

# Content, diagnostic, correlational, and genetic similarities between common measures of childhood aggressive behaviors and related psychiatric traits

Anne M. Hendriks,<sup>1,2,\*</sup> Hill F. Ip,<sup>1,2,\*</sup> Michel G. Nivard,<sup>1,2</sup> Catrin Finkenauer,<sup>1,3</sup>  
Catharina E.M. Van Beijsterveldt,<sup>1</sup> Meike Bartels,<sup>1,2,\*</sup> and Dorret I. Boomsma<sup>1,2,\*</sup>

<sup>1</sup>Department of Biological Psychology, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands; <sup>2</sup>Amsterdam Public Health Research Institute, Amsterdam, The Netherlands; <sup>3</sup>Department of Interdisciplinary Social Sciences: Youth Studies, Utrecht University, Utrecht, The Netherlands

**Background:** Given the role of childhood aggressive behavior (AGG) in everyday child development, precise and accurate measurement is critical in clinical practice and research. This study aims to quantify agreement among widely used measures of childhood AGG regarding item content, clinical concordance, correlation, and underlying genetic construct. **Methods:** We analyzed data from 1254 Dutch twin pairs (age 8–10 years, 51.1% boys) from a general population sample for whom both parents completed the A-TAC, CBCL, and SDQ at the same occasion. **Results:** There was substantial variation in item content among AGG measures, ranging from .00 (i.e., mutually exclusive) to .50 (moderate agreement). Clinical concordance (i.e., do the same children score above a clinical threshold among AGG measures) was very weak to moderate with estimates ranging between .01 and .43 for mother-reports and between .12 and .42 for father-reports. Correlations among scales were weak to strong, ranging from .32 to .70 for mother-reports and from .32 to .64 for father-reports. We found weak to very strong genetic correlations among the measures, with estimates between .65 and .84 for mother-reports and between .30 and .87 for father-reports. **Conclusions:** Our results demonstrated that degree of agreement between measures of AGG depends on the type (i.e., item content, clinical concordance, correlation, genetic correlation) of agreement considered. Because agreement was higher for correlations compared to clinical concordance (i.e., above or below a clinical cutoff), we propose the use of continuous scores to assess AGG, especially for combining data with different measures. Although item content can be different and agreement among observed measures may not be high, the genetic correlations indicate that the underlying genetic liability for childhood AGG is consistent across measures. **Keywords:** Childhood aggressive behavior; item overlap; clinical concordance; genetic correlation.

## Introduction

Childhood aggressive behaviors have a large impact on the child itself, its family members and society as a whole and are associated with adverse outcomes, including high co-occurrence with other behavioral and emotional problems (Bartels et al., 2018), negative consequences for parents (Meltzer, Ford, Goodman, & Vostanis, 2011; Roberts, McCrory, Joffe, de Lima, & Viding, 2017), and high financial costs for society (Rivenbark et al., 2018; Romeo, Knapp, & Scott, 2006). To enlighten clinicians and researchers on the etiology of AGG, we need instruments for measurement that may, for example, be used in epidemiological settings. In order to generalize among epidemiological studies and to combine data, it is important to understand the extent to which different measurement instruments agree on their content, whether the same children would receive a diagnosis, on the rank order of less and more aggressive children and on whether they assess the same underlying genetic construct.

There is not a single gold standard for the assessment of aggressive behaviors. An additional

complexity is that aggressive behavior is a very broad construct that may express itself through overt behaviors such as fighting or disobedience, but also covert behaviors such as gossiping or stealing (Achenbach & Rescorla, 2001; Björkqvist, Lager-spetz, & Kaukiainen, 1992; Goodman, 2001; Vail-lancourt, Brendgen, Boivin, & Tremblay, 2003). Aggressive behavior is considered a behavior problem in its own or may present as a symptom of disruptive behavior disorders. For instance, conduct disorder (CD) and oppositional defiant disorder (ODD) are, according to the fifth edition of the Diagnostic and Statistical Manual of Mental Disorders, characterized by aggressive and antisocial behavior (American Psychiatric Association, 2013). CD and ODD have different symptoms (i.e., violation of social norms and rules vs. defiance and hostility; American Psychiatric Association, 1994; Loeber, Burke, Lahey, Winters, & Zera, 2000), with ODD in many cases preceding CD, and sometimes considered a milder form of CD (e.g., Quay & Stringaris, 2012). CD and ODD have a prevalence around 2.1% and 3.6%, respectively, and are more prevalent in boys than in girls (American Psychiatric Association, 1994; Polanczyk, Salum, Sugaya, Caye, & Rohde, 2015; Quay & Stringaris, 2012; Scott, 2012). The

\*Shared first and shared last authors.

Conflict of interest statement: No conflicts declared.

present study focuses on aggressive behavior and related psychiatric traits (AGG). AGG also presents in childhood disorders such as attention-deficit hyperactivity disorder, autism, or mania, but in these disorders it is not part of the primary disorder (e.g., mania is primarily a mood disorder; American Psychiatric Association, 1994; Hofvander, Ossowski, Lundström, & Anckarsäter, 2009).

A variety of common screening instruments assess AGG or aggression-related disorders (i.e., CD, ODD). The present study focuses on the Autism – Tics, attention-deficit hyperactivity disorder, and other comorbidities (A-TAC; Hansson et al., 2005), the Child Behavior Checklist (CBCL; Achenbach, Ivanova, & Rescorla, 2017; Achenbach & Rescorla, 2001), and the Strengths and Difficulties Questionnaire (SDQ; Goodman, 2001; Goodman & Scott, 1999). These instruments can be applied in large population samples (Anckarsäter et al., 2011; Van Beijsterveldt et al., 2013; Haworth, Davis, & Plomin, 2013; Jaddoe et al., 2012), either through administration of a questionnaire or (telephone) interview (e.g., Achenbach et al., 2017; Halleröd et al., 2010; Michelson, Davenport, Dretzke, Barlow, & Day, 2013; Stone, Otten, Engels, Vermulst, & Janssens, 2010). Many other instruments exist to assess AGG, among which the Behavior Assessment System for Children (Sandoval & Echandia, 1994), the Eyberg Child Behavior Inventory (Eyberg & Ross, 1978), and the Diagnostic Interview Schedule for Children (Shaffer et al., 1993). Different AGG instruments sometimes have been combined in studies with data from multiple research groups, implicitly assuming that they measure the same underlying construct (e.g., Malanchini et al., 2018; Porsch et al., 2016).

In this study, we aim to quantify to what extent different instruments may assess the same construct of childhood AGG. To gain a better understanding of the agreement among AGG measures, we will examine this question on different levels, namely in terms of item content, clinical concordance, correlations among scales, and genetic correlations.

