Title

The Abbreviated Dimensions of Temperament Survey: Factor Structure and Construct Validity Across Three Racial/Ethnic Groups

Journal

Journal of Personality Assessment, 97(5)

ISSN

0022-3891

Authors

Windle, Michael
Wiesner, Margit
Elliott, Marc N
et al.

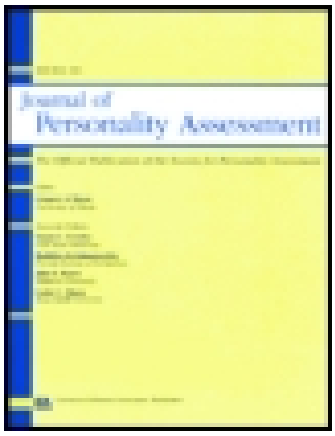
Publication Date

2015-09-03

DOI

10.1080/00223891.2015.1034868

Peer reviewed



Journal of Personality Assessment

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hjpa20

The Abbreviated Dimensions of Temperament Survey: Factor Structure and Construct Validity Across Three Racial/Ethnic Groups

Michael Windle[a], Margit Wiesner[b], Marc N. Elliott[c], Jan L. Wallander[d], David E. Kanouse[c] & Mark A. Schuster[e]

[a] Department of Behavioral Sciences and Health Education, Emory University

[b] Department of Educational Psychology, University of Houston

[c] RAND Corporation, Santa Monica, California

[d] Department of Psychology, University of California-Merced

[e] Division of General Pediatrics, Boston Children's Hospital/Harvard Medical School

Published online: 01 May 2015.

To cite this article: Michael Windle, Margit Wiesner, Marc N. Elliott, Jan L. Wallander, David E. Kanouse & Mark A. Schuster (2015): The Abbreviated Dimensions of Temperament Survey: Factor Structure and Construct Validity Across Three Racial/Ethnic Groups, Journal of Personality Assessment, DOI: 10.1080/00223891.2015.1034868

To link to this article: http://dx.doi.org/10.1080/00223891.2015.1034868

Journal of Personality Assessment, *0*(0), 1–10, 2015

ISSN: 0022-3891 print / 1532-7752 online
DOI: 10.1080/00223891.2015.1034868

Routledge
Taylor & Francis Group

The Abbreviated Dimensions of Temperament Survey: Factor Structure and Construct Validity Across Three Racial/Ethnic Groups

MICHAEL WINDLE,[1] MARGIT WIESNER,[2] MARC N. ELLIOTT,[3] JAN L. WALLANDER,[4] DAVID E. KANOUSE,[3] AND MARK A. SCHUSTER[5]

[1]*Department of Behavioral Sciences and Health Education, Emory University*
[2]*Department of Educational Psychology, University of Houston*
[3]*RAND Corporation, Santa Monica, California*
[4]*Department of Psychology, University of California–Merced*
[5]*Division of General Pediatrics, Boston Children's Hospital/Harvard Medical School*

The factor structure, reliability, and construct validity of an abbreviated version of the Revised Dimensions of Temperament Survey (DOTS–R) were evaluated across Black, Hispanic, and White early adolescents. Primary caregivers reported on 5 dimensions of temperament for 4,701 children. Five temperament dimensions were identified via maximum likelihood exploratory factor analysis and were labeled flexibility, general activity level, positive mood, task orientation, and sleep rhythmicity. Multigroup mean and covariance structures analysis provided partial support for strong factorial invariance across these racial/ethnic groups. Mean level comparisons indicated that relative to Hispanics and Blacks, Whites had higher flexibility, greater sleep regularity, and lower activity. They also reported higher positive mood than Blacks. Blacks, relative to Hispanics, had higher flexibility and lower sleep regularity. Construct validity was supported as the 5 temperament dimensions were significantly correlated with externalizing problems and socioemotional competence. This abbreviated version of the DOTS–R could be used across racial/ethnic groups of early adolescents to assess significant dimensions of temperament risk that are associated with mental health and competent (healthy) functioning.

The Revised Dimensions of Temperament Survey (DOTS–R; Windle, 1999; Windle & Lerner, 1986) was used in this study to assess temperament. The origins of the DOTS–R stem from the highly influential research of Thomas and Chess (1984), who reinvigorated the field of temperament studies in the late 1970s and early 1980s with their pioneering work on the New York Longitudinal Study (NYLS). Thomas and Chess conducted a longitudinal study of young children as they progressed from infancy to childhood to adolescence and then to adulthood. A critical factor in predicting differences in life course trajectories and outcomes with regard to mental health, substance use, and overall life functioning (e.g., educational attainment, quality of interpersonal relationships) was temperament. Temperament referred to stylistic aspects of behavior (i.e., how active a child was rather than was the child active or not). Nine dimensions of temperament were identified based on behavioral observations and quantitative clinical rating scales.

To develop a survey measure of these temperament dimensions that could be used across different age groups, Lerner, Palermo, Spiro, and Nesselroade (1982) developed the Dimensions of Temperament Survey (DOTS). However, it had a few limitations, including low reliability of some of the temperament dimensions, only five of the original nine NYLS factors were represented, and it used a dichotomous response format that might have restricted the range of responses to differentiate individuals. In response to these limitations, Windle and Lerner (1986) developed the DOTS–R that included a 4-point response format for each item, yielded higher reliability estimates, and provided a nine-factor representation of temperament.

Data from the Healthy Passages Study (Schuster et al., 2012; Windle et al., 2004) were used in this article. A high priority of Healthy Passages was to measure multiple health behaviors (e.g., internalizing and externalizing problems, sexual behavior, substance use) and multiple (multilevel) predictors from domains of individual attributes, peer and family factors, and school and neighborhood influences. To accommodate this priority, along with limiting demands on fifth-grade students and their primary caregivers, scale and item selection underwent considerable scrutiny (see Windle et al., 2004). Temperament was viewed as an important individual-level attribute domain but given the competing demands and subject burden considerations, only five of the nine dimensions of the DOTS–R were utilized. The four dimensions not included were activity sleep, which has been more useful in studies of infants rather than children; rhythmicity eating and rhythmicity-daily habits, which have been less consistently predictive of outcomes than rhythmicity sleep (which was included); and approach-withdrawal, which has not been as predictive of outcomes as has behavioral flexibility (which was included). Hence, priority was placed on selecting those DOTS–R scales that would maximally predict the outcomes of the Healthy Passages study. In addition, several items were deleted from the DOTS–R scales that appeared to have substantial content overlap with other items on the scale.

The full-scale DOTS–R has desirable psychometric properties (e.g., high reliability and longitudinal stability, cross-cultural invariance, high heritability, moderate-to-high interrater

Received November 11, 2013; Revised January 29, 2015.

