	1.294	64	<.001
Unconstrained	.073	.957	.951				
Cross-ethnic 1994							
Constrained	.068	.963	.964	126.457	1.276	64	<.001
Unconstrained	.069	.968	.963				

Note: CBCL = Child Behavior Checklist; Constrained = factor pattern coefficients and thresholds constrained to be equal across the three ethnic groups; Unconstrained = factor pattern coefficients and thresholds free to vary across the groups; RMSEA = root mean square error of approximation; CFI = comparative fit index; TLI = Tucker–Lewis index; $\Delta\chi^2$ = chi-square of the difference test using scaling correction; ΔCorr = scaling correction factor difference; Δ*df* = degrees of freedom of the difference; *p* = probability level associated with the χ^2 difference test.

Measurement Invariance Testing of the CBCL-Based Division

Cross-ethnic comparisons. Measurement invariance was defined as the equality of factor loadings and thresholds across the ethnic groups separately at each time point. The constrained model had thresholds and factor loadings set to be equal across ethnic groups, the factor means were set to zero in the first group (White) and free in the others, and the scale factors were set to one in the first group and free in the others (Muthén & Muthén, 2005). The unconstrained models had thresholds and factor loadings free across groups, factor means were set to zero, and scale factors were set to one in all groups. The results, presented in Table 5, suggest that these parameter estimates are equivalent across the groups. Although the chi-square difference tests indicate that there was a significant decrease in model fit (as described earlier, this test statistic is unduly influenced by the sample size), the RMSEA and CFI in all of the comparisons indicate that the measurement invariance model fits the data well (i.e., RMSEA < .10; CFI > .95).

Longitudinal comparisons. Measurement invariance, defined as the equality of factor loadings and thresholds, was also tested across time within each ethnic group in a series of longitudinal models. In the longitudinal models, the constrained model had thresholds and factor loadings constrained to be equal across time, factor means set to zero, and scale factors set to one at Time 1 only. The unconstrained

Table 6
Longitudinal Measurement Invariance Test of the CBCL-Based Division of Scores

	RMSEA	CFI	TLI	$\Delta\chi^2$	ΔCorr	Δ*df*	*p*
White							
Constrained	.049	.965	.965	115.589	1.934	91	.042
Unconstrained	.049	.968	.965				
Black							
Constrained	.041	.970	.970	87.218	1.914	91	.593
Unconstrained	.041	.973	.971				
Hispanic							
Constrained	.048	.965	.965	115.343	1.889	91	.043
Unconstrained	.045	.972	.969				

Note: CBCL = Child Behavior Checklist; Constrained = factor pattern coefficients and thresholds constrained to be equal across the three time points within each ethnic group comparison; Unconstrained = factor pattern coefficients and thresholds free to vary over time; RMSEA = root mean square error of approximation; CFI = comparative fit index; TLI = Tucker–Lewis index; $\Delta\chi^2$ = chi-square of the difference test using scaling correction; ΔCorr = scaling correction factor difference; Δ*df* = degrees of freedom of the difference; *p* = probability level associated with the chi-square difference test.

model had the thresholds and factor loadings free to vary across time, factor means set to zero, and scale factors set to one in all groups (with the exception of the unit loading identification items, which were treated as measurement invariant because they were set to one in all groups). The six latent factors (two at each time point) were allowed to correlate concomitantly, and the error variances of respective items were allowed to correlate across the three time points (Marsh & Grayson, 1994; Vaillancourt, Brendgen, Boivin, & Tremblay, 2003). The results, described in Table 6, suggest that the measurement invariance hypothesis cannot be rejected: The constrained models yielded an excellent fit to the data (i.e., RMSEA < .05; CFI > .95) in the longitudinal comparisons for all three ethnic groups.

