Sex differences and gender-invariance of mother-reported childhood problem behavior

van der Sluis, Sophie; Polderman, Tinca J C; Neale, Michael C; Verhulst, Frank C; Posthuma, Danielle; Dieleman, Gwen C

published in
International Journal of Methods in Psychiatric Research
2017

DOI (link to publisher)
10.1002/mpr.1498

document version
Publisher's PDF, also known as Version of record

International Journal of Methods in Psychiatric Research
Int. J. Methods Psychiatr. Res. 2017;**26**:e1498.
Published online 22 January 2016 in Wiley Online Library
(wileyonlinelibrary.com) **DOI:** 10.1002/mpr.1498

Sex differences and gender-invariance of mother-reported childhood problem behavior

SOPHIE VAN DER SLUIS,[1,2] TINCA J.C. POLDERMAN,[2] MICHAEL C. NEALE,[3] FRANK C. VERHULST,[1] DANIELLE POSTHUMA[1,2] & GWEN C. DIELEMAN[1]

1 Department of Child and Adolescent Psychiatry and Psychology, Erasmus University Medical Center – Sophia Children's Hospital, Rotterdam, The Netherlands
2 Department of Complex Trait Genetics, Center for Neurogenomics and Cognitive Research, Neuroscience Campus Amsterdam, VU/VU Medical Center, Amsterdam, The Netherlands
3 Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, VA, USA

Key words
heterogeneity, measurement invariance, Child Behavior Checklist, sex differences

Correspondence
Gwen Dieleman, Department of Child and Adolescent Psychiatry and Psychology, Erasmus University Medical Center – Sophia Children's Hospital, Dr. Molenwaterplein 60, 3015 GJ Rotterdam, The Netherlands
Telephone (+31) 10-7040209
Email: g.dieleman@erasmusmc.nl

Received 14 April 2015;
revised 10 September 2015;
accepted 2 November 2015

Abstract

Prevalence and severity of childhood behavioral problems differ between boys and girls, and in psychiatry, testing for gender differences is common practice. Population-based studies show that many psychopathology scales are (partially) Measurement Invariance (MI) with respect to gender, i.e. are unbiased. It is, however, unclear whether these studies generalize towards clinical samples. In a psychiatric outpatient sample, we tested whether the Child Behavior Checklist 6–18 (CBCL) is unbiased with respect to gender. We compared mean scores across gender of all syndrome scales of the CBCL in 3271 patients (63.3% boys) aged 6–18. Second, we tested for MI on both the syndrome scale and the item-level using a stepwise modeling procedure. Six of the eight CBCL syndrome scales included one or more gender-biased items (12.6% of all items), resulting in slight over- or under-estimation of the absolute gender difference in mean scores. Two scales, Somatic Complaints and Rule-breaking Behavior, contained no biased items. The CBCL is a valid instrument to measure gender differences in problem behavior in children and adolescents from a clinical sample; while various gender-biased items were identified, the resulting bias was generally clinically irrelevant, and sufficient items per subscale remained after exclusion of biased items. *Copyright © 2016 John Wiley & Sons, Ltd.*

Introduction

The high burden of psychiatric morbidity in children and adolescents (Angold *et al.*, 1996; Ford *et al.*, 2003; Verhulst *et al.*, 1997) is characterized by gender differences (e.g. Grant and Weissman, 2007; Khan *et al.*, 2002; Merikangas *et al.*, 2010; Shear *et al.*, 2007; Widiger, 2007).

Females predominantly display more problems in internalizing behavior, whereas males display more problems in externalizing behavior (Hankin and Abramson, 2001; Heptinstall and Taylor, 2002; Nolen-Hoeksema, 1990; Sterba *et al.*, 2007).

There are many theories why gender differences occur, varying from psychometric (e.g. response bias; de Winter *et al.*, 2005; Rogler *et al.*, 2001) and social (e.g. help seeking behavior, social support, and socio-economic status; Lee *et al.*, 1995; Piccinelli and Wilkinson, 2000; Vlassoff, 1994), to biological factors (e.g. for neurodevelopmental conditions females require more mutations in the genome to reach the clinical threshold; Jacquemont *et al.*, 2014). From a psychometric perspective, gender differences might be augmented through the nature of our measurement models. Because mental disorders cannot be observed directly, they are considered to be latent disorder dimensions measured by individual items that relate to manifest behavioral features. Consequently, current measurement models rely on the underlying item structure to assess mental disorders and gender-specific items might influence prevalence estimates. For example, if an instrument designed to measure conduct problems includes typically male items (e.g. physical violent behavior) while females show more relational aggression (e.g. social exclusion), it could capture conduct problems expressed by females inadequately, inflating the male prevalence rates.

Different prevalence rates across gender are not necessarily problematic because gender-specific norms, cutoffs, or measurement models can remedy gender specific endorsements (for an example of gender- and age-specific factor structures in childhood problem behavior see Achenbach and Edelbrock, 1979). However, it becomes problematic when females, given a certain level of construct severity, have a different probability to answer specific items affirmatively compared to males. For instance, a depression-related item "crying" could show gender-bias if girls in general tend to cry more often than boys, while their depression levels are identical. In other words, girls have a higher probability to endorse this item than boys even if their depression level is similar to that of boys. If this is the case, the item not only reflects depression, but also gender differences that are unrelated to depression level. The presence of such biased items renders composite scores based on these items incomparable across gender, and could thus lead to an over- or under-estimation of the gender difference (e.g. Van der Sluis *et al.*, 2010). Hence, comparison of (subscale) scores on instruments or questionnaires across gender, can only be meaningful if one has established that the subscales measure the same construct in both groups, i.e. whether they are "Measurement Invariance" (MI) with respect to gender. MI thus implies that a measurement instrument measures the same underlying factors or latent traits across groups (Mellenbergh, 1989; Meredith, 1993).

MI has been extensively studied in item response theory (Vandenberg and Lance, 2000), and specific guidelines concerning the development of measurement instruments from the American Psychological Association (APA, 1999) and the International Test Commission (2010) have emphasized the importance of MI. Consequently, MI has now been investigated in a wide variety of measurement instruments across variables like culture (e.g. Belon *et al.*, 2014; Boyraz *et al.*, 2013; Fergus and Wu, 2013; Trent *et al.*, 2013; Van Lieshout *et al.*, 2011), age (e.g. Abdellaoui *et al.*, 2012; Dakanalis *et al.*, 2013; Keefer *et al.*, 2013; Mathyssek *et al.*, 2013; Rosen *et al.*, 2013; Schlotz *et al.*, 2011; Willoughby *et al.*, 2012), time (e.g. Ferro and Boyle, 2013; Mathyssek *et al.*, 2013; Rosen *et al.*, 2013), and patient subgroups (e.g. Ferro *et al.*, 2014; Spinhoven *et al.*, 2014). An overview of recent (2010 onwards) studies testing MI across gender in problem behavior in childhood and adolescence is presented in Table 1. In short, most studies provide evidence for unbiased measurement instruments between boys and girls from a general population in childhood and adolescence, although several report partial violation of MI due to gender-bias in specific items (partial invariance). Typical biased items are "not being loved" and "crying a lot" (boys and girls, respectively, score higher than expected given their latent trait scores) (Lundervold *et al.*, 2013; Verhoeven *et al.*, 2013; Wu and Huang, 2014).