### Item content

Previous work on convergence between AGG measures mainly focused on agreement between scores, but did not test similarity in item content (Goodman & Scott, 1999; Halleröd et al., 2010). The AGG measures in the present study have been developed by different approaches, which also reflects in their item content. The CBCL has an 18-item Aggressive Behavior (AGG) subscale that was derived from factor analysis on large sets of items. It measures symptoms such as arguing, fighting, and disobedience (Achenbach & Rescorla, 2001). The SDQ was developed to be a brief questionnaire and contains a 5-item scale to assess conduct problems (CP), based on criteria from the DSM-IV (Diagnostic and Statistical Manual of Mental Disorders – fourth edition)

and ICD-10 (International Statistical Classification of Diseases and Related Health Problems – 10th). The SDQ-CP scale assesses symptoms such as fighting, disobedience, and lying (Goodman, 2001). The A-TAC is a diagnostic instrument for screening in general populations and contains two subscales to assess CD and ODD based on DSM-IV symptoms. The A-TAC CD scale measures symptoms such as lying, fighting, and stealing; the A-TAC ODD scale measures symptoms such as being angry, arguing, and teasing (Hansson et al., 2005). DMS-IV criteria distinguish between ODD and CD, and if children meet criteria for both, they receive a diagnosis of CD, and thus, the two diagnoses are mutually exclusive (American Psychiatric Association, 1994). Therefore, the A-TAC CD and ODD subscales, directly based on the DSM-IV criteria, measure mutually exclusive disorders, without overlap in item content. The A-TAC CD scale and SDQ-CP scale both assess symptoms related to CD, implying high item content overlap. The CBCL-AGG subscale assesses broad-sense aggression, which suggests neither absence of overlap nor very high overlap with other measures.

### Clinical concordance

Different AGG measures have scale-specific thresholds to distinguish between clinical and nonclinical levels of aggressive behavior, without any certainty, however, that the same child would receive a diagnosis across measures. Nevertheless, the mutual exclusivity of CD and ODD suggests low clinical concordance between the CD and ODD scales from the A-TAC. Prior work found that the CBCL and SDQ discriminated equally well between 4- and 7-year-old children collected at a dental clinic and children referred for externalizing behaviors at psychiatric clinics ( $N = 132$ ; Goodman & Scott, 1999) and between children aged 4 and 16 years from a community sample and psychiatric clinics ( $N = 273$ ; Klasen et al., 2000). These findings suggest good clinical concordance between the CBCL and SDQ AGG measures, for the purpose of distinguishing between different groups, but do not necessarily generalize to concordances within a group of children. To our knowledge, clinical concordance between scales of the A-TAC and other AGG measures has not been examined. Research comparing clinical levels of AGG assessed using the CBCL-AGG scale with DSM-III (Diagnostic and Statistical Manual of Mental Disorders – third edition) diagnoses of CD and ODD, however, found point-biserial correlations of, respectively, .22 and .57 (Gould, Bird, & Jaramillo, 1993), suggesting higher clinical concordance of the CBCL with the DSM-based A-TAC ODD scale than with the DSM-based A-TAC CD scale. Prior work on clinical concordance did not explicitly compare clinical decisions between different AGG measures.

AGG is heterogeneous, which makes it possible that children who score above a clinical cutoff

(within or among measures) have no symptoms in common (e.g., Krueger, Watson, & Barlow, 2005). As a consequence, it is uncertain whether clinical concordance translates into strong correlation among AGG measures. Currently, a shift is taking place from categorical to dimensional diagnoses, which we also take into account through consideration of agreement among AGG measures with regard to correlation.

## Correlations

Previous research established correlations between measures of AGG. For instance, correlations between the A-TAC CD and ODD and the CBCL-AGG scale in a sample of 106 twin pairs aged 9/12 years were, respectively, .48 and .32, indicating moderate convergence (Halleröd et al., 2010). Between the CBCL Externalizing scale (i.e., AGG and Rule Breaking) and the SDQ-CP scale, correlations ranged from .71 to .84, in samples of 132 children aged 4–7 years, 292 children in child welfare aged 3–12 years, and 287 children aged 8–16 years (Goodman & Scott, 1999; Janssens & Deboutte, 2009; Van Widenfelt, Goedhart, Treffers, & Goodman, 2003). The previously found moderate to high correlation across scales suggests agreement among AGG measures on who receives a higher score.

## Genetic architecture

A wide body of literature reports AGG to have a heritability of around 50%, very much regardless of the diagnostic scheme or instrument used (Burt, 2009; Dick, Viken, Kaprio, Pulkkinen, & Rose, 2005; Hudziak, Derks, Althoff, Copeland, & Boomsma, 2005; Odintsova et al., 2019; Tuvblad & Baker, 2011; Waltes, Chiocchetti, & Freitag, 2016). If scores deriving from different instruments are influenced by genes, as indicated by the significant heritability estimates (see also Kerekes et al., 2014; Porsch et al., 2016), we can ask the question if AGG assessed by different instruments reflects a common underlying genetic construct. A main aim of this paper therefore was to estimate the genetic correlations among AGG measures. To address this question, data are needed that come from a genetically informative sample, such as twins. Here, we have at our disposal a large data set collected on 9-year-old twin pairs, whose parents reported on their children's AGG on the same occasion by completing three questionnaires, that is, CBCL, SDQ, and A-TAC. Data were collected in 1254 mono- and dizygotic twin pairs, whose mothers and fathers completed the same set of questionnaires. These data allow us to employ a multivariate genetic model to estimate the heritability and genetic correlations among different scales. Information for estimating heritability comes from the comparison in resemblance for mono- and dizygotic (MZ and DZ) twins, usually summarized in correlations. Likewise,

comparing the cross-trait resemblance in MZ and DZ pairs (e.g., AGG assessed by CBCL in one twin and AGG assessed by SDQ in the co-twin) informs on the genetic correlation between the two scales.

## Method

### Subjects

The sample comprised 2,508 children (1,254 twin pairs) aged 8 to 10 years old (51.1% boys) born between September 2005 and October 2008 from the Netherlands Twin Register (NTR; Van Beijsterveldt et al., 2013), a nation-wide population-based register founded in 1987. In 2016, both parents of these twin pairs were invited to complete a single survey, which included several measures of AGG (i.e., A-TAC, CBSL, SDQ). Mothers reported on at least one measure for 2,405 children, fathers for 1,613 children. Some families had multiple sets of twins; here, we included one twin pair per family, yielding a sample of 1,240 twin pairs of which 486 were monozygotic (MZ) and 754 dizygotic (DZ). For 47% of same-sex pairs, zygosity was based on DNA testing. Data collection was approved by the Medical Ethical Review Committee of the VU University Medical Center Amsterdam; informed consent from participants was appropriately obtained. The research was conducted according to the principles of the Declaration of Helsinki.

### Instruments

**CBCL.** The Aggressive Behavior syndrome (CBCL-AGG) subscale from the CBCL consisted of 18 items, asking parents to report on their children's behaviors in the past six months. Response categories included 0 = 'Not true', 1 = 'Sometimes or somewhat true', or 2 = 'Very true or often true' (Achenbach & Rescorla, 2001). Children with more than three missing items were excluded from analyses. We considered T-scores of 65 or higher as elevated and indicative of a possible clinical diagnosis (Achenbach & Rescorla, 2001).

**SDQ.** The Conduct Problem subscale from the SDQ (SDQ-CP) consisted of five items asking parents to report on their children's behavior. Parents could respond with 0 = 'Not true', 1 = 'Somewhat true', or 2 = 'Certainly true' (Goodman, 1997, 2001). Children with more than two items missing were excluded from analyses. Scores above 3 revealed elevated levels that may indicate clinical diagnosis (Goodman, 1997).

**A-TAC.** Two scales from the A-TAC assessed AGG, namely the CD (A-TAC-CD) scale and ODD (A-TAC-ODD) scale. Both consisted of five items; parents were asked whether their children displayed the problem behaviors more frequently than peers in any period of their life. Response categories were 0 = 'No', 0.5 = 'Yes, to some extent', or 1 = 'Yes' (Hansson et al., 2005). Children with more than a single item missing were excluded from analyses. Scores higher than 1.5 on the A-TAC-CD and 2.5 on the A-TAC-ODD reflected elevated levels, indicative of a possible clinical diagnosis (Kerekes et al., 2014).

