Dimensions of Oppositional Defiant Disorder in Young Children: Model Comparisons, Gender and Longitudinal Invariance

John V. Lavigne · Fred B. Bryant · Joyce Hopkins · Karen R. Gouze

Published online: 13 August 2014

Abstract Identifying the latent structure of Oppositional Defiant Disorder (ODD) may have important clinical and research implications. The present study compared existing dimensional models of ODD for model fit and examined the metric and scalar invariance of the best-fitting model. Study participants were a diverse (38.8 % minority, 49.1 % boys) community sample of 796 children. Parents completed the Child Symptom Inventory and the DISC-YC ODD scales at child ages of 4, 5 and 6–7 years. When comparing single-factor (DSM-IV model), two-factor (oppositional behavior, negative affect), and three-factor models (one with dimensions of oppositional behavior, negative affect, antagonistic behavior; a second with dimensions of irritable, hurtful, and headstrong), the two-factor model showed the best fit. The two-factor model showed configural, metric and scalar invariance across gender and age. Results suggest that, among existing models, ODD is best characterized as two separate dimensions, one behavioral and one affective, which are comparable for both boys and girls in these age groups.

Keywords Oppositional defiant disorder · ODD dimensions · Preschoolers · Developmental psychopathology

Oppositional Defiant Disorder (ODD) is one of the most commonly-occurring disorders in young children (Egger and Angold 2006; Lavigne et al. 2009). Relationships between early ODD and later developing conduct disorder (Burke and Loeber 2010), depression (Burke et al. 2010; Burke et al. 2005; Copeland et al. 2009; Lavigne et al. 2001) and anxiety (Drabick and Kendall 2010; Lavigne et al. 2001) are well established. However, less is known about whether ODD comprises a single construct or consists of multiple dimensions that differentially predict later externalizing or internalizing disorders. Recent studies have provided support for a multi-dimensional structure to ODD in which particular dimensions differentially predict later psychopathology. For example, in a sample of boys, Burke and Loeber (2010) found that ODD consisted of two dimensions, an affective dimension and a behavioral dimension, with the affective dimension associated with later depression. In a sample of girls, a model of ODD with three dimensions was identified; the negative affect dimension was associated with depression, while oppositional and antagonistic behavior were associated with subsequent conduct disorder (Burke et al. 2010). Similarly, other studies have found that an irritability dimension of ODD predicted emotional problems (Stringaris and Goodman 2009a, b) or anxiety (Rowe et al. 2010), while a headstrong dimension predicted later hyperactivity (Stringaris and

This research was supported by National Institute of Mental Health grant MH 063665.

Electronic supplementary material The online version of this article (doi:10.1007/s10802-014-9919-0) contains supplementary material, which is available to authorized users.

J. V. Lavigne (✉) · K. R. Gouze
Department of Child and Adolescent Psychiatry (#10), Ann & Robert H. Lurie Children's Hospital of Chicago, 225 East Chicago Avenue, Chicago, IL 60611, USA
e-mail: jlavigne@luriechildrens.org

J. V. Lavigne · K. R. Gouze
Feinberg School of Medicine, Northwestern University, Chicago, IL, USA

J. V. Lavigne
Mary Ann and J. Milburn Smith Child Health Research Program, Children's Hospital of Chicago Research Center, Chicago, IL, USA

F. B. Bryant
Department of Psychology, Loyola University Chicago, Chicago, IL, USA

J. Hopkins
College of Psychology, Illinois Institute of Technology, Chicago, IL, USA

Goodman 2009a, b) or depression (Rowe et al. 2010). In order for research on the homotypic and heterotypic continuity of ODD (longitudinal relationships with both later ODD/CD and internalizing disorders, respectively) to be meaningful, a critical first step is to identify the best dimensional structure of ODD itself.

Models of the Dimensions of Oppositional Defiant Disorder

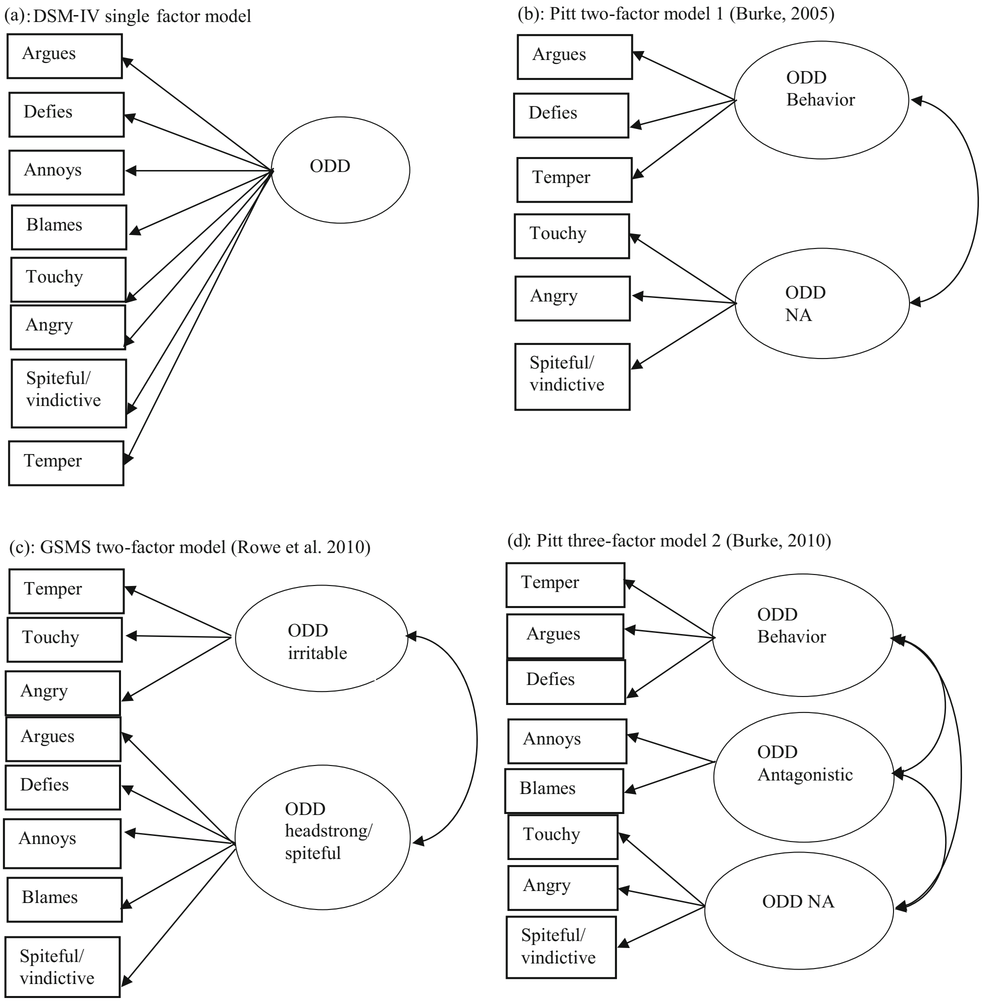
To date, six different models (Figs. 1 and 2) of the dimensions of ODD symptoms have been identified. These include: (a) the single-factor DSM-IV model; (b) a two-factor model (oppositional behavior, negative affect) identified by Burke and colleagues (Burke and Loeber 2010; Burke et al. 2005) with a male sample at the University of Pittsburgh (Pitt-2 model). In the factor analysis used to develop this model, the symptoms of "blames others" and "annoys others" did not load onto either factor and were not included in the model (Burke and Loeber 2010) ; (c) a two-factor model (irritable, headstrong/spiteful) identified by Rowe et al. (Rowe et al. 2010) with the Great Smoky Mountains dataset (GSMS model); (d) a three-factor model (oppositional behavior, negative affect, antagonistic behavior) of ODD dimensions identified by Burke et al. (2010) with a female sample (Pitt-3 model); and (e) a three-factor model (irritable, hurtful, headstrong) developed by Stringaris and Goodman (2009a, b) in the United Kingdom. That model (UK/DSM-5 model) has been adapted for use in DSM-5 with the factor labels changed (in DSM-5 the dimension labels are now angry/irritable, argumentative/ defiant, and vindictiveness, respectfully); and (f) a three-factor model (irritable, headstrong, hurtful) developed by Aebi et al. (2010) with a European/Middle Eastern clinical sample (EUR) model. The three factors identified by Aebi were highly correlated with one another (irritability with headstrong, 0.89; irritability with hurtful, 0.70; headstrong with hurtful, 0.63). The two factors identified by Rowe et al. were correlated 0.55. Correlations between factors were not reported for the other models. Other studies have examined the structure of a broader category of disruptive behavior disorders (Wakschlag et al. 2012) or a specific group of ODD symptoms associated with anger and irritability (Drabick and Gadow 2012) but not ODD per se, and are not considered further herein.

Researchers who identified these six models were primarily concerned with the homotypic or heterotypic continuity of their model with other disorders, and relatively little attention was paid to the validity of the factor structures. Several studies (Aebi et al. 2010; Burke et al. 2005, 2010; Rowe et al. 2010) used exploratory factor analyses (EFAs) to identify the models, but none of these studies replicated the factor structure of their models in subsequent samples using a confirmatory factor analysis (CFA) approach. Stringaris and Goodman (2009a, b) identified the three-factor UK/DSM-5 model on an a priori basis rather than using EFAs or CFAs to support the validity of their model. Although more than one study has identified either two- or three-factor models, the items loading on these factors differed; thus, none of the models identified were replications of one another.