The Child Behavior Checklist (CBCL) is a parent-rated measurement scale designed to assess problem behavior in children and adolescents in a reliable and cost-effective way (Achenbach and Rescorla, 2001; Pauschardt *et al.*, 2010). In clinical settings, the CBCL is broadly used in clinical settings to assess problem behavior in referred children. Therefore it is important to specifically investigate the psychometric properties of the CBCL in a clinical, referred sample. Most studies investigating MI in scales measuring problem behavior in children and adolescents were conducted in population-based samples. The aim of the present study is to test whether the interpretation of the scores of the CBCL scales is similar for boys and girls in a referred clinical sample. Given observed gender differences in presentation of symptoms at the time of referral, it is important to exclude the possibility of gender-related item bias in clinical samples. Further, presentation of symptoms in boys and girls may differ at the time of referral to a child psychiatric clinic. For instance,

Int. J. Methods Psychiatr. Res. 2017;**26**:e1498. DOI: 10.1002/mpr

Table 1. Overview of prior studies investigating measurement invariance (MI) in problem behavior in children and adolescents across gender

Study	Sample	Trait(s)	Instrument (rater)	Invariance level	Biased items/scales
Zimprich and Mascherek (2012)	1107 Swiss college students	Anger expression	State-Trait Anger Expression Inventory (self-rated)	Partial invariance	I am angrier than I am willing to show, I exclaim threats without intending to realize them
Brunet *et al.* (2014)	527 Canadian adolescents	Depressive symptoms	Depressive Symptoms Scale (DSS) (self-rated)	Partial invariance	Latent factor means
Cyders (2013)	1274 US adolescents	Lack of deliberation, lack of perseverance, negative- urgency, positive urgency, sensation seeking.	UPPS-P Impulsive Behavior Scale (self-rated)	Partial invariance	Lack of perseverance, positive urgency, sensation seeking
Lundervold *et al.* (2013)	9702 Norwegian adolescents	Depressive symptoms	Short Mood and Feelings Questionnaire (self-rated)	Partial invariance	I felt miserable or unhappy, I felt so tired I just sat around and did nothing, I cried a lot, I thought nobody really loved me
Verhoeven *et al.* (2013)	2650 Australian adolescents	Depressive symptoms	Center for Epidemiologic Studies Depression Scale (CES-D) (self-rated)	Partial invariance	I had crying spells, I felt that I could not shake of the blues, people were unfriendly, I had poor appetite
Wu (2010a)	2922 Taiwanese college students	Depressive symptoms	Beck Depression Inventory-II (Chinese version)(self-rated)	Partial invariance	Self-dislike, crying, irritation, failure, loss of pleasure, sleep pattern
Wu (2010b)	979 Taiwanese college students	Depressive symptoms	Beck Depression Inventory-II (Chinese version)(self-rated)	Partial invariance	Pessimism, failure, loss of pleasure, self-dislike, crying
Wu and Huang (2014)	827 Taiwanese adolescents	Depressive symptoms	Beck Depression Inventory-II (Chinese version)(self-rated)	Partial invariance	Pessimism, failure, self-dislike, suicidal wish, crying, loss of interest, concentration problems
Yamell *et al.* (2013)	1146 eleven-year-old children from Mauritius	Anxious/Depressed, Withdrawn/Depressed, Somatic Complaints, Attention Problems, Rule-breaking Behavior, Aggressive Behavior	Child Behavior Checklist (parent-rated)	Partial invariance	Cries a lot, daydreams, shy/timid behavior, secretive, demands attention, cruel to others, sulks, tantrums
Preti *et al.* (2013)	649 Italian college students	Affective temperaments	Temperament Evaluation of Memphis, Pisa, Paris and San Diego - Autoquestionnaire (self-rated)	Full invariance	-
Frazier *et al.* (2014)	7921 Autism spectrum disorder (ASD) and non-ASD siblings, and 1012 children and 702 adults	Autistic traits	Social Responsiveness Scale-2 (caregiver- and parent-rated)	Strict invariance	-
Wang *et al.* (2013)	5059 Chinese adolescents	Depressive symptoms		Full invariance	-

(*Continues*)

Table 1498. (Continued)

Study	Sample	Trait(s)	Instrument (rater)	Invariance level	Biased items/scales
			Center for Epidemiologic Studies Depression Scale (CES-D) (self-rated)		
Whisman *et al.* (2013)	7369 US college students	Depressive symptoms	Beck Depression Inventory-II (self-rated)	Full invariance	-
Contractor *et al.* (2013)	6591 US trauma exposed children and adolescents	Post-Traumatic Stress Disorder symptoms	Post-Traumatic Stress Disorder Reaction Index (self-rated)	Full invariance	-
Abdellaoui *et al.* (2012)	2743 Dutch adolescent twins	Thought problems	Adult Self Rating (self-rated)	Full invariance	-
Fonseca-Pedrero *et al.* (2011)	1789 Spanish college students	Cognitive-Perceptual, Interpersonal, Disorganized	Schizotypal Personality Questionnaire-Brief (self-rated)	Full invariance	-
Fonseca-Pedrero *et al.* (2012)	4868 Spanish adolescents	Anxious/Depressed, Withdrawn/Depressed, Somatic Complaints, Social Problems, Thought Problems, Attention Problems, Rule-breaking Behavior, Aggressive Behavior	Youth Self Report (self-rated)	Full invariance	-
Memetovic *et al.* (2014)	1352 Canadian adolescents	Substance abuse dimensions: Impulsivity, Anxiety sensitivity, Sensation seeking, Hopelessness	Substance Use Risk Profile Scale (self-rated)	Strict invariance	-

Note: This table summarizes studies from 2010 onwards testing Measurement Invariance (MI) in problem behavior across gender in children and adolescents. In all of these studies, MI was tested by multi-group Confirmatory Factor Analyses. From these studies, two were conducted in a clinical sample (Contractor *et al.* 2013; Frazier *et al.* 2014). In a full invariance model, no items are biased. In a partial invariance model, one or more items/scales are biased and the instrument often is MI with those biased items removed. In strict invariance, factor loadings, thresholds, and residuals could be constrained to be equal across gender, but it was not tested whether the latent factor means of both gender groups were equal. Abbreviations: Gen. = General; US = United States.

Int. J. Methods Psychiatr. Res. 2017;**26**:e1498. DOI: 10.1002/mpr

in autism spectrum disorders girls in the early years appear to have better social skills than boys, but later on their problems become more obvious, and may actually have a more severe presentation than boys (Kirkovski *et al.*, 2013; Kopp and Gillberg, 1992). Testing MI across gender in a clinical sample in measurement scales designed to assess problem behavior in children and adolescents, such as the CBCL, can provide information on the diagnostic value of items, as more clinical weight can be given to unbiased items. Another advantage of testing for MI in a large clinical sample is that, compared to population-based samples, clinical samples have higher endorsement on scales measuring problem behavior by nature, which allows evaluation of the original item categories and increases specificity of the measurement model. In samples with low endorsement, such as population-based samples, these categories are often collapsed which can result in a loss of information. Therefore, the aim of the present study is to study gender differences in mother-reported problem behavior in a sample of children and adolescents from an outpatient clinic. Specifically, we investigate whether scores obtained with the CBCL are MI with respect to gender in this population. The CBCL has been widely used in population and clinical samples to obtain an indication of eight dimensions of problem behavior. We expect to identify more biased items in comparison with a population-based sample, as clinical samples tend to diverge more than population-based samples.

Method

Sample

Data were collected in a total of 3498 children visiting the outpatient clinic of the department of child and adolescent psychiatry at the Sophia Children's Hospital at the Erasmus Medical Center in Rotterdam from January 2001 until January 2012. At registration at the clinic, children and their parents received a standardized selection of questionnaires from the Achenbach System of Empirically Based Assessment (ASEBA; Achenbach, 1991) as part of routine assessment. The CBCL is part of the ASEBA and was filled in by the mother before the first appointment at the outpatient clinic. The Committee on Research Involving Human Subjects in the Netherlands approved this procedure.

Following a series of exclusion criteria, 227 subjects were excluded (N = 6, aged > 18 years; N = 27, owing to > 8 missing item-scores on the CBCL; N = 194, owing to psychiatric symptoms caused by severe medical conditions), resulting in a final sample size of 3271.

The outpatient clinic is organized in specialized teams (e.g. internalizing) to which a child is assigned depending on the reason for referral, and the child's complaints. Although children were referred to the clinic because of emotional or behavioral problems, not all children received a clinical diagnosis after full diagnostic examination. In the final sample 68% of the children received at least one diagnosis according to the Diagnostic and Statistical Manual of Mental Disorders, fourth edition (DSM-IV; APA, 2000).

Child behavior checklist (CBCL)

To assess problem behavior, the third version of the Child Behavior Checklist for Ages 6–18 (CBCL/6–18, version 2001) was used (Achenbach and Rescorla, 2001). Psychometric properties of the Dutch translation are good with internal consistency (Cronbach's alpha) of internalizing and externalizing problems above 0.78 in childhood and adolescence (Verhulst and Van der Ende, 2013). All CBCL items are scored on a three-point ordinal scale (0 = not true, 1 = somewhat or sometimes true, and 2 = very true or often true). Of the 120 items, 103 can be arranged in eight, culturally independent (Rescorla *et al.*, 2012), syndrome scales: Anxious/Depressed (N_{item}= 13), Withdrawn/Depressed (N_{item}= 8), Somatic Complaints (N_{item}= 11), Social Problems (N_{item}= 11), Thought Problems (N_{item}= 15), Attention Problems (N_{item}= 10), Rule-breaking Behavior (N_{item}= 17), and Aggressive Behavior (N_{item}= 18). The Dutch test–retest reliabilities of the eight syndrome scales are good with correlations varying between 0.74 and 0.87 , within a timeframe of 16 days (Verhulst and van der Ende, 2013).

