Developmental Psychology
2018, Vol. 54, No. 1, 39–50

0012-1649/18/$12.00 http://dx.doi.org/10.1037/dev0000403

Genetic and Environmental Contributions to the Development of Childhood Aggression

Gitta H. Lubke
University of Notre Dame and VU University Amsterdam

Daniel B. McArtor
University of Notre Dame

Dorret I. Boomsma and Meike Bartels
VU University Amsterdam

Longitudinal data from a large sample of twins participating in the Netherlands Twin Register (n = 42,827, age range 3–16) were analyzed to investigate the genetic and environmental contributions to childhood aggression. Genetic auto-regressive (simplex) models were used to assess whether the same genes are involved or whether new genes come into play as children grow up. The authors compared 2 different simplex models to disentangle potentially changing behavioral expressions from changes in genetic and environmental effects. One model provided estimates of genetic and environmental effects at the level of individual aggression questionnaire items, and the other model assessed the effects at the level of an aggression sum score computed from the individual items. The results from both models provided evidence for largely stable genetic effects throughout childhood. The results also highlighted the differential heritability of the different indicators of aggression measured with the Childhood Behavior Checklist, with destruction of property showing a very high genetic component during early childhood and fighting behaviors being more heritable in early adolescence.

Keywords: aggression, genetic simplex, longitudinal twin modeling, genetic stability and genetic innovation, measurement invariance

Supplemental materials: http://dx.doi.org/10.1037/dev0000403.supp

Social impairment resulting from childhood aggression can impose substantial personal and financial burdens on affected children, their caretakers, and their peers. Depending on its operationalization, the prevalence of aggression in children ranges from 2% to 16%, with rates varying across sex and age, and early onset of aggression in children is often predictive of aggressive behavior in adolescence and adulthood (Connor, 2004; Loeber & Hay, 1997). The stability of aggressive behavior in children generally is high, and early aggressiveness is predictive of later serious antisocial behavior and self-reported overt aggression (Huesmann et al., 1984).

A number of behavior genetic studies have investigated the etiology of the stability in aggressive behavior. A recent study comparing the development of aggression between Dutch and English population samples showed that the stability and heritability of aggressive behavioral problems was high for both samples, with longitudinal genetic correlations being the main reason for stability of aggressive behavior (Porsch et al., 2016). In earlier work, van Beijsterveldt et al. (2003) found that genetic factors accounted for most of the stability of aggression across childhood; a genetic longitudinal model suggested a dynamic developmental process consisting of the transmission of existing genetic effects as well as the onset of new genetic influences as children grow up. The authors identified some modification of genetic influences by age and sex. Drawing similar conclusions, Pingault, Rijsdijk, Zheng, Plomin, & Viding (2015) demonstrated the stable effect of genetic factors on conduct problems using a latent growth model and suggested that genetic factors may explain a large amount of the individual variability in developmental trajectories. Furthermore, Lewis and Plomin (2015) observed that genetic factors were the main source of stability in conduct problems between ages 4 and 16, with constant effects of genetic influences throughout childhood. Another longitudinal study of 750 U.S. twin pairs (Niv, Tuvblad, Raine, & Baker, 2013) during childhood (ages 9–10) and

This article was published Online First October 23, 2017.

Gitta H. Lubke, Department of Psychology, University of Notre Dame, and Department of Biological Psychology, VU University Amsterdam; Daniel B. McArtor, Department of Psychology, University of Notre Dame; Dorret I. Boomsma and Meike Bartels, Department of Biological Psychology, VU University Amsterdam.

Gitta H. Lubke, Dorret I. Boomsma, and Meike Bartels are supported by EU FP7-602768. Gitta H. Lubke is also supported by NIDA R37 DA-018673. Daniel B. McArtor is supported by the Royal Netherlands Academy of Science Professor Award (PAH/6635). The Netherlands Twin Register is supported by the following: The Netherlands Organization for Scientific Research (NWO) and MagW/ZonMW Grants 904-61-090, 985-10-002, 904-61-193,480-04-004, 400-05-717, 463-06-001, 451-04-034, Middelgroot-911-09-032, Spinozapremie 56-464-14192, Biobanking and Biomolecular Resources Research Infrastructure (BBMRI –NL, 184.021.007), VU University's Institute for Health and Care Research (EMGO+), the European Research Council (ERC Advanced, 230374), the Avera Institute, Sioux Falls, South Dakota.