In addition, only two studies have conducted comparisons of model fit across multiple models. Ezpeleta et al. (2012) compared several models (single-factor DSM-IV model, UK/DSM-5 model, Rowe's GSMS model , Pitt-3 model) using CFAs. Two fit indices were used with CFI values>0.90 and an RMSEA value<0.06 considered good and a CFI value>0.85 and an RMSEA value<0.10 considered moderately good. For parent-reported questionnaire data, only the Pitt-3 model showed good model fit on both indices while the DSM-IV, UK/DSM-5, and GSMS model showed good fit on the CFI and moderately good fit on the RMSEA. For teacher-reported data, all four models showed good model fit on the CFA and moderately good fit on the RMSEA indices. The authors concluded there was "no compelling reason" (p. 8) to prefer one model to another. In a sample of parent-reported ODD symptoms among Brazilian children ages 6–12 years-old, Krieger et al. (2013) compared goodness of fit of four models (single-factor DSM-IV, UK/DSM-5 model, Pitt-3 model and the GSMS model). Krieger et al. considered CFI and TLI values of>0.95 as preferred and>0.90 acceptable, and RMSEA values of<0.05 preferred and "up to" 0.08 acceptable. Krieger et al. did not describe the rules used to combine the findings for the different fit indices. However, for three of the four models (DSM-IV, GSMS, and Pitt-3), at least one of the three fit indices was not acceptable. Only for the UK/DSM-5 model were all three fit indices good or acceptable. They concluded that the three-factor UK/DSM-5 model best fit the data. Thus, the findings of the two comparative studies were quite different, with one finding all models reasonably acceptable, and the second finding only one to be acceptable. Furthermore, neither the Ezpeleta et al. nor the Krieger et al. studies included the Pitt-2 or EUR models, and these could have been important omissions.

One major limitation to the existing studies of the dimensions of ODD is the lack of attention to the factorial invariance of the proposed models. Factorial invariance concerns the degree to which the items used to measure a construct have the same meaning and measure the construct in the same way across different groups of respondents (Brown 2006; Saban et al. 2010). When invariance is not present, there is the possibility of construct bias in which the meaning of a construct differs across those groups or longitudinally (Kline 2011). If invariance is not present, it is impossible to determine how to interpret differences in risk factors or correlated features of symptoms, prognosis, or treatment outcomes that

Fig. 1 **a** DSM-IV single-factor model **b** Pitt two-factor model 1 (Burke et al. 2005) **c** GSMS two-factor model (Rowe et al. 2010) **d** Pitt three-factor model 2 (Burke et al. 2010)

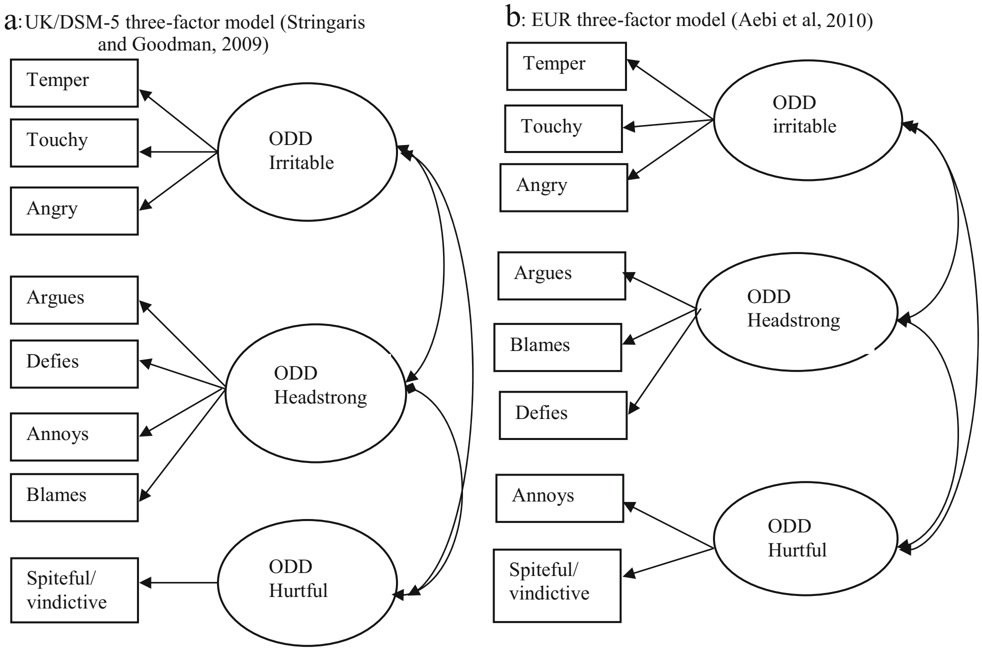

(a): DSM-IV single factor model

(b): Pitt two-factor model 1 (Burke, 2005)

(c): GSMS two-factor model (Rowe et al. 2010)

(d): Pitt three-factor model 2 (Burke, 2010)

Fig. 2 **a** UK/DSM-5 three-factor model (Stringaris and Goodman 2009a, b). **b** EUR three-factor model (Aebi et al. 2010)


a: UK/DSM-5 three-factor model (Stringaris and Goodman, 2009)

b: EUR three-factor model (Aebi et al, 2010) may be associated with defining group differences, such as gender (Burns et al. 2006) or age.

Presently, only one study (Burns et al. 2006) has examined the measurement and structural invariance of ODD symptoms across genders. Using parent reports of ODD in American (ages 3–16 years) and Malaysian (school-age) samples, Burns et al. found support for measurement and scalar invariance across genders in both samples for the single-factor DSM-IV model. In addition, while Burns et al. found measurement and structural invariance across gender for the DSM-IV model, they did not examine age invariance for this model. None of the existing studies have examined gender or age differences among the models positing specific dimensions of ODD symptoms. While Burns et al. found gender invariance for the single dimension DSM-IV model, there are indications that the specific dimensions in two- and three-factor models may not be gender- or age- invariant. Specifically, the studies by Burke and colleagues showing a different number of factors for boys and girls suggest that the structure of ODD is not invariant across gender. Furthermore, Pardini et al. (2010) note that the DSM-IV field trials for ODD included relatively few girls, making it difficult to determine if the ODD construct is the same across genders. Pardini et al. also note that inadequate attention has been paid to longitudinal aspects of the development of ODD, while Burke and Loeber (Burke and Loeber 2010) suggest that longitudinal changes in the prevalence of ODD may be accompanied by corresponding developmental shifts in its factor structure. Because ODD can occur in young children and lead to later internalizing and externalizing disorders even in the early school years, it is particularly important to understand the factor structure of ODD in young children, for whom ODD is the most common psychiatric disorder (Egger and Angold 2006; Lavigne et al. 2009).

Using multi-group CFA methods, it is possible to examine several important aspects of the invariance of a model, including the pattern of factor loadings (configural invariance), the magnitude of factor loadings (metric invariance), and the magnitude of intercepts (scalar invariance). If a model is not invariant, then the heterotypic continuity between ODD dimensions and other behavior problems may differ across ages and genders, so determining the invariance of the best fitting models for the dimensions of ODD has important implications and should be considered before examining heterotypic continuity.

The Present Study

Although DSM-5 adopted a three-factor model of ODD, that model was developed a priori and without adequate attention to the model's gender and longitudinal invariance. Identifying the best model fit and the invariance of the model is important to guide future studies of the predictive ability of the model's dimensions with other disorders. For that reason, the first aim of the present study was to determine which of the six existing measurement models provides a more accurate representation of the ODD construct. These six models are: (a) a single-factor model of oppositional behavior represented in DSM-IV (DSM-IV model); (b) a two-factor model (oppositional behavior, negative affect) of dimensions of ODD developed by Burke and colleagues (Burke et al. 2005) with a male sample at the University of Pittsburgh (Pitt-2 model), (c) a two-factor model developed by Rowe et al. (Rowe et al. 2010) with the Great Smoky Mountains dataset (GSMS model); (d) a three-factor model (oppositional behavior, negative affect, antagonistic behavior) of dimensions of ODD developed by Burke and colleagues (Burke et al. 2010) with a female sample (Pitt-3 model); (e) a three-factor model (irritable, hurtful, headstrong) developed by Stringaris and Goodman (2009a, b) in the United Kingdom (UK/DSM-5 model); and (f) a three-factor model developed by Aebi et al. (2010) with a European/Middle Eastern clinical sample (EUR model). After establishing which model provides the best fit, the second aim was to test: (a) cross-sectional hypotheses about this model's measurement and structural invariance with respect to gender; and (b) longitudinal hypotheses about its measurement and structural invariance with respect to age.

Method

Participants

Data were collected as part of a longitudinal study of risk factors for the development of psychopathology across an important developmental period, ages 4 (preschool), 5 (kindergarten, transition to school), and 6–7 (early school-age). To obtain a diverse sample, 796 4-year-old children and their families were recruited from 23 primary care pediatric clinics throughout Cook County, Illinois and 13 Chicago Public School preschool programs. At the time of the initial interview, eligible children: (a) were 4 years of age; (b) had lived with the parent who participated in the study for at least 6 months; (c) spoke English or Spanish; (d) did not meet criteria for an Autism Spectrum Disorder; (f) obtained a standard score on the Peabody Picture Vocabulary Test≥70 (Dunn and Dunn 1997), were not enrolled in a special education class for the intellectually disabled, and did not have a school IQ test score below 70.

The initial sample of 796 4-year-olds (mean age=4.44) included 391 (49.1 %) boys and 405 (50.9 %) girls. Parent-reported racial/ethnic group membership included: 433 (54.4 %) White, non-Hispanic; 133 (16.7 %) African American; 162 (20.4 %) Hispanic; 19 (2.4 %) Asian; and 35 (4.4 %) multi-racial or "Other." Race/ethnicity was not reported by 14

(1.8 %) parents. All social classes (Hollingshead 1975) were included, with 303 (38.1 %) children in Class I (highest), 290 (36.4 %) in Class II, 79 (9.9 %) in Class III, 63 (7.9 %) in Class IV, and 61 (7.7 %) in Class V. Other details about the age-4 sample are available (Lavigne et al. 2009).

Of the initial sample, 626 children and families (78.6 %) participated in all three waves of data collection. The sample of families and children who completed all three waves of data collection differed from those who did not complete all three waves with respect to: (a) race, with a greater proportion of minority participants dropping out, χ^2(5, N=(796)=77.7, p= 0.001; (b) SES, with a greater proportion of lower SES groups dropping out, χ^2(4, N=(796)=69.61, p=0.001; and (c) age, with those who dropped out being on average 25 days older at study entry, t(773)=2.41, p=0.02. Because imputation is generally preferable to listwise deletion (Graham 2009) missing data were imputed using single imputation with the SPSS V15.0 missing data program. That program uses maximum likelihood procedures utilizing all study variables (child age, sex, race/ethnicity, SES, all ODD symptom items for all 3 age groups) to estimate values. With the imputation, the final sample N was 796.

Measures

Demographics Parents completed a demographic questionnaire to obtain information about child's age, race/ethnicity, and parental education and occupation that was coded for socioeconomic status using the Hollingshead Four-Factor Index of Social Status (Hollingshead 1975).

