Measurement of callous-unemotional traits in very young children: a psychometric
and validity study from 2.5 to 5.0 years

Nicola Wright¹, Helen Sharp¹, Andrew Pickles², Jonathan Hill³

¹Institute of Psychology, Health and Society, University of Liverpool.

²Biostatistics Department at the Institute of Psychiatry, King's College London.

³School of Psychology and Clinical Language Sciences, University of Reading.

Corresponding author: Miss Nicola Wright, School of Psychology, Eleanor Rathbone
Building, Bedford Street South, Liverpool, L69 7SA. nwright@liv.ac.uk, tel: 0151 650 5490

Acknowledgements: The research was funded by grants from the UK Medical
Research Council (G0900654; G0400577). We are grateful to all participating families and
research staff who contributed to this work. We also thank Wirral University Teaching
Hospital NHS Foundation Trust, Cheshire and Wirral Partnership NHS Foundation Trust,
Wirral Community NHS Trust and the National Institute for Health Research (NIHR)
Biomedical Research Centre at South London and Maudsley NHS Foundation Trust.

Abstract

Callous-unemotional (CU) traits are associated with severe and stable antisocial behaviour in childhood and adolescence. In order to understand the earliest origins of CU traits we need first to know whether the construct and measures are valid in young children. This study evaluated the psychometric properties and validity of a CU traits measure at age 2.5 years. The participants ($N = 775$) were members of an epidemiological longitudinal study starting in pregnancy. Items from the Antisocial Process Screening Device and other problem behaviour scales were subjected to exploratory and confirmatory factor analysis. Structural equation modelling was used to test whether age 2.5 CU traits showed incremental validity in predicting aggression at age 5. The CU measure showed acceptable psychometric properties, factorial invariance by sex and good stability. Incremental prediction to later aggression was evident in girls, whereas boys showed strong continuity in aggression not found for girls.

Keywords: Callous-unemotional (CU) traits, physical aggression, preschool, incremental validity, sex differences

Problems of oppositionality and aggression appearing in early childhood confer a substantially increased risk of later antisocial behaviour disorders and a wide range of psychiatric disorders including depression, anxiety and substance misuse (Odgers et al., 2008). While these early onset 'life course persistent' conduct problems share poor long term outcomes, it is likely that there is heterogeneity of risk factors and underlying processes (Hill, 2002). There is much current interest in a possible subgroup of conduct disordered children who show a lack of concern for the feelings of others and lack of guilt or remorse (Frick, 2009). In adults these traits are considered part of the affective dimension of psychopathy, and when applied to children they have been labelled as 'callous-unemotional traits' (CU traits). There is some evidence that there may be distinct developmental processes contributing to the development of conduct problems with and without CU traits. For example, conduct problems in children with CU traits have been found to be more highly heritable (Viding, Jones, Frick, Moffit & Plomin, 2008), less influenced by negative parenting practices (Pasalich, Dadds, Hawes & Brennan, 2012) and less responsive to typical conduct problem interventions (Hawes, Price & Dadds, 2014). Critically, CU traits in childhood and adolescence have been shown to be associated with more severe and enduring antisocial behaviour than conduct problems alone (Frick, Ray, Thornton & Kahn, 2014) supporting the utility of the application of the CU traits construct to childhood.

The majority of CU traits research has focused on samples aged 5-18 years, however, there is a small increasing literature examining whether CU traits can be reliably and validly measured in the pre-school period. Two domains of developmental research suggest that this may be so. First, observational studies have identified that empathy-related behaviours and guilt emerge during the first and second years of life (Kochanska, Gross, Lin & Nicholas, 2002; Vaish, Carpenter & Tomasello, 2009) which suggests that meaningful variations in CU traits may be measureable as early as age 2 years. Second, trajectories of elevated aggressive

behaviour have also been demonstrated to start in the second year of life (Tremblay et al., 2004; Campbell et al., 2010) and if CU traits play a causal role in the development of persistent aggressive behaviour, then CU traits or their precursors may be measurable and operating at this early age too. Identifying the earliest age at which CU traits can be reliably measured has important implications for research examining the developmental pathways to and from CU traits and for the development of potential preventative intervention before severe antisocial behaviour develops.

Two different approaches to measurement of CU traits in young children have been adopted in the field. In the first, standard measures of CU traits developed for older children and adolescents have been modified for preschool use. The most commonly used are the Antisocial Process Screening Device (APSD; Frick & Hare, 2001) which contains a 6 item CU subscale, and the Inventory of Callous-Unemotional traits (ICU; Frick, 2004), a 24 item measure developed by creating multiple items from four of the six APSD items considered to be the most key to the CU construct. Both measures have been modified for preschool use by replacing reference to ‘school-work’ with ‘structured activities’. The second approach involves using a collection of items drawn from existing measures of early child behaviour (hybrid measures). This method has the advantage that items were developed specifically for younger children and so possess more face validity for this age group than items from measures developed for older children. However, they lack the background of validation and replication which characterises the existing CU traits measures.

Factor structure of CU measures in early childhood

An important initial question when considering CU traits measurement in very young children is whether raters can reliably distinguish CU traits behaviours from other problem behaviours (Willoughby, Mills-Koonce, Gottfredson & Wagner, 2014). This has been examined using exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) on

CU traits items and items from other child problem behaviour dimensions. In the first of these studies, Dadds, Hawes, Frost and Fraser (2005) demonstrated that the APSD CU items were separable from the Strengths and Difficulties Questionnaire (SDQ; Goodman, 1997) conduct problems, hyperactivity, anxiety and peer problems in a community sample of 1,359 4-9 year olds using EFA. Three studies of three year olds have subsequently used CFA on items from the Child Behaviour Checklist (CBCL; Achenbach & Rescola, 2001) and demonstrated that a 5 item CU traits subscale was distinct from ODD and ADHD (Waller, Hyde, Grabell, Alves & Olson, 2015; Willoughby, Waschbusch, Moore & Propper, 2011; Willoughby et al., 2014). This question has yet to be examined in children under the age of 3. In this study we test whether mothers can reliably distinguish CU traits from other childhood problem behaviours at age 2 by employing CFA to test whether two separable factors reflecting CU traits and aggression are clearly identifiable.

