6.

CHILDHOOD ODD AND ADHD BEHAVIOR: THE EFFECT OF CLASSROOM SHARING, GENDER, TEACHER GENDER AND THEIR INTERACTIONS

Based on Eveline L. de Zeeuw, Catherina E. M. van Beijsterveldt, Gitta H. Lubke, Tina J. Glasner, Eco J. C. de Geus and Dorret I. Boomsma (2014). Childhood ODD and ADHD behavior: The effect of classroom sharing, gender, teacher gender and their interactions. *Behavioral Genetics.* Accepted for Publication.

One criterion for a Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) diagnosis of Attention Deficit Hyperactivity Disorder (ADHD) and Oppositional Defiant Disorder (ODD) is that symptoms are present in at least two settings, and often teacher ratings are taken into account. The short Conners' Teacher Rating Scales - Revised (CTRS-R) is a widely used standardized instrument measuring ODD and ADHD behavior in a school setting. In the current study CTRS-R data were available for 7, 9 and 12-year-old twins from the Netherlands Twin Register. Measurement invariance (MI) across student gender and teacher gender was established for three of the four scales (Oppositional Behavior (OPP), Hyperactivity (HYP) and ADHD Index (ADHD)) of the CTRS-R. The fourth scale (ATT) showed an unacceptable model fit even without constraints on the data and revision of this scale is recommended. Gene-environment (GxE) interaction models revealed that heritability was larger for children sharing a classroom. There were some gender differences in the heritability of ODD and ADHD behavior and there was a moderating effect of teacher's gender at some of the ages. Taken together, this indicates that there was evidence for GxE interaction for classroom sharing, gender of the student and gender of the teacher.

INTRODUCTION

Attention deficit hyperactivity disorder (ADHD) is characterized by difficulties of both inattention and hyperactivity or impulsiveness that interfere with a child's daily functioning. At school, children have, for example, difficulty remaining in their seats and paying attention for a longer period of time. Oppositional defiant disorder (ODD) is characterized by hostile and defiant behavior towards figures with authority, going beyond normal childhood behavior. Children argue with their teacher and often lose their temper (American Psychiatric Association, 2000). Numerous studies have found a negative association between ADHD and educational achievement (Polderman et al., 2010) and children with ODD receive lower grades at school (Greene et al., 2002). Both children with ADHD and ODD are more likely to attend specialized schools.

The American Psychiatric Association (APA) estimates that 3 to 7 per cent of all school-aged children are diagnosed with ADHD, while estimates of the prevalence of ODD in children range from 2 to 16 per cent (American Psychiatric Association, 2000). It must be noted that more than 50 per cent of the children diagnosed with ADHD also have ODD (Angold, Costello & Erkanli, 1999; Wilens et al., 2002). In the general population, the ratio between boys and girls with ADHD is estimated to be 3:1, while the ratio is higher in a clinical population (Gaub & Carlson, 1997). A potential explanation of the discrepancy in

the ratio between boys and girls on population versus clinical level is bias in the ratings of the teacher (Abikoff et al., 2002; Derks, Hudziak & Boomsma, 2007; Sciutto, Nolfi & Bluhm, 2004), because one criterion for a Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) diagnosis is that symptoms are present in at least two settings and often the evaluation of the teacher is taken into account. In a study focusing on children diagnosed with ADHD (Derks, Hudziak & Boomsma, 2007) teachers reported more disruptive behavior at school for boys than for girls, while there is no difference for mother ratings. For ODD, teachers also report higher prevalence rates in boys than girls while parents do not (Meisel et al., 2013). To further complicate matters, teacher bias may depend on the teacher's gender. An alternative explanation of the discrepancy is that the gender differences in ADHD and ODD behavior are more pronounced in the school environment, which may demand more of a child than the home environment.

When analyzing questionnaire data concerning psychiatric disorders, researchers often use sum scores to combine multiple items of a scale. A meaningful interpretation of a sum score is only possible when a scale measures the same disorder in all specified groups. A meaningful interpretation of a sum score is only possible when a scale measures the same disorder in all specified groups. Mellenbergh et al. (1989) defined measurement invariance (MI) with respect to group as an identical distribution of the observed sum score, conditional on the disorder that the test measures, across groups. The interpretation of group differences with respect to sum scores is only meaningful when the scale is MI (Slof-Op 't Landt MC et al., 2009). MI does not hold for example if boys score on average higher on some of the items than girls without actually scoring higher on the underlying disorder. In this case, a boy and girl, who have the same degree of a disorder, obtain systematically different sum scores. Group differences in the sum score will then reflect measurement bias instead of true underlying differences (Dolan, 2000; Mellenbergh, 1989; Meredith, 1993; Millsap & Yun-Tein, 2004).

Behavioral genetic studies have established that ADHD is amongst the most heritable psychiatric childhood disorders. According to a review of 20 twin studies, the mean estimate of the heritability of ADHD in children is over 75 per cent (Faraone et al., 2005). Estimates for ODD are somewhat lower with a heritability of around 50 per cent (Hudziak et al., 2005). Heritability estimates of problem behavior in primary school children vary widely between twins taught in the same classroom compared to twins with different teachers (Saudino, Ronald & Plomin, 2005). It is a general finding that twin correlations are larger when one teacher rates both children compared to when two teachers each rate one child. One hypothesis is that ratings could be biased due to the same person

rating both children when twins are taught in the same classroom. Each teacher has his or her own perception on behavior, which can make children seem more similar when they have the same teacher (Kan et al., 2013; Simonoff et al., 1998). The second hypothesis is that there is GxE interaction (Eaves, 1984), which holds that the variation in the behavior of children in different classroom environments may depend on their genetic make-up. The classroom environment, teacher characteristics and peers differ when the twins do not share a classroom in primary school, and different environments might trigger different behavior depending on a child's genes. A study of internalizing and externalizing behavior in primary school children concluded that this was not the case, and that the heritability was higher in children sharing a classroom compared to children in different classrooms because of GxE interaction (Lamb et al., 2012). The question is whether this is also true for ODD and ADHD behavior and which differences between classrooms play a role.

In behavioral genetic studies, the absence of MI may have important consequences for heritability estimates. Absence of MI for an environmental factor, for example, gender of the teacher, could lead to differences in heritability estimates between groups (gene-environment (GxE) interaction). Absence of MI for student's gender may lead to what is known as scalar sex limitation, the effect of the genetic and environmental factors may, for example, be larger in boys than girls (Lubke, Dolan & Neale, 2004; Neale, Roysamb & Jacobson, 2006). The short Conners' Teacher Rating Scales - Revised (CTRS-R) is often filled out by teachers to assess ODD and ADHD behavior in a school setting (Conners et al., 1998). The scales of this instrument have been tested for MI in 7-year-old boys and girls (Derks et al., 2007), showing no evidence for measurement bias regarding the gender of the student. However, the study did not take into account possible differences between male and female teachers in the perception of ODD and ADHD behavior nor did it evaluate MI at older ages. Therefore, the first objective of this study is to determine whether the scales of the CTRS-R, measuring ODD and ADHD behavior, are measurement invariant for gender of the student as well as gender of the teacher throughout primary school. When measurement invariance holds, the second objective of this study is to focus on GxE interaction, and investigate whether classroom sharing, gender of the student and gender of the teacher moderate the heritability of teacher-rated ODD and ADHD behavior.

METHODS

PARTICIPANTS

The Netherlands Twin Register (NTR), established around 1987 by the Department of Biological Psychology at the VU University Amsterdam, registers approximately 40 per cent of all multiple births in the Netherlands. A survey about the development of the children is sent to the parents of the twins every two years until the twins are 12 years old (Boomsma et al., 2002; Boomsma et al., 2006; van Beijsterveldt et al., 2013). Since 1999, at approximately age 7, 9 and 12, when the twins attend primary school, parents are asked for their consent to approach the teacher(s) of their children with a survey. The survey sent to the primary school teachers includes items on background information of the teacher, functioning at school, educational achievement and the standardized questionnaires, the Teacher Report Form (TRF) (Achenbach, 1991) and the short version of the Conners' Teacher Ratings Scale - Revised (CTRS-R) (Conners, 2001).

Since 2001 data collection has yielded surveys with information on gender of the teacher for 9365, 8775 and 6649 7, 9 and 12-year-olds, respectively. We excluded children who had a disease or handicap that interfered severely with daily functioning (Age 7: N=97; Age 9: N=128; Age 12: N=95) or attended specialized education, special schools are available for children with extra needs (Age 7: N=109; Age 9: N=237; Age 12: N=226). Surveys were excluded if they were filled out by more than one teacher (Age 7: N=431; Age 9: N=259; Age 12: N=83), filled out by someone other than the regular teacher (Age 7: N=64; Age 9: N=68; Age 12: N=57), or if familiarity with the student was below average (Age 7: N=53; Age 9: N=62; Age 12: N=34). This resulted in a total sample for the measurement invariance analyses of 8611 surveys for 7-year-olds, 8021 surveys for 9-year-olds and 5954 surveys for 12-year-olds.