Peabody Picture Vocabulary Test: 3rdEdition (PPVT-III) Children completed a measure of receptive vocabulary, the PPVT-III (Dunn and Dunn 1997), to assess language skills needed for certain tasks used in the larger study (but relevant to the present report only in terms of exclusion criteria). The PPVT has been shown to have good to excellent concurrent validity (rs=0.63–0.92) with tests of verbal intelligence (Dunn and Dunn 1997).

ODD dimensions Measures of the eight DSM-IV symptoms of ODD were derived from two DSM-IV-coded instruments. The early childhood form of the Child Symptom Inventory (CSI) (Gadow and Sprafkin 1997, 2000) is a parent-completed checklist for which child symptoms are rated on a four-point rating scale ranging from 0 (*never*) to 3 (*very often*). The CSI has been used in prior studies of ODD dimensions (Burke et al. 2010) and the nosology of externalizing problems in girls (Keenan et al. 2010). Internal consistency is good (alpha=0.70).

The Diagnostic Interview Schedule for Children-Parent Scale-Young Child (DISC-YC) version (Fisher and Lucas 2006) is a developmentally-appropriate, structured parent interview that includes items measuring the DSM-IV symptoms of ODD. High levels of agreement are obtained for concrete, observable symptoms, and test-retest reliabilities for the DISC-YC are high. DISC-YC interviewers were clinical psychology graduate students trained to criterion by trained by DISC-YC trainers. Overall reliability of the ODD symptom scale is high, test-retest reliability is 0.88 (C. Lucas, personal communication, September, 2006).

Several alternative approaches were taken to measuring individual ODD symptom. The CSI and DISC-YC each included an item for the eight ODD symptoms (for the CSI, 0= *never*, 1=*sometimes*, 2=*often*, 3=*very often*; for the DISC-YC, 0=*symptom not present*, 1=*symptom present*). Initially, a measurement model was tested in which each CSI and DISC-YC item served as an indicator for the relevant ODD item (e.g., the CSI and DISC-YC "temper tantrum" items were separate indicators for a latent DSM "temper tantrum" item). This approach resulted in an inadmissible solution when the CSI and DISC items were freely estimated, when they were fixed to have equivalent factor loadings, and when errors were allowed to correlate. Subsequently, the one CSI and one DISC-YC item corresponding to each particular ODD DSM-5 symptom was converted to a standard score and the two comparable items (e.g., the CSI and DISC-YC items for temper tantrums) were summed together to create the measure of each ODD symptom (Nunnally and Bernstein 1994). Because this approach resulted in an admissible solution, it was used in subsequent analyses. Items associated with each of the factors in the tested models are illustrated in Figs. 1 and 2.

Procedure

Research assistants approached parents at preschools and pediatric offices and informed them about the study. Subsequently, questionnaires, including the CSI, were mailed to interested parents. At the initial home visit, the PPVT was administered, with children scoring<70 excluded from the study. The use of this exclusion criterion was necessary for completion of other study measures not included in the present study but described elsewhere (Lavigne et al. 2012). The DISC-YC was administered at this visit as well. Graduate research assistants also observed the parent and child interacting in the home environment for approximately 2 h after which they completed a scale noting any symptoms of autistic spectrum disorder that were observed and obtained information on special education programs the child was attending in order to screen for autism and complete study measures not pertaining to this report. Parents were then recontacted approximately 1 year and 2 years after the initial visit for follow-up visits in which the CSI and DISC-YC were re-administered. Written consent to participate was obtained each year. This study was approved by the appropriate Institutional Review Boards.

Data Analysis

Comparing alternative models To assess the appropriateness of each of the six models of the dimensions of ODD, we conducted separate CFAs within each of the three age-groups (ages 4, 5, and 6) using the data for each gender separately, as well as the pooled data of boys and girls, using LISREL 8.8 (Joreskog and Sorbom 2006) to analyze covariance matrices via maximum-likelihood (ML) estimation. To assess goodness-of-fit, we employed: (a) two indices of absolute fit (standardized root mean square residual, SRMR), one of which adjusts for model complexity (root mean square error of approximation, RMSEA); (b) two relative fit indices (non-normed fit index, NNFI; comparative fit index, CFI); and (c) an index of relative information-loss that corrects for sample size and model complexity in comparing measurement models (the Akaike Information Criterion, AIC). There is no universal agreement on verbal descriptors for different fit indices. Marsh et al. (2004), for example suggest that an RMSEA of less that 0.05 is a "close" fit (p. 321), and up to 0.08 is "reasonable" (p. 321) while others apply different descriptors and standards (a fuller discussion of the standards and verbal descriptors of fit indices is available on-line). Because of these differences, Table 2 provides information on the number of fit indices for which each model met the criteria for RMSEA<0.05 and RMSEA≤0.08. Fit based on both of these standards is described as reasonable. For other fit indices, the criteria for a reasonable fit were: NNFI≥0.95; CFI≥0.95 (Browne and Cudeck 1993; Hu and Bentler 1999); and SRMR<0.08 (Brown 2006). Because they were non-nested models, we used AIC to compare goodness-of-fit of competing models, with smaller values representing better fit. As recommended by Brown (2006), we reported the Satorra-Bentler scaled chi-square (SBχ^2) (Satorra and Bentler 1994) but did not use it to assess overall fit because that measure is inflated by large sample sizes (Bollen 1989). To obtain a scaled ML chi-square values, we followed Bryant and Satorra's (2012) guidelines.

Measurement invariance We adopted Vandenberg and Lance's (2000) recommended sequence for conducting tests of measurement invariance. To identify which of the six alternative models showed the best fit across all groups, three types of measurement invariance were examined. Configural invariance is present if the same number of factors and patterns of factor loadings are appropriate for each group (Meredith 1993). Testing configural invariance involves examining the model's goodness-of-fit across groups or time rather than formal statistical null-hypothesis testing. Metric or "weak" invariance (Meredith 1993) exists if a one-unit change in the underlying factor is associated with a comparable change in measurement units for the same given item in each group. Scalar invariance (Meredith 1993) exists if the measurement origins for the items (i.e., item intercepts) are the same across groups in predicting item scores from the latent factors. If, for example, a boy and a girl with the same underlying level of ODD do not obtain the same score on a given observed item, then the item shows "uniform" differential functioning (Teresi 2006) and is biased to produce higher scores for one of the genders even at the same level of the latent trait. Tests of metric and scalar invariance involve assessing the statistical significance of differences in goodness-of-fit chi-square values across nested cross-sectional or longitudinal models. Strong invariance (Meredith 1993) is present if a model shows configural, metric, and scalar invariance. We chose not to assess invariance in item unique error variances and in factor variances-covariances because such analyses: (a) were not essential for meaningful between- and within-group comparisons of levels of ODD (Bontempo and Hofer 2007; Saban et al. 2010); and (b) would have unnecessarily increased the number of statistical tests we conducted, thereby requiring an even stricter Bonferroni adjustment that would predispose the results of null-hypothesis inferential testing toward invariance.

To evaluate the statistical significance of differences in goodness-of-fit between nested CFA models in invariance testing, we used a modified version of the SB χ^2 (Satorra and Bentler 2001) that yields a more accurate scaled difference test for LISREL (Bryant and Satorra 2012). If the SB χ^2 is statistically significant, the parameters in question are not invariant; if nonsignificant, the parameters are invariant. To assess effect size in testing measurement invariance, two indices were used: (a) the difference in CFI values (ΔCFI) between nested models, with ΔCFI≤0.01 considered evidence of measurement invariance (Cheung and Rensvold 2002); (b) the effect size for each probability-based test of invariance expressed in terms of w^2, or the ratio of chi-square divided by N (Cohen 1988), which is analogous to R^2 (i.e., the proportion of explained variance) in multiple regression. A $w^2 \leq 0.01$ is small; $w^2=0.09$, medium; $w^2 \geq 0.25$, large (Cohen 1988).

Because perfectly invariant factors can obscure noninvariant factors and make multivariate global tests of invariance misleading (Bontempo and Hofer 2007), we tested the cross-sectional and longitudinal equivalence of factor loadings and item intercepts separately for each ODD factor. To further reduce the likelihood of capitalizing on chance, we corrected the Type I error rate for probability-based tests of invariance (Cribbie 2007) by imposing a sequentially-rejective Bonferroni adjustment to the generalized p value for each statistical test. Specifically, we used a Sidak step-down adjustment procedure (Holland and Copenhaver 1987; Sidak 1967), to ensure an experiment-wise Type I error rate of $p<0.05$, correcting for the total number of statistical comparisons made (i.e., 20=6 tests of gender metric invariance, 6 tests of gender scalar invariance, 4 tests of age metric invariance, 4 tests of age scalar invariance).

Results

Preliminary Results

Table 1 includes means and standard deviations for each ODD symptom item for each age and gender group. The minimum number of scores for each item was 8, above the 5 category levels that can be used for ordinal data (Newsom, nd).

Model Comparisons

Goodness-of-fit and configural invariance for alternative ODD models across age and gender Figure 1 illustrates conceptual diagrams for the DSM-IV one-factor model (Fig. 1a), Pitt-2 two-factor model (Fig. 1b), the GSMS two-factor model (Fig. 1c), and the Pitt-3 three-factor model (Fig. 1d) of ODD symptoms. Figure 2 illustrates the conceptual diagram for the UK/DMS-5 three-factor model (Fig. 2a) and the EUR three-factor model (Fig. 2b). To retain the three-factor structure of the UK/DSM-5 model while retaining a single spiteful/vindictive item as specified in the DSM, a single manifest indictor was included for the hurtful factor, with the error variance fixed at zero. All other models were exactly as specified by their developers. Table 2 presents goodness-of-fit statistics for each of these measurement models.

