Conners 3-Parent (Short): Measurement Invariance Across Gender, Concurrent and Discriminant Validities

Rapson Gomez[1]* and Alasdair Vance[2]

1 The School of Health and Life Sciences, Federation University Australia, Victoria, Australia

2 Royal Children's Hospital and The University of Melbourne, Victoria, Australia

*Corresponding author: Rapson Gomez

✉ rapson.gomez@federation.edu.au

The School of Health and Life Sciences, Federation University Australia, University Drive, Mt Helen, PO Box 663, Ballarat, Victoria, 3353, Australia.

Abstract

The study examined measurement invariance (configural, factor loadings, thresholds, and error variances), and equivalencies of latent mean scores of the Conners 3-Parent (Short); (C 3-P (S)) across maternal ratings of clinic-referred boys (*N* = 354) and girls (*N* = 151), aged 7 to 17 years. It also examined the concurrent and discriminant validities of the scores for the C 3-P (S). Confirmatory factor analysis (CFA) indicated support for the theorized six-factor model. For this model, there was support for full measurement invariance and equivalencies for the latent mean scores. There was also support for the concurrent and discriminant validities of the scores for the C 3-P (S) scales. The findings are discussed in relation to the use of the C 3-P (S).

Keywords: Conners 3-Parent (Short); Measurement invariance; Gender; Concurrent and discriminant validities

Introduction

The Conners 3-Parent Short (C 3-P (S)) [1] is used as a quick measure for facilitating the diagnosis of Attention Deficit/Hyperactivity Disorder (ADHD) and the more common disorders [in particular Learning Disorder (LD), Conduct Disorder (CD), and Oppositional Defiant Disorder (ODD)) that are comorbid with ADHD in children between 6 and 18 years of age. The C 3-P (S) has a mixture of content and validity scales. The content scales are inattention (IN, 5 items,), hyperactivity/impulsivity (HY, 6 items), learning problems (LP, 5 items), executive functioning (EF, 5 items), aggression (AG, 5 items) and peer relations (PR, 5 items). The present study extended exiting psychometric data for this measure. It used confirmatory factor analysis (CFA) to examine measurement invariance, and equivalencies of the latent factors mean scores across gender, and how the six scales (factors) in the C 3-P (S) were associated with common DSM-IV externalizing and internalizing childhood disorders.

As reported in the Conners 3 (C 3) manual [1], initial validation of the C 3-P (S) using CFA of the items for only the content scales found support for the theorized six-factor oblique model. According to the C 3 manual, the ratings of the C 3-P (S) items are associated with age, gender and race/ethnicity. For gender, the scores for IA, HY and EF are higher for boys than girls.

Consequently, separate normative scores have been provided for boys and girls. However, when providing these scores, it was not established if there is measurement invariance across ratings for boys and girls. This is a serious omission that could compromise the use of these normative scores, as explained next.

Measurement invariance refers to groups reporting the same observed scores when they have the same level of the underlying trait [2]. Invariance would mean that for the groups being compared, the measure in question is using the same measurement and scaling properties. If there is weak or no support for invariance, then it follows that the groups in question cannot be justifiably compared in terms of observed scores as the same observed scores for the groups do not reflect the same levels of the underlying trait. When applied to the C 3-P (S), the absence of gender measurement invariance would mean that we cannot be confident in the use of normative scores provided in the C 3 manual.

Multiple-group CFA is a powerful method for testing measurement invariance [3]. This procedure can test for configural invariance, metric invariance (equal item factor loadings), scaler invariance (equal item intercepts and thresholds for continuous and categorical responses, respectively), and error variances invariance. Support for configural invariance indicates that the

same number of factors and the same patterns of free and fixed parameters hold across groups. Support for metric invariance indicates that the strength of the relationships between the items and their respective factors are equivalent across groups, and that across the groups, the items are measuring their relevant latent factors using the same metric scales. Support for scalar invariance indicates that for the same levels of the latent trait, individuals across the groups will endorse the same observed scores or response categories. Support for error variances invariance indicates equal uniqueness for like items across the groups compared. Metric, scalar and error variances invariance are alternatively referred to as weak, strong and strict invariance [4]. When there is some support for measurement invariance, equivalency for latent factor mean scores can be examined, taking into account the non-invariance in the measurement model.

There are reasons to suspect that there could be lack of measurement invariance across ratings of boys and girls for some of the C 3-P (S) items. There is evidence that developmentally boys show more externalizing and disruptive behaviors than girls [5]. Thus, it can be speculated that parents would generally conceive and expect externalizing and disruptive behaviors to be associated more with boys than with girls [6]. Such expectations could in turn led parents to be more noticeable of and less tolerant of such behaviours among girls than boys. If so, parents are likely to over-report the same levels of severity of externalizing behaviors in girls. Viewed from a measurement invariable viewpoint, this could mean lack of scalar invariance. The same processes could also bias the reporting of other behaviors that could be perceived by parents to be less characteristic of girls than boys, such as academic (in particular, arithmetic) and cognitive abilities. Thus, considering the content items in the C 3-P (S), it can be speculated that there could be lack of scalar invariance across gender for some of its items, especially those in the HY, LP, EF and AG scales.

Another important psychometric property of a clinical measure is discriminative validity or the ability of the scales to identify clinical disorders that the scales were developed to capture. In terms of the C 3-P (S), this would mean especially the ability of the IN/HY, AG and LD scale scores to distinguish those with ADHD, ODD/CD, and LD, respectively, from other clinical disordered and general population groups (for instance, ability of the IN and HY scales to distinguish those with and without ADHD). Consistent with this, the C 3 manual [1] has reported that for the IN and HY scales, an ADHD group had higher scores than ODD/CD, LD, and general population groups; a LD group scored higher than ADHD, ODD/CD and general population groups for the LP scale; and those with ODD/CD scored higher on the AG scale than ADHD, LD, and general population groups.