Discussion

Educators, clinicians, and researchers are all heavily invested in the accurate assessment of behavioral functioning. Only with valid and reliable assessment that takes developmental and cultural factors into consideration can scientific research be accomplished, conceptual accuracy gained, appropriate interventions and preventive efforts developed, and policies aiming to reduce mental health disparities implemented. Behavior rating scores used to compare children from different developmental stages or ethnic/cultural groups are susceptible to measurement bias and must be analyzed in order to ensure such reliable assessment. Measurement bias can affect the interpretation of test scores, prevalence rates, developmental trajectories, and other issues central to the investigation of behavior problems (Tyson, 2004).

We found that the prescribed and frequently used two-factor solution of BPI items (Center for Human Resource Research, 2000; Parcel & Menaghan, 1988) fit the data relatively poorly and lacked configural invariance in the groups under consideration, leading us to conclude that this classification of items exhibits construct bias and a lack of convergent validity. In the revision of the factor structure, we relied on theoretical considerations—a classification of internalizing and externalizing manifestation of dysfunction based on an established and extensively tested measure. Using the CBCL-based division of items, we found support for the internalizing and externalizing factor solution and found configural invariance across the three ethnic groups during the period of middle childhood. We conclude that the same constructs are measured using this factor structure across all groups (i.e., there is a conceptual equivalence in the internalizing and externalizing behavior problem constructs). The proposed CBCL-based structure has an added advantage of being more parsimonious than the original as there was no need to use double loadings or correlated error terms among concurrently measured items. We thus recommend that the researchers use the internalizing and externalizing behavior problem scores derived from the CBCL-based division of items.

Moreover, because direct comparisons among ethnic groups are often the desired foci of research studies, we also tested whether the assessment tool is operating equivalently across these groups. The invariance of factor loadings and thresholds across groups was supported, suggesting that the underlying internalizing and externalizing behavior constructs are measured with equivalent accuracy across groups and over time. We conclude that the scores from the developed 17-item index of internalizing and externalizing behavior problems can be used to make longitudinal cross-ethnic comparisons.

However, measurement invariance, just like reliability, is a property of scores, not of a measure, and should be assessed within the specific sample and testing conditions. The results of this study should be interpreted in light of the fact that the sample followed only school-aged children from three major U.S. ethnic groups during the course of their middle childhood. These results therefore may not be replicated in other samples and should be the subject of further investigation.

Although many powerful and sophisticated techniques for investigation of developmental processes in children from various cultural backgrounds are currently available to social science researchers, the guidelines for data and measurement evaluation that should accompany these analyses have not been uniformly established and followed (e.g., Chan, 1998; Curran & Hussong, 2003). This is unfortunate because answers to many fundamental questions about the nature of change and the variability in this change can be confounded by factors that stem from design flaws or inconsistencies in assessment. Finding measurement bias in behavior problem scores derived from a commonly used measure and factor structure obtained from a large national sample of children, this study highlights the growing concern that measurement invariance across groups under consideration cannot be merely assumed but should be explicitly tested prior to any direct group comparisons. This study offers a practical guide on how to conduct measurement invariance analyses that should precede multigroup comparisons including the increasingly popular longitudinal data analytic techniques such as latent variable growth curve modeling.

References

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

Achenbach, T. M. (1995). "Behavior problems in 5- to 11-year-old children from low-income families": Of norms and cutoffs [Comment]. *Journal of the American Academy of Child & Adolescent Psychiatry*, *34*, 536-537.

Achenbach, T. M., Dumenci, L., & Rescorla, L. A. (2003). Are American children's problems still getting worse? A 23-year comparison. *Journal of Abnormal Child Psychology*, *31*, 1-11.

Achenbach, T. M, & Edelbrock, C. S. (1978). The classification of child psychopathology: A review and analysis of empirical efforts. *Psychological Bulletin*, *85*, 1275-1301.

Achenbach, T. M., & Howell, C. T. (1993). Are American children's problems getting worse? A 13-year comparison. *Journal of the American Academy of Child & Adolescent Psychiatry*, *32*, 1145-1154.

Ackerman, B. P., Brown, E. D., & Izard, C. E. (2004). The relations between contextual risk, earned income, and the school adjustment of children from economically disadvantaged families. *Developmental Psychology*, *40*, 204-216.