Statistical analyses

All descriptive analyses were performed in SPSS v.20. MI analyses were performed in Mplus v.7 (Muthén and Muthén, 2012). First, we describe psychometric qualities of the different syndrome scales for the total sample. Mean scores, standard deviations, scale range, and skewness were computed for the syndrome scales' sum scores. Cronbach's alpha (Cronbach, 1951) was evaluated (values above 0.75 indicate good internal consistency).

In the MI analyses, CBCL scores of boys (N = 2069; mean age 9.85, standard deviation [SD] = 2.69) were compared to the scores of girls (N = 1202; mean age 11.41, SD = 3.24; girls were significantly older compared to boys: $t(2149.97) = -14.10$, $p < 0.001$). Since age is known to have a moderating effect on item prevalence (Contractor *et al.*, 2013), age was included in the factor models as a covariate on all items, with the age effects freely estimated in boys and girls separately. As we

investigated eight syndrome scales, we adopted a Bonferroni corrected α of 0.05/8 = 0.00625.

Measurement invariance (MI) analyses

MI with respect to gender was tested for each of the eight syndrome scales separately in a multi-group Confirmatory Factor Analysis (mgCFA) for ordinal data. For each scale, the baseline models used in the MI investigation were identical to the factor models presented in the manual of Achenbach and Rescorla (2001; see earlier). Specifically, we specified for each syndrome scale only one latent continuous factor on which the scale-specific items loaded. We assumed an underlying normal distribution of liability with two thresholds to model the three original response categories of the observed items. All steps required to test MI with ordinal items are described and discussed elsewhere (Millsap and Yun-Tein, 2004; Widaman and Reise, 1997). Here, we give a short overview of the model constraints that are introduced in a step-wise fashion to test for MI with respect to gender. These steps, implications, and constraints are summarized in Table 2.

First, we tested for configural invariance across gender. In this step, the configuration of factor loadings, and if necessary additional correlated residuals, is required to be the same across gender, yet the values of the estimated model parameters are allowed to vary between boys and girls. In all MI analyses, boys were considered the reference group. For reasons of identification (Millsap and Yun-Tein, 2004), the factor means were fixed to zero in boys, while all item thresholds were estimated freely. In the girls, the first of the two thresholds per item were constrained to be equal to those of boys, and in one item (the reference item), the second threshold was constrained to be equal across gender as well. All remaining thresholds were estimated freely in girls, as were the factor means. Factor variances were fixed to one in both groups, and residual variances were fixed to one in the boys and estimated freely in girls. If fit indices (see later) indicated that the specified one-factor model did not describe the data well (i.e. did not explain all (co)variance in the data), we examined the modification indices calculated by Mplus (i.e. indices of local misfit in the model) to trace the source of the misfit, and introduced additional residual correlations between highly correlating items, to arrive at an adequate baseline model. Such additional correlations were then introduced in both gender groups, to retain an identical configuration of parameters in the baseline model.

Second, metric invariance was tested, i.e. is the function relating the observed items to the latent factor identical in boys and girls. This step entails constraining all factor loadings to be equal in boys and girls. As a result of these constraints, constraining the factor variance in both groups to one for identification purposes becomes superfluous; the factor variance remained fixed to one in boys but was estimated in girls. The fit of this model was compared to the fit of the configural invariance model and a significant deterioration in fit indicated that not all factor loadings were identical across gender, implying ≥1 items did not represent the latent trait equally well in both gender groups. Such items are considered biased.

Third, strong factorial invariance was tested, which implies a restriction on the means structure of the data. Specifically, strong factorial invariance concerns the question whether the mean differences observed between boys and girls in the observed items can all be explained by mean differences on the latent factor. This step thus entails constraining all thresholds to be equal between boys and girls. Consequently, the factor mean could be estimated freely in the girls group, while the factor mean in the boys group remained fixed to zero for reasons of identification. Significant deterioration in the fit of this model compared to the metric invariance model indicates the gender differences on the latent level cannot account for all gender differences in the observed item scores. Items for which the gender differences in observed scores cannot be explained by gender differences on a latent level, are also considered biased.

Fourth, we tested strict factorial invariance by fixing the residual variances (variance not explained by the factor model) to one in both groups. In previous models, residual variances were fixed to one in the reference group, but were estimated freely in girls for model identification. If the assumption of strict factorial invariance holds, then all differences between boys and girls in the means and covariance structure of the data can be accounted for by differences in the latent factor.

Finally, when at least strong factorial invariance proved tenable (i.e. strict factorial invariance is not a prerequisite for testing full MI) we tested full MI by constraining the factor mean of girls to be identical to the factor mean of the boys, which was fixed to zero for reasons of identification. Fixing the factor mean in girls to zero thus tested whether boys and girls differ with respect to their mean latent syndrome level.

We note that when items were identified as biased, the parameters inducing the bias (e.g. factor loadings, thresholds) were subsequently allowed to take on different values in boys and girls. As a consequence, these items were effectively eliminated from the latent model. That is, these items, although still in the model, did not contribute to possible gender differences on a latent level.

Int. J. Methods Psychiatr. Res. 2017;**26**:e1498. DOI: 10.1002/mpr

Table 2. Parameterization of the models testing different forms of measurement invariance

Test	Goal/Implication	Factor loadings	Thresholds	Correlated Residuals	Residuals	Factor means	Factor Variances
Configural	Establish whether the measurement model was the same in boys and girls	Free	First thresholds constrained. All second thresholds free but one item constrained	Constrained	B: fixed@1 G: free	B: fixed@0 G: free	fixed@1
Metric	Relations between the observed items and latent factor are constrained to be identical across gender.	Constrained	First constrained. Second free but one item constrained	Constrained	B: fixed@1 G: free	B: fixed@0 G: free	B: fixed@1 G: free
Strong factorial	Differences between boys and girls on the level of the observed items can all be accounted for by the mean differences on the latent factor.	Constrained	Constrained	Constrained	B: fixed@1 G: free	B: fixed@0 G: free	B: fixed@1 G: free
Strict factorial	In this model, all gender differences were modeled on a latent level, while the measurement part of the model was identical in boys and girls.	Constrained	Constrained	Constrained	fixed@1	B: fixed@0 G: free	B: fixed@1 G: free
Full invariance	Boys and girls are similar with respect to the latent trait of interest, i.e. with respect to the latent syndrome.	Constrained	Constrained	Constrained	fixed@1	fixed@0	B: fixed@1 G: free

Note: If parameters are estimated freely, boy and girl scores are allowed to differ but constrained parameters are fixed to be identical between gender groups. Fixed@ indicates a parameter estimate is fixed to a specific value.
Abbreviations: B = boys; G = girls.

Model evaluation

We evaluated the root mean square error of approximation (RMSEA) and Comparative Fit Index (CFI) as general measures of fit. The RMSEA provides an indication of how well the model fits in the population. Values > 0.10 indicate poor model fit, values between 0.08 and 0.05 indicate adequate model fit, and values of 0.05 or below indicate good fit of the model to the data (Schermelleh-Engel *et al.*, 2003; Yu, 2002). The CFI ranges from zero to one and higher values indicate better fit. It has been shown to be an adequate fit statistic for ordinal data (Yu, 2002) with values larger than 0.95 indicating good fit. Also, we fitted all models using the weighted least squares mean variance (WLSMV) adjusted estimation. As the WLSMV-estimator does not follow a chi-squared distribution, the DIFFTEST option in Mplus was used to make model comparisons using chi-square difference (χ^2_{diff}) tests between models. If these fit indices indicated misfit, we inspected the modification indices to check whether relaxation of parameters resulted in a better model fit (decrease of the RMSEA, increase of the CFI, and decrease of χ^2_{diff}). Mplus applies pair-wise deletion to accommodate missingness in the raw data when WLSMV is used.

Results

The distribution of all children over the different diagnostic teams is shown in Table 3. The distribution of boys and girls across the teams differed significantly ($\chi^2(4, 3271)=479.201$, $p < 0.001$). Boys were mostly referred to the developmental disorders (38%), internalizing (30%), and externalizing (27%) teams. In girls, almost half (47%) of the total patient group was referred to the internalizing disorders team and the remaining girls were equally distributed over externalizing (15%), somatoform (12%), developmental (16%), and eating (10%) disorder teams.