Reliability of CU traits measures in early childhood

A key issue in CU traits measurement in both younger and older samples (e.g. Essua, Sasawaga & Frick, 2006) has been that measures tend to show poor internal consistency reliability. The majority of studies of early childhood have reported unsatisfactory Cronbach's Alpha for their measures. For example, $\alpha = .54$ reported for the APSD from a sample of 2-5 year olds (Kimonis et al., 2006) and $\alpha = .55$ to $.66$ reported for CBCL-based CU traits measure in four samples of 3-4 year olds (Kimonis, Bagner, Linares, Blake & Rodriguez, 2014; Waller et al., 2015; Willoughby et al., 2011; Willoughby et al., 2014). In the only investigation to assess the psychometric properties of a CU traits measure in 2 year olds, Hyde et al. (2013) reported an unsatisfactory Cronbach's Alpha at age 2 years ($\alpha = .57$ for primary caregiver and $\alpha = .47$ for alternative caregiver report), which was slightly improved by age 3 years ($\alpha = .64$ and $\alpha = .66$) and approaching commonly accepted values for both reporters at age 4 ($\alpha = .72$ and $\alpha = .66$) for their 5 item hybrid measure. This led

Hyde et al. to conclude that CU traits may not be sufficiently developed to assess at age 2 years. However, this conclusion may be premature given this is the only study to date to examine CU traits in 2 year olds, and other studies of 3 and 4 year olds using short measures of CU traits have reported similarly low Cronbach's Alpha levels. Further, in this study the measure comprised items assessing deceitful or manipulative behaviours which are not typically included in measures of CU traits in childhood, such as the APSD or the ICU, which focus instead on lack of guilt and empathy and poverty of affect. Of course, lying and manipulateness are relevant behaviours seen in adolescence and adulthood but they are likely to require more advanced cognitive abilities than those normally developed at age 2. This developmental and conceptual issue likely contributed to the poorer performance of their measure at age 2 specifically.

One investigation has employed the 24 item CU traits measure, the ICU, with a sample of 3 year olds with repeated measurement at age 4 ($n = 622$; Ezpeleta, Osa, Granero, Penelo & Domenech, 2013) and reported satisfactory Cronbach's Alpha for the subscales and total score at both ages ($\alpha = .79$ to $\alpha = .93$ for the subscales and total scores). With a measure designed to assess psychopathic traits in children, Colins et al. (2014) reported a Cronbach's Alpha of .95 for a 10 item CU traits subscale in a subsample of 687 3 year olds. Hawes and Dadds (2007) reported improved internal consistency (Cronbach's Alpha = .79) in a mixed age sample of 4-8 year olds after supplementing the APSD with the SDQ prosocial items (reverse coded). It is well known that the Alpha coefficient is affected by the number of items in a scale, with fewer items often yielding lower values (Cortina, 1993). In light of this, in this study we use the APSD as a recognised foundational CU traits measure but we also follow the approach of Dadds and colleagues and supplement it with items from other problem behaviour scales designed for young children in order to establish a measure that adequately assesses the construct of CU traits at age 2 and age 5.0, includes a broader range

of items and shows acceptable psychometric properties, particularly satisfactory internal reliability. We also report an alternative and more appropriate method of assessing internal consistency, Ordinal Alpha. Cronbach's Alpha is based on the Pearson's correlation matrix and so is best suited to items rated on a continuous scale. Ordinal Alpha, which is based on the polychoric correlation matrix, is more appropriate for use with items rated on an ordinal scale (Zumbo, Gaderman & Ziesler, 2007) which most child problem behaviour measures are.

Stability of CU measures across early childhood

CU trait scores have been shown to be moderately stable across childhood and from childhood to adolescence (Frick et al., 2014). To date the findings from studies of early childhood have also supported moderate to high stability in CU traits. Hyde et al. (2013) reported a correlation of .47 between their age 2 and 3 deceitful-callous measures. Ezpeleta et al. (2013) found moderate stability in ICU scores from age 3 to 4 years (ICC = .53 for total and ICC = .51-.53 for subscales). Over a longer period from 3 to 5 years, Willoughby et al. (2011) reported a latent factor correlation of .84. In this study we examine stability from age 2.5 to 5.0.

Validity of CU traits measures in early childhood

The key hypotheses regarding the clinical significance of CU traits focus on their role in disinhibiting violence and physically aggressive behaviour (e.g. the Violence Inhibition Mechanism; Blair, 1995). In particular CU traits are postulated to underpin proactive aggression (Frick, Cornell, Barry, Bodin & Dane, 2003). Whilst studies of the validity of psychopathy and CU traits in adults and adolescents examine criminal outcomes, such as violent versus non-violent crimes, these are not yet relevant in young children and so it is necessary to focus on the prediction of developmentally relevant behaviours, such as physically aggressive behaviour. The key test of the validity of CU traits in predicting

antisocial outcomes is whether they show incremental prediction over and above other known predictors (Frick et al., 2014) such as initial problem behaviour. We examine the incremental prediction of physical aggression at age 5 years accounting for age 2.5 physical aggression.

Associations between CU traits and physical aggression have been reported in at least three early childhood studies, but none have examined the incremental prediction of aggression. Kimonis et al. (2006) found prospective prediction from mother-report on the APSD to teacher-reported aggression 6 months later with a sample of 2-5 year olds. Ezpeleta et al. (2013) demonstrated both cross-sectional and prospective associations from age 3 to 4 years between teacher-reported CU traits and aggression. Willoughby et al. (2014) found age 3 CU traits measured using the CBCL to predict persistent teacher-reported aggression from age 6-12 years, but only in interaction with ODD and disorganised attachment status. Two studies have examined the incremental prediction of CU traits but to broad behaviour problems outcomes only. Hyde et al. (2013) found their age 3 deceitful-callous measure to predict increased behaviour problems from age 2 to 4 years after accounting for initial problem behaviour. Further, Waller et al. (2015) found age 3 mother-reported CU traits predicted teacher-reported externalizing problems at age 6, after accounting for age 3 externalizing problems. No study has yet shown incremental validity for a CU trait measure in 2 year olds in this way.

CU traits measurement and the possible role of sex differences

Whether CU traits show sex differences in associations with outcomes has been given relatively little consideration in the literature, and none of the studies of early childhood have examined for sex differences. It is well-established that boys show higher mean levels of CU traits than girls (Ezpeleta et al., 2013; Fanti & Kimonis, 2013; Marsee, Silverthorne & Frick, 2005). Levels of physical aggression are also lower in girls (Card, Stucky, Sawalani & Little, 2008) even at age 2-3 years (Baillargeon et al., 2007). Some studies have examined

associations between CU traits and aggression separately for boys and girls and found them to be similar (Essau et al., 2006; Kimonis et al., 2006; Silverthorne, Frick & Reynolds, 2001) although significant sex by CU interactions have been reported by others (Fanti & Kimonis, 2013; Thornton, Frick, Crapanzano & Terranova, 2012) Very few studies have sought to establish that CU measures are invariant across sex before examining associations with outcomes, and therefore any sex difference findings reported may be due to measures operating differently in boys and girls. In this study we test whether the CU traits measure that we generate is invariant across sex, and then examine whether there are sex differences in associations between CU traits and aggression. We make no specific hypotheses regarding sex differences.