### Analyses

**Item content.** We examined similarity in item content of AGG measures using the Jaccard index and added the DSM-IV criteria for CD (DSM-CD) and ODD (DSM-ODD) as a benchmark (American Psychiatric Association, 1994). Table S1 displays all items. Together, the scales from the AGG measures and the DSM-IV criteria comprised 55 items assessing 26 different AGG symptoms. Symptoms were coded as present (i.e., 1) or absent (i.e., 0). If multiple items tapped the same symptom, we considered them

as a single item. To examine agreement in item content, we calculated the Jaccard index, ranging from 0 (i.e., not similar) to 1 (i.e., fully similar). This index calculates similarity by dividing the overlap in symptoms between two measures by the total number of symptoms (i.e., number of overlapping items and number of symptoms unique to both measures). In line with Fried (2017), we used the following interpretation: very weak = 0.00–0.19, weak = 0.20–0.39, moderate = 0.40–0.59, strong = 0.60–0.79, very strong = 0.80–1.00.

**Clinical concordance and correlation.** We tested agreement for nonclinical vs. clinical diagnoses among scales with Cohen's Kappa (Landis & Koch, 1977). To assess correlations among AGG measures, we first calculated Pearson's correlations on the continuous scores. Additionally, to analyze the full continuous range of scores while taking the skewed distribution of AGG into account, we calculated Spearman's rank correlations (Spearman, 1904). Bootstrapping with 1,000 repetitions provided 95% confidence intervals for these correlations with the RVAideMemoire package (Herv, 2018). Next, we computed polychoric correlations. These represent the correlations between the two latent normally distributed liabilities that underlie the observed variables. Polychoric correlations were estimated for AGG measures that were categorized into three categories: 0, 0.5/1 (i.e., 0.5 for A-TAC-CD and A-TAC-ODD, and 1 for CBCL-AGG and SDQ-CP), and higher (Table S2 presents frequencies). This categorization was chosen to obtain as many thresholds as possible without encountering numerical problems. We estimated polychoric correlations with the polycor package (Fox, 2016). Confidence intervals were based on standard errors. We interpreted clinical concordance and correlation as following: very weak = 0.00–0.19, weak = 0.20–0.39, moderate = 0.40–0.59, strong = 0.60–0.79, very strong = 0.80–1.00 (Landis & Koch, 1977; Spearman, 1904). All clinical concordance and correlation analyses were performed separately for fathers and mothers and for boys and girls. Missing data were deleted list-wise.

**Genetic analyses.** We performed twin analyses, which leveraged the resemblance between MZ and DZ twins to estimate the contribution of additive genetic factors (A), shared environment (C) common to children from the same family, or nonshared environment (E) to individual differences in AGG (Boomsma, Busjahn, & Peltonen, 2002; Kendler, Neale, Kessler, Heath, & Eaves, 1992). Multivariate twin analysis investigates whether traits are influenced by the same genetic or environmental factors, reflected in genetic and environmental correlations. Analyses were carried out in R version 3.5.1, using the OpenMx package (version 2.11.5; Neale et al., 2016) specifying NPSOL optimizer. Confidence intervals were calculated using MxCI in OpenMx. We fitted a multivariate model to continuous scores of the four AGG measures. Guided by prior work, we allowed for sex differences on the mean, but not on the genetic architecture (Porsch et al., 2016; Vink et al., 2012). Analyses were performed separately for mother and father ratings of AGG. The variance and covariance of the measures were partitioned into components explained by A, C, and E. Because the model assumes that the data follow a multivariate normal distribution, but the measures of AGG are skewed, we may introduce bias (Derks, Dolan, & Boomsma, 2004). We therefore performed multivariate genetic models on the ordinal data as a sensitivity check (see Appendix S1).

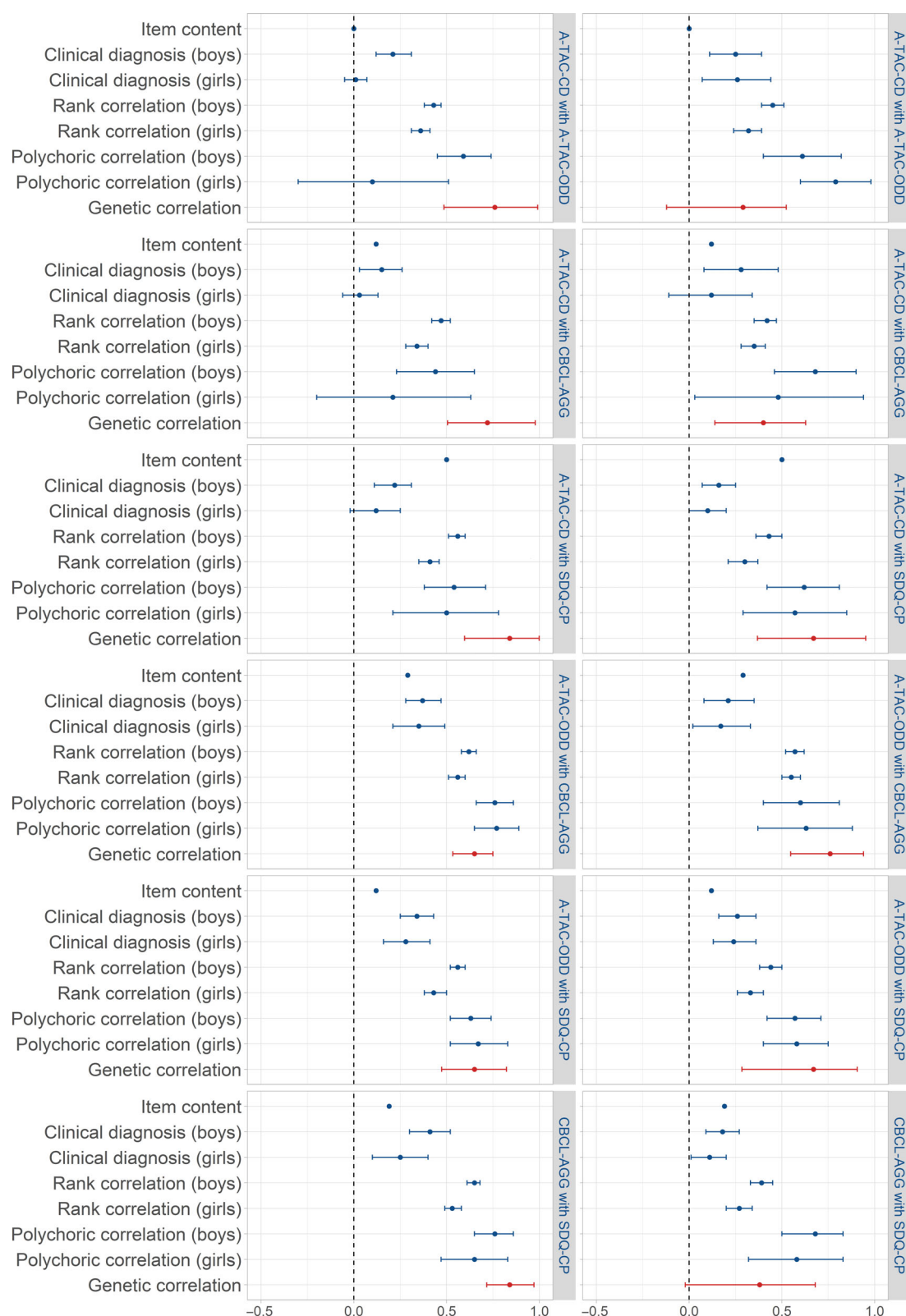
## Results

### Item content

Table 1 summarizes outcomes of the Jaccard analyses. All measures of agreement are displayed in Figure 1. Overlap was absent between DSM-CD and A-TAC-ODD, DSM-ODD and A-TAC-CD, and A-TAC-CD and A-TAC-ODD. Overlap between DSM-CD and A-TAC-CD was weak, and overlap between