Address correspondence to Michael Windle, Department of Behavioral Sciences and Health Education, Emory University, 1518 Clifton Road, NE, Room 514, Atlanta, GA 30322; Email: mwindle@emory.edu

agreement) and both short- and long-term predictive associations with substance use and mental health (Luby & Steiner, 1993; Oniszczenko et al., 2003; Windle & Windle, 2006). The DOTS–R factors also map well onto other major temperament and personality inventories, thereby facilitating cross-study comparisons (Oniszczenko et al., 2003). With regard to heritability, in a twin study of both Polish and German samples, heritability coefficients for the DOTS–R dimensions averaged .39 for the Polish sample and .54 for the German sample for self-reports, and .61 for the Polish sample and .45 for the German sample for peer report data (Oniszczenko et al., 2003).

The DOTS–R has significantly predicted both internalizing and externalizing problems in children and adolescents (Shaw & Steiner, 1997; Tubman & Windle, 1995; Windle et al., 1986), has prospectively predicted psychiatric and substance abuse disorders in young adulthood, and has significantly distinguished alcoholic from nonalcoholic adults (Windle, 1999). Chang, Blasey, Ketter, and Steiner (2003) used the DOTS–R to evaluate differences in the prevalence of psychiatric disorders among children at high risk for bipolar disorder. The DOTS–R temperament dimensions of inflexibility, low positive mood, and low task orientation distinguished children who developed a psychiatric disorder from those who did not develop a disorder. Giancola and Mezzich (2003) also reported that a more difficult temperament as assessed on the DOTS–R (e.g., higher activity, higher distractibility, lower mood quality) was a stronger predictor of substance disorders among adolescent girls than was a neuropsychological measure of executive cognitive functioning. Effect size (ES) estimates for relationships between the DOTS–R dimensions and internalizing and external problems with community samples have typically ranged from .25 to .35. Comparisons between community samples and clinical samples have commonly yielded larger ES estimates (closer to 1.0), such as those between a community sample and an alcoholic inpatient sample (Windle, 1999). Shaw and Steiner (1997) also reported large ES estimates for some temperament dimensions (positive mood, flexibility) when comparing community samples and youth with anorexia.

Method of Analyses

A major focus of this study was the dimensional structure of the 23-item abbreviated DOTS–R used in the Healthy Passages study. The use of confirmatory item-factor analyses with a relatively large number of items and factors has commonly met with difficulties in achieving good model fit using conventional fit indexes and suggested cut points, and this issue can be influenced by larger sample sizes (Bollen, Harden, Ray, & Zavisca, 2014; Marsh et al., 2010; Marsh, Nagengast, & Morin, 2013). Most of the simulation studies used to develop and evaluate alternative goodness-of-fit indexes have relied on much smaller numbers of manifest indicators or items (e.g., three indicators per factor) and factors (e.g., two factors; Hu & Bentler, 1999). Few simulation studies have been completed for the performance of fit indexes for larger, multiple-factor, multiple-item inventories with larger sample sizes; this is problematic because for data with these characteristics, sometimes even minor departures from model fit can yield goodness of model fit indexes that suggest poor or nonoptimal fit (Marsh et al., 2010; Marsh et al., 2013). Some have urged caution in overinterpreting the extant goodness-of-fit indexes, but the literature still appears to hold fast to conventional cutoffs (Bollen et al., 2014; Marsh, Hau, & Wen, 2004).

An alternative approach to estimating large inventory, multiple-factor models, and in our application with a large sample size, is the exploratory structural equation model (ESEM; Asparouhov & Muthén, 2009; Marsh et al., 2013). The logic of the ESEM for applications with data as characterized earlier (i.e., large sample, large number of items, multifactor measures) is that model fit might be negatively impacted by the accumulation of nonsalient factor loadings that are small with respect to interpreting the substantive meaning of the factor representation, but that nevertheless impact goodness-of-fit statistics when such nonsalient loadings are fixed to zero as they would be in standard confirmatory factor analytic applications. For example, if a given item loads highly on its referent factor, it still might have nonsalient loadings of .20 on other factors; in standard confirmatory factor analytic applications, these loadings are fixed to zero. The summation of such nonsalient loadings fixed to zero across a model with a large number of items and a large number of factors might be sufficient to indicate that the specified model does not fit the data well, even though these nontarget loadings are small substantively. For purposes of model identification, the ESEM requires some parameter constraints in the model specification (at least $m - 1$, where m equals the number of factors), but freely estimates the remaining parameters. The ESEM could be viewed as a hybrid model of exploratory and confirmatory approaches, falling along a continuum from exploratory to confirmatory.

In this study we used ESEM to model the item-factor relationships of the abbreviated DOTS–R. Consistent with prior applications of ESEM (Marsh et al., 2010; Marsh et al., 2013), we also ran standard confirmatory model applications for comparison purposes. The ESEM can also be used for multiple-group comparisons (Marsh et al., 2013) and was used in this article to test invariance hypotheses across racial/ethnic groups. In making these comparisons, we hypothesized invariant relationships across groups for the factor structure of the abbreviated DOTS–R. We also examined the associations between the derived temperament factors and two outcome variables (externalizing problems and socioemotional competence) to see if the factors from this abbreviated version of the DOTS–R correspond in direction and magnitude with those that have been reported with the full-scale DOTS–R.

Factor Invariance Tests and Hypotheses

Given the consistency of the factor structure of the DOTS–R across samples that have varied with regard to age, sex, and cultural group (Marcet, Guardia, Almirall, & Lorenzo, 2000; Windle, Iwawaki, & Lerner, 1987), we hypothesized that the five-factor structure of this shortened version would retain its factor integrity. Nevertheless, a range of potential factors (e.g., ordering or sequencing effects) could disrupt the factor integrity of this subset of items. Therefore, we initially evaluated the adequacy of the five-factor representation of the 23 items of the DOTS–R via three methods: eigenvalues equal to or greater than 1.0, parallel analysis, and comparisons of fit

for alternatively specified maximum likelihood exploratory factor solutions.