The present study

In summary, in the current prospective longitudinal study, CU traits and physical aggression were assessed at age 2.5 and again at 5.0 years. The first step was to establish CU traits scales at each age point with satisfactory psychometric properties, including internal reliability and measurement invariance by sex. We examined whether parents could reliably distinguish CU traits and aggression by testing two-factor CU traits and aggression models against one-factor models. CFA was also used to examine whether there were continuities in CU traits from age 2.5 years to 5.0 years. The main analysis used structural equation modelling to examine concurrent and prospective associations between CU traits and physical aggression at 2.5 years and 5.0 years. Our key hypothesis was that age 2.5 CU traits would show incremental validity in predicting physical aggression at age 5.0 years after accounting for age 2.5 physical aggression and all other possible prospective and cross-sectional associations, and that this did not differ by sex of child.

Method

Sample

Participants were mothers and children taking part in the Wirral Child Health and Development Study, a prospective epidemiological cohort study starting in pregnancy designed to investigate the earliest origins of childhood conduct problems. All women gave written informed consent at the point of recruitment in the antenatal clinic. The study used a two stage stratified design in which a consecutive general population sample (the 'extensive' sample) is used to generate a smaller 'intensive' sample stratified by psychosocial risk with more detailed measurement over time and both are followed in tandem (Sharp et al., 2012). Mother's responses to a questionnaire at 20 weeks of pregnancy (recruitment) assessing psychological abuse in their current or recent partner relationship (Moffitt et al., 1997) were used to generate the stratified intensive sample of mothers for more detailed study. The stratification variable was chosen for its known association with a variety of risk factors for early child development. The whole cohort comprised 1233 women of mean age at recruitment of 26.8 years ($SD = 5.8$, range 18-51) and 41.8% of the sample were in the most deprived quintile of UK neighbourhoods (Noble et al., 2004). There were 316 mothers recruited to the intensive sample at 32 weeks pregnancy. This report uses questionnaire data collected from the whole cohort who gave data at 20 weeks pregnancy and again at age 5.0 years ($n = 775$), and from the stratified intensive sub-sample at 2.5 years ($n = 241$). Nonresponse at age 5 was associated with younger maternal age ($U(N = 1233) = 230,692$, $Z = 8.67$, $p < .001$) and living in the most deprived quintile of UK neighbourhoods ($\chi^2(1, N = 1233) = 19.62$, $p < .001$). The mean age of the children at the 2.5 year assessment was 30.86 months ($SD = 2.31$, range = 27 - 42 months) with slightly more girls ($n = 123$) than boys ($n = 118$), and the mean age of all 775 children whose mothers completed questionnaires at 5.0

years was 58.64 ($SD = 3.74$, range = 49 - 73 months) with 402 girls and 373 boys. At age 5.0, 80% of mothers were either married or cohabiting, 4.8% had a partner living elsewhere and 15% were single.

Measures

CU traits. Items were drawn from four different child problem behaviour scales, shown in Table 1. Items were selected based on inclusion in CU traits measures in other studies (Dadds et al., 2005; Hyde et al., 2013; Kimonis et al., 2014; Willoughby et al., 2011) and relevance to the CU traits construct, with a focus on items assessing lack of concern for others, lack of guilt and poverty of affect. All six items from the CU subscale of the APSD were selected for use at both time points (2.5 years and 5.0 years). APSD items are rated on a 3 point scale: 0 = *not true*, 1 = *sometimes true*, 2 = *very true*. Consistent with previous use of the measure with younger samples, the subscale showed somewhat low Cronbach's Alpha ($\alpha = .56$ at age 2.5 years and $\alpha = .60$ at age 5.0 years). Six items were selected from the CBCL at both time points. The item 'doesn't seem to feel guilty after misbehaving' was not included due to similarity to the APSD item 'feels bad or guilty when he/she does something wrong'. CBCL items are rated on a 3-point scale: 0 = *not true/never*, 1 = *somewhat true/sometimes*, 2 = *very true/very often*. At age 5.0, all 5 items from the prosocial subscale of the SDQ were selected, based on the University of New South Wales system of combining items from the APSD with the SDQ prosocial items (Dadds et al., 2005). One item from the Brief Infant Toddler Socio-emotional Assessment (BITSEA; Briggs-Gowan, Carter, Irwin, Wachtel & Cicchetti, 2004) "tries to help when someone is hurt (for example, gives a toy)" was included at age 2.5 based on its similarity to the SDQ prosocial items. A total of 13 items were selected for age 2.5 years and 17 items for age 5.0 years. EFA and CFA were run separately

for age 2.5 and 5.0 years and the CU factor composition was allowed to differ at each age to allow for developmental differences in the manifestation of CU traits.

Physical aggression. Mothers completed a physical aggression questionnaire (Baillargeon et al., 2007) at age 2.5 years and 5.0 years. The questionnaire consists of five items previously shown to yield aggression scores with stability from ages 17 to 29 months (Baillargeon et al., 2007). Each item is rated on a three-point scale: 0 = *not true*, 1 = *somewhat or sometimes true*, 2 = *very true or often true*. The Cronbach's Alpha in the present sample was adequate for age 2.5 years ($\alpha = .67$) and 5.0 years ($\alpha = .82$).