The sample for the GxE interaction analyses included complete phenotype data for most twin pairs (Age 7: N=3793; Age 9: N=3470; Age 12: N=2534). Incomplete data are due to only one of the teachers returning the survey. The sample consisted of 1208, 1102, and 762 twin pairs of opposite sex for respectively age 7, 9 and 12. For the same-sex twin pairs (Age 7: N=2585; Age 9: N=2368; Age 12: N=1772), determination of zygosity status was based on blood or DNA polymorphisms (Age 7: N=224; Age 9: N=331; Age 12: N=393) or on the basis of parental report of items on resemblance in appearance and confusion of the twins by parents and others (Age 7: N=2321; Age 9: N=1987; Age 12: N=1356). This last method established zygosity with an accuracy of approximately 93 per cent (Rietveld et al., 2000). Zygosity was unavailable for some twins and these twin pairs were excluded from the analyses (Age 7: N=40; Age 9: N=50; Age 12: N=23).

MEASUREMENTS

The short Conners' Teacher Rating Scale - Revised (CTRS-R) is a measurement instrument to asses ODD and ADHD behavior at school. Teachers had to indicate whether a child displayed a certain type of behavior currently or in the prior month. The short version of the CTRS-R consists of 28 items scored on a 4 point scale from 0 (not true or never) to 3 (completely true or very often) (Conners et al., 1998; Conners, 2001). The CTRS-R includes 4 scales measuring Oppositional Behavior (OPP: 5 items), Cognitive Problems/Inattention (ATT: 5 items), Hyperactivity (HYP: 7 items) and Attention Deficit Hyperactivity Disorder Index (ADHD: 12 items). One item is included in both the HYP and ADHD scale ('Easily excited, impulsive'). The item 'Inattentive, gets distracted easily' of the ADHD scale was excluded from the MI analyses as it was highly correlated with some of the other items, especially 'Easily distracted or difficulty maintaining attention' (Age 7: r = .812; Age 9: r = .805; Age 12: r = .789) and 'Short attention span' (Age 7: r = .777; Age 9: r = .716; Age 12: r = .745). As a consequence, the more stringent MI models did not converge due to multicolinearity when including this item. For the GxE interaction analyses, a sum score of a scale was computed when there was at most one missing item (OPP, ATT and HYP) or at most two missing items (ADHD) for a scale. Missing items were imputed by the rounded averaged item score of the scale for that child. The sum scores of the scales showed an L-shaped distribution and therefore the data were square root transformed prior to the analyses.

STATISTICALANALYSES

MEASUREMENT INVARIANCE

The factor structure of the four CTRS-R scales was investigated with exploratory factor analyses (EFA) with an Oblimin rotation. The number of latent factors was decided based on the scree plot and eigenvalues (larger than 1) of the factors. To test whether the scales of the CTRS-R were MI across student ('boy' or 'girl') gender and teacher ('male' or 'female') gender, multigroup (4 groups) confirmatory factor analyses (CFA) for ordinal item level data were carried out (Dolan, 2000; Meredith, 1993; Millsap & Yun-Tein, 2004) using Mplus Version 6.1 (Muthén & Muthén, 2010). With ordinal item level data an underlying continuously distributed liability is assumed and thresholds that categorize the disorder are estimated based on the response frequencies (Flora & Curran, 2004). Because of the low frequencies of the most extreme response categories, the highest two response categories were combined. The EFA and CFA models were fitted with the Theta parameterization and the weighted least squares with mean variance adjusted (WLSMV) estimator. Correction for dependency of the observations due to family clustering was done by the 'complex' option. This

'complex' option computes the standard errors and a chi square of model fit taking into account this dependency.

Different levels of MI were tested by constraining the model parameters step by step. The first level is configural invariance (configural MI), where the factor structure is the same across groups. Factor means are fixed to zero for identification purposes while factor variances, thresholds, loadings and residual variances of the continuous latent response variables are group specific. One of the factor loadings is constrained to be equal to 1 for scaling purposes. A stricter model is strong factorial invariance (strong MI), where differences in latent response means are the result of differences in the latent factor means. This model is tested by constraining both the factor loadings and thresholds to be equal across groups. The factor mean of the first group is fixed to zero and freely estimated in the other groups. The last model, strict factorial invariance (strict MI) implies that the differences in the latent response means reflect true differences in the latent factor means and variances. This is tested by constraining the factor loadings, thresholds and residual variances of the continuous latent response variables to be equal across all groups. The factor mean is still fixed to zero in the first group and freely estimated in the other groups (Dolan, 2000; Meredith, 1993; Millsap & Yun-Tein, 2004).

The root mean square error of approximation (RMSEA) and the comparative fit index (CFI) were chosen as indices of model fit. A RMSEA value smaller than .05 indicates a good fit as does a CFI value of .97 or higher (Schermelleh-Engel & Moosbrugger, 2003). The difference in goodness of fit between the nested MI models in chi square values between two nested models when using the WLMSV chi-square values is not distributed as a chi-square and as a consequence regular chi-square testing is not appropriate when using the WLSMV estimator (Muthén & Muthén, 2010). Instead, the 'difftest' option in Mplus can be used to obtain a correct chi-square difference test by using the derivatives of the variables from both models. Due to the large sample sizes these chi-square difference tests models might reject a model on the basis of a significant chi-square difference even though the model actually fit. Interpreting the chi-square as a goodness-of-fit index has been suggested as an alternative for using the chi-square as a formal test statistic. Since there are no absolute standards, a ratio between 2 and 3 is proposed to be indicative of, respectively a good and an acceptable model fit (Schermelleh-Engel & Moosbrugger, 2003). Therefore, a difference in chi-square of more than 3 times the difference in estimated parameters was interpreted as a worsening of the fit of the model. In addition, we looked at the parameter estimates and the magnitude of the modification indices to make reliable decisions on acceptance of MI.

GENE-ENVIRONMENT INTERACTION MODELS

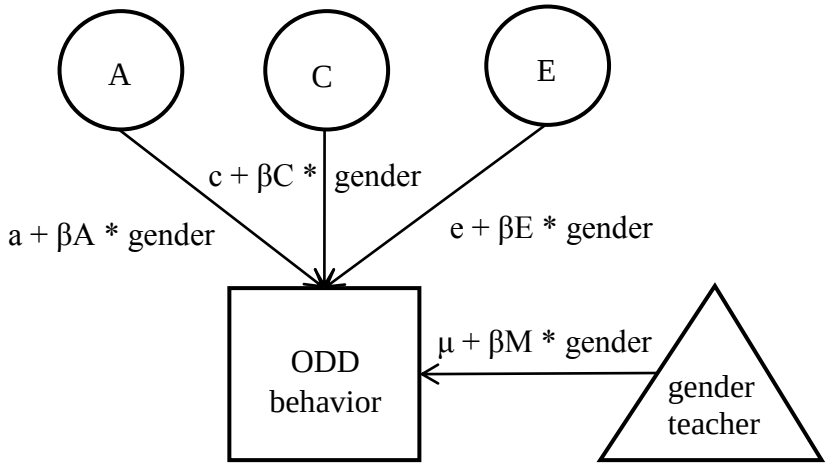
The contribution of genetic and environmental effects to the variance of the CTRS-R scales was estimated in a classical twin model (Boomsma et al., 2002; Plomin R. et al., 2008) in the R (R Core Team, 2014) package OpenMx Version 3.1.0 (Boker S.M. et al., 2011; Boker S.M. et al., 2012) with maximum likelihood estimation. First, a saturated model was fitted to the data in which means, variances and covariances were estimated in the different zygosity-by-gender groups rated by same (ST) and different (DT) teachers. Mean and variance differences between children taught by male and female, between boys and girls, between children sharing a classroom or in different classrooms and across zygosity were tested in the saturated model. It was tested whether the twin correlations could be equated between twins sharing a classroom and twins in different classrooms.

Next, GxE interaction models for gender of the student, classroom sharing and gender of the teacher were fitted to the data. GxE interaction was modelled by using multiple group designs for classroom sharing and gender of the student, and by a moderation model for teacher's gender (Figure 1) (Purcell, 2002). The models included additive genetic effects (A), dominant genetic effects (D) (or common environmental effects (C), shared by twins) and unique environmental effects (E), not shared by twins. To correct for possible confounding by gene-environment correlation (rGE), means were allowed to be different between boys and girls, between twins rated by the same or different teachers and between children rated by male or female teachers (Purcell, 2002). In the first models, differences in heritability between boys and girls were tested by constraining the estimates to be equal over gender of the student. Total variances between boys and girls were allowed to differ. Next, it was tested whether estimates could be constrained to be equal for twins rated by the same and by different teachers. Differences in genetic and environmental variance between the same and different teacher groups could be due to GxE interaction, but may also be the result of rater bias. Therefore, a correlated errors model was applied, which is an extension of the univariate twin model as it allows the unique environmental (E) effects to be correlated for twin pairs rated by the same teacher (Simonoff et al., 1998). In the last models, GxE interaction by gender of the teacher was tested by dropping from the model the moderation of the A, D (C) and E estimates by gender of the teacher.

Difference in goodness of fit of the nested models was assessed with a log-likelihood ratio test (LRT) which calculates the difference in -2log-likelihood (-2LL) between two models and evaluates this χ^2-statistic with the difference in the number of estimated parameters between the models as degrees of freedom. A p-value smaller than 0.01 was considered significant. Constraints were kept,

when a more restrictive model did not significantly decrease the goodness of fit, as a more parsimonious model is preferred.