Although the IN/HY, LP and AG scales of the C 3-P (S) have been shown to be suitable for specifically identifying individuals with ADHD, LD and ODD/CD, respectively, we wish to argue that the discriminative validity or the scores for the C 3-P (S) scales have yet to be comprehensively evaluated. This is because there are data showing that the C 3-P (S) scales are related to psychological syndromes (a group of signs and symptoms that occur together and characterize a particular abnormality) that are closely associated with internalizing anxiety and mood disorders. For instance, the C 3 manual reports that the internalizing syndromes (anxious/depressed, withdrawn, and somatic complaints) of the Child Behavior Checklist (CBCL) [7] are associated with the C 3-P (S) scales. More specifically, the CBCL scales for anxious/depressed, withdrawn and somatic complaints are associated with the C 3-P (S) LP and PR scales; the CBCL scales for anxious/depressed and withdrawn are associated with the C 3-P (S) IN and EF scales; the CBCL anxious/depressed scale is associated the C 3-P (S) HY scale; and CBCL withdrawn scale is associated with C 3-P (S) AG scale. Given this, and the fact that the comorbidity rates for ADHD with mood and anxiety disorders are relatively high (around 22% to 28% for mood disorders, and around 15% to 18% for anxiety disorders) [8], it can be speculated that the C 3-P (S) scales would also be associated with the internalizing disorders. Examination of such associations could also provide insights on the discriminative validity of the C 3-P (S) scales. For example, if we find associations for the C 3-P (S) IN scale with ADHD, but not with other disorders, then it can be interpreted as supportive of the discriminative validity of the IN scale. On the other hand if we find that the C 3-P (S) IN scale is also associated with anxiety and/or depressive disorders, then support for the discriminative validity of the IN scale is diminished. To date this has not been explored.

Given existing limitations and omissions, based on ratings of the C 3-P (S) provided by mothers for large groups of clinic-referred boys and girls, the first aim of the current study was to apply the multiple-group CFA approach to the six-factor oblique model to examine measurement invariance across the gender groups. Related to this aim, we also examined the equivalencies of the latent factor mean scores, taking into consideration non-invariance in the measurement model. The second aim of the study was to examine the concurrent and discriminant validities of the C 3-P (S) scales in terms of their relationships with a range of both DSM-IV externalizing (ADHD, ODD and CD) and internalizing disorders [separation anxiety disorder (SAD), social phobia (SOP), specific phobia (SPP), panic disorder (PD), agoraphobia (AG), generalized anxiety disorder (GAD), obsessive compulsive disorder (OCD), post-traumatic stress disorder (PTSD), dysthymia (DYTH), and major depressive disorder (MDD)]. Based on existing findings and the arguments present earlier (p. 4, paragraph 3) we expected lack of scalar measurement invariance for some of the items in the HY, LP, EF and AG scales. We also expected higher mean scores for boys for the IA, HY and EF latent factors; and stronger associations for C 3-P (S) scales with the externalizing disorders than the internalizing disorders.

Method

Participants

The data for all participants were collected archivally from the Academic Child Psychiatry Unit (ACPU) of the Royal Children's Hospital, Melbourne, Australia. The ACPU is an out-patient psychiatric unit that provides services for children and adolescents with behavioral, emotional, and learning problems. Referrals are generally from other medical services, schools, and

social and welfare organizations. All parents and children were informed that the clinic would provide diagnosis and appropriate treatment, and that assessment will be over two days, covering a range of tests involving the parents, children/adolescents and their teachers. They were informed that all data collected would be kept in an unidentifiable form in a secure database and (if consent was given) used to support future research.

For the current study we used the records of children and adolescents, aged between 6 and 17 years, referred between 2004 and 2017, who had been interviewed for clinical diagnosis. An individual was selected for inclusion in the study if that individual had ratings for the C 3-P (S), completed by his/her mother. Apart from this and the age criteria, no other inclusion/exclusion criterion was applied when selecting participants for the study. In all, there were 505 children and adolescents, comprising 354 (70.1%) boys and 151 (29.95%) girls. The overall mean age of participants was 11.52 years (SD=3.35 years).

Given that measurement invariance was examined for ratings across boys and girls, we initially tested if these groups were equivalent for age and a range of background and demographic information variables, and the percentages of different disorders. As chi-square values are highly sensitive to sample size, the α value was set at .01 to allow for more stringent Type II error control. The mean age (SD) of boys and girls were 11.24 years (SD=3.20 years) and 12.16 years (SD=3.60 years). Although, girls were significantly older than boys, t (503)=2.83, $p<0.01$, the Cohen's d effect size value for the age difference were small at 0.28 [based on Cohen's [9] guidelines for d effect size: small < =0.20, medium > =0.50, and large > =0.80]. **Table 1** presents the results of the comparisons between boys and girls for other background measures. As shown in the table, the frequencies for mother and father employment and educational levels, family income and parental relationships status for different categories showed no group difference. On the whole, most fathers and mothers of participants were employed, and more than two-thirds of participants had fathers and mothers who had attended at least secondary school. In terms of parental relationships, close to 50% were living together and the other 50% were separated or divorced. Slightly more than half the number of participants were from families with income less than $50,000 per year. Apart from GAD (higher frequency among girls) and ODD (higher frequency among boys), there was no difference across gender for the other disorders. Although the groups were not matched for frequencies of GAD and ODD, the phi (equivalent to correlation) values for the differences for GAD and ODD were of small effect size at 0.14 [using guidelines for equivalent d effect size values proposed by Cohen's [9]; small r ≥ 0.10, medium r ≥ 0.24, and large r ≥ 0.37). Thus, although the gender groups differed for age and frequencies of GAD and ODD, these differences were of little importance, and thus the gender groups in the study can be considered sufficiently matched for age, background demographic variables, and clinical disorders.