Analysis plan

First, the CU items were entered into an exploratory factor analysis for ordinal data (using the weighted least squares mean adjusted estimator [WLSM] and promax rotation) in Mplus version 7 (Muthen & Muthen, 2012) separately for age 2.5 and age 5.0 years. Items with a factor loading $>.35$ were retained. Second, a series of multi-group two-factor CFA, estimated using the WLSMV estimator and Theta parameterization, examined measurement invariance across sex in CU traits and aggression. In model 1 (configural model) the pattern of factor loadings were constrained to be the same for boys and girls, testing that the same items formed the CU and aggression factors across sex. In model 2 (metric model) the individual factor loadings were constrained to be the same, testing whether the contribution of individual items varies by sex (weak factorial invariance). In model 3 (scaler model) the thresholds were also constrained to be the same, to examine whether the items performed the same across sex (strong factorial invariance). Full invariance is demonstrated when the placing of additional constraints on the model does not produce a significant worsening in model fit. The *DIFFTEST* command was used to evaluate whether a substantial change in model fit occurred as a result of imposing additional constraints, as well as the CFI change

(Δ CFI). A non-significant chi-square difference test and a small CFI change (in which a decrease is no greater than .01) are considered indicative of invariance (Cheung & Rensvold, 2002). If a significant chi-square difference test is found, the modification indices are examined to determine which items failed the strong factorial variance assumption. In the absence of modification indices the individual items are checked for those showing the largest difference between boys and girls. The thresholds of these items are then allowed to vary freely and model fit is re-examined as a test of partial strong factorial invariance.

Third, two CFA models were estimated and compared using the DIFFTEST command, to test whether a one-factor or two-factor CU and aggression model best fit the data. Finally, two multi-group (by child sex) structural equation models (SEM) for ordinal data were fitted to examine cross-lagged continuity in CU traits and aggression from age 2.5 to 5.0 years, concurrent associations with aggression at each age, and the prospective association between CU traits at age 2.5 and aggression at age 5.0 years, and vice versa. We used the WLMSV estimator in order to include whole cohort at age 5.0 and the intensive sub-sample at age 2.5 years within a single model. To account for sample stratification of the intensive sub-sample we included the measure on which the stratification was based as an auxiliary variable (Graham, 2003). We fitted the model of Figure 1 in which we imposed a common measurement model constraining the factor loadings, measurement errors, and thresholds to be the same across boys and girls and with the factor loading for the first factor fixed at 1 for identification. However, in order to check for sex differences we allowed the means, variances and covariances among the factors to differ by sex of child. The model provided estimates of the effects of time 1 (age 2.5) scores on time 2 (age 5.0) scores, both the simple lagged effects of early CU on later CU traits and aggression on later aggression, and, in order to test the main hypothesis, the cross lagged effects of early CU traits on later aggression, and early aggression on later CU traits. This constrained model gave common

estimates of these effects, assuming that a unit change in age 2.5 CU traits or aggression had the same effects on age 5.0 CU traits and aggression in both boys and girls. A second unconstrained model was run allowing the coefficients for girls to be different from those for boys, and the constrained and unconstrained models were then compared using the DIFFTEST.

The adequacy of all CFA and SEM models was assessed using the Root Mean Square Error (RMSEA) criterion where less than 0.05 is considered a good fit and less than .08 considered reasonable fit, and the Comparative Fit Index (CFI) where values above .95 indicate good fit and .90 reasonable fit (Marsh, Hau, & Wen, 2004).

Results

Exploratory factor analysis on the CU items

Age 2.5 years. The 13 CU items were entered into an EFA. The APSD item “Does not show feelings or emotions” showed problems with empty cells in the cross-tabulation with two CBCL items (“Shows little affection toward people” and “Seems unresponsive to affection”) and thus was removed from the EFA. Eigenvalues for the first three factors were 4.6, 1.6, and 1.1, and the scree plot supported a one factor solution. The CBCL item “Shows too little fear of getting hurt” gave a factor loading <0.35 and was dropped, resulting in 11 items at age 2.5 years.

Age 5 years. The EFA for the 17 CU items gave eigenvalues of 7.1, 1.6 and 1.2, and the scree plot supported a one-factor solution. The item APSD “Does not show feelings or emotions”, gave a factor loading <0.35 and so was omitted, resulting in 16 items at age 5.

Confirmatory factor analyses

Age 2.5 years measurement invariance. The 11 items retained at age 2.5 were then tested for measurement invariance across boys and girls using multi-group CFA. Model 1, the configural model, showed good fit (RMSEA = .05, CFI = .95). The fit improved in model 2 where factor loading invariance was introduced (RMSEA = .04 CFI = .96) with a non-significant chi-square difference test. Threshold invariance was introduced with model 3 and the fit was largely unchanged (RMSEA = .04, CFI = .95) the chi-square tests comparing model 3 to model 1 and 2 were both non-significant, demonstrating strong scalar or strong factorial invariance. The full model fit and comparison results are presented in Table S1.

Age 5 years measurement invariance. The 16 items retained at age 5 were then tested for measurement invariance across sex. Model 1, the configural model, showed good fit (RMSEA = .05, CFI = .95). However, the modification indices indicated that items ‘APSD: does his/her best in structured activities’ and ‘CBCL: selfish’ should cross-load on the aggression factor for boys and girls, and ‘CBCL: shows too little fear’ should cross-load for boys. These three items were not considered central to the CU traits construct, and cross-loading with aggression was undesirable given the aim of examining a purely physical aggression outcome, therefore, the items were removed. A further configural model (Model 1b) was tested on the remaining 13 items and showed improved fit (RMSEA = .03, CFI = .99) with no further modification indices, and so this model was used in further analysis. The introduction of factor loading invariance with model 2 resulted in a further improvement in fit (RMSEA = .02, CFI = .99) with a non-significant chi-square difference test. However, the introduction of threshold invariance in model 3 resulted in a significant chi-square difference test ($p = .003$) and therefore strong factorial invariance was not achieved. There were no modification indices above the minimum value so individual item thresholds were inspected for differences between boys and girls. Items ‘CBCL: hits other children’, ‘CBCL: seems

unresponsive to punishment’, ‘APSD: keeps the same friends’ and ‘SDQ: volunteers to help’ all showed a difference of > 0.4 , with boys showing a lower threshold than girls on all items apart from ‘volunteers to help’, and so the thresholds for those items were freed. The chi-square test for difference testing between the metric and partial scalar models was now non-significant ($p = 0.08$) therefore we found evidence for partial strong or partial scalar invariance at age 5 years. The full model fit and comparison results are presented in Table S2.

One- versus two-factor CU traits and aggression CFA models. We next examined whether mothers could differentiate CU traits and aggression by comparing a one-factor CFA model where all the CU and aggression items loaded on one factor, to the two factor model, using the chi-square DIFFTEST. The model fit statistics and model comparison results are displayed in Table 2. The two-factor model showed the best fit at age 2.5 years (RMSEA = .05, CFI = .95) and 5.0 years (RMSEA = .05, CFI = .94) and the chi-square difference tests indicated that the two-factor models showed significantly better fit ($p < .001$ for both ages). The standardised factor loadings are displayed in Table 1.