FIGURE 1 Gene-environment interaction (GxE) model with moderation by gender of the teacher



RESULTS

MEASUREMENT INVARIANCE

MI of the four scales (OPP, ATT, HYP and ADHD) of the CTRS-R was tested across gender of the student ('boy' or 'girl') and gender of the teacher ('male' or 'female') at age 7 (Age: Mean = 7.44 and SD = .47), age 9 (Age: Mean = 9.92 and SD = .53) and age 12 (Age: Mean = 12.15 and SD = .30), resulting in a 4 group comparison. Information on the gender of the teacher was available for 8611 7-year-olds (boy-male: N=322; boy-female: N=3918; girl-male: N=317; girl-female: N=4054), 8021 9-year-olds (boy-male: N=1050; boy-female: N=2841; girl-male: N=1111; girl-female: N=3019) and 5954 12-year-olds (boy-male: N=1332; boy-female: N=1503; girl-male: N=1381; girl-female: N=1738). Table 1 shows the frequencies of the item responses and the factor loadings of the items for all scales estimated from the exploratory factor analyses (EFA). Factor loadings were overall relatively high. On the basis of the scree plots and eigenvalues, a one-factor solution was chosen for OPP, ATT and HYP and a two-factor solution for ADHD (attention problems (AP) and hyperactivity/impulsivity (HI)) in all age groups (see Table 1).

Results for the tests of the three levels of MI are reported in Table S1. For OPP, HYP and ADHD the configural, strong and strict invariance models all showed

an acceptable to good fit, based on the RMSEA and CFI, for all age groups. Differences in chi-square between the models with increasing equality constraints were rather small and, for the strong MI level, did not exceed more than three times the number of degrees of freedom. However, for the strict MI level, the difference in a chi-square for OPP at age 9 and HYP at age 7 and 12 was somewhat larger than this criterion, but these differences were accompanied by minor changes in RMSEA and CFI. Inspection of the modification indices revealed that they were larger for female teachers compared to male teachers for both boys and girls. Taken together, we could accept MI for the scales OPP, HYP and ADHD, for all ages, with respect to gender of the student and, more tentatively, for gender of the teacher. The fit of the MI models was acceptable to mediocre for ATT in 7-year-olds while the fit of the models was unacceptable for 9 and 12-year-olds. Even the models without constraints on the factor structure did not fit the data very well. Increasing MI levels led to a large decrease in model fit for all ages. Therefore, we could not accept MI across gender of the student and teacher for the ATT scale.

TABLE 1 Frequencies of the item responses and factor loadings as estimated in the EFA

	Age 7					Age 9					Age 12				
	Frequencies of Item Responses			Factor Loadings		Frequencies of Item Responses			Factor Loadings		Frequencies of Item Responses			Factor Loadings	
	0	1	2/3	1	2	0	1	2/3	1	2	0	1	2/3	1	2
Oppositional Behavior															
2 Defiant	.828	.141	.031	.917		.797	.167	.036	.914		.781	.181	.038	.915	
6 Defies	.901	.081	.018	.912		.894	.085	.021	.916		.876	.102	.022	.929	
10 Spiteful	.959	.034	.007	.777		.931	.059	.010	.832		.937	.054	.009	.803	
15 Argues	.862	.117	.021	.879		.841	.130	.029	.917		.817	.152	.031	..938	
20 Explosive	.921	.060	.019	.845		.907	.070	.023	.827		.916	.065	.019	.794	
Inattention/ Cognitive Problems															
4 Forgets things	.698	.225	.077	.880		.645	.260	.095	.857		.668	.248	.084	.854	
8 Poor spelling	.655	.202	.143	.881		.591	.188	.221	.860		.582	.212	.206	.862	
13 Poor reading	.696	.153	.151	.844		.728	.134	.137	.799		.786	.117	.097	.824	
18 Lacks interest	.842	.120	.039	.698		.797	.159	.045	.595		.778	.170	.052	.617	
22 Poor arithmetic	.748	.171	.081	.770		.695	.175	.130	.743		.702	.175	.123	.748	
Hyperactivity															
3 Restless	.680	.221	.099	.766		.706	.209	.085	.743		.766	.176	.058	.757	
7 Always on the go	.856	.098	.046	.830		.859	.098	.043	.794		.875	.093	.033	.794	
11 Leaves seat	.836	.115	.050	.864		.873	.090	.037	.867		.913	.066	.021	.849	
17Difficulty awaiting	.703	.204	.093	.828		.756	.167	.077	.843		.804	.140	.056	.851	
21 Runs about	.937	.047	.016	.876		.950	.038	.012	.878		.964	.028	.008	.884	
24 Difficulty playing	.776	.160	.064	.887		.788	.153	.059	.889		.826	.128	.046	.898	
27 Excitable	.798	.141	.062	.884		.799	.143	.058	.870		.826	.124	.050	.881	

ADHD Index													
Attention Problems													
14 Short attention span	.674	.214	.112	.028	.938	.687	.203	.110	.076	.897	.726	.194	.079
16 Only attention for own interests	.785	.160	.054	.202	.585	.757	.180	.063	.204	.583	.750	.184	.066
19 Distractible	.645	.231	.123	.102	.887	.649	.226	.124	.164	.832	.687	.222	.091
25 Fails to finish	.792	.164	.044	-.045	.908	.797	.163	.040	-.065	.928	.824	.142	.033
26 Not following instructions	.875	.088	.037	-.080	.925	.883	.083	.034	-.094	.949	.895	.080	.024
Hyperactivity													
5 Disturbs other children	.709	.228	.063	.855	.023	.696	.237	.067	.854	.026	.730	.210	.060
9 Cannot remain still	.779	.160	.062	.848	.106	.786	.160	.054	.808	.150	.825	.136	.039
12 Fidgets	.709	.197	.094	.676	.174	.754	.168	.078	.596	.243	.825	.132	.044
23 Interrupts	.750	.191	.059	.920	-.076	.754	.187	.059	.910	-.070	.797	.160	.043
27 Excitable	.798	.141	.062	.893	-.057	.799	.143	.058	.909	-.080	.826	.124	.050
28 Restless	.814	.129	.056	.944	.004	.821	.127	.053	.914	.035	.850	.116	.034

TABLE 2 Means and standard deviations of the untransformed sum scores of the CTRS-R scales at age 7, 9 and 12

	Male teacher								Female Teacher							
	Same Teacher				Different Teacher				Same Teacher				Different Teacher			
	Boys		Girls		Boys		Girls		Boys		Girls		Boys		Girls	
	N	Mean (SD)	N	Mean (SD)	N	Mean (SD)	N	Mean (SD)	N	Mean (SD)	N	Mean (SD)	N	Mean (SD)	N	Mean (SD)
Oppositional Behavior																
Age 7	167	.8 (1.7)	170	.5 (1.3)	109	.7 (1.6)	107	.3 (1.1)	1910	.8 (1.8)	2091	.4 (1.1)	1489	.9 (1.8)	1468	.4 (1.2)
Age 9	557	1.0 (1.9)	594	.5 (1.2)	347	.9 (1.9)	349	.6 (1.6)	1401	1.1 (2.0)	1576	.5 (1.3)	1002	1.0 (2.1)	1039	.5 (1.4)
Age 12	748	1.0 (2.0)	814	.5 (1.1)	381	1.0 (1.9)	365	.6 (1.4)	805	1.2 (2.1)	959	.4 (1.2)	442	1.0 (2.2)	497	.7 (1.6)
Hyperactivity																
Age 7	167	2.5 (3.6)	170	1.5 (2.7)	108	2.3 (3.0)	106	.9 (2.1)	1907	2.7 (3.9)	2093	1.1 (2.2)	1486	2.9 (3.9)	1469	1.2 (2.3)
Age 9	556	2.3 (3.4)	592	1.0 (1.9)	347	2.3 (3.4)	351	1.1 (2.2)	1399	2.5 (3.6)	1578	.9 (1.8)	1000	2.7 (3.8)	1038	1.0 (2.3)
Age 12	752	1.8 (3.0)	815	.8 (1.8)	381	1.8 (2.8)	366	.9 (1.9)	804	2.0 (3.2)	959	.6 (1.5)	442	2.2 (3.6)	496	.9 (2.1)
ADHD Index																
Age 7	167	5.3 (6.6)	170	3.4 (5.4)	108	4.6 (5.0)	107	2.9 (4.4)	1906	5.3 (6.6)	2091	2.9 (4.6)	1485	6.2 (7.1)	1469	3.3 (4.9)
Age 9	553	5.1 (6.4)	589	2.9 (4.6)	348	5.5 (6.9)	351	3.1 (4.6)	139	5.6 (6.7)	1578	2.6 (4.2)	999	6.3 (7.0)	1039	3.0 (4.7)
Age 12	750	4.5 (6.0)	815	2.3 (3.7)	381	4.7 (5.6)	366	2.5 (3.9)	804	4.9 (6.2)	960	1.9 (3.6)	439	5.6 (6.9)	495	2.6 (4.3)

N = number of observations; SD = standard deviation

GENE-ENVIRONMENT INTERACTION MODELS

The results of the variance differences were added to the results section and the paragraph was restructured to improve clarity. 'Table 2 gives the means and standard deviations of the measurement invariant CTRS-R scales for boys and girls with the same or different male or female teachers across the three age groups. The saturated models were used to test for mean and variance differences across these groups. For OPP, there were mean and variance differences between boys and girls at all ages and variance differences across zygosity at age 7, between children sharing a classroom and children in different classrooms at age 12 and between children with the same or different male or female teachers at age 12. For HYP, there were mean and variance differences between boys and girls at all ages, mean differences across zygosity and between children sharing a classroom and children in different classrooms at age 7 and variance differences between children sharing a classroom and children in different classrooms at age 12. For ADHD, there were mean and variance differences between boys and girls at all ages and mean differences between children sharing a classroom and children in different classrooms at all ages.