Bachman, J. G., & O'Malley, P. M. (1984). Yea-saying, nay-saying, and going to extremes: Black-White differences in response style. *Public Opinion Quarterly*, *48*, 291-509.

Barr, R. G. (2000). Excessive crying. In A. Sameroff, M. Lewis, & S. M. Miller, *Handbook of developmental psychopathology* (2nd ed., pp. 327-350). New York: Kluwer Academic.

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

Bentler, P. M., & Wu, E. J. (1995). EQS for Windows user's guide [Computer manual]. Encino, CA: Multivariate Software.

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

Browne, M., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-62). Newbury Park, CA: Sage.

Bureau of Labor Statistics. (2005). *The NLS annotated bibliography*. Retrieved March 27, 2007, from http://www.nlsbibliography.org

Center for Human Resource Research. (2000). *NLSY79 child & young adult data user's guide*. Columbus: Ohio State University.

Chan, D. (1998). The conceptualization and analysis of change over time: An integrative approach incorporating longitudinal mean and covariance structures analysis and multiple indicator latent growth modeling. *Organizational Research Methods*, *1*, 421-483.

Chase-Lansdale, P. L., Mott, F. L., Brooks-Gunn, J., & Phillips, D. A. (1991). Children of the National Longitudinal Survey of Youth: A unique research opportunity. *Developmental Psychology*, *27*, 918-931.

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

Cicchetti, D., & Toth, S. L. (1991). A developmental perspective on internalizing and externalizing disorders. *Rochester Symposium on Developmental Psychopathology*, *2*, 1-20.

Collins, A. W. (1984). Conclusion: The status of basic research on middle childhood. In A. W. Collins (Ed.), *Development during middle childhood: The years from six to twelve* (pp. 398-421). Washington, DC: National Academy Press.

Curran, P. J., & Hussong, A. M. (2003). The use of latent trajectory models in psychopathology research. *Journal of Abnormal Psychology*, *112*, 526-544.

Fan, X., & Thompson, B. (2001). Confidence intervals for effect sizes: Confidence intervals about score reliability coefficients, please. *Educational and Psychological Measurement*, *61*, 517-531.

Flora, D. B., & Curran P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*, 466-491.

Harkness, S., & Super, C. M. (2000). Culture and psychopathology. In A. Sameroff, M. Lewis, & S. Miller (Eds.), *Handbook of developmental psychopathology* (pp. 197-214). New York: Kluwer Academic.

Helstrom, A., Bryan, A., Hutchison, K. E., Riggs, P. D., & Blechman, E. A. (2004). Tobacco and alcohol use as an explanation for the association between externalizing behavior and illicit drug use among delinquent adolescents. *Prevention Science*, *5*, 267-277.

Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development*, *34*, 177-189.

Hinshaw, S. P. (1992). Externalizing behavior problems and academic underachievement in childhood and adolescence: Causal relationships and underlying mechanisms. *Psychological Bulletin*, *111*, 127-155.

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

Hui, C., & Triandis, H. C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology*, *20*, 296-309.

Hutchinson, S. R., & Olmos, A. (1998). Behavior of descriptive fit indexes in confirmatory factor analysis using ordered categorical data. *Structural Equation Modeling*, *5*, 344-364.

King, S. M., Iacono, W. G., & McGue, M. (2004). Childhood externalizing and internalizing psychopathology in the prediction of early substance use. *Addiction*, *99*, 1548-1559.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford.

Lubke, G., & Muthén, B. (2004). Factor-analyzing Likert scale data under the assumption of multivariate normality complicates a meaningful comparison of observed groups or latent classes. *Structural Equation Modeling*, *11*, 514-534.

Lynam, D. R. (1996). The early identification of chronic offenders: Who is the fledgling psychopath? *Psychological Bulletin*, *120*, 209-234.