Internal consistency

Cronbach’s Alpha for the CU traits measure at age 2.5 years was $\alpha = .72$ and age 5.0 years $\alpha = .83$. Ordinal Alpha, a more appropriate index of internal consistency for items rated on an ordinal scale, was $\alpha = .87$ for age 2.5 years and $\alpha = .89$ for 5.0 years.

Sex differences in mean levels

Having established measurement invariance we then used multi-group CFA to test for sex differences in the means on the latent variables. At age 2.5, boys scored significantly higher on CU traits ($-.34, p = .041$) but not aggression ($-.32, p = .198$), whereas at age 5 boys scored significantly higher on both CU traits ($-.32, p = .002$) and aggression ($-.35, p = .006$).

Stability for boys and girls

Factor correlations between age 2.5 and 5.0 year CU traits were similar and substantial for boys and girls (.75, $p < .001$ and .71 $p < .001$) with a somewhat larger association over time for aggression in boys (.59, $p < .001$) than for girls (.35, $p < .001$).

Multivariate model to examine incremental validity

We fitted the model shown in Figure 1, and more fully described in the Methods section, to data from both boys and girls, constraining the four path coefficients between the factors to be the same for boys and girls. The model fitted well (RMSEA = .03 (CI .02 – .03) CFI = .93). We next fitted a model in which the path coefficients were allowed to be different, assessing any improvement in fit using the DIFFTEST. While overall fit statistics remained unchanged the DIFFTEST was highly significant, indicating a clear sex difference ($X^2(4) = 10.95, p = .028$) when all the paths were considered. Figure 1 shows the magnitudes of the standardized coefficients where it is evident that the sex differences were of two kinds. There was continuity of aggression from 2.5 to 5.0 years in boys (.84, $p = .006$) but not girls (.02, $p = .923$). By contrast the cross-lagged path from CU traits at 2.5 years to physical aggression at 5.0 years was substantial in girls (.62, $p < .001$) but entirely non-significant for boys (-.06, $p = .933$). There was no sex difference in continuities from CU traits at 2.5 years to CU traits at 5 years, which was significant for both boys (.74, $p < .001$) and girls (.81, $p < .001$). Cross-sectional correlations between the CU traits and aggression factors at age 2.5 were similar for boys (.71) and girls (.58) and this was the case also at age 5 years (.75 and .76 respectively). Combining the various effects together the model estimated that at age 5 years for boys 43% of the variation in the CU factor and 36% of variation in the aggression factor was explained by the measures at age 2.5. The corresponding values for girls were 50% for CU traits and 60% for aggression.

Discussion

In this study we sought to establish whether CU traits can be measured reliably at 2.5 years, and to assess their validity at this age as evidenced in an incremental prediction of physical aggression from 2.5 to 5.0 years. We measured CU traits in children aged 2.5 years by supplementing a widely used measure of CU traits, the APSD, with items from other problem behaviour scales for young children. This yielded a CU traits scale invariant across sex with satisfactory psychometric properties that showed strong stability from age 2.5 to 5.0 years for both boys and girls. CU traits and aggression showed significant and substantial associations at age 2.5 and 5.0 years for both boys and girls in line with theory. We found that for girls only, CU traits at 2.5 years predicted physical aggression at 5.0 years after accounting for age 2.5 aggression and all other possible cross-sectional and prospective associations.

As outlined earlier, establishing whether CU traits can be identified in young children is a priority if their role in early onset of aggression is to be studied. There are however substantial issues regarding their measurement that were examined in this study. First, in response to the concern that CU traits may not be identifiable as separable from conduct problems, we showed using CFA that mothers' ratings of CU traits were separable from ratings of physical aggression at age 2.5 years. This builds on previous findings which have demonstrated that CU traits are distinct from other problem behaviour dimensions at age 3 years and above (Willoughby et al., 2011; Willoughby et al., 2013; Waller et al., 2015). Second in order to deal with problems of low internal consistency, we supplemented an established measure of CU traits with items from other child problem behaviour scales and this yielded a measure with satisfactory internal consistency. Third given increasing interest in the possibility that there are sex differences both in the origins and consequences of the CU traits, we examined measurement invariance across boys and girls and found strong

invariance for the age 2.5 measure and partial strong invariance for the age 5.0 measure. Very few studies have examined measurement invariance by sex for CU traits measures; our findings are consistent with the only previous study to examine measurement invariance in an early childhood sample, which demonstrated invariance by sex for the ICU at ages 3 and 4 years (Ezpeleta et al., 2013). Finally we examined incremental validity in the prediction of aggression at 5.0 years from CU traits at 2.5 years, making use of the measurement invariance that we had established to test for sex differences. We found age 2.5 CU traits to predict age 5.0 aggression over and above age 2.5 aggression and all other possible cross-sectional and prospective associations, but in girls only.

Strengths of the study included consecutive recruitment from an antenatal clinic serving a defined geographical area, enabling effect estimates applicable to the general population to be generated. In contrast to many studies of CU traits the behavioural outcome was child aggression which is most relevant to key hypotheses regarding the effects of CU traits on behaviour (Blair, 1995). Structural equation modelling (SEM) which has rarely been used in studies of links between CU traits and aggression conferred four main advantages including, first it utilized an item level analysis which takes account of differing contributions of items, second it enabled all possible prospective pathways to be examined simultaneously, third it allowed measures to operate differently over time or by sex, and fourth it provided tests of sex differences. A limitation of SEM with this sample was that, as a result of some sparse cells, it was necessary to collapse some items to binary variables, and for the aggression latent variable, to combine items. A further limitation of the study was that all of the measures were mother-report questionnaires and so the findings may be influenced by common method variance.

The prediction from CU traits at 2.5 years to peer aggression at 5.0 years additional to the prediction from 2.5 years aggression in girls, but not in boys, needs to be considered in

relation to the finding from the same analyses, that the continuity in aggression from 2.5 years to 5.0 years was much stronger in boys than in girls. This is consistent with findings from large scale studies of older children that boys are more likely than girls to show stable aggression (Campbell et al., 2010; Lee et al., 2007), and with similar evidence from smaller scale studies of preschool children (Alink et al., 2006, Cummings, Iannotti & Zahn-Waxler, 1989). Here we report the same phenomenon in a large preschool sample, and specifically in relation to peer aggression, further supporting the possibility that the emergence and maintenance of early childhood aggression may be underpinned by different processes in boys and girls (Crick & Zahn-Waxler, 2003). Thus one possible interpretation of the failure to show an incremental effect of CU traits in boys is that the extent of change in aggression over the period 2.5 years to 5.0 years, either increasing or decreasing, was so limited that there was little for CU traits to explain. The question is then posed as to whether the key risk processes for aggressive behaviours in boys are to be found before age 2.5 years, with a challenge to identify CU traits or their precursors during infancy or the toddler period. There may also be other explanations for the sex difference such as that the translation of CU traits into aggression is dependent on other influences, for example, deficits in behavioural inhibitory processes, so that even in the absence of main effect, effects may be found in interaction with other variables. Equally the CU traits construct may not be valid in young boys because the relevant empathic processes develop later in boys than girls (Rhee et al., 2013), or the measure may not be valid because the behaviours that reflect CU traits are not identified in the items of our existing measures.