Table 1 Means and standard deviations for items for each year and gender: standard scores

	Boys		Girls		Both genders	
	Mean	SD	Mean	SD	Mean	SD
Wave 1						
Temper	0.18	1.68	−0.18	1.70	0	1.70
Argues	0.07	1.72	−0.07	1.65	0	1.68
Defies	0.12	1.59	−0.11	1.59	0	1.59
Touchy	0.12	1.65	−0.12	1.55	0	1.60
Angry	0.09	1.68	−0.08	1.62	0	1.65
Gets even	0.05	1.66	−0.04	1.65	0	1.65
Annoys	0.11	1.79	−0.10	1.75	0	1.77
Blames	−0.05	1.81	0.05	1.78	0	1.79
Wave 2						
Temper	0.15	1.55	−0.15	1.66	0	1.61
Argues	0.15	1.67	−0.14	1.64	0	1.66
Defies	0.22	1.59	−0.21	1.51	0	1.57
Touchy	0.12	1.64	−0.12	1.57	0	1.60
Angry	0.17	1.69	−0.16	1.55	0	1.63
Gets even	0.13	1.66	−0.12	1.54	0	1.60
Annoys	0.20	1.72	−0.20	1.65	0	1.70
Blames	0.08	1.76	−0.08	1.66	0	1.71
Wave 3						
Temper	0.17	1.64	−0.16	1.69	0	1.67
Argues	0.07	1.69	−0.07	1.65	0	1.67
Defies	0.21	1.62	−0.20	1.51	0	1.58
Touchy	0.08	1.70	−0.08	1.49	0	1.60
Angry	0.16	1.66	−0.16	1.52	0	1.59
Gets even	0.11	1.71	−0.11	1.49	0	1.60
Annoys	0.17	1.66	−0.17	1.71	0	1.70
Blames	0.01	1.75	−0.1	1.65	0	1.70

DSM-IV one-factor model For both boys and girls separately at each age, and when both genders were pooled together, the DSM-IV model did not show reasonable fit on any of the fit indices.

Pitt two-factor model (Burke and Loeber 2010; Burke et al. 2005) When the criteria for RMSEA was≤0.08, the nine age x sex groups met the criteria for reasonable fit on all four fit indices for seven groups and for three of four fit indices for the two remaining groups. When the RMSEA criteria was<0.05, the Pitt-2 model met criteria for all four fit indices for one group, and three of four fit indices for the remaining eight groups. None of the Pitt-2 models met criteria for≤2 fit indices.

Pitt three-factor model (Burke et al. 2010) For the Pitt-3 factor model, when the criteria for RMSEA was≤0.08, the Pitt-3 model met criteria on all four fit indices for three groups, and for three of the four fit indices for one other group. For five groups, the Pitt-3 model met criteria on≤2 fit indices. When the criteria for RMSEA was<0.05, the Pitt-3 model did not meet criteria on four fit indices for any of the groups, but did meet criteria on three of four fit indices for one group. For the remaining six groups, the Pitt-3 model met criteria for≤2 of the fit indices.

UK/DSM-5 three-factor model (Stringaris and Goodman 2009a, b) The UK/DSM-5 model did not meet criteria on any of the four fit indices for any of the nine age x sex groups.

GSMS two-factor model (Rowe et al. 2010) When the RMSEA criteria was≤0.08, the GSMS model did not meet criteria on all four fit indices for any of the groups. The GSMS model met criteria for three of the four fit indices for one group, and for≤2 fit indices for the remaining eight groups. The results were the same when the RMSEA criterion was<0.05.

EUR three-factor model (Aebi et al. 2010) When the RMSEA criteria was≤0.08, model fit for the EUR three-factor model met criteria for all four fit indices for one group, but met

Table 2 Goodness-of-fit statistics for alternative CFA models of ODD (combined items)

CFA Models	SB χ^2 (*df*)	RMSEA (90 % CI)	NNFI	CFI	SRMR	Model AIC	Number of the 4 fit indices meeting critieria: RMSEA≤0.08 (RMSEA<0.05)
DSM-IV one-factor model							
Males							
Age 4	100.64(20)***	0.1 (0.09–0.12)	0.94	0.96	0.05	137.74	2 (2)
Age 5	156.56(20)***	0.14 (0.12–0.16)	0.83	0.88	0.08	207.8	0 (0)
Age 6	133.12(20)***	0.12 (0.10–0.14)	0.91	0.93	0.06	171.745	1 (1)
Females							
Age 4	163.63(20)***	0.14 (0.12–0.16)	0.87	0.9	0.07	204.17	1 (1)
Age 5	162.05(20)***	0.14 (0.12–0.16)	0.83	0.88	0.08	205.09	0 (0)
Age 6	123.22(20)***	0.12 (0.10–0.14)	0.9	0.93	0.06	166.52	1 (1)
Both genders pooled							
Age 4	245.57(20)***	0.12 (0.11–0.14)	0.91	0.93	0.06	292.66	1 (1)
Age 5	301.56(20)***	0.14 (0.13–0.15)	0.83	0.88	0.08	368.81	0 (0)
Age 6	234.43(20)***	0.12 (0.11–0.14)	0.9	0.93	0.06	289.93	1 (1)
Pitt two–factor model							
Males							
Age 4	25.71(8)**	0.08 (0.04–0.11)	0.97	0.99	0.04	51.95	4 (3)
Age 5	30.62(8)***	0.08 (0.06–0.12)	0.95	0.97	0.05	55.86	4 (3)
Age 6	35.04(8)***	0.09 (0.06–0.13)	0.95	0.98	0.05	61.06	3 (3)
Females							
Age 4	38.12(8)***	0.09 (0.06–0.12)	0.95	0.97	0.05	59.7	3 (3)
Age 5	32.12(8)***	0.08 (0.05–0.12)	0.95	0.97	0.05	56.95	4 (3)
Age 6	22.55(8)**	0.07 (0.03–0.10)	0.97	0.99	0.04	48.24	4 (4)
Both genders pooled							
Age 4	56.08(8)***	0.08 (0.06–0.11)	0.96	0.98	0.04	78.42	4 (3)
Age 5	56.15(8)***	0.08 (0.06–0.11)	0.95	0.97	0.05	79.96	4 (3)
Age 6	53.74(8)***	0.08 (0.06–0.11)	0.96	0.98	0.04	79.8	4 (3)
Pitt three–factor model							
Males							
Age 4	67.46(17)***	0.09 (0.07–0.11)	0.96	0.97	0.04	109.4	3 (3)
Age 5	70.50(17)***	0.09 (0.07–0.11)	0.92	0.95	0.05	114.16	2 (2)
Age 6	58.77(17)***	0.08 (0.06–0.10)	0.96	0.98	0.05	97.89	4 (3)
Females							
Age 4	100.44(17)***	0.11 (0.09–0.13)	0.94	0.94	0.06	142.92	1 (1)
Age 5	78.98(17)***	0.09 (0.07–0.12)	0.92	0.95	0.06	115.6	2 (2)
Age 6	50.26(17)***	0.07 (0.05–0.09)	0.96	0.98	0.04	89.84	4 (4)
Both genders							
Age 4	151.52(17)***	0.1 (0.09–0.12)	0.94	0.96	0.05	196.93	2 (2)
Age 5	131.60(17)***	0.09 (0.08–0.11)	0.92	0.95	0.05	175.44	2 (2)
Age 6	98.06(17)***	0.08 (0.06–0.10)	0.96	0.97	0.04	140.43	4 (3)
UK/DSM-5 three-factor model							
Males							
Age 4	254.93(20)***	0.19 (0.17–0.21)	0.8	0.86	0.16	340.51	0 (0)
Age 5	289.27(20)***	0.2 (0.18–0.22)	0.65	0.75	0.16	365.8	0 (0)
Age 6	321.93(120)***	0.22 (0.20–0.24)	0.7	0.79	0.17	431.44	0 (0)

Table 2 (continued)

CFA Models	SB χ^2 (*df*)	RMSEA (90 % CI)	NNFI	CFI	SRMR	Model AIC	Number of the 4 fit indices meeting critieria: RMSEA≤0.08 (RMSEA<0.05)
Females							
Age 4	304.33(20)	0.21 (0.19–0.22)	0.7	0.79	0.16	392.95	0 (0)
Age 5	359.98(20)***	0.22 (0.20–0.23)	0.59	0.71	0.17	427.47	0 (0)
Age 6	321.57(20)***	0.21 (0.20–0.23)	0.66	0.76	0.16	425.16	0 (0)
Both genders							
Age 4	550.46(20)***	0.2 .(19–0.21)	0.75	0.82	0.16	696.02	0 (0)
Age 5	630.82(20)***	0.21 (0.20–0.22)	0.63	0.73	0.17	750.41	0 (0)
Age 6	633.24(20)***	0.22 (0.21–0.23)	0.69	0.78	0.16	817.55	0 (0)
GSMS two-factor model							
Males							
Age 4	72.08(19)***	0.09 (0.07–0.11)	0.96	0.97	0.05	109.94	3 (3)
Age 5	138.02(19)***	0.14 (0.12–0.16)	0.84	0.89	0.08	188.94	0 (0)
Age 6	113.05(19)***	0.12 (0.10–0.14)	0.91	0.94	0.06	159.6	1 (1)
Females							
Age 4	130.14(19)***	0.12 (0.10–0.14)	0.89	0.93	0.07	167.29	1 (1)
Age 5	160.89(19)***	0.14 (0.12–0.16)	0.82	0.88	0.08	208.57	0 (0)
Age 6	113.94(19)***	0.12 (0.10–0.14)	0.9	0.93	0.06	162.33	1 (1)
Both genders							
Age 4	187.56(19)***	0.11 (0.10–0.12)	0.93	0.95	0.06	232.14	2 (2)
Age 5	285.33(19)***	0.14 (0.13–0.16)	0.83	0.89	0.07	354.69	1 (1)
Age 6	213.96(19)***	0.12 (0.11–0.14)	0.9	0.93	0.06	275.42	1 (1)
EUR three-factor model							
Males							
Age 4	57.43(17)***	0.08 (0.06–0.10)	0.97	0.98	0.04	97.56	4 (3)
Age 5	126.76(17)***	0.14 (0.12–0.16)	0.84	0.9	0.08	178.59	0 (0)
Age 6	109.97(17)***	0.12 (0.10–0.15)	0.91	0.94	0.06	157.37	1 (1)
Females							
Age 4	114.64(17)***	0.12 (0.10–0.14)	0.9	0.94	0.07	151.47	1 (1)
Age 5	142.17(17)***	0.14 (0.12–0.16)	0.82	0.89	0.07	193.03	1 (1)
Age 6	99.96(17)***	0.11 (0.09–0.13)	0.91	0.94	0.06	142.21	1 (1)
Both genders							
Age 4	157.96(17)***	0.10 (0.09–0.12)	0.93	0.96	0.05	2,001.26	2 (2)
Age 5	255.03(17)***	0.14 (0.13–0.16)	0.83	0.9	0.07	326.03	1 (1)
Age 6	202.66(17)***	0.12 (0.11–0.14)	0.9	0.94	0.06	257.46	1 (1)

SBχ^2 Satorra-Bentler scaled maximum-likelihood chi-square value (Satorra and Bentler 1994), *RMSEA* Root mean square error of approximation, *NNFI* non-normed fit index, *CFI* comparative fit index, *SRMR* Standardized root mean square residual, *GSMS* Great Smoky Mountain Study, *EUR* European, *UK*/*DSM*-5 United Kingdom/DSM-5

criteria on≤2 fit indices for the remaining eight age and gender groups. When the RMSEA criterion was<0.05, the EUR model did not meet criteria for all four fit indices for any age x sex group, met criteria for three of four fit indices for 1 group, and met criteria for≤2 fit indices for the remaining eight groups.