Descriptive statistics and psychometric properties of the different syndrome scales are provided in Table 4. In all syndrome scales, the range of observed scale scores was very similar between boys and girls. Internal consistencies were lower for boys in all syndrome scales except Thought Problems.

Measurement invariance (MI) across gender

Here we provide a concise description of the outcomes from the MI analyses. Details regarding fit statistics and procedures are included in the Supporting Information.

Table 3. Frequencies of referred team for the different gender groups

	Boys	Girls		
	N (%)	*N* (%)	Total *N* (%)	Gender Ratio
Team				
Externalizing disorders	560 (27%)	176 (15%)	736 (23%)	3.18
Internalizing disorders	631 (30%)	562 (47%)	1193 (36%)	1.12
Somatoform disorders	65 (3%)	143 (12%)	208 (6%)	0.45
Developmental disorders	794 (38%)	195 (16%)	989 (30%)	4.07
Eating disorders	19 (1%)	126 (10%)	145 (4%)	0.15
Total *N*	2069	1202	3271	1.72
Primary DSM-IV classification				
Anxiety disorder	275 (13%)	308 (26%)	583 (18%)	0.89
Pervasive developmental disorder	613 (30%)	123 (10%)	736 (23%)	4.98
Attention deficit hyperactivity disorder	401 (19%)	108 (9%)	509 (16%)	3.71
Eating disorder	10 (0.5%)	112 (9%)	122 (4%)	0.09
Behavioral disorder	92 (4%)	35 (3%)	127 (4%)	2.63
Mood disorder	28 (1.5%)	42 (3.5%)	70 (2%)	0.67
Somatoform disorders	18 (1%)	65 (5%)	83 (2.5%)	0.28
Obsessive compulsive disorder	22 (1%)	30 (2.5%)	52 (2%)	0.73

Note: The number of patients per team are displayed; column and row percentages provided between brackets. The number of patients per primary DSM-IV classification are displayed, primary classifications with $N < 50$ are not displayed. The gender ratio reflects the relative frequency of boys over girls in each team, a ratio below one indicates a higher frequency of girls compared to boys.

Int. J. Methods Psychiatr. Res. 2017;**26**:e1498. DOI: 10.1002/mpr

Table 4. Descriptive statistics for the syndrome scales

Syndrome scale		*N*	Mean (SD)	Range	N_{item}	StSkew	α	Clinical score *N* (%)
Anxious/Depressed	Boys*	2039	7.30 (5.09)	24	13	13.15	0.821	500 (25%)
	Girls	1179	8.97 (5.43)	25		6.31	0.825	435 (37%)
Withdrawn/Depressed	Boys*	2057	4.73 (3.42)	15	8	11.65	0.751	688 (33%)
	Girls	1192	5.44 (3.70)	16		6.30	0.779	454 (38%)
Somatic Complaints	Boys*	2027	3.14 (3.11)	18	11	23.41	0.714	298 (15%)
	Girls	1169	4.88 (3.87)	18		10.82	0.733	356 (30%)
Social Problems	Boys*	2047	6.68 (4.07)	19	11	8.17	0.723	652 (32%)
	Girls	1189	6.11 (4.38)	20		8.51	0.788	337 (28%)
Thought Problems	Boys	1972	5.98 (4.40)	25	15	14.4	0.716	807 (41%)
	Girls	1147	4.36 (4.36)	23		11.86	0.703	461 (40%)
Attention Problems	Boys*	2051	9.30 (4.26)	20	10	−2.57	0.779	847 (41%)
	Girls	1187	7.44 (4.77)	19		3.30	0.832	343 (29%)
Rule-breaking Behavior	Boys*	2041	4.04 (3.55)	25	17	25.5	0.741	354 (17%)
	Girls	1177	3.24 (3.59)	25		24.70	0.768	134 (11%)
Aggressive Behavior	Boys*	2036	12.25 (7.76)	36	18	7.43	0.907	667 (33%)
	Girls	1189	9.64 (7.51)	33		11.65	0.916	237 (20%)

Note: The range indicates the highest obtained score in the sample minus the lowest obtained score. As the lowest score is "0" in this dataset, the range reported here is equal to the upper bound.
Abbreviations: *N* = sample size; SD = standard deviation; StSkew = standardized Skewness statistic (i.e. *Z*-scores calculated as skewness divided by its standard error); α = Cronbach's alpha; N_{item} = number of items in the syndrome scale.
*Mean differed significantly ($p < 0.01$) between boys and girls.

Anxious/depressed

A one-factor model did not describe the variance–covariance structure of the 13 Anxious/Depressed items adequately (RMSEA = 0.089, CFI = 0.907). Modification indices showed that the misfit was mainly due to five items that correlated higher than could be accommodated by the one-factor model. Including four correlated residuals (three to accommodate the additional correlations between "feels unloved", "feels worthless", and "talks of suicide", and one between "feels too guilty" and "fears"), resulted in a significant improvement of the fit ($\chi^2_{diff}(4) = 1014.73$, $p < 0.001$), and this extended one-factor model proved an adequately fitting baseline model in both boys and girls (RMSEA = 0.057, CFI = 0.963). Constraining the factor loadings to be equal across gender resulted in a significant deterioration of fit ($\chi^2_{diff}(12) = 92.93$, $p < 0.001$), which was mainly due to the item "cries a lot", which had a higher factor loading in girls than in boys. Freely estimating the factor loading for this item only in both groups resulted in a significant improvement of model fit ($\chi^2_{diff}(1) = 48.86$, $p < 0.001$); this model fit well (RMSEA = 0.055, CFI = 0.963). As this implies that this particular item is gender-biased, all parameters for this item were estimated freely between groups in subsequent models. Constraining thresholds and residuals to be equal across gender (strong and strict factorial invariance, respectively) did not result in significant deterioration of the model fit but constraining the latent factor means did ($\chi^2_{diff}(1) = 37.27$, $p < 0.001$), indicating that girls scored significantly higher than boys on anxious depressed behavior ($d = -0.260$).

Withdrawn/depressed

In the configural model containing eight items, two residual correlations were specified between items "won't talk" – "secretive" and "enjoys little" – "shy or timid" (improvement in model fit $\chi^2_{diff}(2) = 254.55$, $p < 0.001$), resulting in a baseline model with adequate fit (RMSEA = 0.065, CFI = 0.971). The tests for metric and strong invariance showed that three out of eight items were gender-biased (metric: "unhappy, sad or depressed" loaded more strongly in girls than in boys; strong: the first threshold of the items "enjoys little" and "shy or timid" was higher and lower, respectively, in girls). In this abridged model, the latent factor mean, now effectively based on the remaining five items ("rather be alone", "won't talk", "secretive", "lacks energy", and "withdrawn") , did not differ significantly between boys and girls ($\chi^2_{diff}(1) = 3.90$, not significant [ns], $d = -0.080$; RMSEA = 0.050, CFI = 0.976).

Somatic complaints

The configural model contained 11 items and required two residual correlations ("nausea" – "feels dizzy" and "nausea" – "vomiting") to overcome local misfit ($\chi^2_{diff}(2) = 168.40$, $p < 0.001$). The resulting baseline model had

good fit (RMSEA = 0.038, CFI = 0.980) and none of the subsequent MI constraints (metric, strong, strict) resulted in a significant deterioration of the model fit. Constraining the latent factor means to be identical in boys and girls did however result in significant deterioration of model fit ($\chi^2_{diff}(1) = 117.77$, $p < 0.001$): girls scored significantly higher on somatic complaints than boys ($d = 0.518$).

Social problems

The poor fit of the original 11 item ASEBA model (RMSEA = 0.093, CFI = 0.908) was mainly due to four residual correlations ("accident prone" – "clumsy", "clumsy" – "speech problems", "doesn't get along" – "gets teased", "gets teased" – "not liked"): including these improved the fit substantially ($\chi^2_{diff}(4) = 656.98$, $p < 0.001$) and resulted in an adequately fitting model (RMSEA = 0.066, CFI = 0.956). The item "dependent" was not metric invariant ($\chi^2_{diff}(1) = 26.20$, $p < 0.001$, factor loading was higher in girls) and the items "lonely" and "jealous" failed strong MI ($\chi^2_{diff}(2) = 62.67$, $p < 0.001$; for both items, the first threshold was lower in girls). The latent factor means were not equal between boys and girls ($\chi^2_{diff}(1) = 24.74$, $p < 0.001$): girls scored significantly lower than boys ($d = 0.218$).