**Table 1** Jaccard index for item overlap between the different AGG measures and DSM-IV criteria [Colour table can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

	DSM-CD	DSM-ODD	A-TAC-CD	A-TAC-ODD	CBCL-AGG	SDQ-CP
DSM-CD						
DSM-ODD	0					
A-TAC-CD	0.29	0				
A-TAC-ODD	0	0.57	0			
CBCL-AGG	0.17	0.31	0.12	0.29		
SDQ-CP	0.19	0.2	0.5	0.12	0.19	



**Figure 1** Agreement between AGG measures; the left panel is for mother-reports, the right panel for father-reports. Because shared environmental correlations were less stable due to sample size, we did not include them in the figure [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

DSM-ODD and A-TAC-ODD was moderate. SDQ-CP had moderate overlap with A-TAC-CD but very weak overlap with DSM-CD. Overlap between DSM-ODD and SDQ-CP was weak. Overlap between A-TAC-ODD and SDQ-CP was very weak.

CBCL-AGG had very weak to weak overlap with all other scales (i.e., lowest overlap with A-TAC-CD, highest with DSM-ODD). Altogether, overlap in item content ranged from no overlap to moderate overlap.



### Clinical concordance

Prevalences of clinical AGG for boys for mother-report ranged from 3% (A-TAC-CD) to 11% (CBCL-AGG); for father-report, they ranged from 2% (A-TAC-CD) to 9% (CBCL-AGG). For girls, prevalences for mother-report ranged from 1% (A-TAC-CD) to 6% (CBCL-AGG). For father-report on girls, prevalences ranged from 1% (A-TAC-CD) to 7% (CBCL-AGG; see Table S3).

Clinical concordance was very weak to weak between A-TAC-CD and A-TAC-ODD (.01 for mother-report on girls to .26 for father-report on girls). Clinical concordance between A-TAC-CD and CBCL-AGG ranged from .06 (mother-report on girls) to .13 (father-report on boys); between A-TAC-ODD and CBCL-AGG, it ranged from .39 (father-report on boys) to .43 (mother-report on girls). Clinical concordance between A-TAC-CD and SDQ-CP ranged from .12 (mother- and father-report on girls) to .29 (father-report on boys). Clinical concordance between A-TAC-ODD and SDQ-CP ranged from .28 (mother-report on girls) to .40 (father-report on girls). Clinical concordance between CBCL-AGG and SDQ-CP ranged from .28 (mother-report on girls) to .42 (father-report on boys). Overall, clinical concordance between measures of AGG was very weak to moderate (see Table S4).

### Correlation

Pearson correlations between A-TAC-CD and A-TAC-ODD were weak (.34; mother-report, girls) to moderate (.46; father-report, boys). Between A-TAC-CD and CBCL-AGG, correlations were weak (.32; mother-report, girls) to moderate (.43; father-report, boys). Correlations between A-TAC-ODD and CBCL-AGG were strong, ranging from .63 (father-report, boys) to .66 (mother-report, boys). Correlations between A-TAC-CD and SDQ-CP ranged from weak (.35; father-report, girls) to moderate (.52; father-report, boys). Between A-TAC-ODD and SDQ-CP, correlations were moderate, ranging from .49 (mother-report, girls) to .60 (father-report, boys). Correlations between CBCL-AGG and SDQ-CP ranged from moderate (.58; father-report, girls) to strong (.70; mother-report, boys; see Table S5). Overall, correlations among AGG measures were weak to strong.

Spearman rank correlations between A-TAC-CD and A-TAC-ODD were weak (.32; father-report, girls) to moderate (.45; father-report, boys). Rank correlations between A-TAC-CD and CBCL-AGG ranged from .34 (mother-report, girls) to .47 (mother-report, boys). Rank correlations between A-TAC-ODD and CBCL-AGG ranged from .55 (father-report, girls) to .62 (mother-report, boys). For A-TAC-CD and SDQ-CP, correlations ranged from .39 (father-report, girls) to .56 (mother-report, boys); for A-TAC-ODD and SDQ-CP, correlations ranged from .40 (father-report,

girls) to .56 (mother-report, boys). Rank correlations between CBCL-AGG and SDQ-CP ranged from .47 (father-report, girls) to .65 (mother-report, boys; see Table S6). Altogether, rank correlations indicated weak to strong agreement between AGG measures.

Polychoric correlations between A-TAC-CD and A-TAC-ODD ranged from .50 (father-report, girls) to .61 (father-report, boys). For A-TAC-CD with CBCL-AGG, polychoric correlations ranged from .51 (mother-report, girls) to .58 (mother-report, boys). Associations between A-TAC-ODD and CBCL-AGG ranged between .56 (father-report, girls) and .59 (mother-report, girls). Between A-TAC-CD and SDQ-CP, associations ranged from .60 (father-report, girls) to .70 (mother-report, boys). Between A-TAC-ODD and SDQ-CP, polychoric correlations ranged from .40 (father-report, girls) to .61 (mother-report, boys). Finally, polychoric correlations between CBCL-AGG and SDQ-CP ranged from .50 (father-report, girls) to .66 (mother-report, boys; see Table S7). Overall, polychoric correlations indicated moderate to strong agreement between AGG measures.

### Genetic analyses

Cross-twin cross-instrument correlations for MZ and DZ twins between A-TAC-CD, A-TAC-ODD, CBCL-AGG, and SDQ-CP for mother- and father-report are presented in Table S8. Table S9 contains estimates of the means and variances, and sex differences in the means. For mother-report, the contribution of additive genetic factors to the variances for A-TAC-CD, A-TAC-ODD, CBCL-AGG, and SDQ was, respectively, 34%, 42%, 61%, and 42%. Common environment explained, respectively, 4%, 19%, 13%, and 9% and unique environment 62%, 39%, 26%, and 50%. For father-report, the contribution of A to the variances for A-TAC-CD, A-TAC-ODD, CBCL-AGG, and SDQ was, respectively, 39%, 34%, 45%, and 33%; C explained 11%, 26%, 25%, and 14%; E contributed 49%, 40%, 30%, and 53%. For mother-report, the covariance between the different measures was moderately to strongly accounted for by genetic factors, namely 50% (A-TAC-ODD and SDQ-CP) to 83% (A-TAC-CD and CBCL-AGG). For father-reports, genetic covariance ranged from 26% (A-TAC-CD and A-TAC-ODD) to 57% (A-TAC-CD and SDQ-CP). For mother-reports, common environmental factors explained up to 22% of the covariance (A-TAC-ODD and CBCL-AGG) between the different measures. For father-reports, the common environmental covariance ranged from 21% (A-TAC-CD and SDQ-CP) to 37% (A-TAC-CD and A-TAC-ODD). Nonshared environment weakly explained covariance for mother-reports between different measures, namely between 20% (A-TAC-CD and CBCL-AGG) and 32% (A-TAC-ODD and SDQ-CP). For father-reports, nonshared environment explained between 22% (A-TAC-CD and SDQ-CP) and 37% (A-TAC-CD and A-TAC-ODD) of

the covariance among AGG measures (see Table 2 and Figure 1).