In conclusion, this study showed that a measure of CU traits can be constructed that meets current concerns about poor psychometric properties of measures of CU traits in very young children, and that measures a construct that is distinct from aggression. Studies of sex differences at this age can be conducted with confidence that items have the same

relationship to the latent variable in boys and girls. The marked sex difference, both in stability of aggression and prediction from CU traits, may reflect different pathways to aggression in boys and girls, or differences in validity of the CU traits construct or measure in boys at 2.5 years. Designs of early intervention studies to prevent CU traits will need to account for the possibility that therapeutic approaches may need to vary by sex of the child, and that measurement of early CU traits outcomes may not mean the same in boys and girls.

References

- Achenbach, T. M., & Rescorla, L. A. (2000) *Manual for the ASEBA preschool forms & profiles*. Burlington, VT: University of Vermont, Research Center for Children, Youth and Families.
- Alink, L. R. A., Mesman, J., van Zeijl, J., Stolk, M. N., Juffer, F., Koot, H. M., . . . van IJzendoorn, M. H. (2006). The early childhood aggression curve: Development of physical aggression in 10-to 50-month-old children. *Child Development*, 77(4), 954-966. doi:10.1111/j.1467-8624.2006.00912.x
- Baillargeon, R. H., Zoccolillo, M., Keenan, K., Côté, S., Pérusse, D., Wu, H., . . . Tremblay, R. E. (2007). Gender differences in physical aggression: A prospective population-based survey of children before and after 2 years of age. *Developmental Psychology*, 43(1), 13-26. doi: 10.1037/0012-1649.43.1.13.
- Blair, R. J. R. (1995). A cognitive developmental-approach to morality - investigating the psychopath. *Cognition*, 57(1), 1-29. doi:10.1016/0010-0277(95)00676-P
- Briggs-Gowan, M. J., Carter, A. S., Irwin, J. R., Wachtel, K., & Cicchetti, D. V. (2004). The Brief Infant -Toddler Social and Emotional Assessment: screening for social-emotional problems and delays in competence. *Journal of Pediatric Psychology*, 29(2), 143-55. doi: <https://doi.org/10.1093/jpepsy/jsh017>
- Campbell, S. B., Spieker, S., Vandergrift, N., Belsky, J., Burchinal, M., & Nichd Early Child Care Res Network. (2010). Predictors and sequelae of trajectories of physical aggression in school-age boys and girls. *Development and Psychopathology*, 22(1), 133-150. doi:10.1017/S0954579409990319
- Card, N. A., Stucky, B. D., Sawalani, G. M., & Little, T. D. (2008). Direct and indirect aggression during childhood and adolescence: A meta-analytic review of gender

- differences, intercorrelations, and relations to maladjustment. *Child Development*, 79(5), 1185-1229. doi:10.1111/j.1467-8624.2008.01184.x
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9, 233–255.
- Colins, O. F., Andershed, H., Frogner, L., Lopez-Romero, L., Veen, V., & Andershed, A. K. (2014). A new measure to assess psychopathic personality in children: The Child Problematic Traits Inventory. *Journal of psychopathology and behavioral assessment*, 36(1), 4-21.
- Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78, 98-104.
- Crick, N. R., & Zahn-Waxler, C. (2003). The development of psychopathology in females and males: Current progress and future challenges. *Development and Psychopathology*, 15(3), 719-742. doi:10.1017/S095457940300035X
- Cummings, E. M., Iannotti, R. J., & Zahnwaxler, C. (1989). Aggression between peers in early-childhood - individual continuity and developmental-change. *Child Development*, 60(4), 887-895. doi:10.1111/j.1467-8624.1989.tb03521.x
- Dadds, M.R., Hawes, D. J., Frost, A., & Fraser, J. (2005). Disentangling the underlying dimensions of psychopathy and conduct problems in childhood: A community study. *Journal of Consulting and Clinical Psychology*, 73(3), 400-410. doi:10.1037/0022-006X.73.3.400
- Essau, C. A., Sasagawa, S., & Frick, P. J. (2006). Callous-unemotional traits in a community sample of adolescents. *Assessment*, 13(4), 454-469. doi:10.1177/1073191106287354
- Ezpeleta, L., Osa, N. d. l., Granero, R., Penelo, E., & Domènech, J. M. (2013). Inventory of callous-unemotional traits in a community sample of preschoolers. *Journal of Clinical Child & Adolescent Psychology*, 42(1), 91-105. doi:10.1080/15374416.2012.734221

- Fanti, K. A., & Kimonis, E. R. (2013). Dimensions of juvenile psychopathy distinguish "bullies," "bully-victims," and "victims". *Psychology of Violence*, 3(4), 396-409. doi:10.1037/a0033951
- Frick, P. J., & Hare, R. D. (2001). *The antisocial process screening device (APSD)*. Toronto: Multi-Health Systems.
- Frick, P. J. (2004). *Inventory of Callous-Unemotional Traits*. Unpublished rating scale, University of New Orleans; 2004.
- Frick, P.J., Cornell, A. H., Barry, C. T., Bodin, S. D., & Dane, H. E. (2003). Callous-unemotional traits and conduct problems in the prediction of conduct problem severity, aggression, and self-report of delinquency. *Journal of Abnormal Child Psychology*, 31(4), 457-470. doi:10.1023/A:1023899703866
- Frick, P. J. (2009). Extending the construct of psychopathy to youth: Implications for understanding, diagnosing, and treating antisocial children and adolescents. *Canadian Journal of Psychiatry-Revue Canadienne De Psychiatrie*, 54(12), 803-812. doi: 10.1177/070674370905401203
- Frick, P. J., Ray, J. V., Thornton, L. C., & Kahn, R. E. (2014). Can callous-unemotional traits enhance the understanding, diagnosis, and treatment of serious conduct problems in children and adolescents? A comprehensive review. *Psychological Bulletin*, 140(1), 1-57. doi:10.1037/a0033076
- Goodman, R. (1997). The strengths and difficulties questionnaire: A research note. *Journal of Child Psychology and Psychiatry*, 38, 581-586.
- Graham, J. W. (2003). Adding missing-data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling*, 10, 80-100.