Thought problems

In the 15-item configural model, the items "shows sex parts in public" and "shows sex parts too much" correlated 1.00; we therefore excluded the first item from further analyses. After introducing residual correlation between the items "sleeps less" and "trouble sleeping" ($\chi^2_{diff}(1) = 405.83$, $p < 0.001$), the baseline model fitted adequately (RMSEA = 0.049, CFI = 0.939). In metric testing, only one item ("twitching") was gender-biased ($\chi^2_{diff}(1) = 35.44$, $p < 0.001$; higher factor loading in boys). Strong invariance testing did not result in a significant worsening of the model fit (RMSEA = 0.044, CFI = 0.942). However, the latent factor means of thought problems could not be constrained to be equal across gender ($\chi^2_{diff}(1) = 7.63$, $p = 0.0057$): girls scored slightly, yet significantly, higher than boys ($d = -0.136$).

Attention problems

After introducing three additional correlations between the items "confused", "daydreams", and "stares" ($\chi^2_{diff}(3) = 1198.98$, $p < 0.001$), the fit of the 10-item baseline model was adequate (RMSEA = 0.059, CFI = 0.980). One item ("sitting still") violated the metric invariance test ($\chi^2_{diff}(1) = 21.38$, $p < 0.001$; higher factor loading in girls). The latent factor means were not equal across gender ($\chi^2_{diff}(1) = 54.42$, $p < 0.001$): girls scored significantly lower than boys ($d = 0.294$).

Rule-breaking behavior

The configural 17 item model fitted adequately after introducing four additional residual correlations ("drinks alcohol" – "uses drugs", "lies/cheats" – "uses drugs", "sex problems" – "thinks of sex too much", "steals at home" – "steals outside home": $\chi^2_{diff}(4) = 325.39$, $p < 0.001$, RMSEA = 0.035, CFI = 0.966). Metric and strong testing did not result in significant misfit. The latent scores were not equal across gender ($\chi^2_{diff}(1) = 70.52$, $p < 0.001$): girls scores significantly lower than boys ($d = 0.367$).

Aggressive behavior

The one-factor model fitted the data poorly (RMSEA = 0.098, CFI = 0.925), but introduction of eight additional residual correlations resulted in adequate fit (RMSEA = 0.067, CFI = 0.966; "mean" – "teases a lot", "destroys own things" – "destroys others things", "disobedient at home" – "disobedient at school", "stubborn" – "mood changes", "stubborn" – "sulks", "mood changes" – "sulks", "sulks" – "suspicious", "screams a lot" – "loud"). Three of the eighteen items failed the metric invariance test ("gets in fights", "mood changes/tantrums" had lower factor loading in girls, "suspicious" had a higher factor loading in girls; $\chi^2_{diff}(3) = 98.89$, $p < 0.001$), and one item failed the test for strong invariance ("sulks"; $\chi^2_{diff}(1) = 55.75$, $p < 0.001$; second threshold was higher in boys). The latent factor means were not equal across gender ($\chi^2_{diff}(1) = 63.23$, $p < 0.001$): girls scored significantly lower compared to boys ($d = 0.315$).

Effect size calculation with and without biased items

To investigate whether biased items caused under- or over-estimation of the effect size of the difference in problem behaviors between boys and girls, we calculated the effect size *d* based on the regular sum score differences (including all items) and the effect size *d* from the unbiased sum score (without gender-biased items). Table 5 displays effect sizes of gender effects as calculated on the latent factor means (effectively based on unbiased items only), sum scores based on all items, sum scores based on only the unbiased items, and effect size differences between biased an unbiased sum scores (*d*). Note that the effect size of the gender differences is much more pronounced when calculated on the latent factor means, which illustrates how removing measurement noise increases signals. In three syndrome scales, the inclusion of biased items would result

Int. J. Methods Psychiatr. Res. 2017;**26**:e1498. DOI: 10.1002/mpr

Table 5. Means and effect sizes (Cohen's *d*) for latent means, sum scores with and sum scores without gender-Biased items across groups

	Latent means			Regular Sum			Unbiased sum				
	Boys	Girls	*d*	Boys	Girls	*d*	Boys	Girls	N_{biased}	*d*	Δd
Internalizing											
Anxious/Depressed	0 (1.00)	0.259 (0.992)	−0.260	7.30 (5.09)	8.97 (5.43)	−0.321	6.86 (4.85)	8.30 (5.09)	1	−0.292	0.029
Withdrawn/Depressed	0 (1.00)	0.080 (1.00)	−0.080	4.73 (3.42)	5.44 (3.69)	−0.201	2.79 (2.36)	3.20 (2.47)	3	−0.171	0.030
Somatic Complaints	0 (1.00)	0.501 (0.909)	−0.518	3.14 (3.11)	4.88 (3.87)	−0.510	—	—	0	—	—
Social Problems	0 (1.00)	−0.264 (1.51)	0.218	6.68 (4.07)	6.11 (4.38)	0.136	4.75 (3.19)	3.84 (3.29)	3	0.282	−0.146
Thought Problems	0 (1.00)	0.136 (1.00)	−0.136	5.88 (4.31)	5.99 (4.32)	−0.026	5.30 (4.04)	5.51 (4.03)	1	−0.052	−0.026
Attention Problems	0 (1.00)	−0.345 (1.42)	0.294	9.30 (4.26)	7.44 (4.77)	0.418	8.15 (3.87)	6.70 (4.32)	1	0.359	0.059
Externalizing											
Rule-breaking Behavior	0 (1.00)	−0.442 (1.49)	0.367	4.04 (3.55)	3.24 (3.59)	0.220	—	—	0	—	—
Aggressive Behavior	0 (1.00)	−0.354 (1.31)	0.315	12.25 (7.76)	9.64 (7.52)	0.340	9.79 (6.23)	7.25 (5.97)	4	0.414	−0.074

Note: The latent means are effectively based on the unbiased items only. The unbiased sum score excluded all gender-biased items and included items with low prevalence or high inter-item correlations (Thought Problems includes item 59; "show's sex parts in public"). Standard deviation provided in parentheses.
Abbreviations: N_{biased} = number of gender-biased items in the scale; Δd = difference in effect sizes of regular sum and unbiased sum scores.
*$p < 0.01$;
**$p < 0.001$.

in a small over-estimation of the effect size of the gender difference; Anxious/Depressed $d = 0.029$, Withdrawn/Depressed $d = 0.03$, and Attention Problems $d = 0.059$. In Social Problems, Thought Problems, and Aggressive Behavior, the biased items would result in under-estimation of the effect size of the gender difference ($d = -0.146$, -0.026, and -0.074, respectively). The Somatic Complaints and Rule-breaking Behavior scales did not contain biased items.

Discussion

This study used MI analysis to examine the possible presence of gender-bias in items of the CBCL. In two (Somatic Complaints/Rule-breaking Behavior) of the eight CBCL syndrome scales, the full scale proved fully MI across gender, whereas in the other six several gender-biased items were identified, consistent with a partially invariant model (13 of 103 items; 12.6%). On these biased items (e.g. "crying", "shy/timid behavior", "mood changes", "lonely", "jealous", "can't sit still", and "getting into fights"), boys and girls differed more than would be expected based on their latent scores. Remarkably, in Withdrawn/Depressed problems, the latent factor mean of boys was not significantly different from girls after excluding three gender-biased items, which is contrary to the finding reported in population-based studies (e.g. Grant and Weissman, 2007) that girls score higher on depression. This actually illustrates that results concerning gender differences in population-based samples do not necessarily generalize to clinical populations.

Both internalizing and externalizing scales included biased items (4/32 and 4/35 items, respectively). Overall, differences in the effect size of gender differences based on scale scores either including or excluding biased items were negligible to moderate, varying between 0.026 and 0.146, illustrating the varying effect of the presence of biased items on composite scores.

Our results partially confirm earlier studies reporting M between boys and girls from population-based studies (see Table 1). Specifically, several gender-biased items identified in this study overlap with biased items reported earlier (e.g. "crying", "shy/timid behavior", and "mood changes"; Verhoeven *et al.*, 2013; Wu and Huang, 2014; Yarnell *et al.*, 2013), adding to the construct validity of earlier studies where items were dichotomized. Some items such as "can't concentrate" or "feelings of not being liked or loved" were found to be MI in the current study, but resulted in gender-biased scale scores in another study (Lundervold *et al.*, 2013). Compared to a previous study on the psychometric properties of the CBCL in a population-based sample (Yarnell *et al.*, 2013), several items were found to be MI in our study (e.g. "secretive

behavior", "demands attention", "daydreams"), whereas other items were non-MI in our study (e.g. "enjoys little", "sad"). We cautiously speculate that some items are MI in our sample in contrast to the general population study of Yarnell *et al.* (2013), because of the severity of problem behavior in clinical samples. For instance, the item "demands attention" is MI in our study, whereas in the study of Yarnell *et al.* (2013) girls had a lower item loading than boys. This could be due to the fact that, regardless of gender, children with psychiatric problems demand more parental attention, whereas in a healthy general population boys tend to demand more attention than girls. Further, in this study "sad" was non-MI, with girls having a higher factor loading compared to boys, while "sad" was MI in the general population sample of Yarnell *et al.* (2013). This could imply that in a clinical sample the presentation of psychiatric problems in girls is accompanied with more expressed sadness compared to boys. However, this lack of MI for the item "sad" could also be due to referral bias since boys are overrepresented in pervasive developmental disorders, and girls are overrepresented in anxiety disorders.