Measures

The measures included in this study were the parent version of the Anxiety Disorders Interview Schedule for Children (ADISC-IV)

Table 1: Demographics information for boys and girls.

	Boy	Girl	Chi-square (df)
Number	354	151	
Mother Employment (Percentage)			
Employed	44.4	50.3	11.86 (4)
Home duties	39.9	27.2	
Pensioner	5.3	11.6	
Unemployment	3.2	4.8	
Others	7.3	6.1	
Mother Education (Percentage)			
Primary	.03	1.4	5.07 (4)
Some secondary	28.4	34.9	
Completed secondary	14.9	11.9	
Technical	25.4	21.9	
Tertiary	31.0	20.9	
Father Employment (Percentage)			
Employed	79.1	65.9	11.86 (4)
Home duties	2.2	3.0	
Pensioner	4.4	9.1	
Unemployment	6.9	14.4	
Others	7.5	7.5	
Father Education (Percentage)			
Primary	1.6	3.9	2.65 (4)
Some secondary	35.6	35.7	
Completed secondary	10.1	9.3	
Technical	25.6	22.7	
Tertiary	27.1	28.7	
Family income (Percentage)			
0 - < $30,000	33.3	38.5	6.64 (4)
$30,000 - < $40,000	12.9	8.4	
$40,000 - <$50,000	7.8	11.2	
>$50,000	45.9	4.3	
Parental relationship (Percentage)			
Living together	53.6	44.9	3.46 (4)
Separated	27.2	30.0	
Divorced	13.3	14.3	
Death of one parent	2.6	3.4	
Other	3.2	3.4	
Disorders (Percentage)			
Separation Anxiety	24.7	26.8	0.25 (1)
Social Phobia	42.7	50.3	2.44 (1)
Specific Phobia	30.7	40.9	4.92 (1)
Panic	11.9	17.4	2.72 (1)
Agoraphobia	8.2	12.8	2.46 (1)
Generalized Anxiety	44.0	59.5	9.92 (1)**
Obsessive Compulsive	21.9	24.2	0.31 (1)
Post-Traumatic Stress	14.2	22.1	4.72 (1)
Dysthymic	37.8	33.87	0.72 (1)
Major Depressive	21.4	20.3	0.75 (1)
Conduct	46.3	43.2	0.39 (1)
Opposition Defiant	71.9	57.7	9.58 (1)**
Attention Deficit Hyperactivity	82.1	73.8	4.42 (1)

**$p<0.01$.

[10] that was used for clinical diagnosis, and the C 3-P (S) [1]. The C 3-P (S) was not used for facilitating diagnosis.

Anxiety Disorders Interview Schedule for Children, Parent Version (ADISC-IV-P): The ADISC-IV-P was used for diagnosis [10]. The ADISC-IV-P is a semi-structured interview, based on the DSM-IV-TR diagnostic system (American Psychiatric Association, 2000). It has been designed to facilitate the diagnosis of major childhood disorders. The ADISC-IV-P guidelines for diagnosis are that the child/adolescent be given a diagnosis of all disorders meeting the diagnostic criteria, and not in terms of primary and secondary disorders. Thus, all disorders that an individual qualified for were seen as equally applicable to that individual. The scores of ADISC-IV-P have sound psychometric properties [11]. Test-retest reliabilities for the ADISC-IV-P scores over a 7 to 14-day interval have shown good to excellent reliabilities. Kappa values for interview with children between 7 and 16 years ranged from 0.61-0.80 [11].

Conners 3rd Edition-Parent Short (C 3-P (S)): As the C 3-P (S) was described comprehensively in the introduction, this section will only provide additional information not provided in the introduction [1]. For this measure, respondents indicate the degree or frequency of each behavior described in the item on a scale of 0 (not true at all), 1 (just a little true true), or 2 (pretty much true), or 3 (very much true). The rating period is 1 month. For the sample in the current study, the Cronbach's alpha values for the IN, HY, LP, EF, AG, and PR scales were 0.91, 0.90, 0.80, 0.82, 0.89, and 0.88, respectively. All these values are well above 0.70 that is generally considered the minimum level for acceptable internal consistency reliability [12].

Procedure

Children and parents participated in separate interviews and testing sessions with breaks over two days. Information was also obtained from teachers using various checklists and questionnaires. In all cases, parental and child consent forms were completed prior to the assessment. The consent forms from both parents and children gave permission for all relevant data collected by the ACPU of the RCH or provided by others to be used in future research and was approved by the RCH ethics committee as part of the ACPU's comprehensive examination of psychopathology in children and adolescents. The data collected covered a comprehensive demographic, medical (primarily neurological and endocrinological), educational, psychological, familial, and social assessment of the child and his or her family. All psychological data were collected by research assistants, who were advanced doctoral students in clinical psychology, and under the supervision of two registered clinical psychologists.

The research assistants were provided with extensive supervised training and practice by the two psychologists prior to them collecting data. Training of the ADISC-IV-P included observations of it being administered by the psychologists. The research assistants commenced administering the ADISC-IV-P only after they attained competence in its administration, as assessed by the two registered psychologists. There was adequate inter-rater reliability for the diagnoses made between the research assistants

and the psychologists, and between research assistants (average kappa value across all diagnoses =0.88).

Standard procedures were used for the administration of all measures. Approximately 85% of the parent ADISC-IV-P interviews involved mothers only, and the rest involved fathers only or both fathers and mothers together. Using the categorical data from the parent ADISC-IV-P, clinical diagnosis was determined by two consultant child and adolescent psychiatrists who independently reviewed the data. The inter-rater reliability for diagnoses of the two psychiatrists was high (kappa =0.90).