TABLE 3 Twin correlations for the CTRS-R scales rated by the same teacher or different teachers at age 7, 9 and 12

	Oppositional Behavior		Hyperactivity		ADHD Index	
	ST	**DT**	**ST**	**DT**	**ST**	**DT**
Age 7						
MZm	.772	.495	.842	.479	.820	.555
DZm	.360	.280	.347	.289	.437	.292
MZf	.617	.394	.749	.492	.770	.514
DZf	.404	.233	.310	.211	.342	.217
DOS	.294	.112	.301	.176	.339	.250
Age 9						
MZm	.763	.334	.790	.465	.792	.447
DZm	.405	.211	.342	.208	.353	.296
MZf	.635	.442	.712	.407	.793	.497
DZf	.498	.081	.302	.145	.379	.270
DOS	.244	.133	.296	.242	.327	.254
Age 12						
MZm	.719	.518	.792	.434	.818	.546
DZm	.350	.282	.297	.310	.283	.301
MZf	.606	.500	.681	.361	.751	.414
DZf	.338	.297	.315	.282	.276	.245
DOS	.232	.185	.234	.205	.265	.233

ST = same teacher; DT = different teacher; MZm = monozygotic boys; DZm = dizygotic boys; MZf = monozygotic girls; DZf = dizygotic girls; DOS = dizygotic of opposite sex

Twin correlations for each gender by zygosity group rated by the same teacher or by different teachers are given in Table 3. For all scales, MZ correlations were higher, sometimes more than twice as high, than DZ correlations, suggesting additive (and in some cases dominant) genetic effects. Only for the OPP scale were DZ correlations larger than half the MZ correlations, suggesting common environmental effects. The GxE interaction model fitting results are reported in the online supplementary materials for the OPP (Table S2), HYP (Table S3) and ADHD (Table S4) scales of the CTRS-R. The standardized estimates (Table 4) and the contribution of the variance components (Figure 2) are given for the most parsimonious and best fitting models.

CLASSROOM SHARING

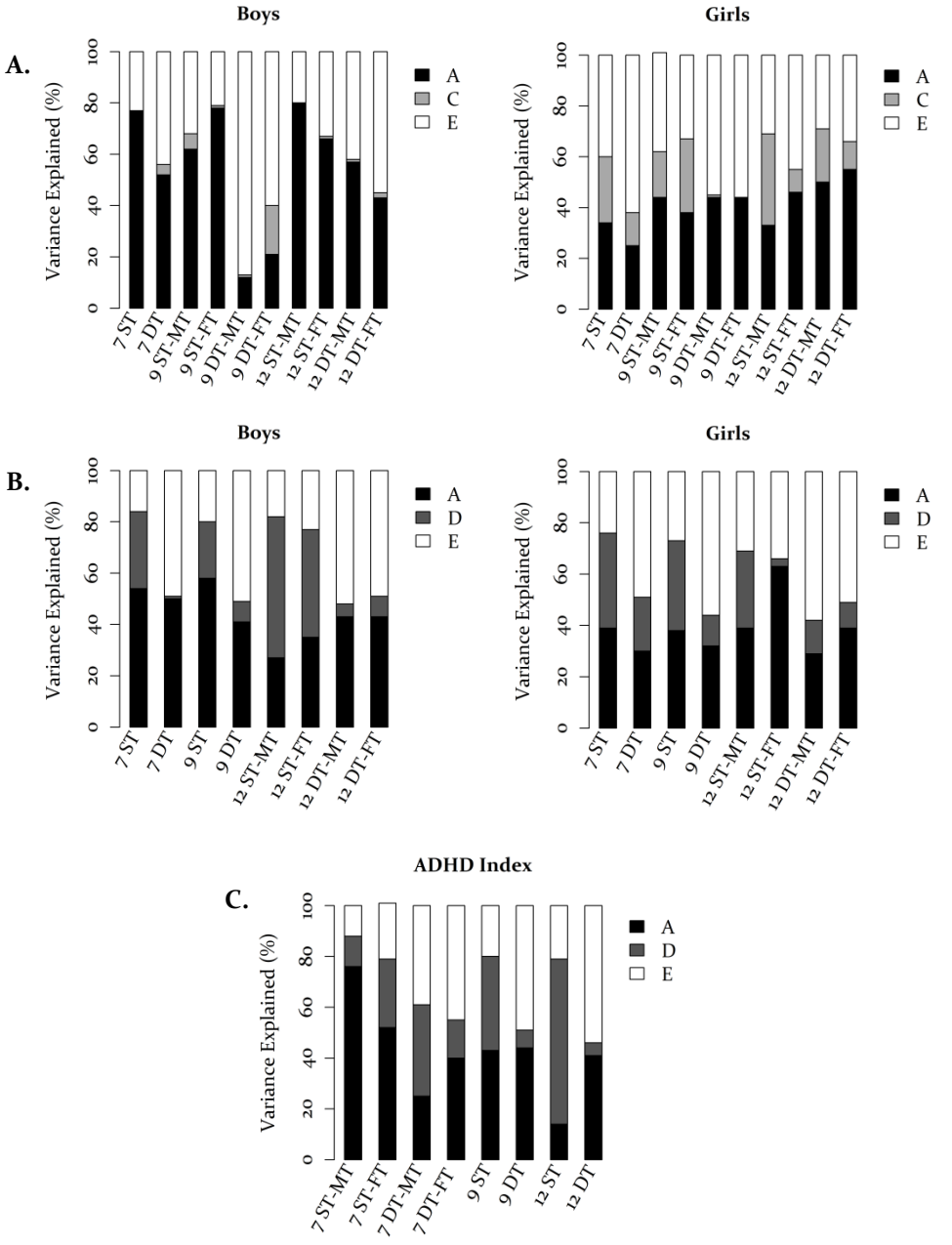
Correlations between twins rated by the same teacher could not be constrained to be equal to correlations between twins with different teachers. Constraining the variance components to be equal across same and different teachers also resulted in a significant deterioration of the model fit. A model with correlated errors was fitted to the data to check whether the differences between the same teacher and different teacher groups could be explained by rater bias. For none of the scales did the correlated errors model provide a better fit. In general, the proportion of the variance explained by genetic effects (heritability) was higher, at all ages, for children taught by the same teacher (ST) than for children rated by different teachers (DT) for OPP in boys (ST: 62-80%; DT: 12-57%) and girls (ST: 33-46%; DT: 25-55%), HYP in boys (ST: 76-84%; DT: 48-51%) and girls (ST: 66-75%; DT: 43-51%) and ADHD (ST: 78-88%; DT: 46-61%).

TABLE 4 Standardized estimates [95% Confidence intervals] of the total genetic (G), additive genetic (A), dominant genetic (D), common environmental (C) and unique environmental (E) effects on the four CTRS-R scales for 7, 9 and 12-year-olds in the best-fitting models