For mother-report, genetic correlations ranged from .65 (95%CI = .53–.75; A-TAC-ODD and CBCL-AGG) to .84 (95%CI = .60–1.00; CBCL-AGG and SDQ-CP). For father-reports, genetic correlations ranged from .31 (95%CI = –0.08–0.55; A-TAC-CD and A-TAC-ODD) to .87 (95%CI = 0.56–0.98; A-TAC-ODD and SDQ-CP). Correlations between common environmental influences on AGG scores for mother-reports ranged from –.16 (95%CI = –1.00–1.00; A-TAC-CD with CBCL-AGG) to .90 (95%CI = 0.63–1.00; A-TAC-ODD with the CBCL-AGG). For father-reports, they ranged from .66 (95%CI = 0.14, 1.00; A-TAC-ODD and SDQ-CP) to .97 (95%CI = 0.67–1.00; A-TAC-CD and A-TAC-ODD). When the contribution of C to variance is small, its contribution to the covariance between measures is also small and estimates of correlations may be imprecise. Correlations between the nonshared environmental influences on the scale scores for mother-report varied between .19 (95%CI = 0.09–0.28; A-TAC-CD with A-TAC-ODD) and .55 (95%CI = 0.48–0.61; CBCL-AGG with SDQ-CP). For father-report, they varied between .20 (95%CI = 0.10–0.30; A-TAC-CD and SDQ-CP) and .44 (95%CI = 0.34–0.51; CBCL-AGG and SDQ-CP; see Table 2 and Figure 1). To check for bias, we also conducted categorical twin analyses; results are presented in Appendix S1 and Tables S10 and S11. Generally, genetic correlations were of a similar

strength or stronger compared to the continuous analyses, confirming the results of the continuous analyses.

## Discussion

We aimed to quantify the agreement among four different measures of AGG. To this end, we examined convergence of item content, concordance at the recommended clinical cutoff, correlation between the different scales, and the extent to which they measure the same underlying genetic mechanisms.

Overlap in item content across AGG measures ranged from absent (i.e., mutually exclusive) to moderate. Absence of overlap between A-TAC CD and ODD scales confirmed mutual exclusivity of these psychiatric disorders. Between the different measures, overlap was highest (i.e., moderate) between A-TAC CD scale and SDQ conduct problems (CP); SDQ-CP scale weakly overlapped with the other measures. As expected, CBCL-AGG scale weakly overlapped with all other measures with the strongest overlap for A-TAC ODD scale. This indicated that based on their content, different AGG measures cannot be used interchangeably.

Clinical concordance for the different AGG measures was very weak to moderate. Although the CBCL and SDQ, in prior research, discriminated equally well between children from general population samples and clinical samples (Goodman &

**Table 2** Standardized variance and covariance decomposition into contribution of genetic (A), shared environmental (C), and nonshared environmental (E) factors are presented in the lower triangles

Measure		A-TAC-CD	A-TAC-ODD	CBCL-AGG	SDQ-CP
Mothers					
A	A-TAC-CD	.34 [0.15, 0.46]	.76 [0.49, 0.99]	.72 [0.50, 0.97]	.84 [0.60, 0.99]
	A-TAC-ODD	.74 [0.40, 1.04]	.42 [0.27, 0.58]	.65 [0.53, 0.75]	.65 [0.47, 0.82]
	CBCL-AGG	.83 [0.56, 1.07]	.51 [0.34, 0.68]	.61 [0.49, 0.69]	.84 [0.72, 0.97]
	SDQ-CP	.70 [0.42, 0.93]	.50 [0.28, 0.72]	.63 [0.47, 0.77]	.42 [0.26, 0.54]
C	A-TAC-CD	.04 [0.00, 0.19]	.12 [–1.00, 1.00]	–.16 [–1.00, 1.00]	.31 [–1.00, 1.00]
	A-TAC-ODD	.03 [–0.19, 0.29]	.19 [0.07, 0.32]	.90 [0.63, 1.00]	.78 [0.13, 1.00]
	CBCL-AGG	–.03 [–0.19, 0.18]	.22 [0.08, 0.37]	.13 [0.05, 0.24]	.45 [–0.42, 0.95]
	SDQ-CP	.04 [–0.10, 0.24]	.19 [0.05, 0.36]	.07 [–0.03, 0.21]	.09 [0.01, 0.21]
E	A-TAC-CD	.62 [0.55, 0.89]	.19 [0.09, 0.27]	.19 [0.10, 0.28]	.21 [0.13, 0.30]
	A-TAC-ODD	.23 [0.11, 0.36]	.39 [0.34, 0.44]	.55 [0.48, 0.61]	.40 [0.32, 0.47]
	CBCL-AGG	.20 [0.10, 0.30]	.27 [0.22, 0.33]	.26 [0.23, 0.30]	.55 [0.49, 0.61]
	SDQ-CP	.26 [0.16, 0.38]	.32 [0.25, 0.40]	.30 [0.25, 0.36]	.50 [0.45, 0.56]
Fathers					
A	A-TAC-CD	.39 [0.21, 0.55]	.31 [–0.08, 0.55]	.42 [0.14, 0.65]	.73 [0.46, 0.89]
	A-TAC-ODD	.26 [–0.05, 0.57]	.34 [0.17, 0.52]	.77 [0.56, 0.95]	.87 [0.56, 0.98]
	CBCL-AGG	.44 [0.12, 0.75]	.48 [0.27, 0.69]	.45 [0.30, 0.62]	.77 [0.53, 0.97]
	SDQ-CP	.57 [0.25, 0.85]	.53 [0.26, 0.77]	.49 [0.25, 0.73]	.33 [0.14, 0.52]
C	A-TAC-CD	.11 [0.02, 0.24]	.97 [0.67, 1.00]	.70 [0.16, 1.00]	.78 [0.17, 1.00]
	A-TAC-ODD	.37 [0.13, 0.61]	.26 [0.11, 0.40]	.73 [0.42, 0.96]	.66 [0.14, 1.00]
	CBCL-AGG	.29 [0.04, 0.55]	.29 [0.11, 0.47]	.25 [0.10, 0.39]	.76 [0.22, 1.00]
	SDQ-CP	.21 [0.01, 0.45]	.23 [0.02, 0.44]	.23 [0.03, 0.43]	.14 [0.00, 0.29]
E	A-TAC-CD	.49 [0.42, 0.58]	.37 [0.27, 0.46]	.28 [0.17, 0.38]	.20 [0.10, 0.30]
	A-TAC-ODD	.37 [0.26, 0.49]	.40 [0.34, 0.46]	.43 [0.34, 0.51]	.30 [0.21, 0.39]
	CBCL-AGG	.27 [0.16, 0.39]	.23 [0.17, 0.30]	.30 [0.26, 0.35]	.44 [0.35, 0.52]
	SDQ-CP	.22 [0.10, 0.35]	.25 [0.17, 0.35]	.28 [0.21, 0.37]	.53 [0.46, 0.61]

Genetic, shared environmental, and nonshared environmental correlations are presented in the upper triangles. Results for mother-reports are in the top half and for father-reports in the lower half. Because the parameter estimates on the lower triangles are standardized, the total of A, C, and E adds up to 1.