In our study only the mother rated the problems of her child. However, Derks *et al.* (2004) found in a large community twin sample that, although the major part of the variance in problem behavior is explained by aspects of the child's behavior that are perceived similarly by mothers and fathers, part of the variance is indeed explained by unique perceptions of the child's behavior. It is possible that gender-invariance is under- or over-rated for some items in our sample, because mothers might hold different expectations regarding normal behavior for boys and girls than fathers or children themselves.

Compared to studies using the Youth Self Report (Abdellaoui *et al.*, 2012; Fonseca-Pedrero *et al.*, 2012), we identified more biased items. This difference might be sample-specific: clinical samples tend to diverge more than population-based samples (Hartman *et al.*, 1999). Another explanation is that the Youth Self Report is a self-report instrument, while the CBCL is a parent-report instrument.

In general, girls are less often referred to the clinic than boys (Sourander *et al.*, 2008) and presentation of symptoms in boys and girls can differ at the time of referral to a child psychiatric clinic (Kirkovski *et al.*, 2013; Kopp and Gillberg, 1992). Given the observed gender difference in presentation of symptoms at the time of referral, one could hypothesize that in a referred clinical sample more items might be non-MI. In contrast, however, our study demonstrates that the influence of referral-bias on MI is limited, given the significant overlap in biased-items between our study in a referred clinical sample and previous studies in population-based non-referred samples. Importantly, however, bias and absence of MI are relative concepts that are defined with respect to a certain scale/instrument in a certain population. The results of this study do, thus, not necessarily generalize to other instruments or to study cohorts of a different composition (e.g. non-clinical, different representation of clinical groups).

Implications

From this work, we know that CBCL scales contain several gender-biased items. Removal of these items yields scales that measure the same latent construct across gender once these biased items are excluded. Current gender-specific norms can be considered useful when deciding whether behavior is in the normal or clinical range because: (1) we observed large mean differences between gender groups which justifies gender specific norms; (2) the presence of gender-biased items demonstrates that gender-specific evaluation is sensible. The MI analyses, however, illustrate that some items cannot be used reliably to compare boys to girls because there is an unequal probability for one group towards a particular response. Although the overwhelming majority (87.4%) of the CBCL's items proved MI in our study, the comparisons in Table 5 clearly support the recommendation to exclude biased items when testing for gender differences (especially for the scales Social Problems and Aggressive Behavior), to assure that the comparison is fair and based only on items that have comparable psychometric characteristics in both gender groups.

A strength of this study is the use of the full CBCL, including the dimensions Social Problems and Thought Problems, which were excluded in a prior CBCL study (Yarnell *et al.*, 2013). Future research could investigate whether the items found to be biased in this study remain so under a different rater. In our models we corrected for age and did not compare measurement models between children and adolescents. While previous longitudinal studies support the stability of the MI assumption over development (Mathyssek *et al.*, 2013; Willoughby *et al.*, 2012), we did not specifically address that question in this study. Finally, we specified residual correlations between pairs of items to obtain acceptable fit, which is in violation with the assumption of local independence (Millsap and Yun-Tein, 2004). This problem presents a trade-off between choosing the statistical optimum (multidimensional models) and generalizability though attaining the widely used ASEBA structure. We opted for the latter to stay as close as possible to clinical practice (i.e. the use of scale-based sum scores) and follow earlier studies (e.g. Contractor *et al.*, 2013; Cyders, 2013). Another

Int. J. Methods Psychiatr. Res. 2017;**26**:e1498. DOI: 10.1002/mpr

limitation is that only one measure of the child's problem behavior was available. Parental perceived burden as resulting from their children's problems is a predictor of the use of specialty mental health service (Angold *et al.*, 1998) and parental problems can influence parent ratings of their child's problems (Maoz *et al.*, 2014). Therefore, assessment of psychopathology in children and adolescents should include gathering data from multiple informants and cannot solely rely on parent reports (Comer and Kendall, 2004).

In summary, we showed that the CBCL is a valid instrument to measure gender differences of mother-reported problem behavior in children and adolescents from a clinical sample; while various gender-biased items were identified, the resulting bias was generally small (except for the scales Social Problems and Aggressive Behavior), and sufficient items per subscale remained after exclusion of biased items. These results, combined with results of previous studies on MI, support the use of the CBCL to measure gender differences in problem behavior in both clinical and non-clinical samples.

Acknowledgements

The authors thank all the participating patients and their families. The authors thank M.P. Roeling for his contribution to this research. This work was supported by Sophia Stichting voor Wetenschappelijk Onderzoek (SSWO grant #593). Sophie van der Sluis is financially supported by the Netherlands Scientific Organization (Nederlandse Organisatie voor Wetenschappelijk Onderzoek, gebied Maatschappij-en Gedragswetenschappen: NWO/MaGW VIDI-452-12-014).

Declaration of interest statement

Prof Dr Verhulst publishes the Dutch translations of ASEBA from which he receives remuneration.

References

Abdellaoui A., de Moor M.H., Geels L.M., van Beek J.H., Willemsen G., Boomsma D.I. (2012) Thought problems from adolescence to adulthood: measurement invariance and longitudinal heritability. *Behavior Genetics*, **42**(1), 19–29. DOI:10.1007/s10519-011-9478-x.

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

Achenbach T.M., Edelbrock C.S. (1979) Child-Behavior Profile. 2. Boys aged 12–16 and girls aged 6–11 and 12–16. *Journal of Consulting and Clinical Psychology*, **47**(2), 223–233. DOI:10.1037/0022-006x.47.2.223.

Achenbach T.M., Rescorla L.A. (2001) Manual for the ASEBA School-Age Forms & Profiles, Burlington, VT: University of Vermont, Research Center for Children, Youth & Families.

American Psychological Association (APA) (1999) Standards for Educational and Psychological Testing, Washington, DC: APA.

American Psychological Association (APA) (2000) Diagnostic and Statistical Manual of Mental Disorders, Washington, DC: APA.

Angold A., Erkanli A., Costello E.J., Rutter M. (1996) Precision, reliability and accuracy in the dating of symptom onsets in child and adolescent psychopathology. *Journal of Child Psychology and Psychiatry*, **37**(6), 657–664.

Angold A., Messer S.C., Stangl D., Farmer E.M., Costello E.J., Burns B.J. (1998) Perceived parental burden and service use for child and adolescent psychiatric disorders. *American Journal of Public Health*, **88**(1), 75–80.

Belon K.E., McLaughlin E.A., Smith J.E., Bryan A.D., Witkiewitz K., Lash D.N., Winn J.L. (2014) Testing the measurement invariance of the eating disorder inventory in nonclinical samples of Hspanic and Caucasian Women. *International Journal of Eating Disorders*, **48**(3), 262–270. DOI:10.1002/eat.22286.

Boyraz G., Lightsey O.R., Jr., Can A. (2013) The Turkish version of the Meaning In Life Questionnaire: assessing the measurement invariance across Turkish and American adult samples. *Journal of Personality Assessment*, **95**(4), 423–431. DOI:10.1080/00223891.2013.765882.

Brunet J., Sabiston C.M., Chaiton M., Low N.C., Contreras G., Barnett T.A., O'Loughlin J.L. (2014) Measurement invariance of the depressive symptoms scale during adolescence. *BMC Psychiatry*, **14**, 95. DOI:10.1186/1471-244X-14-95.

Comer J.S., Kendall P.C. (2004) A symptom-level examination of parent–child agreement in the diagnosis of anxious youths. *Journal of the American Academy of Child and Adolescent Psychiatry*, **43**(7), 878–886.