Statistical procedures

All the CFA models in the study were conducted using Mplus (Version 7) software [13]. As there are four order response categories for all the C 3-P (S) items, we used the mean and variance-adjusted weighted least squares (WLSMV) extraction for all the CFA analyses [14]. Multiple-group CFA measurement invariance was tested using the procedure proposed by Millsap and Yun-Tein [15] for the WLSMV estimator with theta parameterization. Details of this procedure are not provided here because of word limitation. For details, the reader is referred to Millsap and Yun-Tein [15].

The goodness-of-fit of the CFA models was examined using WLSMVχ^2. Like all other χ^2 values, WLSMVχ^2 values are inflated by large sample sizes. In addition to the WLSMVχ^2, the fit of the models was examined using the approximate fit values of root-mean-square error of approximation (RMSEA), the comparative fit index (CFI), and Tucker-Lewis Index (TLI). The guidelines suggested by Hu and Bentler [16] are that RMSEA values of 0.06 or below be taken as good fit, values >0.06 to 0.08 be considered moderate fit, values >0.08 to 0.10 be considered marginal fit, and values >0.10 be considered poor fit. For the CFI and TLI, values of 0.95 or above are taken as indicating good model-data fit, values of >0.90 and <0.95 are taken as acceptable fit, and values less than 0.90 as poor fit. Despite the widespread use of these values, it is worth noting that a simulation study by Nye and Drasgow [17] concluded that appropriate cut-off values for WLSMV estimation can vary across conditions. For measurement invariance, the difference between models can be tested using ΔWLSMVχ^2 values. However, as $\Delta\chi^2$ values (including ΔWLSMVχ^2 values) are also highly sensitive to large sample sizes, researchers have also used difference in approximate fit indices. Based on simulation studies involving maximum likelihood estimation, it has been proposed that ΔCFI >0.01 and ΔRMSEA > -0.015 can be interpreted as lack of support for invariance [18,19]. A recent study by Sass, Schmitt, and Marsh [20] concluded that although these values could be used when WLSMV estimation is applied, there is need for caution, especially with misspecified models. Given these concerns, we examined measurement invariance using both difference in approximate fit indices (ΔRMSEA > -0.015 and ΔCFI >0.0) and ΔWLSMVχ^2 values. For the latter, the α value was set at 0.01 to allow for more stringent Type II error control.

To examine the concurrent and discriminant validities of the C 3-P (S) scales, these scales were correlated with clinical diagnoses as established by the ADISC-IV-P. The sizes of the correlations were interpreted using guidelines for equivalent d effect size values proposed by Cohen's [9] (small r ≥ 0.10, medium r ≥ 0.24, and large r ≥ 0.37).

Results

Missing data

Out of a total of 15,655 scores for the C 3-P (S) (31 items x 505 participants), there were 308 scores missing (i.e., around 2%). For WLSMV estimator, M*plus* uses pairwise deletion (i.e. includes everybody who answers both items in an item pair to estimate the covariance for that pair) to deal with missing values.

Goodness-of-fit for the single groups and reliabilities of the factors for the C 3-P (S)

The fit values for the six-factor oblique model for boys were WLSMVχ^2 (df = 419) = 966.84, p <0.001; RMSEA =0.061 (90% confidence interval =0.056 to 0.066); CFI=0.958; and TLI=0.953. The values for girls were WLSMVχ^2 (df=419) = 501.59, p<0.001; RMSEA=0.036 (90% confidence interval =0.022 to 0.048); CFI=0.988; and TLI=0.986. Thus, the CFI, TLI and RMSEA values indicated good fit for both boys and girls.

Although the Cronbach's alphas indicated support for internal consistency reliabilities for the C 3-P (S) factors, within a CFA measurement model, more desirable measures of internal consistency reliabilities or more specifically convergent validity, are composite reliability (CR) and average variance extracted (AVE) [21]. The CR estimates the extent to which a set of latent construct indicators share in their measurement of a construct, while the AVE is the amount of common variance among latent construct indicators [22]. Fornell and Larcker [21] have also proposed that for a CFA measurement model, the discriminant validity of the constructs can be examined by comparing the square root of the AVE of a construct with its correlations with other constructs in the model. According to Hair et al., CR 0.70 or more, and/or AVE 0.50 or more are supportive of convergent validity, and if the square root of the AVE of a construct is higher than its correlations with other constructs then discriminant validity for that construct can be assumed. Given this, these values were also computed, based on the standardized factor loadings of the proposed 6-factor oblique model for the C 3-P (S). The fit indices for this model that involved both boys and girls together were WLSMVχ^2 (df=419) = 1043.12, p<0.001, CFI=0.969, TLI=0.966, and RMSEA=0.054 (90 CI=0.052 to 0.062). The CFI, TLI and RMSEA values indicate good fit. **Table 2** shows the range of factor loadings with each of the six factors, and the CR, the AVE, and the square soot of the AVE for the different factors, and correlations between six latent factors in the model. As shown, the CR and AVE for each of the construct were all above 0.70 and 0.50, respectively, thereby indicating support for their convergent validities. Also, for each of the construct, the square root of its AVE was higher than the correlations for the construct with other constructs, thereby supporting the discriminant validities of all constructs. For indicators/items, loadings of 0.70 or more can be interpreted as acceptable reliability [22]. Only one item ("Forgets to turn in completed work" – an item belonging to the executive functioning scale) had a loading of less than 0.70. The value was however close to 0.70 at 0.68.