Model comparisons Of the six models, only the Pitt-2 model met criteria for at least three of four fit indices for all 9 age x sex groups when the RMSEA criteria was≤0.08 and at least three of four fit indices for all nine groups when the RMSEA criteria was<0.05. For no model did the Pitt-2 model meet criteria on≤2 fit indices. In comparison, the Pitt-3 model was closest to the Pitt-2 model in the number of fit indices meeting the "reasonable" criteria, but that model met criteria on≤2 fit indices when RMSEA criterion was<0.08 for seven of nine groups and for eight of nine groups when the RMSEA criterion was<0.05.

When AICs are used to compare models, lower values indicate better fit. For each age x sex group, the AIC for the Pitt-2 model was lower than that of all alternative models. Thus, the Pitt-2 model is preferred in comparison to each of the other models.

Correlations between Pitt-2 dimensions If ODDB and ODDNA factors of the Pitt-2 model are very highly correlated, that would suggest the two factors are conceptually redundant, so the strength of the correlation between ODDB and ODDNA at each age is of conceptual interest. Squaring the within-age factor correlations (factor correlations: age 4 boys, 0.82; age 5 boys, 0.56; age 6 boys, 0.67; age 4 girls, 0.67; age 5 girls, 0.59; age 6 girls, 0.69; age 4 combined sexes, 0.75; age 5 combined sexes, 0.59; age 6 combined sexes, 0.64) reveals that the two ODD factors share the following percentages of their variance at each age: for boys: age 4 (67.2 %), age 5 (31.4 %), age 6 (44.9 %); for girls: age 4 (44.9 %), age 5 (34.8 %), age 6 (47.6 %) for both sexes combined: age 4 (56.2 %), age 5 (34.7 %), and 6 (40.1 %). These results indicate that the two ODD dimensions are not so highly related as to be conceptually redundant, supporting the discriminant validity of the factors in the Pitt-2 model (see table in supplemental material, available on line, for the correlations among Pitt-2 factors.

Areas of local ill fit for the Pitt-2 model Goodness-of-fit statistics provide a global index of model fit. While the global fit indices may be acceptable, it is possible that there are specific areas of ill fit or strain within each model (Brown 2006). In this study, there were 9 individual models of the best-fitting Pitt-2 model to consider. In examining the 153 standardized residuals (SRs) in the nine models, we considered residuals greater than the absolute value of 2.58, which corresponds to a p value of 0.01, to be significant because of the large sample size (Brown 2006); Bonferroni corrections were not applied in this supplemental analysis. Across the nine Pitt-2 models, there were a total of 27 areas of ill fit based on the standardized residuals. While there were areas of local ill fit in each one, no combination of factors showed significant residuals in all 9 groups, the most common problems involved the covariance of get even with defies (8 of 9 models), angry/argues (6 of 9 models), and defies/temper (5 of 9 models). A more extensive discussion of specific areas of ill fit is included on-line.

Examining modification indices may provide clues about ways in which the measurement models could be improved. Modification indices (MIs) can provide suggestions about specific estimated parameters that might be added to a model to improve model fit. MIs>3.84 could possibly improve a model at a statistically significant level ($p<0.05$). Modification indices, however, are sensitive to sample size—it is possible that estimating the parameter associated with a significant MI could result in a factor loading that is very small and of little value. For these reasons, examining MIs to gain insight into areas of poor model fit should also include examination of completely standardized expected parameter change (EPC) scores (See on-line supplementary tables for these values).

Overall, there were 6 significant MIs for the 3 models for boys, 11 significant MIs for girls, and 10 significant MIs for the combined sex groups when considering item cross-loadings. These results suggest that adding the factor loadings associated with these MIs would improve model fit overall for all or most of those models. However, low or moderate factor loadings would be eliminated because they were far below the desired standard for factor loadings of 0.70. As a result, only one large factor loading (age 4 boys, ODDNA→ temper) might be retained, but doing so would have the disadvantage of eliminating configural invariance for boys with the Pitt-2 model. For these reasons, adding that cross-loading would not be desirable.

Gender Invariance of the Pitt-2 Model

Testing gender invariance After establishing configural invariance for the Pitt-2 model (i.e., the same pattern of factor loadings for gender x age group), we examined measurement invariance for that model. For all multiple-group CFA models, we defined the variance units of the latent variable by fixing an unstandardized factor loading of one item at 1.0 for each factor. Because using a referent item that functions differently across groups can either mask or exacerbate nonequivalence in other items (Johnson et al. 2009), we selected referent items for which loadings were as comparable as possible across the single-group solutions. These referent items were "temper tantrums" for ODDB and "touchy" for ODDNA.

Metric invariance: gender within age groups Table 3 presents the results of tests of metric invariance for the Pitt-2 ODD model with respect to gender within age groups. The loadings of both ODD factors in the Pitt-2 model were invariant with respect to gender within each age group: (1) *Age 4*: gender invariant loadings for ODDB, SB $\chi^2(2)=1.91$, Bonferroni $p=0.9996$, $\Delta\text{CFI}=0.0006$, $\omega^2=0.002$; gender invariant loadings for ODDNA, SB $\chi^2(2)=1.47$, Bonferroni $p=0.9999$, $\Delta\text{CFI}=0.0003$, $\omega^2=0.002$; (2) *Age 5*: gender invariant loadings for ODDB, SB $\chi^2(2)=4.42$, Bonferroni $p=0.9655$, $\Delta\text{CFI}=0.0018$, $\omega^2=0.006$; gender invariant loadings for ODDNA, SB $\chi^2(2)=2.72$, Bonferroni $p=0.9983$, $\Delta\text{CFI}=0.0007$, $\omega^2=0.003$; and (3) *Age 6*: gender invariant loadings for ODDB, SB

Table 3 Testing metric invariance for the Pitt-2 ODD model with respect to gender within age groups

CFA Model	Comparative Statistics								
	SBχ^2	*df*	Contrast with Model #	SB$\Delta\chi^2$	Δdf	Unadj. *p*<	Bonf. Adj. *p* <	ΔCFI	w^2
Testing gender invariance of factor loadings at age 4:									
1. Baseline model: Pitt-2 CFA model with no equality constraints for boys & girls at age 4	68.862	16	- -	- -	- -	- -	- -	- -	- -
2. Gender invariant loadings for Behavior factor	70.549	18	1	1.908	2	0.386	0.998	0.0006	0.002
3. Gender invariant loadings for Negative Affect factor	70.479	18	1	1.470	2	0.480	0.999	0.0003	0.002
Testing gender invariance of factor loadings at age 5:									
4. Baseline model: Pitt-2 CFA model with no equality constraints for boys & girls at age 5	64.746	16	- -	- -	- -	- -	- -	- -	- -
5. Gender invariant loadings for Behavior factor	68.836	18	4	4.422	2	0.110	0.900	0.0018	0.006
6. Gender invariant loadings for Negative Affect factor	67.632	18	4	2.717	2	0.258	0.993	0.0007	0.003
Testing gender invariance of factor loadings at age 6–7:									
7. Baseline model: Pitt-2 CFA model with no equality constraints for boys & girls at age 6–7	59.789	16	- -	- -	- -	- -	- -	- -	- -
8. Gender invariant loadings for Behavior factor	60.051	18	7	0.309	2	0.857	0.999	0.0003	0.001
9. Gender invariant loadings for Negative Affect factor	62.677	18	7	2.606	2	0.272	0.993	0.0003	0.003

N=796 (males: n=391; females: n=405). *SB*χ^2 Satorra-Bentler scaled maximum-likelihood chi-square value (Satorra and Bentler 1994). $\Delta SB\chi^2$ maximum-likelihood scaled difference test for LISREL (Bryant and Satorra 2012). *Unadj. p* unadjusted generalized per comparison *p*-value. *Bonf. Adj. p* Bonferroni adjusted *p*-value, Δ*CFI* difference in comparative fit indices (Cheung and Rensvold 2002). w^2 χ^2/N, an index of effect size (0.01=small, 0.09=medium, 0.25=large; (Cohen 1988)

$\chi^2(2)=0.319$, Bonferroni $p=0.9999$, ΔCFI=0.0003, $\omega^2=0.001$; gender invariant loadings for ODDNA, SB $\chi^2(2)=2.61$, Bonferroni $p=0.9983$, ΔCFI=0.0003, $\omega^2=0.003$. Therefore, we concluded that ODDB and ODDNA have the same meaning for 4-, 5-, and 6-year-old boys and girls.