We also specified and tested a series of invariance hypotheses (Little, 1997; Marsh et al., 2013; Meredith, 1993) about the factor structure of the abbreviated DOTS–R across three major racial/ethnic groups. On the basis of prior research about the similarity of factor loadings across samples (Marcet et al., 2000; Windle et al., 1987), it was hypothesized that a five-factor model of weak factorial invariance would be supported (i.e., parameter estimates corresponding to factor loadings would be equivalent across groups). In addition, we hypothesized that a five-factor model of strong factorial invariance would be supported (i.e., parameter estimates corresponding to factor loadings and intercepts would be equivalent across groups). Due to limited previous temperament research on mean comparisons with early adolescents, analyses were also conducted to evaluate the equivalence of means across racial/ethnic groups. Finally, internal consistency estimates were computed and compared with those reported in prior DOTS–R full-scale studies, as were correlations between the abbreviated DOTS–R scores and the factors of externalizing problems and socioemotional competence that have been supported in prior research with full-scale scores (Windle et al., 1986). These latter analyses were conducted to evaluate the internal consistency and construct validity, respectively, of the abbreviated DOTS–R. Although no universal minimally accepted values exist for reliability and validity, we followed recommendations of approximately .70 as an indicator of low but acceptable reliability and .20 to .40 for Pearson correlations as supportive of construct validity associations. These low-to-moderate numeric values for construct validity are consistent with prior values of reliability and validity studies of the full-scale DOTS–R (Windle et al., 1986; Windle & Lerner, 1986), as well as consistent with theoretical models that posit that temperament interacts with other features of the environment (e.g., parents, peers, teachers) to predict important aspects of healthy and unhealthy functioning (Thomas & Chess, 1984; Windle et al., 1986). Hence, low to moderate values are anticipated for construct validity of the temperament dimensions and serve to support the theoretical nomological net of hypothesized relationships.

Methods

Participants

Healthy Passages is a longitudinal study of a cohort of 5,147 fifth-graders and their parents that explores health behaviors, outcomes, and related risk and protective factors using a multilevel approach (Schuster et al., 2012; Windle et al., 2004). In this study, baseline data were collected between 2004 and 2006 from primary caregivers of non-Hispanic Black, Hispanic, and non-Hispanic White children ($N = 4{,}701$). Of these 4,701 children, 36.6% ($n = 1{,}721$) were Black, 36.9% ($n = 1{,}733$) were Hispanic, and 26.5% ($n = 1{,}247$) were White. The sex distribution was approximately equal, with 2,383 girls (50.7%) and 2,318 boys (49.3%), and average age was 10.63 years ($SD = 0.64$). Sex did not differ significantly across racial/ethnic groups, $\chi^2(2) = 5.78$, *ns*. The average age of the primary caregivers completing the DOTS–R was 38.7 years ($SD = 7.4$). Approximately 56.7% ($n = 2{,}665$) were currently married, 7.2% ($n = 338$) were living with a partner, 69.2% ($n = 3{,}253$) were working part time or full time, 24.8% ($n = 1{,}166$) had not graduated from high school, 20.0% ($n = 940$) had a GED or high school degree but had not attended college, and 55.2% ($n = 2{,}595$) attended some college.

Procedures

All three Healthy Passages research sites used standardized data collection materials and protocols, including training manuals, field manuals, and validation procedures. Both computer-assisted personal interviews (CAPI) and audio computer-assisted interviews (A-CASI) were used to collect data from participants. Institutional review boards at each study site and the Centers for Disease Control and Prevention approved the study. On average, it took about 3 hr for the field interviews to be completed. Primary caregivers were paid $50 and children were given a $20 gift card from a national chain store as reimbursement for completing the interview.

Measures

Revised Dimensions of Temperament Survey abbreviated form. The DOTS–R (Windle & Lerner, 1986) form used in this study assessed five factors with 23 of the original 33 items to assess these five dimensions. Findings on the full-scale DOTS–R have provided evidence of a replicated factor structure, moderate to high reliability for the derived dimensions, high parent–child concordance, and concurrent and prospective validity (Luby & Steiner, 1993; Merikangas, Swendsen, Preisig, & Chazan, 1998; Windle & Windle, 2006). The following are the five temperament dimensions assessed and the corresponding number of items per dimension in parentheses: behavioral flexibility (4), general activity level (5), sleep rhythmicity (5), positive mood (4), and task orientation (5). Cronbach's alphas, in sequence, for these dimensions in the original Windle and Lerner (1986) study were .62, .75, .69, .80, and .70. The DOTS–R was administered at Wave 1 only during the A-CASI with the primary caregiver, and primary caregivers completed measures on externalizing behaviors and socioemotional competence.

Externalizing behaviors. The presence of symptoms of conduct disorder (8 items) and oppositional defiant disorder (10 items) in the past year was assessed by primary caregivers with 18 items adapted from the Diagnostic Interview Schedule for Children Predictive Scales (DPS; Leung et al., 2005; Lucas et al., 2001). The DPS is a widely used screening tool that is based on the Diagnostic Interview Schedule for Children (K. W. Chen, Killeya-Jones, & Vega, 2005; Leung et al., 2005). Primary caregivers rated the presence of each symptom on a dichotomous scale (1 = *yes,* 0 = *no*) during the A-CASI portion of the interview. Subscale scores were calculated by summing affirmative responses across the items. Symptoms of conduct disorder and oppositional defiant disorder were summed to form a measure of externalizing behaviors. The internal consistency estimate for this measure was $\alpha = .81$, and the intraclass correlation for externalizing behaviors in relation to the school cluster variable was .029. The intraclass correlation represents the proportion of variance in the outcome variable (externalizing behaviors) that is between groups

(i.e., the Level-2 units, which in this study refers to schools that were sampled). The larger these effects are (as indicated by a higher intraclass correlation), the greater is the clustering effect of schools (i.e., children with highly similar externalizing behaviors are more likely to be in the same school). Lower values, such as those provided for the externalizing behaviors, reflect minimal school selection effects.

Socioemotional competence. The Social Skills scale from the Social Skills Rating System (Gresham & Elliott, 1990) was completed by primary caregivers in relation to their children. Each of 26 behavior-based items was rated on a 3-point scale ranging from 1 (*never occurred*) to 3 (*very often*). This scale measures aspects of communication, empathy, cooperation, and assertion (Gresham & Elliott, 1990). Individuals scoring high on this scale have higher levels of socioemotional competence in interacting with others. The internal consistency estimate for this measure was $\alpha = .85$, and the intraclass correlation for socioemotional competence in relation to the school cluster variable was .087. Information on sociodemographic characteristics was mostly gathered during the CAPI with the primary caregiver.

Statistical Analysis

As described previously, initial analyses focused on evaluating the number of factors underlying this 23-item abbreviated version of the DOTS–R. Scree plots, parallel analysis, and findings from maximum likelihood factor analyses were used for this purpose. Then, guided by previous research on specifying, testing, and evaluating hypotheses about invariance relations across groups (Meredith, 1993; Millsap, 2011; Widaman & Reise, 1997), a sequential testing procedure was used to evaluate invariance relations across racial/ethnic groups. These models included a simultaneous multigroup model to evaluate configural invariance across groups (i.e., did the same number of factors represent the three racial/ethnic groups similarly?). Then, group-equality constraints were imposed on all factor loadings to evaluate the weak factorial invariance model. Additional group-equality constraints were imposed on all indicator intercepts to test the strong factorial invariance model.