Multiple group CFA analyses for measurement invariance across boys and girls, based on difference in χ^2 test

Table 3 shows the results of the analyses for invariance testing, based on difference in χ^2 test. As shown, the RMSEA, CFI and TLI values for the configural invariance model (M1 in **Table 3**) indicated good fit for configural invariance. **Table 3** shows that there was a difference between the configural invariance model and the full metric invariance model (M2 in **Table 3**), ΔWLSMVχ^2 (df=25) = 52.82, p<0.001. Further analyses showed lack of invariance for factor loadings of item numbers 15, 35 and 4 (M2.3 in **Table 3**). There was also a difference between the final partial metric invariance model (M2.3 in **Table 3**) and the full scalar invariance model (M3 in **Table 3**), ΔWLSMVχ^2 (df=50) = 101.08, p<0.001. Further analyses showed lack of invariance for the following thresholds: threshold number 1 of item number 8 (8\$1), threshold number 1 of item number 1 (1\$1), threshold number 1 of item number 34 (34\$1), and threshold number 3 of item number 1 (1\$3) (M3.4 in **Table 3**). There was no difference between the final partial scalar invariance model (M3.4 in **Table 3**) and the full error variances invariance model (M4 in **Table 3**), ΔWLSMVχ^2 (df=31), ns. Thus, there was support for the full error variances invariance model. These findings indicate support for only partial invariance for the measurement model.

Given support for at least partial invariance for the measurement

Table 2 Range of Factor Loadings, Composite Reliability (CR), Average Variance Extracted (AVE), Square Root of the AVE (√AVE) and Correlations between Scales.

Scales	Loadings	Reliability		√AVE (bold)/correlation (off-diagonal)					
	Range	CR	AVE	1	2	3	4	5	6
Inattention (1)	0.92-0.79	0.94	0.77	**0.88**					
Hyperactivity/impulsivity (2)	0.91-0.78	0.94	0.73	0.72	**0.85**				
Learning problems (3)	0.70-0.89	0.89	0.62	0.73	0.43	**0.79**			
Executive functioning (4)	0.68-0.82	0.87	0.57	0.82	0.54	0.62	**0.87**		
Aggression (5)	0.72-0.91	0.93	0.72	0.42	0.58	0.25	0.54	**0.85**	
Peer relations (6)	0.72-0.91	0.91	0.72	0.39	0.37	0.35	0.43	0.45	**0.82**

model, further analysis was conducted for equivalency in latent mean scores. As shown in **Table 1**, this analysis showed no support for equivalency for the factor mean scores model (M5 in **Table 3**), as this model differed from the final partial scalar invariance model (M3.4 in **Table 3**), ΔWLSMVχ² (*df*=6) = 52.74, *p*<0.001. Additional analyses showed that for the criteria used here (*p*<0.01) the groups differed for all latent factors (inattention, hyperactivity/impulsivity, learning problems,

executive functioning, and aggression), except peer relations **(Table 3)**.

Table 4 shows the unstandardized estimates for boys and girls for the non-invariant parameters. As shown in **Table 4**, girls had higher factor loadings for s EF item number 15 and PR item number 4, while boys had higher loading for EF item number 35. For all four the non-invariant thresholds (threshold number 1 of LP item number 8, threshold number 1 of EF item number 1,

Table 3 Results of tests for invariance across boys and girls based on differences in χ² values.

Models (M)	χ²	*844*	Model Fit RMSEA (90% CI)	CFI	TLI	Model Difference ΔM	Δ *df*	Δχ²
M1: Configural invariance	1449.46***	844	0.053 (0.049-0.058)	0.968	0.965	-	-	-
M2: Metric invariance	1486.04***	869	0.053 (0.48-0.058)	967	0.965	M2-M1	25	52.82***
M2.1. M2 with loadings for item 15 free	1481.26***	868	0.053 (0.48-0.057)	968	0.965	M2.1-M1	24	46.58**
M2.2. M2 with loadings for items 15 & 35 free	1479.17***	867	0.053 (0.48-0.057)	968	0.965	M2.2-M1	23	43.31**
M2.3. M2 with loadings for items 15, 35 & 4 free	1473.67***	866	0.053 (0.48-0.057)	0.968	0.965	M2.3-M1	22	35.42
M3: Scalar invariance :	1545.10***	916	0.052 (0.048-0.057)	0.967	0.966	M3-M2.3	50	101.08***
M3.1. M3 with threshold 8$1 free	1537.44***	915	0.052 (0.047-0.056)	0.967	0.966	M3.1-M2.3	49	89.45***
M3.2. M3 with thresholds 8$1 & 1$1 free	1531.63***	914	0.052 (0.047-0.056)	0.967	0.967	M3.2-M2.3	48	81.55**
M3.3. M3 with thresholds 8$1, 1$1 & 34$1 free	1528.14***	913	0.052 (0.047-0.056)	0.967	0.967	M3.3-M2.3	47	75.93**
M3.4. M3 with thresholds 8$1, 1$1, 34$1 & 1$3 free	1522.46***	912	0.051 (0.047-0.056)	0.968	0.967	M3.4-M2.3	46	65.36
M4. M3.4 with all error variances constrained equal	1507.63***	943	0.049(0.044-0.0540	0.970	0.970	M4 – M3.4	31	46.01
M5: Invariance for the means of the latent factors	1711.88***	918	0.059 (0.054-0.063)	0.958	0.957	M5-M3.4	6	52.74***
M5.1: Invariance for the mean of Factor for IA	1555.90***	913	0.053 (0.048-0.057)	0.965	0.965	M5.1-M3.4	1	11.99***
M5.2: Invariance for the mean of Factor for HY	1657.54***	913	0.057 (0.052-0.061)	0.961	0.960	M5.2-M3.4	1	25.97***
M5.3: Invariance for the mean of Factor for LP	1537.68***	913	0.052 (0.048-0.057)	0.967	0.966	M5.3-M3.4	1	7.03**
M7.4: Invariance for the mean of Factor for EF	1539.12***	913	0.052 (0.048-0.057)	0.967	0.966	M5.4-M3.4	1	7.67**
M5.5: Invariance for the mean of Factor for AG	1167.38***	913	0.053 (0.049-0.058)	0.966	0.965	M5.5-M3.4	1	12.24***
M5.6: Invariance for the mean of Factor for PR	1058.04***	913	0.052 (0.047-0.056)	0.967	0.967	M5.6-M3.4	1	4.95

Note: χ²= weighted least square with mean and variance adjusted chi-square (WLSMVχ²), RMSEA= root mean square error of approximation; CFI= comparative fit index, TLI = Tucker-Lewis Index. All WLSMVχ² values were significant (p<0.001).
**p < .01, ***p < .001.