- Hawes, D. J., & Dadds, M. R. (2007). Stability and malleability of callous-unemotional traits during treatment for childhood conduct problems. *Journal of Clinical Child and Adolescent Psychology*, 36(3), 347-355.
- Hawes, D. J., Price, M. J., Dadds, M. R. (2014). Callous-unemotional traits and the treatment of conduct problems in childhood and adolescence: A comprehensive review. *Clinical Child and Family Psychology Review*, 17(3), 248-267. doi: 10.1007/s10567-014-0167-1
- Hill, J. (2002). Biological, psychological and social processes in the conduct disorders. *Journal of Child Psychology and Psychiatry*, 43(1), 133-64. doi: 10.1111/1469-7610.00007
- Hyde, L. W., Shaw, D. S., Gardner, F., Cheong, J., Dishion, T. J., & Wilson, M. (2013). Dimensions of callousness in early childhood: Links to problem behavior and family intervention effectiveness. *Development and Psychopathology*, 25(2), 347-363. doi:10.1017/S0954579412001101
- Kimonis, E. R., Bagner, D. M., Linares, D., Blake, C. A., & Rodriguez, G. (2014). Parent training outcomes among young children with callous–unemotional conduct problems with or at risk for developmental delay. *Journal of child and family studies*, 23(2), 437-448.
- Kimonis, E. R., Frick, P. J., Boris, N. W., Smyke, A. T., Cornell, A. H., Farrell, J. M., & Zeanah, C. H. (2006). Callous-unemotional features, behavioral inhibition, and parenting: Independent predictors of aggression in a high-risk preschool sample. *Journal of Child & Family Studies*, 15(6), 741-752. doi:10.1007/s10826-006-9047-8

- Kochanska, G. K., Gross, J. N., Lin, M. H., & Nichols, K. E. (2002). Guilt in young children: Development, determinants, and relations with a broader system of standards. *Child Development*, 73, 461–482. doi: 10.1111/1467-8624.00418
- Lee, K., Baillargeon, R. H., Vermunt, J. K., Wu, H., & Tremblay, R. E. (2007). Age differences in the prevalence of physical aggression among 5-11-year-old canadian boys and girls. *Aggressive Behavior*, 33(1), 26-37. doi:10.1002/ab.20164
- Marsee, M. A., Silverthorn, P., & Frick, P. J. (2005). The association of psychopathic traits with aggression and delinquency in non-referred boys and girls. *Behavioral Sciences & the Law*, 23(6), 803-817. doi:10.1002/bsl.662
- Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis- testing approaches to setting cutoff values for fit indices and dangers to overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, 11, 320–341. doi: 10.1207/s15328007sem1103_2
- Moffitt, T., Caspi, A., Krueger, R., Magdol, L., Margolin, G., Silva, P., et al. (1997). Do partners agree about abuse in their relationship? A psychometric evaluation of interpartner agreement. *Psychological Assessment*, 9(1), 47-56.
- Muthén, L. K., Muthén, B. O. (2012). *Mplus User's Guide. Seventh Edition*. Los Angeles, CA: Muthén & Muthén.
- Noble, M., Wright, G., Dibben, C., Smith, G., McLennan, D., Anttila, C., . . . Lloyd, M. (2004). *The English Indices of Deprivation 2004 (revised)*. Report to the Office of the Deputy Prime Minister. London: Neighbourhood Renewal Unit.
- Odgers, C. L., Moffitt, T. E., Caspi, A., Broadbent, J. M., Dickson, N., Hancox, R. J., . . . Sears, M. R. (2008). Female and male antisocial trajectories: From childhood origins to adult outcomes. *Development and Psychopathology*, 20(2), 673-716. doi:10.1017/S0954579408000333

- Pasalich, D. S., Dadds, M. R., & Hawes, D. J. (2014). Cognitive and affective empathy in children with conduct problems: Additive and interactive effects of callous-unemotional traits and autism spectrum disorders symptoms. *Psychiatry Research*, 219(3), 625-630. doi: 10.1016/j.psychres.2014.06.025
- Rhee, S. H., Friedman, N. P., Boeldt, D. L., Corley, R. P., Hewitt, J. K., Knafo, A., . . . Zahn-Waxler, C. (2013). Early concern and disregard for others as predictors of antisocial behavior. *Journal of Child Psychology and Psychiatry*, 54(2), 157-166. doi:10.1111/j.1469-7610.2012.02574.x
- Sharp, H., Pickles, A., Meaney, M., Marshall, K., Tibu, F., & Hill, J. (2012). Frequency of infant stroking reported by mothers moderates the effect of prenatal depression on infant behavioural and physiological outcomes. *Plos One*, 7(10) doi:10.1371/journal.pone.0045446
- Silverthorn, P., Frick, P. J., & Reynolds, R. (2001). Timing of onset and correlates of severe conduct problems in adjudicated girls and boys. *Journal of Psychopathology and Behavioral Assessment*, 23(3), 171-181. doi:10.1023/A:1010917304587
- Thornton, L. C., Frick, P. J., Crapanzano, A. M., & Terranova, A. M. (2012). The incremental utility of callous-unemotional traits and conduct problems in predicting aggression and bullying in a community sample of boys and girls. *Psychological Assessment*, doi:10.1037/a0031153
- Tremblay, R. E., Nagin, D. S., Seguin, J. R., Zoccolillo, M., Zelazo, P. D., Boivin, M., . . . Japel, C. (2004). Physical aggression during early childhood: Trajectories and predictors. *Pediatrics*, 114(1), E43-E50. doi:10.1542/peds.114.1.e43
- Vaish, A., Carpenter, M., & Tomasello, M. (2009). Sympathy through affective perspective taking and its relation to prosocial behavior in toddlers. *Developmental Psychology*, 45(2), 534-543. doi:10.1037/a0014322