Analyses were conducted using the statistical software program M*plus* 7.2 (Muthén & Muthén, 1998–2012). All analyses were performed with design weights (to account for differential probabilities of selection of students according to their school) and a cluster variable (to account for clustering of students within schools). A two-level model was used to estimate intraclass correlations for the school cluster variable and ranged across DOTS–R items from .001 to .073, with $M = .02$. All models were tested using the robust maximum likelihood (MLR) estimator, which accounts both for nonnormality and dependence due to the clustering of students within schools. Specifically, MLR uses the pseudomaximum likelihood (PML) asymptotic covariance matrix and a scaled test statistic ($MLR\chi^2$) that is asymptotically equivalent to the Yuan–Bentler T2* test statistic (Asparouhov, 2005). Within M*plus,* the MLR estimation procedure is recommended for use with complex sampling designs such as the one used in this study (Muthén & Muthén, 1998–2012). It also assisted in that Mardia's estimate of multivariate kurtosis was 80.55, suggesting violations of multivariate normality. MLR estimation provides robust estimates even if the multivariate normality assumption is violated.

Evaluation of the specified models was based on multiple criteria that considered statistical, practical, and substantive fit. Following the recommendations of Hu and Bentler (1999), we used the comparative fit index (CFI; Bentler, 1990), the root mean square error of approximation (RMSEA; Steiger, 1990), and the standardized root mean square residual (SRMR; Hu & Bentler, 1999). The CFI ranges in value from zero to one; values equal to or greater than .95 typically reflect good model fit (Hu & Bentler, 1999). The RMSEA is a measure of a model's approximate fit in the population. Values less than .05 indicate good fit, and values as high as .08 represent acceptable errors of approximation in the population (Browne & Cudeck, 1993; Steiger, 1990). Finally, the SRMR is the average standardized residual value derived from fitting the hypothesized variance–covariance matrix to that of the observed data. It ranges from zero to one, with a value less than .08 indicating good model fit (Hu & Bentler, 1999).

In making comparisons among nested hierarchical models, it has been suggested that the chi-square difference test often detects inconsequentially small differences in loadings when sample sizes are large (Cheung & Rensvold, 2002; Little, 1997; Marsh, Hau, & Wen, 2004), thereby limiting its usefulness as a practical criterion when testing invariance constraints with large samples such as the one used in this study. For this reason, we additionally inspected the practical fit of models with group-equality constraints (for a discussion, see Byrne & Stewart, 2006; Little, 1997). However, simulation studies have indicated that many of the practical fit indexes have similar properties (F. Chen, 2007; Hu & Bentler; 1999) when testing is completed with large samples. Simulation studies on goodness-of-fit indexes to examine model invariance hypotheses have been conducted and general recommendations formulated to assist in model evaluation and fit (F. Chen, 2007; Cheung & Rensvold, 2002).

To determine whether the fit of more restrictive invariance models deteriorates significantly, we also used criteria regarding the magnitude of observed differences suggested by F. Chen (2007), along with substantive considerations (Marsh et al., 2004). Accordingly, metric noninvariance (i.e., for the weak invariance model) was indicated by a change larger than –.01 in CFI, supplemented by a change larger than .015 in RMSEA or a change larger than .03 in SRMR compared with the configural invariance model. Regarding scalar invariance (i.e., for the strong invariance model) noninvariance was indicated by a change larger than –.01 in CFI, supplemented by a change larger than .015 in RMSEA or a change larger than .01 in SRMR compared with the metric invariance model. Importantly, we evaluated all three difference fit indexes and included consistency across fit indexes in determining the adequacy of the proposed invariance hypotheses.

RESULTS

Number of Factors

Three methods were used to evaluate the adequacy of the number of factors to represent the DOTS–R items. First, based on an exploratory factor analysis, five eigenvalues exceed 1.0

TABLE 1.—Model fit of exploratory factor models ranging from one to six factors.

No. Factors	$ML\chi^2$	*df*	CFI	RMSEA	SRMR
One-factor	9,576.51	230	.341	.093	.122
Two-factor	5,889.26	208	.599	.076	.083
Three-factor	2,940.12	187	.806	.056	.050
Four-factor	1,697.60	167	.892	.044	.032
Five-factor	931.13	148	.945	.034	.021
Six-factor	617.11	130	.966	.028	.017

Note. All models were adjusted for sample weights and clusters. $ML\chi^2$ = maximum likelihood chi-square test statistic; CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual.

(the values of these five were 3.85, 3.05, 2.22, 1.63, and 1.22) and accounted for 52% of the variance. Second, parallel analysis was performed, and five factors from the actual data with eigenvalues that were greater than those from random data were retained (Horn, 1965). Third, maximum likelihood factor analytic models were conducted with factors ranging from 1 to 6. As summarized in Table 1, the fit indexes for the five-factor model accounted well for the data, and item-factor relations of salient and nonsalient factor loadings were consistent with the hypothesized five-factor structure of the measure. The six-factor representation resulted in a fractionation of items across factors that made them difficult to interpret (i.e., a mixture of items identified an additional factor but the substantive meaning of this factor was not evident). Hence, the five-factor model was retained for subsequent analyses. Parameter estimates corresponding to factor loadings for the five-factor model are provided in Table 2 and support the proposed five-factor structure with salient loadings corresponding to items that were anticipated to load on their referent factor. Factor intercorrelations, provided at the bottom of Table 2, indicate that higher flexibility was associated with lower general activity level but that the remainder of the correlations were relatively low in magnitude.

ESEM Versus Standard Confirmatory Factor Model

A five-factor representation consistent with the structure identified in the exploratory model was specified and estimated for a standard confirmatory factor analytic model and for an ESEM to evaluate the adequacy of this representation for the full sample. For the ESEM, 25 constraints (five per factor) were imposed on factor loadings that had the lowest estimated value in the full exploratory factor analytic solution. For the confirmatory factor model, the salient loadings (shown in bold in Table 2) were freely estimated and the remaining item-factor relations were fixed to zero. Information regarding fit for each of these models is provided in Table 3, along with the fit statistics for the five-factor exploratory model and a modified ESEM. The confirmatory model did not provide an adequate fit to data, and the difference in CFI between this model (M2) and M1 was .203, thereby exceeding the criteria for adequate fit. The initial ESEM (M3) provided an adequate fit relative to M1, and the difference in CFI was .003, consistent with the criteria outlined previously for evaluating nested models. Nevertheless, the CFI value was .941, somewhat below the desired .95; to accommodate this finding, on the basis of modification indexes, two within-factor correlated errors (Items 5 and 11; Items 17 and 18) were freely estimated and this improved the CFI to .96 (M4). This modified model provided support of overall fit for the five-factor

TABLE 2.—Factor loadings and factor intercorrelations of five-factor exploratory model.