Table 4 Comparisons of unstandardized estimates for the non-invariant parameters across gender.

Parameter	Boy	Girl
Factor Loading		
Item 15 (trouble getting started on tasks or projects)	1.13 (0.18)	1.35 (0.23)
Item 35 (messy or disorganized)	1.37 (0.22)	1.02 (0.18)
Item 4 (last to be picked for teams/games)	1.00 (0.00)	1.44 (0.27)
Threshold		
8$1 (cannot grasp arithmetic.)	-0.59 (0.11)	-1.41 (0.31)
1$1 (forgets to turn in completed work)	-0.51 (0.38)	0.03 (0.15)
34$1 (inattentive, easily distracted)	-4.16 (0.38)	-3.25 (0.40)
1$3 (Forgets to turn in completed work)	0.38 (0.09)	1.02 (0.27)
Latent Mean		
Inattention	0.00 (0.00)	-0.50**(0.15)
Hyperactivity/Impulsivity	0.00 (0.00)	-0.77***(0.16)
Learning Problems	0.00 (0.00)	-0.35 (0.15)
Executive Functioning	0.00 (0.00)	-0.33**(0.12)
Aggression	0.00 (0.00)	-0.77**(0.25)
Peer Relations	0.00 (0.00)	0.20(0.109)

Note: Values in parenthesis are standard errors.

threshold number 1 of IN item number 34, and threshold number 3 of EF item number 1), girls had higher scores. Also, for all the non-invariant latent mean scores, girls had lower scores.

Multiple group CFA analyses for measurement invariance across boys and girls, based on the ΔCFI and ΔRMSEA values

Table 5 shows the results of the analyses for invariance testing, based on difference in CFI and RMSEA values. As shown, the RMSEA, CFI and TLI values for the configural invariance model (M1 in **Table 5**) indicated good fit for configural invariance. As will be recalled, decrease of ≥ 0.01 in CFI values *and* an increase of ≥ 0.015 or the RMSEA values were interpreted as indicative of non-invariance. As shown in the table, there was no difference between the configural model (M1 in **Table 5**) and the full metric invariance model (M2 in **Table 5**); the full metric invariance model and the full scalar invariance model (M3 in **Table 5**), and the full scalar invariance model and the full error variances invariance model (M4 in **Table 5**). Also, there was also no difference between the full scalar invariance model and the equivalency for latent mean scores model (M5 in **Table 5**). Thus, the findings, based on the ΔCFA and ΔRMSEA values indicated good support for full measurement invariance (metric, scalar and error variances

invariance) for all C 3-P (S) items, and equivalencies for latent factor mean scores for the C 3-P (S) across the gender groups.

Correlations of the C 3-P (S) factors with DSM-IV childhood disorders

Table 6 shows the correlations of the C 3-P (S) factors with the externalizing and internalizing disorders, derived *via* ADISC-IV-P. As shown in the table, SAD, SOP, SPP, PD, AG, and PTSD showed no statistically significant associations with any of the C 3-P (S) factors (based on the criteria used for interpreting statistically significant, *p*<0.01). GAD correlated significantly and negatively with only IN. OCD correlated significantly and positively with only PR. DYTH correlated significantly and positively with EF and AG, and MDD correlated significantly and positively with AG. ADHD and CD correlated significantly and positively with all the C 3-P (S) factors, and ODD correlated significantly and positively with all the C 3-P (S) factors, except LP. In terms of effect sizes, based on Cohen's [9] guidelines for interpreting r effect sizes (small ≥ 0.10, medium ≥ 0.24, and large ≥ 0.37), the correlations for IN. HY, and AG with ADHD were of large effect sizes, and the correlations of LP and AG with ADHD were of medium effect sizes. The correlations for CD and ODD with AG were both of large effect sizes. The correlations of HY with ODD and CD were of medium effect sizes,

Table 5 Results of Tests for Invariance across Boys and Girls Based on Differences in RMSEA and CFI Values.

Models (M)	χ^2	*df*	Model Fit			Model Difference		
			RMSEA (90% CI)	CFI	TLI	ΔM	ΔRMSEA	ΔCFI
M1: Configural invariance	1449.46***	844	0.053 (0.049-0.058)	0.968	0.965	-	-	-
M2: Metric invariance	1486.04***	869	0.053 (0.48-0.058)	0.967	0.965	M2-M1	0.000	-0.001
M3: Thresholds invariance	1553.81***	919	0.052 (0.048-0.057)	0.966	0.966	M3-M2	-0.001	-0.001
M4. Invariance for error variances	1524.98***	934	0.050 (0.045-0.055)	0.969	0.969	M4-M3	-0.002	0.003
M5: Invariance for the means of the latent factors	1748.20***	925	0.059 (0.055-.064)	0.956	0.956	M5-M3	0.007	0.010

Note: χ^2= weighted least square with mean and variance adjusted chi-square (WLSMVχ^2), RMSEA= root mean square error of approximation; CFI= comparative fit index, TLI = Tucker-Lewis Index. All WLSMVχ^2 values were significant (p<0.001).