Contractor A.A., Layne C.M., Steinberg A.M., Ostrowski S.A., Ford J.D., Elhai J.D. (2013) Do gender and age moderate the symptom structure of PTSD? Findings from a national clinical sample of children and adolescents. *Psychiatry Research*, **210**(3), 1056–1064. DOI:10.1016/j.psychres.2013.09.012.

Cronbach L.J. (1951) Coefficient alpha and the internal structure of tests. *Psychometrika*, **16**, 297–334.

Cyders M.A. (2013) Impulsivity and the sexes: measurement and structural invariance of the UPPS-P Impulsive Behavior Scale. *Assessment*, **20**(1), 86–97. DOI:10.1177/1073191111428762.

Dakanalis A., Zanetti M.A., Clerici M., Madeddu F., Riva G., Caccialanza R. (2013) Italian version of the Dutch Eating Behavior Questionnaire. Psychometric proprieties and measurement invariance across sex, BMI-status and age. *Appetite*, **71**, 187–195. DOI:10.1016/j.appet.2013.08.010.

de Winter A.F., Oldehinkel A.J., Veenstra R., Brunnekreef J.A., Verhulst F.C., Ormel J. (2005) Evaluation of non-response bias in mental health determinants and outcomes in a large sample of pre-adolescents. *European Journal of Epidemiology*, **20**(2), 173–181.

Derks E.M., Hudziak J.J., van Beijsterveldt C.E., Dolan C.V., Boomsma D.I. (2004) A study of genetic and environmental influences on maternal and paternal CBCL syndrome scores in a large sample of 3-year-old Dutch twins. *Behavior Genetics*, **34**(6), 571–583. DOI:10.1007/s10519-004-5585-2.

Fergus T.A., Wu K.D. (2013) The intolerance of uncertainty scale: measurement invariance, population heterogeneity, and its relation with worry among self-identifying White and Black respondents. *Assessment*, **20**(5), 555–564. DOI:10.1177/1073191112460272.

Ferro M.A., Boyle M.H. (2013) Longitudinal invariance of measurement and structure of

global self-concept: a population-based study examining trajectories among adolescents with and without chronic illness. *Journal of Pediatric Psychology*, **38**(4), 425–437. DOI:10.1093/jpepsy/jss112.

Ferro M.A., Boyle M.H., Scott J.G., Dingle K. (2014) The child behavior checklist and youth self-report in adolescents with epilepsy: testing measurement invariance of the attention and thought problems subscales. *Epilepsy & Behavior*, **31**, 34–42. DOI:10.1016/j.yebeh.2013.11.009.

Fonseca-Pedrero E., Paino M., Lemos-Giraldez S., Sierra-Baigrie S., Muniz J. (2011) Measurement invariance of the Schizotypal Personality Questionnaire – Brief across gender and age. *Psychiatry Research*, **190**(2–3), 309–315. DOI:10.1016/j.psychres.2011.05.021.

Fonseca-Pedrero E., Sierra-Baigrie S., Lemos-Giraldez S., Paino M., Muniz J. (2012) Dimensional structure and measurement invariance of the youth self-report across gender and age. *Journal of Adolescent Health*, **50**(2), 148–153. DOI:10.1016/j.jadohealth.2011.05.011.

Ford T., Goodman R., Meltzer H. (2003) The British Child and Adolescent Mental Health Survey 1999: the prevalence of DSM-IV disorders. *Journal of the American Academy of Child and Adolescent Psychiatry*, **42**(10), 1203–1211. DOI:10.1097/00004583-200310000-00011.

Frazier T.W., Ratliff K.R., Gruber C., Zhang Y., Law P.A., Constantino J.N. (2014) Confirmatory factor analytic structure and measurement invariance of quantitative autistic traits measured by the Social Responsiveness Scale – 2. *Autism*, **18**(1), 31–44. DOI:10.1177/1362361313500382.

Grant B.F., Weissman M.M. (2007) Gender and the prevalence of psychiatric disorders. In Age and Gender Considerations in Psychiatric Diagnosis: A Research Agenda for DSM-V, pp. 31–45, Arlington, VA: American Psychiatric Publishing.

Hankin B.L., Abramson L.Y. (2001) Development of gender differences in depression: an elaborated cognitive vulnerability-transactional stress theory. *Psychological Bulletin*, **127**(6), 773–796.

Hartman C.A., Hox J., Auerbach J., Erol N., Fonseca A.C., Mellenbergh G.J., Novik T.S., Oosterlaan J., Roussos A.C., Shalev R.S., Zilber N., Sergeant J.A. (1999) Syndrome dimensions of the child behavior checklist and the teacher report form: a critical empirical evaluation. *Journal of Child Psychology and Psychiatry*, **40**(7), 1095–1116.

Heptinstall E., Taylor E. (2002) Sex Differences and their Significance, Cambridge: Cambridge University Press.

International Test Commission. (2010) International Test Commission Guidelines for Translating and Adapting Tests. https://www.intestcom.org/

Jacquemont S., Coe B.P., Hersch M., Duyzend M.H., Krumm N., Bergmann S., Beckmann J.S., Rosenfeld J.A., Eichler E.E. (2014) A higher mutational burden in females supports a "female protective model" in neurodevelopmental disorders. *American Journal of Human Genetics*, **94**(3), 415–425. DOI:10.1016/j.ajhg.2014.02.001.

Keefer K.V., Holden R.R., Parker J.D. (2013) Longitudinal assessment of trait emotional intelligence: measurement invariance and construct continuity from late childhood to adolescence. *Psychological Assessment*, **25**(4), 1255–1272. DOI:10.1037/a0033903.

Khan A.A., Gardner C.O., Prescott C.A., Kendler K.S. (2002) Gender differences in the symptoms of major depression in opposite-sex dizygotic twin pairs. *American Journal of Psychiatry*, **159**(8), 1427–1429.

Kirkovski M., Enticott P.G., Fitzgerald P.B. (2013) A review of the role of female gender in autism spectrum disorders. *Journal of Autism and Developmental Disorders*, **43**(11), 2584–2603.

Kopp S., Gillberg C. (1992) Girls with social deficits and learning problems: autism, atypical Asperger syndrome or a variant of these conditions. *European Child and Adolescent Psychiatry*, **1**, 89–99.

Lee P.R., Moss N., Krieger N. (1995) Measuring social inequalities in health. Report on the Conference of the National Institutes of Health. *Public Health Report*, **110**(3), 302–305.

Lundervold A.J., Breivik K., Posserud M.B., Stormark K.M., Hysing M. (2013) Symptoms of depression as reported by Norwegian adolescents on the Short Mood and Feelings Questionnaire. *Frontiers in Psychology*, **4**, 613. DOI:10.3389/fpsyg.2013.00613.

Maoz H., Goldstein T., Goldstein B.I., Axelson D.A., Fan J., Hickey M.B., Monk K., Sakolsky D., Diler R.S., Brent D., Kupfer D.J., Birmaher B. (2014) The effects of parental mood on reports of their children's psychopathology. *Journal of the American Academy of Child and Adolescent Psychiatry*, **53**(10), 1111–1122.

Mathyssek C.M., Olino T.M., Hartman C.A., Ormel J., Verhulst F.C., Van Oort F.V. (2013) Does the Revised Child Anxiety and Depression Scale (RCADS) measure anxiety symptoms consistently across adolescence? The TRAILS study. *International Journal of Methods in Psychiatric Research*, **22**(1), 27–35. DOI:10.1002/mpr.1380.

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

Memetovic J., Ratner P.A., Richardson C.G. (2014) Gender-based measurement invariance of the Substance Use Risk Profile Scale. *Addictive Behaviors*, **39**(3), 690–694. DOI:10.1016/j.addbeh.2013.10.016.

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

Merikangas K.R., He J.P., Burstein M., Swanson S.A., Avenevoli S., Cui L., Benjet C., Georgiades K., Swendsen J. (2010) Lifetime prevalence of mental disorders in U.S. adolescents: results from the National Comorbidity Survey Replication – adolescent supplement (NCS-A). *Journal of the American Academy of Child and Adolescent Psychiatry*, **49**(10), 980–989. DOI:10.1016/j.jaac.2010.05.017.

Millsap R.E., Yun-Tein J. (2004) Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, **39**(3), 479–515. DOI:10.1207/S15327906mbr3903_4.

Muthén L.K., Muthén B. (2012) Mplus Users's Guide, 7th edn, Los Angeles, CA: Muthén & Muthén.

Nolen-Hoeksema S. (1990) Sex Differences in Depression, Stanford, CA: Stanford University Press.

Pauschardt J., Remschmidt H., Mattejat F. (2010) Assessing child and adolescent anxiety in psychiatric samples with the child behavior checklist. *Journal of Anxiety Disorders*, **24**(5), 461–467.