Scalar invariance Table 4 presents the results of tests of scalar invariance for the Pitt-2 ODD model with respect to gender within age groups. The item intercepts of both ODD factors in the Pitt-2 model (behavior and negative affect) were invariant with respect to gender within each of the three age groups, as follows: (a) *Age 4*: gender invariant intercepts for ODDB, SB $\chi^2(2)=2.765$, Bonferroni $p=0.99$, ΔCFI=0.0005, $\omega^2=0.004$; gender invariant intercepts for ODDNA, SB $\chi^2(2)=0.87$, Bonferroni $p=0.9999$, ΔCFI=0.0003, $\omega^2=0.001$; (b) *Age 5*: gender invariant intercepts for ODDB, SB $\chi^2(2)=3.94$, Bonferroni $p=0.9983$, ΔCFI=0.0011, $\omega^2=0.005$; gender invariant intercepts for ODDNA factor, SB $\chi^2(2)=0.35$, Bonferroni $p=0.9999$, ΔCFI=0.0021, $\omega^2=0.0004$; and (c) *Age 6*: gender invariant intercepts for ODDB, SB $\chi^2(2)=7.74$, Bonferroni $p=0.5315$, ΔCFI=0.0076, $\omega^2=0.0097$; gender invariant intercepts for ODDNA, SB $\chi^2(2)=1.78$, Bonferroni $p=0.9996$, ΔCFI=0.0001, $\omega^2=0.002$. Thus, we concluded that the behavior and negative affect items function equivalently in assessing ODD for 4-, 5-, and 6-year-old boys and girls. Considered together, these findings indicate that the Pitt-2 model shows strong gender invariance (Meredith 1993) within all three age groups.

Age Invariance of the Pitt-2 Model

Metric invariance: age within gender Having established configural, metric, and scalar invariance for the Pitt-2 model across gender at ages 4, 5, and 6, we next examined the measurement invariance of the Pitt-2 model with respect to age longitudinally within each gender. To test age invariance in ODD for boys and girls, we estimated longitudinal CFA models in which we specified the two Pitt-2 factors at ages 4, 5, and 6 as six correlated latent variables separately for each gender. We defined the variance units of the latent variables at each wave by fixing at 1.0 the factor loadings of the referent items for ODDB and ODDNA. Following common practice in longitudinal measurement modeling (Brown 2006), all three-wave CFA models included autocorrelated measurement errors, reflecting temporally stable indicator-specific variance, i.e., method effects (Brown 2006), through which the unique variance in each of the six ODD items at each wave was allowed to correlate with the unique variance in the same item at the other two waves. We also allowed all ODD factors to correlate with one another both within and across waves. This six-factor model provided an acceptable fit to the longitudinal ODD data of both boys, SB ML $\chi^2(102, N=391)=175.41$, RMSEA=0.041, CFI=0.99, NNFI=0.99, AIC=305.685, and girls, SB ML $\chi^2(102, N=405)=149.13$, RMSEA=0.031, CFI=0.99, NNFI=0.99, AIC=280.689.

Table 4 Testing scalar invariance for the Pitt-2 ODD model with respect to gender within age groups

CFA Model	Comparative Statistics								
	SBχ^2	*df*	Contrast with Model #	SB$\Delta\chi^2$	Δdf	Unadj. *p*<	Bonf. Adj. *p* <	ΔCFI	w^2
Testing gender invariance of item intercepts at age 4:									
10. Baseline model: Pitt-2 CFA model with gender invariant loadings for both ODD factors at age 4	72.130	20	- -	- -	- -	- -	- -	- -	- -
11. Gender invariant intercepts for Behavior factor	75.011	22	13	2.765	2	0.251	0.993	0.0005	0.004
12. Gender invariant intercepts for Negative Affect factor	73.148	22	13	0.858	2	0.652	0.999	0.0003	0.001
Testing gender invariance of item intercepts at age 5:									
13. Baseline model: Pitt-2 CFA model with gender invariant loadings for both ODD factors at age 5	72.016	20	- -	- -	- -	- -	- -	- -	- -
14. Gender invariant loadings for Behavior factor	75.872	22	17	3.938	2	0.285	0.993	0.0011	0.005
15. Gender invariant loadings for Negative Affect factor	72.369	22	17	0.353	2	0.839	0.999	0.0021	0.0004
Testing gender invariance of item intercepts at ages 6–7:									
16. Baseline model: Pitt-2 CFA model with gender invariant loadings for both ODD factors at ages 6–7	62.946	20	- -	- -	- -	- -	- -	- -	- -
17. Gender invariant loadings for Behavior factor	70.377	22	21	7.742	2	0.021	0.546	0.0076	0.0097
18. Gender invariant loadings for Negative Affect factor	64.714	22	21	1.785	2	0.410	0.998	0.0001	0.002

N=796 (males: *n*=391; females: *n*=405). *SB*χ^2 Satorra-Bentler scaled maximum-likelihood chi-square value (Satorra and Bentler 1994). $\Delta SB\chi^2$ maximum-likelihood scaled difference test for LISREL (Bryant and Satorra 2012). *Unadj. p* unadjusted generalized per comparison *p*-value, *Bonf. Adj. p* Bonferroni adjusted *p*-value, Δ*CFI* difference in comparative fit indices (Cheung and Rensvold 2002). $w^2=\chi^2/N$, an index of effect size (0.01=small, 0.09=medium, 0.25=large; (Cohen 1988)

Table 5 presents the results of tests of metric invariance for the Pitt-2 ODD model with respect to age within both genders. The loadings of the two ODD factors in the Pitt-2 model were invariant with respect to age for both boys and girls: (a) *Boys*: age invariant loadings for ODDB, SB $\chi^2(4)=2.90$, Bonferroni $p=0.9999$, ΔCFI=0.0001, $\omega^2=0.007$; age invariant loadings for ODDNA, SB $\chi^2(4)=1.73$, Bonferroni $p=0.9999$, ΔCFI=0.0001, $\omega^2=0.004$; (b) *Girls*: age invariant loadings for ODDB, SB $\chi^2(4)=3.04$, Bonferroni $p=0.9999$, ΔCFI=0.0001, $\omega^2=0.007$; age invariant loadings for ODDNA, SB $\chi^2(4)=7.54$, Bonferroni $p=0.9655$, ΔCFI=0.0005, $\omega^2=0.019$. Thus, we concluded that oppositional behavior and negative affect have the same meaning across ages 4, 5, and 6 for both boys and girls.

Table 5 Testing metric invariance for the Pitt-2 ODD model with respect to age within boys and girls

CFA Model	Comparative Statistics								
	SBχ^2	*df*	Contrast with Model #	SB$\Delta\chi^2$	Δdf	Unadj. *p*<	Bonf. Adj. *p* <	ΔCFI	w^2
Testing age invariance of factor loadings for boys:									
19. Baseline model: Longitudinal Pitt-2 CFA model with no equality constraints across ages 4, 5, and 6–7	181.301	102	- -	- -	- -	- -	- -	- -	- -
20. Age invariant loadings for Behavior factor	183.857	106	19	2.896	4	0.576	0.999	0.0001	0.007
21. Age invariant loadings for Negative Affect factor	182.817	106	19	1.733	4	0.785	0.999	0.0001	0.004
Testing age invariance of factor loadings for girls:									
22. Baseline model: Longitudinal Pitt-2 CFA model with no equality constraints across ages 4, 5, and 6–7	156.304	102	- -	- -	- -	- -	- -	- -	- -
23. Age invariant loadings for Behavior factor	158.762	106	22	3.036	4	0.552	0.999	0.0001	0.007
24. Age invariant loadings for Negative Affect factor	165.211	106	22	7.545	4	0.110	0.891	0.0005	0.019
Testing age invariance of factor loadings for boys:									

N=796 (males: *n*=391; females: *n*=405). *SB*χ^2 Satorra-Bentler scaled maximum-likelihood chi-square value (Satorra and Bentler 1994). $\Delta SB\chi^2$ maximum-likelihood scaled difference test for LISREL (Bryant and Satorra 2012). *Unadj. p* unadjusted generalized per comparison *p*-value, *Bonf. Adj. p* Bonferroni adjusted *p*-value, Δ*CFI* difference in comparative fit indices (Cheung and Rensvold 2002). $w^2=\chi^2/N$, an index of effect size (0.01=small, 0.09=medium, 0.25=large; (Cohen 1988)

Scalar invariance Table 6 presents the results of tests of scalar invariance for the Pitt-2 ODD model with respect to age within both genders. The item intercepts of both ODD factors in the Pitt-2 model were invariant with respect to age for both boys and girls:(a) *Boys*: age invariant intercepts for ODDB, SB $\chi^2(4)=2.81$, Bonferroni $p=0.9999$, ΔCFI=0.0002, $\omega^2=0.007$; age invariant intercepts for ODDNA, SB $\chi^2(4)=1.92$, Bonferroni $p=0.9999$, ΔCFI=0.0003, $\omega^2=0.005$; and (b) *girls*: age invariant intercepts for ODDB, SB $\chi^2(4)=3.27$, Bonferroni $p=0.9983$, ΔCFI=0.0001, $\omega^2=0.008$; age invariance intercepts for ODDNA, SB $\chi^2(12)=1.56$, Bonferroni $p=0.9983$, ΔCFI=0.0003, $\omega^2=0.004$. Thus, we concluded that the behavior and negative affect items function equivalently in assessing ODD for both boys and girls at ages 4, 5, and 6 years. Considered together, these findings indicate that the Pitt-2 model shows strong age invariance (Meredith 1993),i.e., configural, metric, and scalar invariance across ages 4, 5, and 6, for both boys and girls.

Table 7 presents the gender- and age-invariant CFA factor loadings, squared multiple correlations, and Cronbach's alphas for the Pitt-2 CFA model. Factor loadings were gender and age invariant, and squared multiple correlations were highly comparable across gender for both ODDB factor (boys: median=0.53; girls: median=0.51) and ODDNA (boys: median=0.54; girls: median=0.56). The ODDB subscale had reasonable internal consistency reliability at each age for both boys (age 4: $\alpha=0.76$; age 5: $\alpha=0.74$; age 6: $\alpha=0.77$) and girls (age 4: $\alpha=0.73$; age 5: $\alpha=0.76$; age 6: $\alpha=0.78$). The ODDNA subscale had reasonable internal consistency reliabilities at age 4 (boys: $\alpha=0.77$; girls: $\alpha=0.73$) and age 6 (boys: $\alpha=0.74$; girls: $\alpha=0.71$), but Cronbach's alphas were lower for this subscale at age 5 for both boys ($\alpha=0.68$) and girls ($\alpha=0.66$). While a cutoff of 0.70 is often used for assessing the adequacy of alpha, lower scores may be acceptable when the measure has other desirable measurement properties (Schmitt 1996), as in the present model.