DOTS–R: Item description	Flexibility	Activity level	Rhythmicity sleep	Task orientation	Positive mood
1. Time to adjust to new thing in home	**.46**	−.09	.02	−.09	.05
7. Time to adjust to new schedules	**.54**	−.19	.10	−.03	.00
19. When things out of place, difficulty get used to	**.70**	.03	−.06	.02	−.02
21. Resists changes in routines	**.60**	−.01	−.08	.04	.01
2. Can't stay still for long	.02	**.72**	.01	−.01	−.02
6. Gets restless if has to stay in one place	−.04	**.60**	−.02	.02	−.05
9. Stays still for long periods	.22	**.45**	−.02	−.32	−.01
12. Gets fidgety	−.09	**.66**	.00	−.01	.01
15. Never seems to stop moving	−.03	**.71**	.05	−.04	.05
3. Wakes up at different times	.15	−.29	**.29**	−.15	−.03
13. Gets same amount of sleep each night	.02	−.01	**.58**	.03	.08
16. Gets sleepy at same time every night	−.06	−.02	**.63**	−.02	.10
17. When away from home, wakes up at same time	.02	.02	**.58**	.05	−.03
18. No matter goes to sleep, wakes up at same time	−.08	.02	**.48**	.08	−.05
4. Once involved, cannot distract	.02	.04	−.01	**.54**	−.03
5. Persists at tasks until finished	.10	−.02	.04	**.56**	.09
8. Something else occurring will not disturb focus	−.10	.08	−.04	**.51**	−.03
10. Other things happening will not distract	−.07	.09	−.01	**.59**	−.02
11. Once child initiates a task, stays with it	.14	−.02	.08	**.48**	.09
14. Smiles often	.05	.11	.10	.05	**.59**
20. Generally cheerful	.01	−.02	.06	.02	**.65**
22. Laughs frequently	−.01	.04	−.04	−.02	**.74**
23. Happy	−.03	−.07	−.05	−.01	**.74**
Flexibility	1.0				
Activity level	−.45	1.0			
Rhythmicity sleep	.07	−.14	1.0		
Task orientation	−.21	−.06	.21	1.0	
Positive mood	.19	−.05	.32	.18	1.0

Note. Bold indicates highest factor loadings on each factor. DOTS–R = Revised Dimensions of Temperament Survey.

TABLE 3.—Fit statistics for exploratory, confirmatory, and exploratory structural equation models for full sample.

Model(M)	$MLR\chi^2$	*df*	CFI	RMSEA	SRMR	Model comparison	ΔCFI
M1: Exploratory model	931.13	148	.945	.034	.021	—	—
M2: Confirmatory model	4,461.55	267	.742	.058	.075	2 vs. 1	.203
M3: ESEM	985.85	154	.941	.034	.024	3 vs. 1	.003
M4: ESEM-2 correlated errors	711.26	152	.960	.028	.021	—	—

Note. All models were adjusted for sample weights and clusters. $MLR\chi^2$ = robust maximum likelihood chi-square test statistic; CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual; ESEM = exploratory structural equation model.

representation across the full sample and this model was used in subsequent simultaneous multigroup analyses to evaluate invariance hypotheses across the three racial/ethnic groups.

Multigroup Racial/Ethnic Models

A summary of the fit statistics to evaluate the sequential invariance hypotheses across groups is shown in Table 4. Findings from the independent racial/ethnic group models indicated acceptable fit indexes for all groups, although the CFI for the Hispanic sample was somewhat lower than for the other two samples. The configural invariance hypothesis was tested using the model specification of the ESEM (M4). For this test of configural invariance, the same number of factors and the same factor loading pattern were specified across racial/ethnic groups; however, there were no parameter constraints relating to the equality of estimated parameters across groups. The configural invariance hypothesis was supported, with all fit indexes meeting the fit criteria that had been described previously. Similarly, the weak invariance hypothesis, in which factor loadings were constrained to equivalence across racial/ethnic groups, was supported as indicated both by the model fit statistics and the difference in the CFI between the configural invariance model and weak invariance model, which was equal to .005, well within levels of acceptability (F. Chen, 2007).

The strong invariance hypothesis, in which both the factor loadings and the intercepts were constrained to equivalence across groups, was weakly supported; the change in CFI (.023) exceeded the criterion established by F. Chen (2007), although it met the criteria for differences in the other two model fit indexes (RMSEA and SRMR). Also, the CFI was .927, somewhat lower than desired. To accommodate these minor misfit issues, we sequentially freed five intercepts and the resulting model (Table 4, M4) provided more adequate fit both in terms of the change in CFI value (.009) and overall fit (CFI = .941). These modifications (i.e., freeing five intercepts) indicated differential item functioning for these five items (Items 6, 12, 15, 16, 18) in that racial/ethnic group differences at the level of the intercepts could not be fully explained in terms of latent means, and thus only partial invariance of intercepts was supported in the model (Marsh et al., 2013). Further racial/ethnic group comparisons among these five intercepts (three associated with general activity and two with sleep rhythmicity) by using constrained group analyses indicated that Blacks and Hispanics differed from Whites on Items 12, 16, and 18 (Whites scored lower, whereas Blacks and Hispanics did not differ significantly from one another); Blacks differed from Hispanics and Whites on Item 6 (Blacks scored higher, whereas Hispanics and Whites did not differ significantly from one another); and Hispanics differed from Blacks and Whites on Item 15 (Hispanics scored higher, whereas Blacks and Whites did not differ significantly from one another).

The range of standardized factor loadings for M4 was also very similar across the three groups and highly similar to those reported for the five-factor model of the full sample in Table 2. For example, in the multigroup model, the factor loadings for flexibility ranged from .45 to .64 ($M = .53$); for the full sample model, the factor loadings for flexibility ranged from .46 to .70 ($M = .57$). Similarly, for general activity in the multigroup model, the factor loadings ranged from .45 to .72 ($M = .63$); in the full sample model, factor loadings ranged from .34 to .68 ($M = .57$). Similar relationships for multigroup and full-sample models were indicated for the other factors in the model. Furthermore, similar to the full-sample model, the factor loadings on nonreferent factors for the multigroup model were of low magnitude. For these reasons, the assumption of factor loading invariance was retained in subsequent analyses.