Table 6 Correlations of the C 3-P (S) factors with DSM-IV childhood disorders derived *via* ADISC-IV-P.

Disorder	C 3-P (S) Factors					
	IN	HY	LP	EF	AG	PR
Separation Anxiety (SAD)	0.02	0.00	0.06	-0.02	0.01	0.09
Social Phobia (SOP)	-0.09	-0.09	-0.04	0.00	0.06	0.11
Specific Phobia (SPP)	0.06	0.06	0.11	0.04	0.03	0.07
Panic (PD)	-0.05	-0.04	-0.09	0.00	-0.05	0.03
Agoraphobia (AG)	-0.05	0.01	0.02	-0.04	0.03	0.10
Generalized Anxiety (GAD)	-0.14**	-0.07	-0.04	-0.02	-0.11	0.05
Obsessive Compulsive (OCD)	0.05	0.07	0.10	0.03	0.05	0.19***
Post-Traumatic Stress (PTSD)	0.00	0.00	0.03	0.04	0.06	0.08
Dysthymic (DYST)	0.06	0.00	0.00	0.13**	0.22***	0.08
Major Depressive (MDD)	-0.03	-0.03	-0.05	0.11	0.16**	0.08
Conduct (CD)	0.15**	0.28***	0.14**	0.16**	0.55***	0.19***
Opposition Defiant (ODD)	0.18***	0.29***	0.09	0.26***	0.59***	0.26***
Attention Deficit Hyperactivity ADHD)	0.49***	0.41***	0.33***	0.39***	0.30***	0.21***

Note: IN = intention, HY = hyperactivity/impulsivity, LP = learning problems, EF = executive functioning, AG = aggression, PR = peer problems. **p<0.01, ***p<0.001.

and correlations of EF and PR with ODD were of medium effect sizes. All other statistically significant correlations were of small effect sizes.

Discussion

Consistent with the findings reported in the C3 manual [1], our findings indicated good fit for the proposed oblique six-factor model for the C 3-P (S) for boys and girls. For this model, our findings showed support for partial measurement invariance (metric, and scalar), based on the difference in χ^2 test. There was support for full measurement invariance for all error variances. Also, all but PR, showed differences across the sex for latent mean scores. More specifically, girls had higher factor loadings for EF item number 15 ("trouble getting started on tasks or projects") and PR item number 4 ("last to be picked for teams/games"), while boys had higher loading for EF item number 35 ("messy or disorganized"). For all four non-invariant thresholds (threshold number 1 of LP item number 8 ("cannot grasp arithmetic"), threshold number 1 of EF item number 1 ("forgets to turn in completed work"), threshold number 1 of IN item number 34 ("inattentive, easily distracted"), and threshold number 3 of EF item number 1 ("forgets to turn in completed work"), girls had higher scores. Also, for all the non-invariant latent mean scores, girls had lower scores. In contrast to the findings based on the difference in χ^2 test, we found support for strict full measurement invariance (configural, metric, scalar, and error variances), and equivalencies across the gender groups for mean scores for all six latent factors, based on the difference in CFI and RMSEA values. In terms of concurrent and discriminant validities of the C 3-P (S) scales, our findings showed no significant associations for SAD, SOP, SPP, PD, AG, and PTSD with any of the C 3-P (S) scales. Although GAD, OCD, DYTH and MDD correlated significantly with one or more of the C 3-P (S) scales, the magnitudes of these correlation were of small effect sizes. In contrast, apart from the correlation involving ODD and LP (that was not significant), the correlations involving all the other scales with ADHD, ODD and CD were significant. In terms of effect sizes, based on Cohen's [9] guidelines for interpreting r effect sizes (small ≥ 0.10, medium ≥ 0.24, and large ≥ 0.37), the correlations for IN. HY, and AG with ADHD were of large effect sizes, and the correlations of LP and AG with ADHD were of medium effect sizes. The correlations for CD and ODD with AG were both of large effect sizes. The correlations of HY with ODD and CD were of medium effect sizes, and correlations of EF and PR with ODD were of medium effect sizes. Taken together, these findings can be interpreted as supporting the concurrent and discriminant validities of the C 3-P (S) scales.

Our findings have implications for the use of the C 3-P (S). First, the support for the six-factor model is consistent with the model recommended for the C 3-P (S) in the C3 manual [1]. Additionally, with the exception of one item ("Forgets to turn in completed work" - an item belonging to the executive functioning scale), there was support for the item reliabilities of all the other 30 items. Even the exceptional item had a factor loading of 0.68 which is close to 0.70, used as the cut-off for acceptable reliability. Furthermore, there was support for the convergent validities of