			Same Teacher				Different Teacher			
			G	A	C/D	E	G	A	C/D	E
Oppositional Behavior										
Age 7		Boys	.77 [.71-.80]	.77 [.71-.80]	.00 [.00-.05]	.23 [.20-.27]	.52 [.32-.62]	.52 [.32-.62]	.04 [.00-.20]	.45 [.37-.53]
		Girls	.34 [.15-.54]	.34 [.15-.54]	.26 [.07-.43]	.40 [.35-.46]	.25 [.00-.42]	.25 [.00-.42]	.13 [.00-.34]	.62 [.53-.72]
Age 9	Male	Boys	.62 [.41-.74]	.62 [.41-.74]	.06 [.00-.25]	.32 [.24-.41]	.12 [.00-.29]	.12 [.00-.29]	.01 [.00-.13]	.87 [.71-.99]
	Teacher	Girls	.44 [.20-.66]	.44 [.20-.66]	.18 [.00-.38]	.39 [.30-.50]	.44 [.30-.56]	.44 [.30-.56]	.01 [.00-.09]	.56 [.43-.70]
	Female	Boys	.78 [.69-.82]	.78 [.69-.82]	.01 [.00-.09]	.21 [.18-.25]	.21 [.00-.25]	.21 [.00-.25]	.19 [.04-.36]	.60 [.49-.72]
	Teacher	Girls	.38 [.21-.55]	.38 [.21-.55]	.29 [.13-.44]	.34 [.29-.29]	.44 [.31-.54]	.44 [.31-.54]	.00 [.00-.08]	.56 [.46-.69]
Age 12	Male	Boys	.80 [.72-.84]	.80 [.72-.84]	.00 [.00-.07]	.20 [.16-.25]	.57 [.34-.69]	.57 [.34-.69]	.01 [.00-.18]	.42 [.31-.55]
	Teacher	Girls	.33 [.13-.56]	.33 [.13-.56]	.36 [.15-.54]	.31 [.25-.38]	.50 [.27-.70]	.50 [.27-.70]	.21 [.04-.43]	.29 [.19-.42]
	Female	Boys	.66 [.53-.73]	.66 [.53-.73]	.01 [.00-.20]	.33 [.26-.41]	.43 [.22-.55]	.43 [.22-.55]	.02 [.00-.20]	.55 [.44-.68]
	Teacher	Girls	.46 [.27-.60]	.46 [.27-.60]	.09 [.00-.25]	.44 [.36-.54]	.55 [.35-.69]	.55 [.35-.69]	.11 [.00-.30]	.34 [.25-.46]
Hyperactivity										
Age 7		Boys	.84 [.81-.86]	.54 [.20-.86]	.30 [.00-.64]	.16 [.14-.19]	.51 [.43-.59]	.50 [.17-.59]	.01 [.00-.36]	.49 [.41-.57]
		Girls	.75 [.72-.78]	.39 [.16-.78]	.37 [.00-.60]	.25 [.21-.28]	.51 [.42-.58]	.30 [.04-.55]	.21 [.00-.49]	.49 [.42-.58]
Age 9		Boys	.80 [.76-.83]	.58 [.20-.82]	.22 [.00-.60]	.20 [.17-.24]	.49 [.39-.58]	.41 [.10-.57]	.08 [.00-.42]	.51 [.42-.61]
		Girls	.72 [.68-.76]	.38 [.15-.75]	.35 [.00-.58]	.28 [.24-.32]	.44 [.33-.53]	.32 [.06-.51]	.12 [.00-.41]	.56 [.47-.67]
Age 12	Male	Boys	.82 [.77-.86]	.27 [.00-.62]	.55 [.20-.82]	.18 [.14-.23]	.48 [.34-.60]	.43 [.09-.59]	.05 [.00-.40]	.52 [.40-.66]
	Teacher	Girls	.68 [.61-.74]	.39 [.00-.70]	.30 [.00-.69]	.32 [.26-.39]	.43 [.24-.57]	.29 [.03-.50]	.13 [.00-.43]	.57 [.43-.76]
	Female	Boys	.76 [.70-.81]	.35 [.03-.68]	.42 [.09-.74]	.24 [.19-.30]	.51 [.37-.62]	.43 [.11-.60]	.08 [.00-.41]	.49 [.38-.63]
	Teacher	Girls	.66 [.58-.73]	.63 [.34-.72]	.03 [.00-.33]	.34 [.27-.42]	.49 [.27-.42]	.39 [.11-.58]	.10 [.00-.38]	.51 [.39-.65]
ADHD Index										
Age 7	Male		.88 [.83-.92]	.76 [.35-.90]	.12 [.00-.53]	.12 [.08-.17]	.61 [.48-.72]	.25 [.07-.63]	.36 [.00-.62]	.39 [.28-.52]
	Teacher									
	Female		.78 [.76-.81]	.52 [.34-.69]	.27 [.10-.44]	.22 [.19-.24]	.55 [.50-.60]	.40 [.18-.56]	.15 [.00-.38]	.45 [.40-.50]
	Teacher									
Age 9			.80 [.77-.82]	.43 [.25-.60]	.37 [.20-.55]	.20 [.18-.23]	.50 [.43-.57]	.44 [.21-.55]	.07 [.00-.31]	.50 [.43-.57]
Age 12			.79 [.76-.81]	.14 [.00-.34]	.65 [.44-.80]	.21 [.19-.24]	.46 [.38-.53]	.41 [.11-.52]	.05 [.00-.38]	.54 [.46-.62]

G = genetic effects (summation of additive and dominant genetic effects); A = additive genetic effects; C = common environmental effects; E = unique environmental effects; MT = male teacher; FT = female teacher

FIGURE 2 The relative contribution of the additive genetic, dominant genetic, common environmental and unique environmental effects for the most parsimonious and best fitting models for Oppositional Behavior (A), Hyperactivity (B) and Attention Deficit Hyperactivity Disorder Index (C)



ST = different teacher; ST = same teacher; FT = female teacher; MT = male teacher

GENDER OF THE STUDENT

For the scales OPP and HYP, the contribution of the variance components differed between boys and girls at all ages, while this was not the case for the ADHD scale. Heritability of OPP was higher for boys (ST: 62-80%; DT: 12-57%) than girls (ST: 33-46%; DT: 25-55%). The influence of common environmental effects was, at most ages, negligible in boys (ST: 0-6%: DT: 1-19%) while it had some influence in girls (ST: 9-36%; DT: 0-21%). Heritability of HYP was slightly higher for boys (ST: 76-84%; DT: 48-51%) than girls (ST: 66-75%; DT: 43-51%). Differences between boys and girls on this scale could mainly be attributed to differences in the influence of dominant genetic effects.

GENDER OF THE TEACHER

Moderation by gender of the teacher was significant for OPP at age 9 and 12, HYP at age 12 and ADHD at age 7. For OPP at age 9, the relative influence of genetic effects was larger in boys with female teachers (ST: 78%; DT: 21%) than with male teachers (ST: 62%; DT: 12%) while it was somewhat larger for girls with male teachers (ST: 44%; DT: 44%) compared to with female teachers (ST: 38%; DT: 44%). For OPP at age 12, the opposite was true; heritability was larger in boys with male teachers (ST: 80%; DT: 57%) than with female teachers (ST: 66%; DT: 43%) while heritability was somewhat larger when girls were taught by a female teacher (ST: 46%: DT 55%) compared to when they were taught by a male teacher (ST: 33%; DT: 50%). For HYP at age 12, heritability was almost equal in boys and girls with male and female teachers, but the extent to which dominant genetic effects played a role differed across gender of the teacher. For ADHD at age 7, heritability was larger for children with male teachers (ST: 88%; DT: 61%) compared to with female teachers (ST: 78%; DT: 55%).

DISCUSSION

Three (Oppositional Behavior (OPP), Hyperactivity (HYP) and Attention Deficit Hyperactivity Disorder Index (ADHD)) of the four scales of the short Conners' Teacher Ratings Scale - Revised (CTRS-R) (Conners, 2001), used in a school setting to assess ODD and ADHD behavior, were measurement invariant across gender of the student and teacher. This means that gender differences in means and variances may be interpreted as reflecting true differences on the underlying disorder. In contrast, measurement invariance did not hold for the Inattention/Cognitive Problems (ATT) scale. Explanations for the absence of measurement invariance could be the low factor loadings and the moderate test-retest reliability of this scale. Problems with the item content have been

previously suggested (Conners et al., 1998). In our sample, the internal reliability of the Inattention/Cognitive Problems scale of the short CTRS-R ranged from .78 to .82. The results of the measurement invariance analyses strongly question the reliability of this scale and its use in clinical practice. Revision of this scale is recommended as the ratings might reflect a bias instead of true differences.

Heritability of ODD and ADHD behavior, measured with the Oppositional Behavior (OPP), Hyperactivity (HYP) and Attention Deficit Hyperactivity Disorder Index (ADHD) scales of the CTRS-R is substantial. Common environmental effects had some influence on ODD behavior while dominant genetic effects had an influence on ADHD behavior. The finding of common environmental effects is consistent with earlier studies of ODD behavior using parental ratings (Burt et al., 2001; Tuvblad et al., 2009). The influence is larger in girls which may be explained by the fact that girls appear to be more sensitive to reprimands from the teacher than boys. Earlier research already concluded that girls more often feel the pressure from peers or others to behave prosocially (Roberts & Strayer, 1996). Girls might be more inclined to adapt their behavior when they are called upon by the teacher. In younger girls the common environment also has an influence when they do not share a classroom. Factors in the home environment that have been proposed to have an influence on ODD behavior are, for example, parental discipline and parental involvement (Frick et al., 1992) and the influence of these factors could depend on the gender of a child and decrease when a child grows older. The finding of dominant genetic effects for ADHD behavior, especially in children sharing a classroom, could also be due to rater contrast effects. Only when one teacher rates both children of a twin pair can the behavior of the children be contrasted and result in negative interaction effects. A higher rating for ADHD behavior in one of the children of a twin pair could lead to a lower rating for ADHD behavior in the co-twin. However, the variance in ADHD behavior is not significantly smaller in MZ twin pairs compared to DZ twin pairs, which disconfirms the presence of this type of rater bias. This is in accordance with the results of a study looking into mother and teacher ratings of hyperactivity. A contrast effect was found for the maternal ratings while the teacher ratings did not show this form of rater bias (Simonoff et al., 1998).

Heritability estimates for ADHD behavior are comparable to those found in studies taking differences between same and different teachers into account. For example, Merwood et al., (2013) also found differences in heritability between 12-year-old children sharing a classroom (76%) and not sharing a classroom (49%). One study included only twin pairs sharing a classroom and observed a heritability of 74 per cent (Hartman et al., 2007) while another included only twins not sharing a classroom and estimated a heritability of 46 per cent

(Towers et al., 2000). GxE interaction was the most plausible explanation for internalizing and externalizing problems, assessed with the Teacher Report Form, in 7 to 12-year-old twin pairs of which approximately 60 per cent shared a classroom (Lamb et al., 2012). Other studies looking into GxE interaction for ADHD in 11 to 12-year-olds (Merwood et al., 2013), and hyperactivity in 7-year olds (Saudino, Ronald & Plomin, 2005) also observed that heritability was larger when children shared a classroom. On the other hand, a study in 7-year-olds did not observe a difference between children sharing a classroom and children in different classrooms in the heritability of ODD and ADHD behavior (Derks et al., 2007), but it could be that this study did not have enough power to detect these differences in the heritability (Derks, Dolan & Boomsma, 2004).