Table 5 summarizes differences in the observed means across the three racial/ethnic groups. The findings indicated

TABLE 4.—Testing for invariance of five-factor model across White ($n = 1,247$), Black ($n = 1,721$), and Hispanic ($n = 1,733$) early adolescents.

Model (M)	$MLR\chi^2$	*df*	CFI	RMSEA	SRMR	Model comparison	ΔCFI
Independent group							
Hispanic	499.65	156	.938	.036	.029	—	—
Black	421.20	156	.956	.031	.042	—	—
White	431.45	156	.951	.038	.061	—	—
M1E: Configural invariance	1,352.34	468	.949	.035	.044	—	—
M2: Weak factor invariance (FI)	1,498.69	637	.950	.029	.039	2 vs. 1	.005
M3: Strong FI	1,945.38	673	.927	.035	.042	3 vs. 2	.023
M4: Strong FI + 5 intercepts freed	1,689.30	663	.941	.031	.039	4 vs. 2	.009

Note. All models were adjusted for sample weights and clusters. $MLR\chi^2$ = robust maximum likelihood chi-square test statistic; CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual.

TABLE 5.—Observed mean level comparisons of temperament across racial and ethnic groups.

DOTS–R factors	Hispanic (H)	Black (B)	White (W)	F-statistic	Post-hoc tests and effect sizes
Flexibility	11.27	12.25	12.86	123.04***	W > B, H (.12, .55); B > H (.34)
General activity	12.70	12.30	11.34	46.24***	H > B, W (.10, .35); B >W (.25)
Sleep rhythmicity	15.35	14.81	15.97	49.79***	W > H, B (.17, .19); H > B (.36)
Positive mood	14.55	14.52	14.72	4.27*	W > B (.10)
Task orientation	13.00	12.90	12.72	2.83 (*ns*)	*ns*

Note. Post-hoc tests were conducted via Bonferroni multiple comparisons. DOTS–R = Revised Dimensions of Temperament Survey.
*$p < .05$. ***$p < .001$.

that White, relative to Black and Hispanic early adolescents, were rated as more flexible, as having lower activity levels, and as sleeping with greater regularity. White, relative to Hispanic early adolescents, were rated as more flexible, as having lower general activity levels, and as sleeping with more regularity. White, relative to Black early adolescents, were also rated as having a more positive mood. Black, relative to Hispanic, early adolescents were rated as more flexible and with lower regularity of sleeping. Using Cohen's (1988) cutoffs for ES, all but one pairwise group comparison yielded small ESs (i.e., $\leq .30$). The exception to this was a moderate ES for the difference for flexibility between Hispanic and White early adolescents, where the ES was .55.

Reliability of Abbreviated Version of the DOTS–R

The internal consistency reliability coefficients (Cronbach's alphas) for the five temperament dimensions are provided in Table 6 for each of the racial/ethnic groups, the total sample, and the original sample used in developing the DOTS–R (Windle & Lerner, 1986).

The numeric values of these coefficients largely fell within the acceptable .70 target value with the exception of rhythmicity sleep for the Hispanic sample, which was .53. Across dimensions, the average reliability for Hispanics was .66, for Blacks it was .72, and for Whites it was .78. Overall, these α coefficients illustrate a relatively high level of similarity across the racial/ethnic groups and in relation to the original sample (average reliability across dimensions was .72). Hence, with some allowance for the lower reliability of the rhythmicity sleep dimension for Hispanics, there is support that this abbreviated version of the DOTS–R used in Healthy Passages yields acceptable levels of reliability for these five dimensions. Further statistical comparisons of racial/ethnic group differences among these alpha levels using the Fisher–Bonett approach (Kim & Feldt, 2008) for the Healthy Passages samples indicated that alphas for flexibility, general activity, and positive mood differed significantly ($p < .01$) for White early adolescents relative to Hispanic and Black early adolescents. White early adolescents also differed significantly ($p < .01$) from Hispanic early adolescents on task orientation. Black early adolescents differed significantly ($p < .01$) from Hispanic early adolescents on general activity.

TABLE 6.—Reliability coefficients (αs) for five temperament factors across racial and ethnic groups.

Temperament dimension	Hispanic[a]	Black[b]	White[c]	Total[d]	Original sample[e]
Flexibility	.67	.67	.78	.71	.62
General activity	.68	.80	.85	.78	.75
Sleep rhythmicity	.53	.69	.73	.65	.69
Positive mood	.76	.74	.84	.77	.80
Task orientation	.65	.68	.71	.67	.70

[a]$n = 1{,}733$. [b]$n = 1{,}721$. [c]$n = 1{,}247$. [d]$N = 4{,}701$. [e]$n = 224$; reliability data (Cronbach's alpha) for child sample from Windle and Lerner (1986).

Construct Validity of Abbreviated Version of the DOTS–R

The two factors of externalizing behaviors and socioemotional competence were selected to evaluate the construct validity of the DOTS–R. Consistent with prior theory and research (Chang et al., 2003; Thomas & Chess, 1984; Windle et al., 1986), temperament dimensions derived from the abbreviated DOTS–R were generally significantly correlated with these two factors in the anticipated direction (see Table 7). The magnitude of these correlations was also consistent with our low-to-moderate anticipated levels of construct validity in the .2 to .4 range. Lower flexibility (more inflexibility), higher activity levels, lower task orientation, less regularity of sleeping, and a less positive mood were significantly associated with externalizing behaviors. Higher flexibility, lower general activity, higher sleep regularity, higher task orientation, and positive mood were significantly associated with higher socioemotional competence.

DISCUSSION

This study used ESEMs with a large, multiracial community sample of fifth-graders to examine psychometric characteristics of an abbreviated version of the DOTS–R. A series of analyses were conducted, including an evaluation of the number of factors, the invariance of the factor structure, internal consistency estimates, and construct validity analyses. Findings indicated that the hypothesized five-factor structure of

TABLE 7.—Correlations between temperament factors, externalizing behaviors, and socioemotional competence.