Marin, G., Gamba, R. J., & Marin, B. V. (1992). Extreme response style and acquiescence among Hispanics: The role of acculturation and education. *Journal of Cross-Cultural Psychology*, *23*, 498-509.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit-indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin*, *103*, 391-410.

Marsh, H. W., & Grayson, D. (1994). Longitudinal confirmatory factor analysis: Common, time-specific, item-specific, and residual-error components of variance. *Structural Equation Modeling*, *1*, 116-145.

McCulloch, A., Wiggins, R. D., Joshi, H. E., & Sachdev, D. (2000). Internalizing and externalizing children's behaviour problems in Britain and the U.S.: Relationships to family resources. *Children & Society*, *14*, 368-383.

McLeod, J. D., & Kaiser, K. (2004). Childhood emotional and behavioral problems in educational attainment. *American Sociological Review*, *69*, 636-658.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*, 127-143.

Moffitt, T. E. (1993). Adolescence-limited and life-course-persistent antisocial behavior: A developmental taxonomy. *Psychological Review*, *100*, 674-701.

Murray, K. T., & Kochanska, G. (2002). Effortful control: Factor structure and relation to externalizing and internalizing behaviors. *Journal of Abnormal Child Psychology*, *30*, 503-514.

Muthén, B., du Toit, S. H. C., & Spisic, D. (in press). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. *Psychometrika.*

Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of nonnormal Likert variables. *British Journal of Mathematical and Statistical Psychology*, *38*, 171-189.

Muthén, L. K., & Muthén, B. O. (2005). *Mplus user's guide* (4th ed.). Los Angeles, CA: Authors.

Parcel, T. L., & Menaghan, E. G. (1988). *Measuring behavior problems in a large cross sectional survey: Reliability and validity for children of the NLS youth*. Columbus: Ohio State University, Department of Sociology.

Peterson, J. L., & Zill, N. (1986). Marital disruption, parent-child relationships, and behavior problems in children. *Journal of Marriage and Family*, *48*, 295-307.

Raadal, M., Milgrom, P., Cauce, A. M., & Mancl, L. (1994). Behavior problems in 5- to 11-year-old children from low-income families. *Journal of the American Academy of Child & Adolescent Psychiatry*, *33*, 1017-1025.

Raffalovich, L. E., & Bohrnstedt, G. W. (1987). Common, specific, and error variance: Components of factor models. *Sociological Methods & Research*, *15*, 285-405.

Reinherz, H. Z., Giaconia, R. M., Pakiz, B., Silverman, A. B., Frost, A. K., & Lefkowitz, E. S. (1993). Psychosocial risks for major depression in late adolescence: A longitudinal community study. *Journal of the American Academy of Child & Adolescent Psychiatry*, *32*, 1155-1163.

Santrock, J. W. (2000). *Children* (6th ed.). New York: McGraw-Hill.

Spencer, M., Fitch, D., Grogan-Kaylor, A., & McBeath, B. (2005). The equivalence of the Behavior Problem Index across U.S. ethnic groups. *Journal of Cross-Cultural Psychology*, *36*, 573-589.

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

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

Tyson, E. H. (2004). Ethnic differences using behavior rating scales to assess the mental health of children: A conceptual and psychometric critique. *Child Psychiatry and Human Development*, *34*, 167-201.

Vaillancourt, T., Brendgen, M., Boivin, M., & Tremblay, R. E. (2003). A longitudinal confirmatory factor analysis of indirect and physical aggression: Evidence of two factors over time? *Child Development*, *74*, 1628-1638.

Vandenberg, 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*, 4-69.

van de Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology*, *54*, 119-135.

van de Vijver, F. J., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment*, *13*, 29-37.

Wood, J. J., Cowan, P. A., & Baker, B. L. (2002). Behavior problems and peer rejection in preschool boys and girls. *Journal of Genetic Psychology*, *163*, 72-88.

Zill, N. (1985). *Behavior problem scales developed from the 1981 Child Health Supplement to the National Health Interview Survey*. Washington, DC: Child Trends.