Studies towards the heritability of teacher-rated ODD behavior are scarce. The findings of gender differences and common environmental effects were in accordance with the results of a study by Hudziak et al. (2005) that was based on a subsample of the present study. Heritability estimates for both boys (38%) and girls (21%) were somewhat different. However, this study did not take into account whether the children were rated by the same or different teachers (Hudziak et al., 2005). In contrast with current findings, none of the heritability estimates of the maternal-rated ODD behavior differed between boys and girls (Dick et al., 2005; Tuvblad et al., 2009). The differences between parent and teacher ratings of ODD behavior could be due to the fact that children can express different behavior in the classroom than they do at home. The OPP scale of the CTRS-R takes these differences into account by including different items for the teacher survey. A study observed that, although parents rated children rather similar over time, teachers with different teaching styles rated the same children very different across grades, suggesting that behavior differed in response to different teaching styles (Vitaro, Tremblay & Gagnon, 1995). Another explanation is that teachers have highly informed views on general childhood behavior for both boys and girls and are better able to assess which behavior is normative for a child of a certain age and gender.

Heritability of ODD and ADHD behavior was larger in children who shared a classroom compared to those who did not. The correlated errors model did not provide a better explanation for the differences in correlations between children rated by the same and different teachers, excluding teacher bias as an explanation, and therefore these findings are in line with GxE interaction for classroom sharing. In general, the heritability of ODD and ADHD behavior was lower in children not sharing a classroom leading to a larger impact of the environment which suggests that different behavior is elicited by different classroom environments. The children are taught by different teachers, with different rules and teaching methods and have different peers. All these factors

could contribute to differences between children. For example, how teachers handle disruptive behavior is related to the behavior of a child (Rydell & Henricsson, 2004). The unique environmental variance also contains measurement error which might be increased when different teachers rate the two children of a twin pair as rater variance ends up in the measurement error (Hoyt, 2000). An important question is which differences between classroom environments play a role. Peer problems are related to ODD and ADHD behavior (Paap et al., 2013). Genetic variance in childhood aggression is moderated by peer victimization and might also moderate the heritability of ODD and ADHD (Brendgen et al., 2008). A study towards differences between monozygotic twins in their perception of the classroom environment identified, for example, the perception of a student about the relationship with the teacher as a unique environmental factor that differed between the genetically identical twins and was linked to hyperactivity as rated by the teacher (Somersalo, Solantaus & Almqvist, 2002).

For one teacher characteristic, gender, we investigated whether it moderated genetic effects on behavior in the classroom. The expression of a child's genetic vulnerability for displaying ODD and ADHD behavior at school depended in some cases on the gender of the teacher. The direction of the difference in heritability may provide an indication for one of two hypotheses. Male teachers and female teachers could provide a different learning and classroom environment with regard to, for example, structure and rules. The bioecological model (Bronfenbrenner & Ceci, 1994) predicts that the heritability of a phenotype will be lower in an adverse environment because risk environments will prevent the amplification of underlying genetic differences between children while the diathesis-stress model suggests that heritability will be higher in an adverse environment due to the expression of a genetic vulnerability that is triggered by a risk environment (Rende & Plomin, 1992). A same-gender teacher might be seen as a supportive environment as it is suggested to have a positive influence on the behavior and educational achievement of a child (Carrington, Tymms & Merrell, 2008). According to the bioecological model, genetic variation will be higher when children are taught by a same-gender teacher while the diasthesis-stress model predicts that heritability will be lower. However, in our study, the direction of the effects of gender of the teacher was not consistent which makes interpreting the GxE interaction findings difficult.

To summarize, three of the four scales of the short CTRS-R measuring teacher-rated ODD and ADHD behavior in 7, 9 and 12-year-olds were measurement invariant for student gender and teacher gender. Revision of the fourth scale (ATT) is highly recommended in order to be useable in clinical practice. The heritability of ODD and ADHD behavior was lower for children in different

classrooms compared to children sharing a classroom, suggesting that different behavior is elicited by different classroom environments. Apparently, teachers, the classroom and/or peers are important environmental factors that influence the expression of ODD and ADHD behavior in primary school. The direction of the moderation of the heritability of ODD and ADHD behavior by gender of the teacher was not consistent, which makes interpretation difficult. Finding environmental factors with a moderating influence on the heritability ODD and ADHD might help improve learning environments at school to prevent manifestation of ODD and ADHD behavior in children with an increased genetic vulnerability for these disorders.

TABLE S1 Model fitting results for measurement invariance tested in three age groups across gender of the teacher and gender of the student

		N	ep	RMSEA	χ^2	CFI	χ^2 Difference Test	df	p
Oppositional Behavior									
Age 7	EFA	8552	60	.058	150.195	.994			
	Configural	8552	60	.060	173.850	.993			
	Strong	8552	36	.034	188.452	.994	50.395	24	.001
	Strict	8552	21	.039	202.633	.994	25.778	15	.040
Age 9	EFA	7962	60	.073	215.595	.993			
	Configural	7962	60	.074	237.804	.993			
	Strong	7962	36	.044	214.997	.994	33.557	24	.093
	Strict	7962	21	.042	263.845	.993	58.267	15	<.001
Age 12	EFA	5904	60	.065	130.095	.996			
	Configural	5904	60	.065	143.429	.996			
	Strong	5904	36	.041	152.748	.996	45.131	24	.006
	Strict	5904	21	.037	180.625	.996	33.185	15	.004
Cognitive Problems/ Inattention									
Age 7	EFA	8551	60	.094	382.373	.986			
	Configural	8551	60	.091	376.516	.987			
	Strong	8551	36	.079	633.634	.979	303.322	24	<.001
	Strict	8551	21	.073	723.741	.976	126.082	15	<.001
Age 9	EFA	7963	60	.145	840.426	.956			
	Configural	7963	60	.140	799.807	.963			
	Strong	7963	36	.130	1528.966	.930	765.792	24	<.001
	Strict	7963	21	.119	1721.781	.921	250.020	15	<.001
Age 12	EFA	5904	60	.147	645.088	.956			
	Configural	5904	60	.147	660.227	.961			
	Strong	5904	36	.131	1150.344	.932	530.606	24	<.001
	Strict	5904	21	.119	1291.816	.925	166.737	15	<.001
Hyperactivity									
Age 7	EFA	8552	84	.044	242.830	.995			
	Configural	8552	84	.041	261.458	.995			
	Strong	8552	48	.035	329.143	.994	100.176	36	<.001
	Strict	8552	27	.033	383.403	.993	77.061	21	<.001

Age 9	EFA	7959	84	.043	221.821	.994			
	Configural	7959	84	.043	267.452	.994			
	Strong	7959	48	.033	288.498	.994	75.832	36	<.001
	Strict	7959	27	.031	330.362	.993	59.778	21	<.001
Age 12	EFA	5904	84	.038	134.893	.995			
	Configural	5904	84	.041	194.261	.993			
	Strong	5904	48	.029	208.982	.994	50.365	36	.056
	Strict	5904	27	.032	281.340	.992	75.149	21	<.001
ADHD Index									
Age 7	EFA	8552	136	.086	2205.268	.984			
	Configural	8552	136	.070	1983.366	.986			
	Strong	8552	82	.060	1948.785	.987	100.227	54	<.001
	Strict	8552	49	.050	1661.353	.989	47.255	33	.052
Age 9	EFA	7961	136	.082	1868.673	.985			
	Configural	7961	136	.073	1979.756	.984			
	Strong	7961	82	.063	2012.996	.984	155.307	54	<.001
	Strict	7961	49	.054	1757.824	.986	60.369	33	.003
Age 12	EFA	5904	136	.078	1270.317	.985			
	Configural	5904	136	.064	1214.061	.986			
	Strong	5904	82	.054	1201.933	.987	81.171	54	.010
	Strict	5904	49	.048	1143.753	.988	90.742	33	<.001

N = number of observations; ep = estimated parameters; RMSEA = root mean square error of approximation; X2 = chi square; CFI = comparative fit index; df = degrees of freedom; EFA = exploratory factor analysis

TABLE S2 Genetic modeling results for the oppositional behavior (OPP) scale

	ep	-2ll	df	model	χ2	Δdf	p
Age 7							
0 Saturated	52	14503.83	7379	-	-	-	-
1 Saturated: ST = DT	37	14614.79	7394	0	110.96	15	<.001
2 ACE	23	14583.00	7408	0	79.17	29	<.001
3 ACE: Boys = Girls	15	14673.59	7416	2	90.58	8	<.001
4 ACE: ST = DT	17	14656.08	7414	2	73.08	6	<.001
5 ACE: Correlated Errors	18	14592.64	7413	0	87.79	33	<.001
6 ACE: FT = MT	*17*	*14587.70*	*7414*	*2*	*4.70*	*6*	*.583*
Age 9							
0 Saturated	52	14271.56	6713	-	-	-	-
1 Saturated: ST = DT	37	14417.89	6728	0	146.33	15	<.001
2 ACE	*23*	*14302.35*	*6742*	*0*	*30.79*	*29*	*.375*
3 ACE: Boys = Girls	15	14385.60	6750	2	83.25	8	<.001
4 ACE: ST = DT	17	14428.55	6748	2	126.19	6	<.001
5 ACE: Correlated Errors	18	14349.08	6747	0	75.60	33	<.001
6 ACE: FT = MT	17	14322.82	6748	2	20.47	6	.002
Age 12							
0 Saturated	52	10447.34	4913	-	-	-	-
1 Saturated: ST = DT	37	10509.68	4928	0	62.34	15	<.001
2 ACE	*23*	*10461.64*	*4942*	*0*	*14.30*	*29*	*.990*
3 ACE: Boys = Girls	15	10538.20	4950	2	76.56	8	<.001
4 ACE: ST = DT	17	10509.94	4948	2	48.30	6	<.001
5 ACE: Correlated Errors	18	10515.73	4947	0	59.45	33	.003
6 ACE: FT = MT	17	10498.14	4948	2	36.50	6	<.001