	Hispanic	Black	White
Externalizing			
Flexibility	−.03	−.28***	−.20***
Activity level	.20***	.35***	.26***
Rhythmicity sleep	−.15***	−.12***	−.21***
Task orientation	−.15***	−.26***	−.20***
Positive mood	−.22***	−.24***	−.40***
Socioemotional competence			
Flexibility	.29***	.30***	.36***
Activity level	−.15***	−.28***	−.27***
Rhythmicity sleep	.20***	.24***	.25***
Task orientation	.07**	.36***	.21***
Positive mood	.26***	.29***	.46***

**$p < .01$. ***$p < .001$.

the DOTS–R provided an acceptable representation of the data across the full sample, and largely across all three racial/ethnic groups. For the multigroup comparisons, configural invariance was supported, indicating that the five-factor model provided an equally good representation for each of the three racial/ethnic groups. The weak factorial invariance model was also supported, indicating that the factor loadings could be constrained to equivalence across groups without significant departures in goodness of model fit. The strong factorial invariance model was also partially supported, indicating that the factor loadings and intercepts could be constrained to equivalence across groups without significant departures in goodness of model fit. However, to achieve optimal levels of model fit for the strong factorial invariance model, it was necessary to free five intercepts across groups; thus, there was only partial invariance for the intercepts across the three racial/ethnic groups as the intercepts for 18 items could be constrained to equivalence, but the intercepts for 5 items could not be constrained. This finding suggests some minor differential item functioning for these items across groups, with three of these items from the general activity factor and two from the rhythmicity sleep factor. Nevertheless, the overall factor structure for abbreviated factor structure of the DOTS–R was best represented by a five-factor model and invariant relationships for most parameters (i.e., all factor loadings and 18 of 23 intercepts) were supported across three racial/ethnic groups. This suggests that the underlying factor structure of the abbreviated DOTS–R as rated by primary caregivers was largely equivalent for Black, Hispanic, and White children.

To our knowledge, this is the first study to apply an invariance testing strategy in the form of mean and covariance structure analysis to the abbreviated DOTS–R to examine the comparability of its hypothesized factor structure across different racial/ethnic groups. The overall findings of this study were consistent with prior U.S. and international research (Marcet et al., 2000; Windle et al., 1987) in supporting the multidimensional structure of the DOTS–R with regard to the factor loading pattern, although more rigorous statistical testing was conducted in this study relative to descriptive indexes (e.g., congruence coefficients, pattern similarity index) used in prior studies.

In addition to tests regarding the number of factors and invariance issues across racial/ethnic groups, several other analyses were completed. Observed mean level analyses indicated that the three racial/ethnic groups did not differ with regard to task orientation. However, White early adolescents were rated by their primary caregivers as having higher levels of behavioral flexibility and sleep rhythmicity than Black or Hispanic early adolescents. They were also rated as having lower levels of general activity than Black early adolescents. Black early adolescents were also rated by their primary caregivers as having greater flexibility and lower sleep rhythmicity than Hispanic early adolescents. Hence, there were some overall differences in mean levels across racial/ethnic groups, especially for flexibility, general activity, and sleep rhythmicity, although effect sizes were typically small. The internal consistency estimates for the temperament dimensions of the abbreviated DOTS–R were somewhat higher for Whites and somewhat lower for Hispanic early adolescents, although the overall levels were highly similar to those reported with the full-scale DOTS–R, with levels appropriate for research purposes.

The construct validity findings for the abbreviated DOTS–R were consistent with prior theorizing and prior empirical findings with regard to similarities of low-to-moderate correlations between temperament scores and externalizing problems and socioemotional competence (Chang et al., 2003; Thomas & Chess, 1984; Windle et al., 1986). Specifically, a difficult temperament profile characterized by inflexibility, high activity levels, irregular sleep, low task orientation, and low positive mood has been associated with externalizing problems in numerous studies (Giancola & Mezzich, 2003; Merikangas et al., 1998). Similarly, a profile characterized by higher flexibility, greater regularity of sleep, higher task orientation, a more positive mood, and lower general activity has been associated with higher levels of social and emotional competence (Windle et al., 1986). These findings are consistent with past research that has emphasized how temperament dimensions serve as risk or protective factors in influencing associations with factors such as externalizing problems and socioemotional competence (Giancola & Mezzich, 2003; Thomas & Chess, 1984; Windle et al., 1986).

A noteworthy feature of this study was that primary caregivers were used as reporters to minimize potential monomethod reporter effects when investigators want to study relationships between temperament and other variables reported by children. Maternal reports have been used in the temperament literature for infants and children, but less often for children in, or nearing, early adolescence. Extant literature has demonstrated high rates of parent–child agreement for the DOTS–R (Luby & Steiner, 1993), and this study further supports that primary caregivers can reliably report on their child's temperament. This is not to suggest that other biases (e.g., halo effects, maternal depression) might not occur with primary caregiver ratings, only that the primary caregivers were able to provide reliable ratings. The size of the reliability coefficients (i.e., the alpha levels) for the temperament dimensions were generally similar across racial/ethnic groups in this study and with those reported in previous studies that used the full, versus abbreviated, DOTS–R. This suggests that the abbreviated version provides an equally reliable assessment of the DOTS–R compared with the full form.

Although findings of this study were largely supportive of the invariance of the abbreviated DOTS–R across three racial/ethnic groups, they should be viewed in the context of study limitations. First, the DOTS–R is a questionnaire (self- or other report) measure, and, as with all report instruments, responses might have been influenced by confounds such as informant bias (halo effects by primary caregivers regarding their children) and biases associated with psychopathological characteristics of the rater (e.g., maternal depression). Behavioral observation data or physiological assessments would enhance the value of temperament assessment and would facilitate further validation of this abbreviated measure. Second, the limited group sizes of some racial/ethnic groups (e.g., Asian Americans, subgroups of Hispanics from South American countries, Cuba, or Puerto Rico vs. our sample of primarily Mexican heritage) precluded systematic testing of weak and strong factorial invariance of the DOTS–R for these groups. Third, the item response options for the DOTS–R contain 4-point options, and it is possible that these data might be

equally or better modeled as categorical with an alternative estimator such as robust weighted least squares. The impact of this issue in this application is reduced to some extent by the estimator we chose because it produces robust estimates even if data are nonnormally distributed. Furthermore, other characteristics of the data (e.g., sample size, symmetric vs. asymmetric data distribution) might affect the relative value of modeling data such as these as continuous versus categorical variables. Fourth, this was a study of reports by primary caregivers regarding their children. Although prior research has indicated high levels of interrater agreement between primary caregivers and their offspring with the DOTS–R (Luby & Steiner, 1993), and between self- and peer ratings (Oniszczenko et al., 2003), in this study with the abbreviated form we do not know if child self-assessments would have corresponded well with primary caregiver assessments.

Nevertheless, this study also had several strengths, including the large sample size and data from three racial/ethnic groups, which facilitated our efforts to address issues related to measurement invariance that are critical for making group comparisons. Further research on this issue should include similar analyses for other racial/ethnic groups, such as Asian Americans, the collection of validity data via other methods of assessment (e.g., physiological or behavioral assessments), and longitudinal associations between the DOTS–R dimensions and important health-related behaviors such as substance use, depression, and delinquency.