Discussion

Results of analyses comparing model fit of the 6 different models proposed to date showed that the two-factor ODD model (Pitt-2) identified by Burke et al. (2005) best fit the data, for both genders separately and when genders were combined, for all three age groups (4, 5, and 6). In addition, the results indicated: (a) there is configural invariance (Brown 2006) for both boys and girls across ages for the Pitt-2 model because the two-factor structure showed the best fit to the data for each age x gender group; (b) there is metric invariance with respect to age and gender, i.e., the factor loading of each measured indicator on its underlying ODD dimension was equivalent across age and gender groups; and (c) there is scalar invariance, with the ODD items producing equivalent scores for children with the same underlying level of ODD, regardless of gender or age. Thus, studies of homotypic and heterotypic continuity of the Pitt-2 ODD factors with other disorders in this age range can be conducted with clear evidence that the dimensions of ODD do not show developmental differences in structural form, factor loadings, or value of scale items. The results support Burke et al.'s conclusion that ODD is best characterized as being composed of separate

Table 6 Testing scalar invariance for the Pitt-2 ODD model with respect to age within boys and girls

CFA Model	Comparative Statistics								
	SBχ^2	*df*	Contrast with Model #	SBΔχ^2	Δ*df*	Unadj. *p*<	Bonf. Adj. *p* <	ΔCFI	*w*^{2}
Testing age invariance of item intercepts for boys:									
25. Baseline model: Longitudinal Pitt-2 CFA model with age invariant loadings for both ODD factors	185.378	110	- -	- -	- -	- -	- -	- -	- -
26. Age invariant intercepts for Behavior factor	188.114	114	25	2.810	4	0.590	0.999	0.0002	0.007
27. Age invariant intercepts for Negative Affect factor	187.261	114	25	1.917	4	0.752	0.999	0.0003	0.005
Testing age invariance of item intercepts for girls:									
28. Baseline model: Longitudinal Pitt-2 CFA model with age invariant loadings for both ODD factors	167.663	110	- -	- -	- -	- -	- -	- -	- -
29. Age invariant intercepts for Behavior factor	170.847	114	29	3.272	4	0.314	0.993	0.0001	0.008
30. Age invariant intercepts for Negative Affect factor	169.207	114	29	1.564	4	0.816	0.999	0.0003	0.004

N=796 (males: *n*=391; females: *n*=405). *SB*χ^2 Satorra-Bentler scaled maximum-likelihood chi-square value (Satorra and Bentler 1994). Δ*SB*χ^2 maximum-likelihood scaled difference test for LISREL (Bryant and Satorra 2012). *Unadj. p* unadjusted generalized per comparison *p*-value, *Bonf. Adj. p* Bonferroni adjusted *p*-value, Δ*CFI* difference in comparative fit indices (Cheung and Rensvold 2002). $w^2=\chi^2/N$, an index of effect size (0.01=small, 0.09=medium, 0.25=large; (Cohen 1988)

Table 7 Gender and age invariant CFA factor loadings, squared multiple correlations, and Cronbach's alphas for the Pitt-2 ODD model for boys (N=391) and girls (N=405) at ages 4, 5, and 6–7

	ODD Behavior							ODD Negative Affect						
	R^2							R^2						
CSI Item	λ	Boys			Girls			λ	Boys			Girls		
		Age 4	Age 5	Age 6–7	Age 4	Age 5	Age 6–7		Age 4	Age 5	Age 6–7	Age 4	Age 5	Age 6–7
Temper	1	0.54	0.53	0.58	0.50	0.54	0.56	- -	- -	- -	- -	- -	- -	- -
Argues	1.00	0.51	0.53	0.55	0.49	0.55	0.56	- -	- -	- -	- -	- -	- -	- -
Defies	0.88	0.49	0.43	0.46	0.44	0.47	0.51	- -	- -	- -	- -	- -	- -	- -
Touchy	- -	- -	- -	- -	- -	- -	- -	1	0.59	0.49	0.54	0.57	0.43	0.47
Angry	- -	- -	- -	- -	- -	- -	- -	1.11	0.67	0.54	0.68	0.67	0.59	0.56
Spiteful	- -	- -	- -	- -	- -	- -	- -	0.78	0.36	0.28	0.32	0.30	0.56	0.32
α	- -	0.76	0.74	0.77	0.73	0.76	0.78	- -	0.77	0.68	0.74	0.73	0.66	0.71

λ=unstandardized factor loading, or regression coefficient in predicting each measured variable from its latent ODD factor. Tx item intercept, R^2 squared multiple correlation, or the proportion of variance in each CSI item that the given ODD factor explains. α=Cronbach's alpha. Cronbach's alphas are based on unit-weighted subscales formed from summing the standardized component items for each ODD factor. Tabled results are from a multigroup confirmatory factor analysis (CFA) using robust maximum-likelihood estimation via LISREL 8.8 (Joreskog and Sorbom 2006), Satorra-Bentler scaled χ^2 (254, N=796)=392.27, RMSEA=0.033, SRMR=0.048, CFI=0.99, NNFI=0.99. The factor loading for each CSI item was constrained to be equal for boys and girls within age and equal across age within each gender. Loadings of 1 were fixed at unity to identify the CFA model. Blank loadings were fixed at 0.0 in the CFA model. Each item intercept was also constrained to be gender and age invariant in this CFA model. Gender -and age-invariant item intercepts (Txs) and standard errors (SEs) were as follows: Temper (Tx=−0.0012, SE=0.0473); Argues (Tx=−0.0051, SE=0.0472), Defies (Tx=−0.0095, SE= 0.0437), Touchy (Tx=−0.0133, SE=0.0423), Angry (Tx=−0.0154, SE=0.0433), Spiteful (Tx=−0.0136, SE=0.0442). None of the item intercepts were significantly different from zero (all Zs<0.36, ps> 0.72) processes of behavioral and affective dysregulation, rather than being a single distinct disorder.

These results have important implications for the structure of ODD in the recently-released DSM-5 (American Psychiatric Association 2013). First, because neither DSM-IV nor DSM-5 proposed gender or developmental differences in the structure of the symptoms of ODD, there is an implicit assumption that the ODD dimensions are invariant for gender and age. With the exception of the Pitt-2 model, age and gender invariance were not present for any of the other models, including the UK/DSM-5 model that forms the basis of the DSM-5 dimensions. These results support the use of the Pitt two-factor model as a tool for understanding, as well as for diagnosing ODD in children ages 4–7 rather than the three-factor model adopted in the DSM-5.

The results of the present study differ somewhat from those of prior comparisons, chiefly because neither the Krieger et al. (2013) nor the Ezpeleta et al. (2012) studies include the Pitt-2 model which, in the present study, showed the best fit overall in each of the separate and combined gender groups at all three ages. Furthermore, while multiple studies identified three-factor structures such as the one adopted by DSM-5, none of these studies were replications of one another because the factor loadings of items differed across models.

One characteristic of the Pitt-2 model is that the ODD symptoms of “annoys” and “blames others” did not load on either factor in the EFA conducted by Burke and Loeber (2010). The factor loadings of these two items differ across the other multidimensional models of ODD. Both items load on the same factor in the GSMS model (headstrong/spiteful), the Pitt three-factor model (antagonistic), and the UK/DSM-5 model (headstrong), while they load on different factors in the EUR model. Given these inconsistencies, these items could be described as “other” ODD symptoms and possibly eliminated as critical to diagnosis in the future. The decision to either retain or eliminate those items would depend on future research on the structural invariance of ODD with older children and the ability of the different ODD dimensions to predict heterotypic comorbidity of those dimensions with other disorders. In a separate report (2014), the Pitt-2 factors without the items “annoys” and “blames others” were found to be associated with subsequent depression.

Further research will also be needed to address possible limitations to the Pitt-2 model. While better than the alternatives, there is room to improve overall global fit as well as specific areas of local strain in the young child age group, and it remains to be seen whether the model fit and specific areas of strain are problems in older children. If the Pitt-2 model continues to show the best, but imperfect fit, across age and gender groups, further improvement in measuring the ODD and the Pitt-w model may require the development of measures that retain the same core indicators of ODD, but include multiple measures of each item to improve model fit, or allow for more fine-grained responses than the four-point scales often used on behavior problem checklists. Such changes, however, increase the number of parameters in the model and may also require cross-loadings between factors.

The few existing studies of ODD dimensions have been conducted with a diverse set of participant samples. Studies have been done with both clinical and community samples. Because high levels of comorbidity that could affect the internal structure of ODD symptoms are likely to be present in clinical samples (Caron and Rutter 1991) examining the structure of ODD in community samples is particularly important.

Outside of the U.S., community samples have been utilized in the United Kingdom and in Barcelona, Spain. The UK sample was highly representative of the population, but as noted earlier, no CFA of the three-factor model was conducted with that sample. In the Barcelona sample, 89.5 % of participants were white, and 78.7 % were high or middle SES. No information was provided on how representative this sample was of Barcelona or Spain. Presently, there are four studies that have examined dimensions of ODD in the U.S. One of these studies (Burke et al. 2005) was conducted with a referred sample of boys only. Another was of a community sample of low income girls from Pittsburgh that was 45 % African American and 50 % Caucasian. Clearly, these samples were not representative of their geographical areas even based on gender or race/ethnicity, and they included few Hispanics. The GSMS sample (Rowe et al. 2010) included 25 % Native Americans in the initial wave. The sample in the Rowe et al. 2010 report included 8 % African American, and<1 % Hispanic. Neither the number of Native Americans nor SES information for the final sample was reported. Thus, none of the existing studies are truly representative of the U.S. population. Lacking national registries, it is likely that a series of studies of different community samples in the U.S. will be needed to understand the most representative version of the structure of ODD. It is also important to compare models within a variety of different samples as was done in the present study, to make direct comparisons among the competing models.

The current study has several limitations. First, findings are limited by the use of parent report of symptoms. Although parent report is the most common way in which symptom reports are obtained in young children, studies comparing reports of symptoms of ODD for teacher and parents suggest that they are source-specific (Drabick et al. 2011; Lavigne et al. 2014). Thus, it will be important to determine whether the two-factor structure of oppositional behavior and negative affect are invariant when measures from other sources (e.g. teachers, observers) are used. In addition, the findings of this study are clearly limited to the developmental period between preschool and formal school entry, and may differ for older children and adolescents. It is possible, as well, that these relationships are different in a clinical rather than in a community sample.