Scott, 1999; Klasen et al., 2000), their clinical concordance in the present study was weak. This indicates that, although prior work suggested good clinical concordance, they do not agree very well on which children receive a diagnosis in a general population sample. In line with prior work (Gould et al., 1993), clinical concordance of the CBCL-AGG scale with the A-TAC ODD scale was higher (i.e., weak) than with the A-TAC CD scale (i.e., very weak). Similarly, clinical concordance of the SDQ-CP scale was higher with the A-TAC ODD scale (i.e., weak) than with the A-TAC CD scale (i.e., very weak). Despite the mutual exclusivity of the A-TAC CD scale and the A-TAC ODD scale, there was very weak clinical concordance between these measures. These findings revealed that different AGG instruments may result in different clinical decisions with respect to inclusion, exclusion, referral, or treatment.

Pearson correlations and rank correlations among continuous scores of AGG measures were weak to strong, suggesting stronger agreement when considering continuous scores compared to clinical cutoffs. The association between the A-TAC ODD scale and the CBCL-AGG was highest (i.e., moderate); the association between the A-TAC CD scale and the A-TAC ODD scale was lowest (i.e., moderate), yet not absent. Thus, there is overlap between different AGG measures, but they also provide distinct information. Polychoric correlations revealed moderate to strong agreement between the different AGG measures. Strongest agreement was between the A-TAC CD scale and the SDQ-CP scale. Agreement between the A-TAC ODD scale and the SDQ-CP scale was weakest, but still moderate at least. These results reveal that agreement between the different AGG measures based on continuous scores yields higher agreement than clinical cutoff scores.

Our results demonstrate in the assessment of AGG, and diagnosis largely depends on the measure, whereas measures converge moderately to strongly on who receives a higher AGG score. There are several other arguments in favor of a continuous approach to the assessment of AGG. For instance, fluctuations above and below clinical thresholds across development may cause children to not receive treatment although they might score above-threshold at another age (Biederman, Mick, Faraone, & Burback, 2001). In addition, similar to clinical AGG, subthreshold AGG associates with adverse outcomes, and therefore, it is beneficial to detect heightened, yet subthreshold, levels of AGG (Fatori et al., 2018). An earlier diagnosis is associated with better longitudinal outcomes, suggesting additional benefits from the detection of subthreshold AGG (Campbell, Lundstrom, Larsson, Lichtenstein, & Lubke, 2018). Therefore, we propose use of continuous scores to assess AGG, especially when combining data with different measures.

Genetic correlations generally indicated substantial overlap (i.e., strong to very strong for mother-

report ranging between .65 and .84, and weak to very strong for father-report, ranging between .31 and .87) in underlying genetic liability among the AGG measures. This was especially the case for mother-reports, for which there also was a larger sample size. These findings are important for ongoing research into the genetic basis of AGG, for example in meta-analyses of genome-wide association results across multiple cohorts because they suggest that different measures of AGG may be combined in such genetic association analyses.

Our results, especially from the correlation and genetic analyses, suggest overlap in the constructs assessed by the different AGG measures. These findings are in line with prior work, suggesting that disruptive behavior problems partly share an underlying liability (Burt, 2009; Dick et al., 2005). Despite their mutual exclusivity by definition, ODD is predictive of CD later in life (Burke, Loeber, Lahey, & Rathouz, 2005) and research found a correlation around .50 between these disorders in a general population sample (Bartels et al., 2018). Depending on the research question (e.g., comparing or predicting diagnoses vs. prediction of higher scores or understanding the genetic architecture), research using these different AGG measures can be combined.

### *Strengths and limitations*

Because participants were twins, the correlations among measures could be decomposed into parts explained by genetic and environmental factors. Additionally, we considered possible bias induced by the skewness of AGG by conducting sensitivity analyses.

Our study also has limitations. First, the instruments in the present study to measure AGG assess a heterogeneous set of constructs (i.e., CD, ODD, AGG, CP), and it is uncertain to what extent the results in the present study may generalize to other AGG instruments. Second, the order of items was the same for all participants (Brace, 2008). Parents may interpret questions in light of prior questions, which may cause them to structurally respond more positively or negatively in the beginning compared to the end of the questionnaire. Prior work found varying covariances among measures when assessed in different orders (Weinberger, Darkes, Del Boca, Greenbaum, & Goldman, 2006). Changing order, however, might induce random error. Nonetheless, analyses in the present study may under- or overestimate agreement between AGG measures due to the same order of items for all participants. In addition, we analyzed data for one age-group and cannot make inferences on all of childhood. Because AGG expresses itself differently with age, agreement among measures may vary across development. Nonetheless, stability in the underlying genetic mechanisms of AGG (Porsch et al., 2016; Wichers



et al., 2013) suggests stability of genetic correlations among AGG measures across development.

## Conclusion

We considered several definitions of agreement to compare AGG instruments. Across definitions, conclusions as to whether the instruments measure the same construct differ. For example, item content suggests limited overlap, whereas genetic analyses suggest shared etiology among instruments. Whether researchers regard agreement between instruments as satisfactory depends on the application. It is recommended to consider multiple metrics of similarity to decide whether different measures assess the same. By leveraging a genetically informative design and several commonly used instruments, we attempted to provide a holistic perspective on nuances involved in measurement of AGG in childhood.

## Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article:

**Appendix S1.** Description sensitivity analysis.

**Table S1.** Item content for all included AGG measures and DSM-IV criteria, organized per symptom.

**Table S2.** Frequencies of categorical responses for the polychoric correlations and genetic analyses on categorical data.

**Table S3.** Prevalences of clinical levels of AGG across all measures, for boys and girls. Results for mother-reports are in the top half, for father-reports in the lower half.

**Table S4.** Cohen's Kappa to assess clinical concordance among AGG measures with 95% confidence intervals.

**Table S5.** Pearson's correlation to assess correlation among AGG measures with 95% confidence intervals.

**Table S6.** Spearman's rank correlation to assess correlation among AGG measures while accounting for skewness with 95% confidence intervals.

**Table S7.** Polychoric correlations between AGG scores categorized into three categories to assess correlation among AGG measures while accounting for skewness with 95% confidence intervals.

**Table S8.** Monozygotic (MZ) and dizygotic (DZ) twin correlations of the saturated model for mother-report on the top half and for father-report on the lower half with 95% confidence intervals.

**Table S9.** Means, variances, and mean differences in AGG between boys and girls from the saturated model reported by mothers in the top half and reported by fathers in the lower half.

**Table S10.** Monozygotic (MZ) and dizygotic (DZ) correlations from the saturated categorical twin models, reported by mothers on the top half and by fathers on the lower half, with 95% confidence intervals.

**Table S11.** Results from genetic analyses on categorical data.

## Acknowledgements

This work was supported by the 'Aggression in Children: Unraveling gene-environment interplay to inform Treatment and Intervention strategies' project (ACTION). ACTION receives funding from the European Union Seventh Framework Program (FP7/2007- 2013) under grant agreement no 602768. The authors have declared that they have no competing or potential conflicts of interest.

## Correspondence

Anne M. Hendriks, Department of Biological Psychology, Vrije Universiteit Amsterdam, Van der Boerhorststraat 7, 1081 BT Amsterdam, The Netherlands; Email: a.m.hendriks@vu.nl

## Key points

- For interpretation of prior research findings and future collaboration projects, it is important to gauge convergence between different measures for childhood aggressive behavior.
- Results reveal great variation in item content. Agreement in clinical scores among measures is weak; correlations are moderate to strong. Despite differences, genetic overlap is strong, suggesting that different measures for childhood aggressive assess a similar genetic construct.
- Higher agreement among continuous scores suggests that decisions regarding referral for treatment or inclusion/exclusion for research are more robust among measures when using a continuous score instead of an indication based on a cutoff.
- Agreement between measures of childhood aggressive behavior depends on the metric of agreement (i.e., item content, clinical concordance, correlation, genetic overlap), which needs to be considered in future research.