FT = female teacher; MT = male teacher; DT = different teacher; ST = same teacher; ep = estimated parameters; df = degrees of freedom; -2ll = -2loglikelihood; A = additive genetic effects; C = common environmental effects; E = unique environmental effects

TABLE S3 Genetic modeling results for the hyperactivity (HYP) scale

	ep	-2ll	df	model	χ2	Δdf	p
Age 7							
0 Saturated	52	20030.50	7374	-	-	-	-
1 Saturated: ST = DT	37	20187.51	7389	0	157.01	15	<.001
2 ACE	23	20063.58	7403	0	33.08	29	.275
3 ACE: Boys = Girls	15	20102.04	7411	2	38.46	8	<.001
4 ACE: ST = DT	17	20199.68	7409	2	136.10	6	<.001
5 ACE: Correlated Errors	18	20085.22	7408	0	54.00	33	.012
6 ACE: FT = MT	*17*	*20078.37*	*4709*	*2*	*14.79*	*6*	*.022*
Age 9							
0 Saturated	52	17649.84	6709	-	-	-	-
1 Saturated: ST = DT	37	17783.59	6724	0	133.76	15	<.001
2 ACE	23	17681.08	6738	0	31.24	29	.354
3 ACE: Boys = Girls	15	17707.93	6746	2	26.84	8	.001
4 ACE: ST = DT	17	17793.64	6744	2	112.56	6	<.001
5 ACE: Correlated Errors	18	17704.15	6743	0	53.63	33	.013
6 ACE: FT = MT	*17*	*17697.01*	*6744*	*2*	*15.92*	*6*	*.014*
Age 12							
0 Saturated	52	12142.50	4917	-	-	-	-
1 Saturated: ST = DT	37	12258.51	4932	0	117.01	15	<.001
2 ACE	*23*	*12176.31*	*4946*	*0*	*33.81*	*29*	*.246*
3 ACE: Boys = Girls	15	12219.40	4954	2	43.10	8	<.001
4 ACE: ST = DT	17	12249.56	4952	2	73.26	6	<.001
5 ACE: Correlated Errors	18	12216.43	4951	0	74.35	33	<.001
6 ACE: FT = MT	17	12204.84	4952	2	28.53	6	<.001

FT = female teacher; MT = male teacher; DT = different teacher; ST = same teacher; ep = estimated parameters; df = degrees of freedom; -2ll = -2loglikelihood; A = additive genetic effects; C = common environmental effects; E = unique environmental effects

TABLE S4 Genetic modeling results for the ADHD index (ADHD) scale

	ep	-2ll	df	model	χ2	Δdf	p
Age 7							
0 Saturated	52	24482.63	7369	-	-	-	-
1 Saturated: ST = DT	37	24614.63	7384	0	132.00	15	<.001
2 ACE	23	24513.40	7398	0	30.77	29	.376
3 ACE: Boys = Girls	*15*	*24533.14*	*7406*	*2*	*19.73*	*8*	*.011*
4 ACE: ST = DT	12	24640.09	7409	3	106.95	3	<.001
5 ACE: Correlated Errors	13	24549.49	7408	3	59.55	38	.014
6 ACE: FT = MT	12	24546.27	7409	3	13.13	3	.004
Age 9							
0 Saturated	52	22137.31	6703	-	-	-	-
1 Saturated: ST = DT	37	22271.04	6718	0	133.72	15	<.001
2 ACE	23	22159.55	6732	0	22.24	29	.810
3 ACE: Boys = Girls	15	22174.92	6740	2	15.37	8	.052
4 ACE: ST = DT	12	22274.78	6743	3	99.85	3	<.001
5 ACE: Correlated Errors	13	22197.56	6742	0	60.25	38	.012
6 ACE: FT = MT	*12*	*22176.08*	*6743*	*3*	*1.15*	*3*	*.765*
Age 12							
0 Saturated	52	15589.30	4912	-	-	-	-
1 Saturated: ST = DT	37	15704.31	4927	0	115.02	15	<.001
2 ACE	23	15624.83	4941	0	35.53	29	.188
3 ACE: Boys = Girls	15	15638.42	4949	2	13.59	8	.093
4 ACE: ST = DT	12	15733.73	4952	3	95.30	3	<.001
5 ACE: Correlated Errors	13	15679.13	4951	0	89.60	38	<.001
6 ACE: FT = MT	*12*	*15645.36*	*4952*	*3*	*6.94*	*3*	*.074*

FT = female teacher; MT = male teacher; DT = different teacher; ST = same teacher; ep = estimated parameters; df = degrees of freedom; -2ll = -2loglikelihood; A = additive genetic effects; C = common environmental effects; E = unique environmental effects

REFERENCES

Abikoff, H.B., Jensen, P.S., Arnold, L.L., Hoza, B., Hechtman, L., Pollack, S., Martin, D., Alvir, J., March, J.S., Hinshaw, S., Vitiello, B., Newcorn, J., Greiner, A., Cantwell, D.P., Conners, C.K., Elliott, G., Greenhill, L.L., Kraemer, H., Pelham, W.E., Jr., Severe, J.B., Swanson, J.M., Wells, K., & Wigal, T. (2002). Observed classroom behavior of children with ADHD: relationship to gender and comorbidity. *Journal of Abnormal Child Psychology, 30 (4)*, p. 349-359.

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

American Psychiatric Association (2000). Diagnostic and Statistical Manual of Mental Disorders: 4th ed., text rev. DSM-IV-TR. Washington, DC: American Psychiatric Association.

Angold, A., Costello, E.J., & Erkanli, A. (1999). Comorbidity. *Journal of Child Psychology and Psychiatry, 40 (01)*, p. 57-87.

Boker S.M., Neale, M.C., Maes, H.H.M., Wilde M.J., Spiegel M., Brick T.R., Estabrook R., Bates T.C., Mehta P., von Oertzen T., Gore R.J., Hunter M.D., Hackett D.C., Karch J., & Brandmaier A. (2012). OpenMx 1.2 User guide.

Boker S.M., Neale, M.C., Maes, H.H.M., Wilde M.J., Spiegel M., Brick T.R., Spies J., Estabrook R., Kenny S., Bates T.C., Mehta P., & Fox J. (2011). An open source extended structural equation modeling framework. *Psychometrika, 76 (2)*, p. 306-317.

Boomsma, D.I., de Geus, E.J., Vink, J.M., Stubbe, J.H., Distel, M.A., Hottenga, J.J., Posthuma, D., van Beijsterveldt, T.C., Hudziak, J.J., Bartels, M., & Willemsen, G. (2006). Netherlands Twin Register: from twins to twin families. *Twin Research and Human Genetics, 9 (6),* p. 849-857.

Boomsma, D.I., Vink, J.M., van Beijsterveldt, T.C., de Geus, E.J., Beem, A.L., Mulder, E.J., Derks, E.M., Riese, H., Willemsen, G.A., Bartels, M., van den, B.M., Kupper, N.H., Polderman, T.J., Posthuma, D., Rietveld, M.J., Stubbe, J.H., Knol, L.I., Stroet, T., & van Baal, G.C. (2002). Netherlands Twin Register: a focus on longitudinal research. *Twin Research, 5 (5)*, p. 401-406.

Brendgen, M., Boivin, M., Vitaro, F., Girard, A., Dionne, G., & Perusse, D. (2008). Gene-environment interaction between peer victimization and child aggression. *Developmental Psychopathology, 20 (2)*, p. 455-471.

Bronfenbrenner, U. & Ceci, S.J. (1994). Nature-nurture reconceptualized in developmental perspective: a bioecological model. *Psychological Reviews, 101 (4)*, p. 568-586.

Burt, S.A., Krueger, R.F., McGue, M., & Iacono, W.G. (2001). Sources of covariation among attention-deficit/hyperactivity disorder, oppositional defiant disorder, and conduct disorder: the importance of shared environment. *Journal of Abnormal Psychology, 110 (4)*, p. 516-525.

Carrington, B., Tymms, P., & Merrell, C. (2008). Role models, school improvement and the gender gap: do men bring out the best in boys and women the best in girls? *British Educational Research Journal, 34 (3)*, p. 315-327.

Conners, C.K. (2001). Conners' rating scales - revised. New York, NY: Multi-Health Systems, Inc.

Conners, C.K., Sitarenios, G., Parker, J.D., & Epstein, J.N. (1998). Revision and restandardization of the Conners Teacher Rating Scale (CTRS-R): factor structure, reliability, and criterion validity. *Journal of Abnormal Child Psychology, 26 (4)*, p. 279-291.

Derks, E.M., Dolan, C.V., & Boomsma, D.I. (2004). Effects of censoring on parameter estimates and power in genetic modeling. *Twin Research, 7 (6)*, p. 659-669.

Derks, E.M., Dolan, C.V., Hudziak, J.J., Neale, M.C., & Boomsma, D.I. (2007). Assessment and etiology of attention deficit hyperactivity disorder and oppositional defiant disorder in boys and girls. *Behavior Genetics, 37 (4)*, p. 559-566.