- Viding, E., Jones, A. P., Frick, P. J., Moffitt, T. E., & Plomin, R. (2008). Heritability of antisocial behaviour at 9: Do callous-unemotional traits matter? *Developmental Science*, 11(1), 17-22. doi: 10.1111/j.1467-7687.2007.00648.x
- Waller, R., Hyde, L. W., Grabel, A. S., Alves, M. L., & Olson, S. L. (2015). Differential associations of early callous-unemotional, oppositional, and ADHD behaviors: Multiple domains within early-starting conduct problems? *Journal of Child Psychology and Psychiatry*, 56(6), 657-666. doi:10.1111/jcpp.12326
- Willoughby, M. T., Mills-Koonce, W. R., Gottfredson, N. C., & Wagner, N. J. (2014). Measuring callous unemotional behaviors in early childhood: Factor structure and the prediction of stable aggression in middle childhood. *Journal of Psychopathology and Behavioral Assessment*, 36(1), 30-42. doi: 10.1007/s10862-013-9379-9
- Willoughby, M. T., Waschbusch, D. A., Moore, G. A., & Propper, C. B. (2011). Using the ASEBA to screen for callous unemotional traits in early childhood: Factor structure, temporal stability, and utility. *Journal of Psychopathology and Behavioral Assessment*, 33(1), 19-30. doi:10.1007/s10862-010-9195-4.
- Zumbo, B. D., Gadermann, A. M., & Zeisser, C. (2007). Ordinal versions of coefficients alpha and theta for Likert rating scales. *Journal of modern applied statistical methods*, 6(1), 4.

Table 1

Standardised Factor Loadings from the Two-Factor CU traits and Aggression Models at Age 2.5 and 5.0 years

Items	Age 2.5	Age 5
<u>CU traits items</u>		
APSD 1: Concerned about the feelings of others (R)	.48	.47
APSD 2: Seems motivated to do his/her best in structured activities (R)	.61	
APSD 3: Is good at keeping promises (R)	.54	.47
APSD 4: Feels bad or guilty when he/she does something wrong (R)	.43	.62
APSD 5: Keeps the same friends (R)	.36	.61
APSD 6: Does not show emotions		
CBCL 14: Cruel to animals	.99	.60
CBCL 58: Punishment doesn't change his/her behavior	.62	.72
CBCL 67: Seems unresponsive to affection	.77	.77
CBCL 69: Selfish or won't share	.42	
CBCL 70: Shows little affection toward people	.48	.84
CBCL 72: Shows too little fear of getting hurt		
BITSEA 22: Tries to help if someone is hurt (R)	.69	
SDQ 1: Considerate of other people's feelings (R)		.75
SDQ 4: Shares readily with other children (R)		.53
SDQ 9: Helpful if someone is hurt, upset or feelings ill (R)		.57

SDQ 17: Kind to younger children (R) .60

SDQ 20: Often volunteers to help others (R) .46

Aggression items

Hits other children .76 .87

Bites other children/Kicks other children .75 .87

Gets in many fights/Physically attacks others .87 .90

Note. CBCL = Child Behavior Checklist (CBCL), APSD = Anti-Social Processes Screening Device, BITSEA = Brief Infant Toddler Social and Emotional Assessment (BITSEA), SDQ = Strengths and Difficulties Questionnaire (SDQ) .

Table 2

Fit statistics and results of DIFFTEST for one versus two factor CU and aggression models

	χ^2 (df)	CFI	RMSEA	χ^2 DIFFTEST
<hr/> Age 2.5 years				
1 factor	104.91(57)***	.92	.06	
2 factor CU and aggression	83.25(56)*	.95	.05	16.01(1)***
 Age 5 years				
1 factor	677.29(152)***	.91	.07	
2 factor CU and aggression	275.08(151)***	.95	.05	54.94(1)***

***p<.001, *p<.05.

Figure 1: Standardised path estimates for the multivariate model (covariances not shown).

Coefficient presented for boys/girls.

Supplementary material S1: additional information for the CFA and SEM models

To avoid numerical problems associated with sparse data in the multivariate model, where endorsement rates were $< 1.5\%$, scores of 1 and 2 were collapsed to create binary variables. This was applied to the CU traits items ‘cruel to animals’ ‘selfish, and ‘shows little affection’ at 2.5 years and 5 years, and to ‘unresponsive to affection’, ‘volunteers to helps others’ (reverse coded), ‘kind to others’ (reverse coded) and ‘considerate to others’ (reverse coded) at 5 years only. Similarly all of the physical aggression items at both ages, except ‘gets in many fights’ at 2.5 years months which had not received a ‘2’ response, were collapsed to create binary variables. Although this generated adequate cell sizes in the sample as a whole, the analytic approach required adequate numbers in both males and females. In females cell sizes were small for ‘gets in many fights’ at age 2.5 years and ‘bites other children’ at age 5 years. “Gets in many fights” was combined with the similar item “physically attacks others”, and “bites other children” was combined with the next rarest item “kicks other children”. The items were combined at both ages to ensure consistency in the physical aggression latent variable across the two ages. Inter-item correlations were added to the CFA and SEM models for items from the same measure to attempt to account for method effects. The item ‘cruel to animals’ showed a negative residual variance at age 2.5 in the multivariate model and so the variance was fixed to .01.

Table S1: Age 2.5 years CFA models testing measurement invariance across sex

	Parameters	Chi2(df)	<i>p</i>	RMSEA	RMSEA 90% C. I	CFI
Model 1: configural	114	141.70(110)	.023	.05	.02 - .07	.95
Model 2: metric	100	148.17(124)	.069	.04	.00 - .06	.96
Model 3: scalar	83	167.28(141)	.065	.04	.00 - .06	.95
Model 1 vs Model 2		14.89(14)	.444			
Model 1 vs Model 3		31.88(31)	.422			
Model 2 vs Model 3		14.76(17)	.612			

Table S2: Age 5.0 years CFA models testing measurement invariance across sex

	Parameters	Chi2(df)	<i>p</i>	RMSEA	RMSEA 90% C. I	CFI
Model 1a: configural	166	471.88(232)	.001	.05	.05 - .06	.95
Model 1b: configural (modified)	126	199.68(160)	.018	.03	.01 - .04	.99
Model 2: metric	113	195.97(113)	.111	.02	.00 - .03	.99
Model 3a: scalar	92	237.69(194)	.018	.02	.01 - .03	.99
Model 3b: Scalar (modified)	190	218.99	.073	.02	.00 - .03	.99
Model 1b vs Model 2		9.38(13)	.744			
Model 1b vs Model 3a		50.22(21)	.003			
Model 1b vs Model 3b		28.89(3)	.523			
Model 2 vs Model 3b		25.76(17)	.079			