the items within the six constructs in terms of their CR and AVE values, and also the discriminating validity of the six constructs as the square root of their AVE values were all higher than their correlations with other constructs. Thus, the six-factor model can be seen as a robust and valuable model for research and clinical applications. Second, the support for full measurement invariance based on the difference in CFI and RMSEA values indicate that at the practical level, the C 3-P (S) has the same measurement and scaling properties when applied to parent ratings of boys and girls, and that these groups can be directly and justifiably compared in terms of observed scores. The support for only partial measurement invariance (metric, and scalar), suggests that girls had higher factor loadings for EF item number 15 ("trouble getting started on tasks or projects") and PR item number 4 ("last to be picked for teams/games"), while boys had higher loading for EF item number 35 ("messy or disorganized"). Also, girls had higher scores for four thresholds [threshold number 1 of LP item number 8 ("cannot grasp arithmetic"), threshold number 1 of EF item number 1 ("forgets to turn in completed work"), threshold number 1 of IN item number 34 ("inattentive, easily distracted"), and threshold number 3 of EF item number 1 ("forgets to turn in completed work"). Also, boys had higher latent scores for all but the PR factors. These findings suggest that the normative scores for boys and girls provided in the C 3 manual [1] are confounded by differences in measurement and scaling properties for these groups, and they cannot be used confidently for interpretation. However, it is possible that as our findings did not control for age and ethnicity, our findings may be confounded by these factors as they have been shown to be associated with parent ratings of the C 3-P (S). Thus, we recommend that clinicians exercise a lot of caution when using the different normative scores for the gender groups that are provided in the C 3 manual [1]. Our discriminant validity findings suggest that all the C 3-P (S) scales were either unrelated or only weakly related to DSM-IV internalizing disorders (SAD, SOP, SPP, PD, AG, PTSD, GAD, OCD, DYTH and MDD). In contrast, for DSM-IV externalizing disorders, the correlations for IN. HY, and AG with ADHD were of large effect sizes, and the correlations of LP and AG with ADHD were of medium effect sizes. The correlations for CD and ODD with AG were both of large effect sizes. The correlations of HY with ODD and CD were of medium effect sizes, and correlations of EF and PR with ODD were of medium effect sizes. Taken together, these findings can be interpreted as supporting the concurrent and discriminant validities of the C 3-P (S) scales.

In concluding, the findings and interpretations made in the study needs to be viewed with some limitations in mind. First, since age, ethnicity and socioeconomic status were not controlled in the current study, it is possible that the findings here may be confounded by these variables. Second, the findings reported here are based on one sample, on archival data, using diagnostic determinations based on the ADISC IV. Thus, our findings may not be generalized and warrant further investigation and cross-validation on other well-diagnosed samples before they can be used with confidence. Third, as all the participants in this study were from the same clinic, it is possible that this may constitute an additional bias for the sample examined. Fourth, as this study

used a clinic sample, the findings here may not be applicable to the general community. Fifth, as the sample examined was highly heterogeneous and comorbid for a range of disorders, these may have confounded findings. Sixth, as this study was based on DSM-IV diagnoses, the relevance of the findings for DSM-5 is not directly clear. Given the limitations highlighted here, our findings may not be generalized. There is certainly a need for further investigation and cross-validation on other well-diagnosed samples before they can be used with confidence. Our findings indicate this would be worthy of future research.

References

1 Conners CK (2008) Conners rating scales (3rd Edn). Toronto, ON, Canada: Multi-Health Systems.

2 Reise SP, Widaman KF, Paugh PH (1993) Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. Psychol Bull 114: 552-566.

3 Vandenberg RJ, Lance CE (2000) A review and synthesis of the management invariance literature: suggestions, practices, and recommendations for organizational research. Organ Res Methods 3: 4-69.

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

5 Loeber R, Capaldi DM, Costello E (2013) Gender and the development of aggression, disruptive behavior, and early delinquency from childhood to early adulthood. Disruptive Behavior Disorders 1: 137-160.

6 Endendijk JJ, Groeneveld MG, Bakermans-Kranenburg MJ, Mesman J (2016) Gender-differentiated parenting revisited: meta-analysis reveals very few differences in parental control of boys and girls. PLoS One 11: e0159193.

7 Achenbach TM, Rescorla LA (2001) Manual for the ASEBA school-age forms & profiles. Burlington, VT: University of Vermont, Research Center for Children, Youth, and Families, p: 238.

8 Taurines R, Schmitt J, Renner T, Conner AC, Warnke A, et al. (2010) Developmental comorbidity in attention-deficit/hyperactivity disorder. Atten Defic Hyperact Disord 2: 267-289.

9 Cohen J (1992) A power primer. Psychol Bull 112: 155-159.

10 Silverman WK, Albano AM (1996) Manual for the ADIS-IV C/P. New York: Psychological Corporation.

11 Silverman WK, Saavedra LM, Pina AA (2001) Test–retest reliability of anxiety symptoms and diagnoses with the anxiety disorders interview schedule for DSM-IV: child and parent versions. J Am Acad Child Adolesc Psychiatry 40: 937-944.

12 Nunnaly J (1978) Psychometric theory. New York: McGraw-Hill.

13 Muthen LK, Muthen BO (2013) Mplus user's guide (7th Edn). Los Angeles, CA: Muthen & Muthen.

14 Rhemtulla M, Brosseau-Liard PÉ, Savalei V (2012) When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. Psychol Methods 17: 354-373.

15 Millsap RE, Yun-Tein J (2004) Assessing factorial invariance in ordered-categorical measures. Multivariate Behav Res 39: 479-511.

16 Hu LT, Bentler PM (1998) Fit indices in covariance structure modeling: sensitivity to under parameterized model misspecification. Psychol Methods 3: 424-453.

17 Nye CD, Drasgow F (2011) Effect size indices for analyses of measurement equivalence: understanding the practical importance of differences between groups. J Appl Psychol 96: 966-980.

18 Chen FF (2007) Sensitivity of goodness-of-fit indexes to measurement invariance. Struct Equ Model 14: 464-504.

19 Cheung GW, Rensvold RB (2002) Evaluating goodness-of-fit indexes for testing measurement invariance. Struct Equ Model 9: 233-255.

20 Sass DA, Schmitt TA, Marsh HW (2014) Evaluating model fit with ordered categorical data within a measurement invariance framework: a comparison of estimators. Struct Equ Model 21: 167-180.

21 Fornell C, Larcker DF (1981) Evaluating structural equation models with unobservable variables and measurement error. J Mark Res 1: 39-50.

22 Hair JF, Black WC, Babin BJ, Anderson RE (2010) Multivariate data analysis (7th Edn). New Jersey: Prentice Hall, p: 729.