Derks, E.M., Hudziak, J.J., & Boomsma, D.I. (2007). Why more boys than girls with ADHD receive treatment: a study of Dutch twins. *Twin Reearch and Human Genetics, 10 (5)*, p. 765-770.

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 (2)*, p. 219-229.

Dolan, C.V. (2000). Investigating Spearman's hypothesis by means of multi-group confirmatory factor analysis. *Multivariate Behavioral Research, 35 (1)*, p. 21-50.

Eaves, J.L. (1984). The resolution of genotype x environment interaction in segregation analysis of nuclear families. *Genetic Epidemiology, 1,* p. 215-228.

Faraone, S.V., Perlis, R.H., Doyle, A.E., Smoller, J.W., Goralnick, J.J., Holmgren, M.A., & Sklar, P. (2005). Molecular genetics of attention-deficit/hyperactivity disorder. *Biological Psychiatry, 57 (11)*, p. 1313-1323.

Flora, D.B. & Curran, P.J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9 (4)*, p. 466-491.

Frick, P.J., Lahey, B.B., Loeber, R., Stouthamer-Loeber, M., Christ, M.A., & Hanson, K. (1992). Familial risk factors to oppositional defiant disorder and conduct disorder: parental psychopathology and maternal parenting. *Journal of Consulting and Clinical Psychology, 60 (1)*, p. 49-55.

Gaub, M. & Carlson, C.L. (1997). Gender differences in ADHD: a meta-analysis and critical review. *Journal of the American Academy of Child and Adolsecent Psychiatry, 36 (8)*, p. 1036-1045.

Greene, R.W., Biederman, J., Zerwas, S., Monuteaux, M.C., Goring, J.C., & Faraone, S.V. (2002). Psychiatric comorbidity, family dysfunction, and social impairment in referred youth with oppositional defiant disorder. *American Journal of Psychiatry, 159 (7)*, p. 1214-1224.

Hartman, C.A., Rhee, S.H., Willcutt, E.G., & Pennington, B.F. (2007). Modeling rater disagreement for ADHD: are parents or teachers biased? *Journal of Abnormal Child Psychology, 35 (4)*, p. 536-542.

Hoyt, W.T. (2000). Rater bias in psychological research: when is it a problem and what can we do about it? *Psychological Methods, 5 (1)*, p. 64-86.

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. *J.Am.Acad.Child Adolesc.Psychiatry, 44 (9)*, p. 907-914.

Kan, K.J., Dolan, C.V., Nivard, M.G., Middeldorp, C.M., van Beijsterveldt, C.E., Willemsen, G., & Boomsma, D.I. (2013). Genetic and environmental stability in attention problems across the lifespan: evidence from the Netherlands twin register. *Journal of the American Academy of Child and Adolsecent, 52 (1)*, p. 12-25.

Lamb, D.J., Middeldorp, C.M., van Beijsterveldt, C.E., & Boomsma, D.I. (2012). Gene-environment interaction in teacher-rated internalizing and externalizing problem behavior in 7- to 12-year-old twins. *Journal of Child Psychology and Psychiatry, 53 (8)*, p. 818-825.

Lubke, G.H., Dolan, C.V., & Neale, M.C. (2004). Implications of absence of measurement invariance for detecting sex limitation and genotype by environment interaction. *Twin Research, 7 (3)*, p. 292-298.

Meisel, V., Servera, M., Cardo, E., & Garcia-Banda, G. (2013). Prevalence of oppositional defiant disorder in a sample of Spanish schoolchildren. *Spanish Journal of Psychology, 16* p. E63.

Mellenbergh, G.J. (1989). Item bias and item response theory. *Educational Research, 13,* p. 127-143.

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

Merwood, A., Greven, C.U., Price, T.S., Rijsdijk, F., Kuntsi, J., McLoughlin, G., Larsson, H., & Asherson, P.J. (2013). Different heritabilities but shared etiological influences for parent, teacher and self-ratings of ADHD symptoms: an adolescent twin study. *Psychological Medicine, 43 (9)*, p. 1973-1984.

Millsap, R.E. & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research, 39 (3)*, p. 479-515.

Muthén, L.K. & Muthén, B.O. (2010). Mplus user's guide. *6th Edition*. Los Angeles, CA: Muthén L.K. & Muthén.

Neale, M.C., Roysamb, E., & Jacobson, K. (2006). Multivariate genetic analysis of sex limitation and G x E interaction. *Twin Research and Human Genetics, 9 (4)*, p. 481-489.

Paap, M.C., Haraldsen, I.R., Breivik, K., Butcher, P.R., Hellem, F.M., & Stormark, K.M. (2013). The link between peer relations, prosocial behavior, and ODD/ADHD symptoms in 7-9-year-old children. *Psychiatry Journal, 2013,* p. 319874.

Plomin R., DeFries J.C., McClearn G.E., & McGuffin P.s (2008). Behavioral Genetics. *5th Edition.* New York, NY: Worth Publishers.

Polderman, T.J.C., Boomsma, D.I., Bartels, M., Verhulst, F.C., & Huizink, A.C. (2010). A systematic review of prospective studies on attention problems and academic achievement. *Acta Psychiatrica Scandinavica, 122 (4)*, p. 271-284.

Purcell, S. (2002). Variance components models for gene-environment interaction in twin analysis. *Twin Research, 5 (6)*, p. 554-571.

R Core Team (2014). R: A language and environment for statistical computing. Vieanna, Austria: R Foundation for Statistical Computing.

Rende, R. & Plomin, R. (1992). Diathesis-stress models of psychopathology: A quantitative genetic perspective. *Applied and Preventive Psychology, 1 (4)*, p. 177-182.

Rietveld, M.J., van, D., V, Bongers, I.L., Stroet, T.M., Slagboom, P.E., & Boomsma, D.I. (2000). Zygosity diagnosis in young twins by parental report. *Twin Research, 3 (3)*, p. 134-141.

Roberts, W. & Strayer, J. (1996). Empathy, emotional expressiveness, and prosocial behavior. *Child Development, 67 (2)*, p. 449-470.

Rydell, A.M. & Henricsson, L. (2004). Elementary school teachers' strategies to handle externalizing classroom behavior: a study of relations between perceived control, teacher orientation and strategy preferences. *Scandinavian Journal of Psychology, 45 (2)*, p. 93-102.

Saudino, K.J., Ronald, A., & Plomin, R. (2005). The etiology of behavior problems in 7-year-old twins: substantial genetic influence and negligible shared environmental influence for parent ratings and ratings by same and different teachers. *Journal of Abnormal Child Psychology, 33 (1)*, p. 113-130.

Schermelleh-Engel, K. & Moosbrugger, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online, 8,* p. 23-74.

Sciutto, M.J., Nolfi, C.J., & Bluhm, C. (2004). Effects of child gender and symptom type on referrals for ADHD by elementary school teachers. *Journal of Emotional and Behavioral Disorders, 12 (4)*, p. 247-253.

Simonoff, E., Pickles, A., Hervas, A., Silberg, J.L., Rutter, M., & Eaves, L. (1998). Genetic influences on childhood hyperactivity: contrast effects imply parental rating bias, not sibling interaction. *Psychological Medicine, 28 (4)*, p. 825-837.

Slof-Op 't Landt MC, Dolan, C.V., Rebollo-Mesa, I., Bartels, M., Van Furth, E.F., van Beijsterveldt, C.E., Meulenbelt, I., Slagboom, P.E., & Boomsma, D.I. (2009). Sex differences in sum scores may be hard to interpret: the importance of measurement invariance. *Assessment, 16 (4)*, p. 415-423.

Somersalo, H., Solantaus, T., & Almqvist, F. (2002). Classroom climate and the mental health of primary school children. *Nordic Journal of Psychiatry, 56 (4)*, p. 285-290.

Towers, H., Spotts, E., Hetherington, E.M., Plomin, R., & Reiss, D. (2000). Genetic and environmental influences on teacher ratings of the Child Behavior Checklist. *International Journal of Behavioral Development, 24,* p. 373-381.

Tuvblad, C., Zheng, M., Raine, A., & Baker, L.A. (2009). A common genetic factor explains the covariation among ADHD ODD and CD symptoms in 9-10 year old boys and girls. *Journal of Abnormal Child Psychology, 37 (2)*, p. 153-167.

van Beijsterveldt, C.E., Groen-Blokhuis, M., Hottenga, J.J., Franic, S., Hudziak, J.J., Lamb, D., Huppertz, C., de Zeeuw, E., Nivard, M., Schutte, N.,

Swagerman, S., Glasner, T., van Fulpen, M., Brouwer, C., Stroet, T., Nowotny, D., Ehli, E.A., Davies, G.E., Scheet, P., Orlebeke, J.F., Kan, K.J., Smit, D., Dolan, C.V., Middeldorp, C.M., de Geus, E.J., Bartels, M., & 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 (1)*, p. 252-267.

Vitaro, F., Tremblay, R.E., & Gagnon, C. (1995). Teacher ratings of children's behaviors and teachers' management styles: a research note. *Journal of Child Psychology and Psychiatry, 36 (5)*, p. 887-898.

Wilens, T.E., Biederman, J., Brown, S., Tanguay, S., Monuteaux, M.C., Blake, C., & Spencer, T.J. (2002). Psychiatric comorbidity and functioning in clinically referred preschool children and school-age youths with ADHD. *Journal of the American Academy of Child & Adolescent Psychiatry, 41 (3)*, p. 262-268.