Piccinelli M., Wilkinson G. (2000) Gender differences in depressionCritical review. *British Journal of Psychiatry*, **177**, 486–492.

Preti A., Vellante M., Gabbrielli M., Lai V., Muratore T., Pintus E., Pintus M., Sanna S., Scanu R., Tronci D., Corrias I., Petretto D.R., Carta M.G. (2013) Confirmatory factor analysis and measurement invariance by gender, age and levels of psychological distress of the short TEMPS-A. *Journal of Affective Disorders*, **151**(3), 995–1002. DOI:10.1016/j.jad.2013.08.025.

Rescorla L., Ivanova M.Y., Achenbach T.M., Begovac I., Chahed M., Drugli M.B., Emerich D.R., Fung D.S., Haider M., Hansson K., Hewitt N., Jaimes S., Larsson B., Maggiolini A., Markovic J., Mitrovic D., Moreira P., Oliveira J.T., Olsson M., Ooi Y.P., Petot D., Pisa C., Pomalima R., da Rocha M.M., Rudan V., Sekulic S., Shahini M., de Mattos Silvares E.F., Szirovicza L., Valverde J., Vera L.A., Villa M.C., Viola L., Woo B.S., Zhang E.Y. (2012) International epidemiology of child and adolescent psychopathology II:

Int. J. Methods Psychiatr. Res. 2017;**26**:e1498. DOI: 10.1002/mpr

integration and applications of dimensional findings from 44 societies. *Journal of the American Academy of Child and Adolescent Psychiatry*, **51**(12), 1273–1283. DOI:10.1016/j.jaac.2012.09.012.

Rogler L.H., Mroczek D.K., Fellows M., Loftus S.T. (2001) The neglect of response bias in mental health research. *Journal of Nervous and Mental Disease*, **189**(3), 182–187.

Rosen L.H., Beron K.J., Underwood M.K. (2013) Assessing peer victimization across adolescence: measurement invariance and developmental change. *Psychological Assessment*, **25**(1), 1–11. DOI:10.1037/a0028985.

Schermelleh-Engel K., Moosbrugger H., Müller H. (2003) Evaluating the fit of structural equation models: test of significance and descriptive goodness-of-fit measures. *International Journal of Methods in Psychological Research*, **8**(2), 23–74.

Schlotz W., Yim I.S., Zoccola P.M., Jansen L., Schulz P. (2011) The Perceived Stress Reactivity Scale: measurement invariance, stability, and validity in three countries. *Psychological Assessment*, **23**(1), 80–94. DOI:10.1037/a0021148.

Shear K., Halmi K.A., Widiger T.A., Boyce C. (2007) Sociocultural factors and gender. In Age and Gender Considerations in Psychiatric Diagnosis: A Research Agenda for DSM-V, pp. 65–79, Arlington, VA: American Psychiatric Publishing.

Sourander A., Niemelä S., Santalahti P., Helenius H., Piha J. (2008) Changes in psychiatric problems and service use among 8-year-old children: a 16-year population-based time-trend study. *Journal of the American Academy of Child and Adolescent Psychiatry*, **47**(3), 317–327.

Spinhoven P., Penninx B.W., Hickendorff M., van Hemert A.M., Bernstein D.P., Elzinga B.M. (2014) Childhood trauma questionnaire: factor structure, measurement invariance, and validity across emotional disorders. *Psychological Assessment*, **26**(3), 717–729. DOI:10.1037/pas0000002.

Sterba S.K., Prinstein M.J., Cox M.J. (2007) Trajectories of internalizing problems across childhood: heterogeneity, external validity, and gender differences. *Development and Psychopathology*, **19**(2), 345–366. DOI:10.1017/S0954579407070174.

Trent L.R., Buchanan E., Ebesutani C., Ale C.M., Heiden L., Hight T.L., Damon J.D., Young J. (2013) A measurement invariance examination of the Revised Child Anxiety and Depression Scale in a Southern sample: differential item functioning between African American and Caucasian youth. *Assessment*, **20**(2), 175–187. DOI:10.1177/1073191112450907.

Van der Sluis S., Vinkhuyzen A.A.E., Boomsma D.I., Posthuma D. (2010) Sex differences in adults' motivation to achieve. *Intelligence*, **38**(4), 433–446.

Van Lieshout R.J., Cleverley K., Jenkins J.M., Georgiades K. (2011) Assessing the measurement invariance of the Center for Epidemiologic Studies Depression Scale across immigrant and non-immigrant women in the postpartum period. *Archives of Women's Mental Health*, **14**(5), 413–423. DOI:10.1007/s00737-011-0236-0.

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**, 4–70. DOI:10.1177/109442810031002.

Verhoeven M., Sawyer M.G., Spence S.H. (2013) The factorial invariance of the CES-D during adolescence: are symptom profiles for depression stable across gender and time? *Journal of Adolescence*, **36**(1), 181–190. DOI:10.1016/j.adolescence.2012.10.007.

Verhulst F.C., Van der Ende J. (2013) *Handleiding ASEBA. Vragenlijsten voor leeftijden 6 tot en met 18 jaar*, Rotterdam: ASEBA Nederland.

Verhulst F.C., van der Ende J., Ferdinand R.F., Kasius M.C. (1997) The prevalence of DSM-III-R diagnoses in a national sample of Dutch adolescents. *Archives of General Psychiatry*, **54**(4), 329–336.

Vlassoff C. (1994) Gender inequalities in health in the Third World: uncharted ground. *Social Science & Medicine*, **39**(9), 1249–1259.

Wang M., Armour C., Wu Y., Ren F., Zhu X., Yao S. (2013) Factor structure of the CES-D and measurement invariance across gender in Mainland Chinese adolescents. *Journal of Clinical Psychology*, **69**(9), 966–979. DOI:10.1002/jclp.21978.

Whisman M.A., Judd C.M., Whiteford N.T., Gelhorn H.L. (2013) Measurement invariance of the Beck Depression Inventory – second edition (BDI-II) across gender, race, and ethnicity in college students. *Assessment*, **20**(4), 419–428. DOI:10.1177/1073191112460273.

Widaman K.F., Reise S.P. (1997) Exploring the measurement invariance of psychological instruments: applications in the substance use domain. In Bryant K.J., Windle M.D., West S.G. (eds) The Science of Prevention: Methodological Advances from Alcohol and Substance Abuse Research, pp. 281–324, Washington, DC: American Psychological Association.

Widiger T.A. (2007) DSM's approach to gender: history and controversies. In Age and Gender Considerations in Psychiatric Diagnosis: A Research Agenda for DSM-V, pp. 19–29, Arlington, VA: American Psychiatric Publishing.

Willoughby M.T., Wirth R.J., Blair C.B., Family Life Project I (2012) Executive function in early childhood: longitudinal measurement invariance and developmental change. *Psychological Assessment*, **24**(2), 418–431. DOI:10.1037/a0025779.

Wu P.C. (2010a) Differential functioning of the Chinese version of Beck Depression Inventory – II in adolescent gender groups: use of a multiple-group mean and covariance structure model. *Social Indicators Research*, **96**(3), 535–550. DOI:10.1007/s11205-009-9491-0.

Wu P.C. (2010b) Measurement invariance and latent mean differences of the Beck Depression Inventory II across gender groups. *Journal of Psychoeducational Assessment*, **28**(6), 551–563. DOI:10.1177/0734282909360772.

Wu P.C., Huang T.W. (2014) Gender-related invariance of the Beck Depression Inventory II for Taiwanese adolescent samples. *Assessment*, **21**(2), 218–226. DOI:10.1177/1073191112441243.

Yarnell L.M., Sargeant M.N., Prescott C.A., Tilley J.L., Farver J.A., Mednick S.A., Venables P.H., Raine A., Luczak S.E. (2013) Measurement invariance of internalizing and externalizing behavioral syndrome factors in a non-Western sample. *Assessment*, **20**(5), 642–655. DOI:10.1177/1073191113498114.

Yu C.Y. (2002) Evaluating Cutoff Criteria of Model Fit Indices For Latent Variable Models with Binary and Continuous Outcomes, Los Angeles, CA, University of California. http://www.statmodel.com/download/Yudissertation.pdf

Zimprich D., Mascherek A. (2012) Anger expression in Swiss adolescents: establishing measurement invariance across gender in the AX scales. *Journal of Adolescence*, **35**(4), 1013–1022. DOI:10.1016/j.adolescence.2012.02.008.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.