## References

- Achenbach, T.M., Ivanova, M.Y., & Rescorla, L.A. (2017). Empirically based assessment and taxonomy of psychopathology for ages 1½–90+ years: Developmental, multi-informant, and multicultural findings. *Comprehensive Psychiatry*, 79, 4–18.
- Achenbach, T.M., & Rescorla, L. (2001). *Manual for the ASEBA school-age forms & profiles*. Burlington, VT: University of Vermont, Research Center for Children, Youth, & Families.
- American Psychiatric Association (1994). *DSM-IV diagnostic and statistical manual of mental disorder* (4th edn, vol. 33). Philadelphia: Author.
- American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders* (5th edn.). Arlington, TX: Author.
- Anckarsäter, H., Lundström, S., Kollberg, L., Kerekes, N., Palm, C., Carlström, E., ... & Lichtenstein, P. (2011). The child and adolescent twin study in Sweden (CATSS). *Twin Research and Human Genetics*, 14, 495–508.
- Bartels, M., Hendriks, A., Mauri, M., Krapohl, E., Whipp, A., Bolhuis, K., ... & Boomsma, D.I. (2018). Childhood aggression and the co-occurrence of behavioural and emotional problems: results across ages 3–16 years from multiple raters in six cohorts in the EU-ACTION project. *European Child and Adolescent Psychiatry*, 27, 1105–1121.
- Biederman, J., Mick, E., Faraone, S.V., & Burbach, M. (2001). Patterns of remission and symptom decline in conduct disorder: A four-year prospective study of an ADHD sample. *Journal of the American Academy of Child and Adolescent Psychiatry*, 40, 290–298.
- Björkqvist, K., Lagerspetz, K.M.J., & Kaukiainen, A. (1992). Do girls manipulate and boys fight? Developmental trends in regard to direct and indirect aggression. *Aggressive Behavior*, 18, 117–127.
- Boomsma, D.I., Busjahn, A., & Peltonen, L. (2002). Classical twin studies and beyond. *Nature Reviews Genetics*, 3, 872–882.
- Brace, I. (2008). *Questionnaire design: How to plan, structure and write survey material for effective market research*. London, UK: Kogan Page Limited.
- Burke, J.D., Loeber, R., Lahey, B.B., & Rathouz, P.J. (2005). Developmental transitions among affective and behavioral disorders in adolescent boys. *Journal of Child Psychology and Psychiatry*, 46, 1200–1210.
- Burt, S.A. (2009). Are there meaningful etiological differences within antisocial behavior? Results of a meta-analysis. *Clinical Psychology Review*, 29, 163–178.
- Campbell, I., Lundström, S., Larsson, H., Lichtenstein, P., & Lubke, G. (2018). The relation between the age at diagnosis of problem behaviors related to aggression and distal outcomes in Swedish children. *European Child and Adolescent Psychiatry*, 28, 899–911.
- Derks, E.M., Dolan, C.V., & Boomsma, D.I. (2004). Effects of censoring on parameter estimates and power in genetic modeling. *Twin Research*, 7, 659–669.
- Dick, D.M., Viken, R.J., Kaprio, J., Pulkkinen, L., & Rose, R.J. (2005). Understanding the covariation among childhood externalizing symptoms: Genetic and environmental influences on conduct disorder, attention deficit hyperactivity disorder, and oppositional defiant disorder symptoms. *Journal of Abnormal Child Psychology*, 33, 219–229.
- Eyberg, S.M., & Ross, A.W. (1978). Assessment of child behavior problems: The validation of a new inventory. *Journal of Clinical Child Psychology*, 7, 113–116.
- Fatori, D., Salum, G., Itria, A., Pan, P., Alvarenga, P., Rohde, L.A., ... & Graeff-Martins, A.S. (2018). The economic impact of subthreshold and clinical childhood mental disorders. *Journal of Mental Health*, 27, 588–594.
- Fox, J. (2016). Polychoric and polyserial correlations: polycor. Version 0.7-9. Available from <http://cran.r-project.org/package=polycor%0D> [last accessed 14 February 2019].
- Fried, E. (2017). The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. *Journal of Affective Disorders*, 208, 191–197.
- Goodman, R. (1997). The strengths and difficulties Questionnaire: A research note. *Journal of Child Psychology and Psychiatry*, 38, 581–586.
- Goodman, R. (2001). Psychometric properties of the strengths and difficulties questionnaire. *Journal of the American Academy of Child and Adolescent Psychiatry*, 40, 1337–1345.
- Goodman, R., & Scott, S. (1999). Comparing the strengths and difficulties questionnaire and the child behavior checklist: Is small beautiful? *Journal of Abnormal Child Psychology*, 27, 17–24.
- Gould, M.S., Bird, H., & Jaramillo, B.S. (1993). Correspondence between statistically derived behavior problem syndromes and child psychiatric diagnoses in a community sample. *Journal of Abnormal Child Psychology*, 21, 287–313.
- Halleröd, S.L.H., Larson, T., Ståhlberg, O., Carlström, E., Gillberg, C., Anckarsäter, H., ... & Gillberg, C. (2010). The autism-Tics, AD/HD and other comorbidities (A-TAC) telephone interview: Convergence with the child behavior checklist (CBCL). *Nordic Journal of Psychiatry*, 64, 218–224.
- Hansson, S.L., Røjvall, A.S., Rastam, M., Gillberg, C., Gillberg, C., & Anckarsäter, H. (2005). Psychiatric telephone interview with parents for screening of childhood autism - Tics, attention-deficit hyperactivity disorder and other comorbidities (A-TAC): Preliminary reliability and validity. *British Journal of Psychiatry*, 187, 262–267.
- Haworth, C.M.A., Davis, O.S.P., & Plomin, R. (2013). Twins Early Development Study (TEDS): A genetically sensitive investigation of cognitive and behavioral development from childhood to young adulthood. *Twin Research and Human Genetics*, 16, 117–125.
- Herv, M. (2018). RVAideMemoire: Testing and plotting procedures for biostatistics. Version 0.9-69-3. Available from <https://rdrr.io/cran/RVAideMemoire/> [last accessed 22 August 2018].
- Hofvander, B., Ossowski, D., Lundström, S., & Anckarsäter, H. (2009). Continuity of aggressive antisocial behavior from childhood to adulthood: The question of phenotype definition. *International Journal of Law and Psychiatry*, 32, 224–234.
- Hudziak, J.J., Derks, E.M., Althoff, R.R., Copeland, W., & Boomsma, D.I. (2005). The genetic and environmental contributions to oppositional defiant behavior: A multi-informant twin study. *Journal of the American Academy of Child and Adolescent Psychiatry*, 44, 907–914.
- Jaddoe, V.W.V., van Duijn, C.M., Franco, O.H., van der Heijden, A.J., van Ijzendoorn, M.H., de Jongste, J.C., ... & Hofman, A. (2012). The Generation R Study: Design and cohort update 2012. *European Journal of Epidemiology*, 27, 739–756.
- Janssens, A., & Deboutte, D. (2009). Screening for psychopathology in child welfare: The Strengths and Difficulties Questionnaire (SDQ) compared with the Achenbach System of Empirically Based Assessment (ASEBA). *European Child and Adolescent Psychiatry*, 18, 691–700.
- Kendler, K., Neale, M., Kessler, R., Heath, A., & Eaves, L. (1992). Major depression and generalized anxiety disorder: Same genes, (partly) different environments? *Archives of General Psychiatry*, 49, 716–722.
- Kerekes, N., Lundström, S., Chang, Z., Tajnia, A., Jern, P., Lichtenstein, P., ... & Anckarsäter, H. (2014). Oppositional defiant- and conduct disorder-like problems: Neurodevelopmental predictors and genetic background in boys and girls, in a nationwide twin study. *PeerJ*, 2, e359.
- Klasen, H., Woerner, W., Wolke, D., Meyer, R., Overmeyer, S., Kaschnitz, W., ... & Goodman, R. (2000). Comparing the German versions of the Strengths and Difficulties Questionnaire (SDQ-Deu) and the Child Behavior Checklist. *European Child and Adolescent Psychiatry*, 9, 271–276.