Funding

The Healthy Passages study was funded by the Centers for Disease Control and Prevention (Cooperative Agreements U48DP000046, U48DP000057, and U48DP000056). The findings and conclusions in this report are those of the author (s) and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

References

Asparouhov, T. (2005). Sampling weights in latent variable modeling. *Structural Equation Modeling*, *12*, 411–434.

Asparouhov, T., & Muthen, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling*, *16*, 397–438.

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

Bollen, K. A., Harden, J. J., Ray, S., & Zavisca, J. (2014). BIC and alternative information criteria in the selection of structural equation models. *Structural Equation Modeling*, *21*, 1–19.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Byrne, B. M., & Stewart, S. M. (2006). The MACS approach to testing for multigroup invariance of a second-order structure: A walk through the process. *Structural Equation Modeling*, *13*, 287–321.

Chang, K. D., Blasey, C. M., Ketter, T. A., & Steiner, H. (2003). Temperament characteristics of child and adolescent bipolar offspring. *Journal of Affective Disorders*, *77*, 11–19.

Chen, F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, *14*, 464–504.

Chen, K. W., Killeya-Jones, L. A., & Vega, W. A. (2005). Prevalence and co-occurrence of psychiatric symptom clusters in the U.S. adolescent population using DISC predictive scales. *Clinical Practice and Epidemology in Mental Health*, *1*, 22.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Giancola, P. R., & Mezzich, A. C. (2003). Executive functioning, temperament, and drug use involvement in adolescent females with a substance use disorder. *Journal of Child Psychology and Psychiatry*, *44*, 857–866.

Gresham, F., & Elliott, S. N. (1990). *Social skills rating system manual*. Circle Pines, MN: American Guidance Service.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *32*, 179–185.

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

Kim, S., & Feldt, L. S. (2008). A comparison of tests for equality of two or more independent alpha coefficients. *Journal of Educational Measurement*, *45*, 179–193.

Lerner, R. M., Palermo, M., Spiro, A., & Nesselroade, J. (1982). Assessing the dimensions of temperament individuality across the life-span: The Dimensions of Temperament Survey. *Child Development*, *53*, 149–159.

Leung, P. W., Lucas, C. P., Hung, S. F., Kwong, S. L., Tang, C. P., Lee, C. C., . . . Shaffer, D. (2005). The test–retest reliability and screening efficiency of DISC Predictive Scales–version 4.32 (DPS–4.32) with Chinese children/youths. *European Child and Adolescent Psychiatry*, *14*, 461–465.

Little, T. D. (1997). Means and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research*, *32*, 53–76.

Luby, J. L., & Steiner, H. (1993). Concordance of parent and child temperament ratings in a clinical sample of adolescent girls. *Child Psychiatry & Human Development 23*, 297–305.

Lucas, C. P., Zhang, H., Fisher, P. W., Shaffer, D., Regier, D. A., Narrow, W. E., . . . Friman, P. (2001). The DISC Predictive Scales (DPS): Efficiently screening for diagnoses. *Journal of the American Academy of Child and Adolescent Psychiatry*, *40*, 443–449.

Marcet, C., Guardia, J., Almirall, H., & Lorenzo, U. (2000). Factorial structure of the Spanish version of the DOTS–R questionnaire. *European Journal of Psychological Assessment*, *16*, 59–65.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's findings. *Structural Equation Modeling*, *11*, 320–341.

Marsh, H. W., Ludtke, O., Muthen, B., Asparouhov, T., Morin, A. J. S., Trautwein, U., & Nagengast, B. (2010). A new look at the big-five factor structure through exploratory structural equation modeling. *Psychological Assessment*, *22*, 471–491.

Marsh, H. W., Nagengast, B., & Morin, A. J. S. (2013). Measurement invariance of Big-Five Factors over the life span: ESEM tests of gender, age, plasticity, maturity, and La Dolce Vita effects. *Developmental Psychology*, *49*, 1194–1218.

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

Merikangas, K. R., Swendsen, J. D., Preisig, M. A., & Chazan, R. Z. (1998). Psychopathology and temperament in parents and offspring: Results of a family study. *Journal of Affective Disorders*, *5*, 163–74.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance.* New York, NY: Routledge.

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Oniszczenko, W., Zawadzki, B., Strelau, J., Reiman, R., Angleitner, A., & Spinath, F. M. (2003). Genetic and environmental determinants of temperament: A comparative study based on Polish and German samples. *European Journal of Personality*, *17*, 207–220.

Schuster, M. A., Elliott, M. N., Kanouse, D. E., Wallander, J. L., Tortolero, S. R., Ratner, J. A., . . . Banspach, S. W. (2012). Racial and ethnic health disparities among fifth-graders in three cities. *New England Journal of Medicine*, *367*, 735–745.

Shaw, R. J., & Steiner, H. (1997). Temperament in juvenile eating disorders. *Psychosomatics*, *38*, 126–131.

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

Thomas, A., & Chess, S. (1984). Genesis and evolution of behavioral disorders: From infancy to early adult life. *American Journal of Psychiatry*, *141*, 1–9.

Tubman, J. G., & Windle, M. (1995). Continuity of difficult temperament in adolescence: Relations with depression, life events, family support, and substance use across a one year period. *Journal of Youth and Adolescence*, *24*, 133–153.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention* (pp. 281–324). Washington, DC: American Psychological Association.

Windle, M. (1999). Temperament and psychopathology: Alternative models and developmental pathways. In J. Mervielde, I. Dreary, F. DeFruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe* (Vol. 7, pp. 159–173). Tilburg, The Netherlands: Tilburg University Press.

Windle, M., Grunbaum, J. A., Elliott, M., Tortelero, S., Berry, S., Gilliland, J., . . . Schuster, M. A. (2004). Healthy Passages: A multilevel, multimethod longitudinal study of adolescent health. *American Journal of Preventive Medicine*, *27*, 164–172.

Windle, M., Hooker, K., Lenerz, K., East, P. L., Lerner, J. V., & Lerner, R. M. (1986). Temperament, perceived competence, and depression in early- and late-adolescents. *Developmental Psychology*, *22*, 384–392.

Windle, M., Iwawaki, S., & Lerner, R. (1987). Cross-cultural comparability of temperament among Japanese and American early and late adolescents. *Journal of Adolescent Research*, *2*, 423–446.

Windle, M., & Lerner, R. M. (1986). Reassessing the dimensions of temperamental individuality across the life span: The Revised Dimensions of Temperament Survey (DOTS–R). *Journal of Adolescent Research*, *1*, 213–229.

Windle, M., & Windle, R. C. (2006). Adolescent temperament and lifetime psychiatric and substance abuse disorders assessed in young adulthood. *Personality and Individual Differences*, *41*, 15–25.