Nevertheless, this study has important research and clinical implications for understanding and treating ODD in children. By clearly establishing the best model for understanding ODD in young children it provides a framework for moving forward with research on the relationships between early occurring ODD, the most common early childhood disorder, and later externalizing and internalizing disorders in children. Clinically, this provides significant information about how to treat early childhood ODD. Presently, for example, parent management training is the most effective treatment for ODD in preschoolers (Webster-Stratton et al. 2004), but we also know that approximately 30 % of children do not benefit from this treatment. Possibly, the different dimensions of ODD may be important moderators of the effectiveness of parent management training for ODD. In addition, this study has implications for the ODD diagnostic criteria adopted for use in the DSM-5. The clinical results of this study suggest that the structure of ODD adopted for use in DSM-5 does not show invariance over gender and age in preschool and early school-age children, while an alternative two-factor model does. Since it is largely the DSM-5 which will drive future conceptualizations of ODD for both research and clinical purposes, recognizing that these dimensions might not best represent ODD symptoms in children is critical to future work on ODD. This disorder is highly prevalent in young children and has implications for the development of psychopathology over time. Understanding the dimensional aspects of ODD especially in the context of homotypic and heterotypic continuity over time is critical to developing the best possible early interventions for this disorder.

Conflicts of Interest The authors have no conflicts of interest to report.

References

Aebi, M., Muller, U. C., Asherson, P., Banachewski, T., Buitelaar, J., Ebstein, R. P., & Steinhausen, H. C. (2010). Predictability of oppositional defiant disorder and symptom dimensions in children and adolescents with ADHD combined type. *Psychological Medicine, 40*, 2089–2100. doi:10.1017/S0033291710000590.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington: American Psychiatric Association.

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

Bontempo, D.E., & Hofer, S.M. (2007). Assessing factorial invariance in cross-sectional and longtitudinal studies. In A. D. Ong & M. H. M. van Dulmen (Eds.), *Oxford handbook of methods in positive psychology* (pp. 153–175). New York: Oxford.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford.

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

Bryant, F. B., & Satorra, A. (2012). Principles and practice of scaled difference chi-square testing. *Structural Equation Modeling, 19*, 372–398. doi:10.1080/10705511.2012.687671.

Burke, J. D., & Loeber, R. (2010). Oppositional defiant disorder and the explanation of the comorbidity between behavioral disorders and depression. *Clinical Psychology: Science and Practice, 17*, 319–326.

Burke, J. D., Loeber, R., Lahey, B. B., & Rathouz, P. J. (2005). Developmental transitions among affective and behavioral disorders in adolescent boys. *Journal of Child Psychology and Psychiatry, 46*, 1200–1210.

Burke, J. D., Hipwell, A. E., & Loeber, R. (2010). Dimensions of oppositional defiant disorder as predictors of depression and conduct disorder in preadolescent girls. *Journal of the American Academy of Child and Adolescent Psychiatry, 49*, 484–492.

Burns, G. L., Walsh, J. A., Gomez, R., & Hafetz, N. (2006). Measurement and structural invariance of parent reatings of ADHD and ODD symptoms across gender for American and Malaysian children. *Psychological Assessment, 18*, 452–457. doi:10.1037/1040-3590.18.4.452.

Caron, C., & Rutter, M. (1991). Comorbidity in child psychopathology: concepts, issues and research strategies. *Journal of Child Psychology and Psychiatry, 32*, 1063–1080.

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: Erlbaum.

Copeland, W. E., Shanahan, L., Costello, E. J., & Angold, A. (2009). Childhood and adolescent psychiatric disorders as predictors of young adult disorders. *Archives of General Psychiatry, 66*, 764–772.

Cribbie, R. A. (2007). Multiplicity control in structural equation modeling. *Structural Equation Modeling, 14*, 98–112.

Drabick, D. A. G., & Gadow, K. D. (2012). Deconstructing oppositional defiant disorder: clinic-based evidence for an anger/irritability phenotype. *Journal of the American Academy of Child and Adolescent Psychiatry, 51*, 384–393.

Drabick, D. A. G., & Kendall, P. C. (2010). Developmental psychopathology and the diagnosis of mental health problems among youth. *Clinical Psychology: Science and Practice, 17*, 272–280.

Drabick, D. A. G., Bubier, J. L., Chen, D., Price, J., & Lanza, H. (2011). Source-specific oppositional defiant disorder among inner-city children: prospective prediction and moderation. *Journal of Clinical Child & Adolescent Psychology, 40*, 23–35.

Dunn, L., & Dunn, L. (1997). *The Peabody Picture Vocabulary Test* (3rd ed.). Circle Pines: American Guidance Service.

Egger, H. E., & Angold, A. (2006). Common emotional and behavioral disorders in preschool children: presentation, nosology, and epidemiology. *Journal of Child Psychology and Psychiatry, 47*, 313–337.

Ezpeleta, L., Granero, R., de la Osa, N., Penelo, E., & Domenech, J. M. (2012). Dimensions of oppositional defiant disorder in 3-year-old preschoolers. *Journal of Child Psychology and Psychiatry, 53*, 1128–1138. doi:10.1111/j.1469-7610.2012.02545.x.

Fisher, P., & Lucas, C. (2006). *Diagnostic Interview Schedule For Children (DISC-IV)-Young Child*. New York: Columbia University.

Gadow, K. D., & Sprafkin, J. (1997). *Early Childhood Inventory 4 Norms Manual*. Stonybrook: Checkmate Plus.

Gadow, K. D., & Sprafkin, J. (2000). *Early Childhood Inventory 4 Screening Manual*. Stonybrook: Checkmate Plus.

Graham, J. W. (2009). Missing data analysis: making it work in the real world. *Annual Review of Psychology, 60*, 549–576.

Holland, B. S., & Copenhaver, M. D. (1987). An improved sequentially rejective Bonferroni test procedure. *Biometrics, 43*, 417–423.

Hollingshead, A. B. (1975). *Four-factor Index of Social Position*. New Haven: Yale University Department of Sociology.

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.

Johnson, E. C., Meade, A. W., & DuVernet, A. M. (2009). The role of referent indicators in tests of measurement invariance. *Structural Equation Modeling, 16*, 642–657.

Joreskog, K. G., & Sorbom, D. (2006). *LISREL for Windows*. Chicago: Scientific Software International.

Keenan, K., Wroblewski, K., Hipwell, A. E., Loeber, R., & Stouthamer-Loeber, J. (2010). Age of onset, symptom threshold, and expansion of the nosology of conduct disorder for girls. *Journal of Abnormal Psychology, 119*, 689–699.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York: Guilford.

Krieger, F. V., Polanczyk, G. V., Goodman, R., Rohde, L. A., Graeff-Martins, A. S., Salum, G., & Stringaris, A. (2013). Dimensions of oppositionality in a Brazilian community sample: testing the DSM-5 proposal and etiological links. *Journal of the American Academy of Child and Adolescent Psychiatry, 52*, 389–400.

Lavigne, J. V., Cicchetti, C., Gibbons, R. D., Binns, H. J., Larsen, L., & DeVito, C. (2001). Oppositional defiant disorder with onset in preschool years: longitudinal stability and pathways to other disorders. *Journal of the American Academy of Child and Adolescent Psychiatry, 40*, 1393–1400.

Lavigne, J. V., LeBailly, S. A., Hopkins, J., Gouze, K. R., & Binns, H. J. (2009). The prevalence of ADHD, ODD, depression and anxiety in a community sample of 4-year-olds. *Journal of Clinical Child and Adolescent Psychology, 38*, 315–328.

Lavigne, J. V., Gouze, K. R., Hopkins, J., Bryant, F. B., & LeBailly, S. A. (2012). A multi-domain model of risk factors for ODD symptoms in a community sample of 4-year-olds. *Journal of Abnormal Child Psychology, 40*(5), 741–757. doi:10.1007/s10802-011-9603-6.

Lavigne, J.V., Dahl, K.P., Gouze, K.R., LeBailly, S.A., & Hopkins, J. (2014). Multi-domain predictors of oppositional defiant disorder symptoms in preschool children: cross-informant differences. *Child Psychiatry and Human Development*.

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.

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

Newsom, J. (nd). Practical approaches to dealing with nonnormal and categorical variables. www.upa.pdx.edu/IOW/newsom/semclass/ho_estimate2.pdf.

Nunnally, J., & Bernstein, I. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Pardini, D. A., Frick, P. J., & Moffitt, T. E. (2010). Building an evidence base for DSM-5 conceptualizations of oppositional defiant disorder and conduct disorder: Introduction to the special section. *Journal of Abnormal Psychology, 119*, 683–688.

Rowe, R., Costello, E. J., Angold, A., & Copeland, W. E. (2010). Developmental pathways in oppositional defiant disorder and conduct disorder. *Journal of Abnormal Psychology, 119*, 726–738. doi: 10.1037/a0020798.

Saban, K.L., Bryant, F.B., Reda, D.J., Stroupe, K.T., & Hynes, D.M. (2010). Measurement invariance of the kidney disease and quality of life instrument (KDQOL-SF) across veterans and non-veterans. *Health and Quality of Life Outcomes,* www.hqlo.com/content/8/1/120, 120.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399–419). Thousand Oaks: Sage.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66*, 507–514.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*, 350–353.

Sidak, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association, 62*, 626–633.

Stringaris, A., & Goodman, R. (2009a). Longitudinal outcomes of youth oppositionality: irritable, headstrong, and hurtful behaviors have distinctive predictions. *Journal of the American Academy of Child and Adolescent Psychiatry, 48*, 404–412.

Stringaris, A., & Goodman, R. (2009b). Three dimensions of oppositionality in youth. *Journal of Child Psychology and Psychiatry, 50*, 216–233.

Teresi, J. (2006). Overview of quantitative methods: equivalence, invariance and differential item functioning in health applications. *Medical Care, Suppl. 3*(44), S39–S49.

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.

Wakschlag, L. S., Henry, D. B., Tolan, P. H., Carter, A. S., Burns, J. L., & Briggs-Gowan, M. J. (2012). Putting theory to the test: modeling a multidimensional, developmentally-based approach to preschool disruptive behavior. *Journal of the American Academy of Child and Adolescent Psychiatry, 51*, 593–604.

Webster-Stratton, C. S., Reid, M. J., & Hammond, M. (2004). Treating children with early-onset conduct problems: intervention outcomes for parent, child, and teacher training. *Journal of Clinical Child & Adolescent Psychology, 33*, 105–124.