- Krueger, R.F., Watson, D., & Barlow, D.H. (2005). Introduction to the special section: Toward a dimensionally based taxonomy of psychopathology. *Journal of Abnormal Psychology*, 114, 491–493.
- Landis, J.R., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159–174.
- Loeber, R., Burke, J.D., Lahey, B.B., Winters, A., & Zera, M. (2000). Oppositional defiant and conduct disorder: A review of the past 10 years, Part I. *Journal of the American Academy of Child & Adolescent Psychiatry*, 39, 1468–1484.
- Malanchini, M., Smith-Woolley, E., Ayorech, Z., Rimfeld, K., Krapohl, E., Vuoksima, E., ... & Plomin, R. (2018). Aggressive behaviour in childhood and adolescence: The role of smoking during pregnancy, evidence from four twin cohorts in the EU-ACTION consortium. *Psychological Medicine*, 49, 646–654.
- Meltzer, H., Ford, T., Goodman, R., & Vostanis, P. (2011). The burden of caring for children with emotional or conduct disorders. *International Journal of Family Medicine*, 2011, 1–8.
- Michelson, D., Davenport, C., Dretzke, J., Barlow, J., & Day, C. (2013). Do evidence-based interventions work when tested in the “real world?” A systematic review and meta-analysis of parent management training for the treatment of child disruptive behavior. *Clinical Child and Family Psychology Review*, 16, 18–34.
- Neale, M.C., Hunter, M.D., Pritikin, J.N., Zahery, M., Brick, T.R., Kirkpatrick, R.M., ... & Boker, S.M. (2016). OpenMx 2.0: Extended structural equation and statistical modeling. *Psychometrika*, 81, 535–549.
- Odintsova, V.V., Roetman, P.J., Ip, H.F., Pool, R., Van der Laan, C.M., Tona, D.K., ... & Boomsma, D.I. (2019). Genomics of human aggression: Current state of genome-wide studies and an automated systematic review tool. *Psychiatric Genetics*, 29, 170–190.
- Polanczyk, G.V., Salum, G.A., Sugaya, L.S., Caye, A., & Rohde, L.A. (2015). Annual research review: A meta-analysis of the worldwide prevalence of mental disorders in children and adolescents. *Journal of Child Psychology and Psychiatry*, 56, 345–365.
- Porsch, R.M., Middeldorp, C.M., Cherny, S.S., Krapohl, E., van Beijsterveldt, C.E.M., Loukola, A., ... & Bartels, M. (2016). Longitudinal heritability of childhood aggression. *American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics*, 171, 697–707.
- Quy, K., & Stringaris, A. (2012). Oppositional defiant disorder. In J.M. Rey (Ed.), *IACAPAP e-Textbook of child and adolescent mental health*. Geneva, Switzerland: International Association for Child and Adolescent Psychiatry and Allied Professions.
- Rivenbark, J.G., Odgers, C.L., Caspi, A., Harrington, H.L., Hogan, S., Houts, R.M., ... & Moffitt, T.E. (2018). The high societal costs of childhood conduct problems: Evidence from administrative records up to age 38 in a longitudinal birth cohort. *Journal of Child Psychology and Psychiatry*, 59, 703–710.
- Roberts, R., McCrory, E., Joffe, H., de Lima, N., & Viding, E. (2017). Living with conduct problem youth: Family functioning and parental perceptions of their child. *European Child and Adolescent Psychiatry*, 27, 1–10.
- Romeo, R., Knapp, M., & Scott, S. (2006). Economic cost of severe antisocial behaviour in children—and who pays it. *The British Journal of Psychiatry: The Journal of Mental Science*, 188, 547–553.
- Sandoval, J., & Echandia, A. (1994). Behavior assessment system for children. *Journal of School Psychology*, 32, 419–425.
- Scott, S. (2012). Conduct disorders. In J.M. Rey (Ed.), *IACAPAP e-Textbook of child and adolescent mental health*. Geneva, Switzerland: International Association for Child and Adolescent Psychiatry and Allied Professions.
- Shaffer, D., Schwab-Stone, M., Fisher, P., Cohen, P., Piacenti, J.C., Davies, M., ... & Regier, D. (1993). The Diagnostic Interview Schedule for Children-Revised version (DISC-R): I. Preparation, field testing, interrater reliability, and acceptability. *Journal of American Academy of Child & Adolescent Psychiatry*, 32, 643–650.
- Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, 15, 72–101.
- Stone, L.L., Otten, R., Engels, R.C.M.E., Vermulst, A.A., & Janssens, J.M.A.M. (2010). Psychometric properties of the parent and teacher versions of the strengths and difficulties questionnaire for 4- to 12-Year-olds: A review. *Clinical Child and Family Psychology Review*, 13, 254–274.
- Tuvblad, C., & Baker, L.A. (2011). Human aggression across the lifespan: Genetic propensities and environmental moderators. *Advances in Genetics*, 75, 171–214.
- Vaillancourt, T., Brendgen, M., Boivin, M., & Tremblay, R.E. (2003). A longitudinal confirmatory factor analysis of indirect and physical aggression: Evidence of two factors over time? *Child Development*, 74, 1628–1638.
- Van Beijsterveldt, C.E.M., Groen-Blokhuis, M., Hottenga, J.J., Franić, S., Hudziak, J.J., Lamb, D., ... & Boomsma, D.I. (2013). The Young Netherlands Twin Register (YNTR): Longitudinal twin and family studies in over 70,000 children. *Twin Research and Human Genetics*, 16, 252–267.
- Van Widenfelt, B.M., Goedhart, A.W., Treffers, P.D.A., & Goodman, R. (2003). Dutch version of the Strengths and Difficulties Questionnaire (SDQ). *European Child and Adolescent Psychiatry*, 12, 281–289.
- Vink, J.M., Bartels, M., van Beijsterveldt, T.C.E.M., van Dongen, J., van Beek, J.H.D.A., Distel, M.A., ... & Boomsma, D.I. (2012). Sex differences in genetic architecture of complex phenotypes? *PLoS ONE*, 7, e47371.
- Waltes, R., Chiocchetti, A.G., & Freitag, C.M. (2016). The neurobiological basis of human aggression: A review on genetic and epigenetic mechanisms. *American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics*, 171, 650–675.
- Weinberger, A.H., Darkes, J., Del Boca, F.K., Greenbaum, P.E., & Goldman, M.S. (2006). Items as context: Effects of item order and ambiguity on factor structure. *Basic and Applied Social Psychology*, 28, 17–26.
- Wichers, M., Gardner, C., Maes, H.H., Lichtenstein, P., Larsson, H., & Kendler, K.S. (2013). Genetic innovation and stability in externalizing problem behavior across development: A multi-informant twin study. *Behavior Genetics*, 43, 191–201.

Accepted for publication: 3 January 2020