Correspondence concerning this article should be addressed to Gitta H. Lubke, 110 Haggar Hall, University of Notre Dame, Notre Dame, IN 46656. E-mail: glubke@nd.edu

adolescence (ages 14–15) indicated that some of the genetic factors that influence antisocial behavior (a latent trait that combines aggression and rule-breaking behavior) in adolescence overlap with the genetic influences in childhood, but some of the genetic influences found at ages 14–15 were not observed in younger children. Using data from 9- to 10- and 11- and 14-year-old twins, Tuvblad, Raine, Zheng, and Baker (2009) focused on the differences between reactive and proactive aggression and found that especially proactive aggression was largely genetically stable. Lacourse et al. (2014) studied genetic and environmental stability of overt aggression in young children (20–50 months old) and found moderate genetic stability. In line with these results, a recent meta-analysis showed that aggression is more stable in middle childhood than in early childhood, but a predominant genetic influence on stability was observed across both age groups (Hannigan, Walaker, Waszczuk, McAdams, & Eley, 2017).

In this study we focus on overt (or physical) aggression for two reasons. First, overt aggression is part of the diagnoses of oppositional defiant disorder (ODD) and conduct disorder (CD). Within ODD, overt aggression is considered to be predictive of CD (see, e.g., Althoff, Kuny-Slock, Verhulst, Hudziak, & van der Ende, 2014). Overt aggression therefore forms an important shared aspect of these disorders. Second, our choice is based on earlier work with data from the Netherlands Twin Register, which showed that overt aggression and relational aggression were both influenced by a single underlying set of shared environmental factors, but only partly impacted by the same genes (Ligthart et al., 2005). This less than perfect genetic correlation implies partly independent genetic influences, which is likely due the fact that overt and relational aggression do not form a unidimensional construct.

The primary goal of this study is to quantify the extent to which genetic and environmental factors influence overt aggression as children grow up. More specifically, we want to investigate whether mainly the same genes affect overt aggressive behaviors across childhood ("genetic stability") or whether different genes become active throughout childhood and early adolescence ("genetic innovation"). Achieving this goal is not straightforward, however, because the questionnaire items that measure aggression change as children progress from early childhood into adolescence as a result of aggression manifesting differently across age groups. The secondary goal of this study is therefore methodological; we propose using the framework of the genetic simplex model (Boomsma & Molenaar, 1987) to disentangle the effects of changes in measurement from the genetic and environmental bases of aggression across childhood.

In addition to increasing the general understanding of genetic and environmental influences on aggression, the quantification of its genetic stability versus its genetic innovation also provides the necessary basis for the design of successful gene-finding studies. For instance, if the same genes are relevant in explaining aggression throughout childhood, data collected across multiple age groups could be aggregated to increase sample sizes and therefore statistical power. In case of substantial genetic innovation, on the other hand, the aggregation of data across different ages would introduce heterogeneity and therefore potentially decrease power.

An important question beyond that of gene finding concerns the nature of environmental influences, specifically, to what extent environmental influences on aggression are shared by children who grow up in the same family. In genetic models based on the classical twin design, these shared environmental influences (C: common/shared environment) can be distinguished from the environmental influences that act on an individual level (E: nonshared environment), provided that the sample size is large enough to provide sufficient power (Posthuma & Boomsma, 2000). It is important to investigate the extent to which both of these environmental influences are stable across childhood as well as the extent to which they change.

The genetic simplex model is a statistical tool for longitudinal twin data designed to assess not only the stability of genetic and environmental effects as children grow up, but also the extent to which new effects take root starting at particular ages. This is accomplished by conducting a variance decomposition of the observed scores at the different measurement occasions. Most commonly, sum scores (SS) of multiple questionnaire items are used in a genetic simplex model.[1] Alternatively, the model can be applied at the item-level such that changes of genetic and environmental effects can be studied for each questionnaire item. In either case, the model is used to estimate the extent to which genetic and environmental variance components of the underlying trait are transmitted from one age to the next (i.e., stability effects) and the extent to which new sources of variance take effect starting at a given age point (i.e., innovation effects). The ability to estimate the stability and innovation of each variance component makes the genetic simplex an ideal tool for assessing the extent to which the genetic and environmental factors influencing aggression change as children develop. This feature differentiates the genetic simplex from other common longitudinal twin models like the genetic growth-curve model, which focuses instead on using genetic and environmental differences to explain changes in mean structure over time.

In the case of childhood aggression, the task of partitioning innovated genetic and environmental variance components from transmitted components is not trivial because questionnaire items designed to measure aggressive behaviors can change in relevance throughout childhood. The item, "threatens people," for example, is not expected to be a strong indicator of aggression in young children, but its relevance is expected to increase throughout childhood and into adolescence. Changes in measurement properties such as increasing relevance over time (as quantified by increasing factor loadings) imply that measurement invariance (MI) does not hold over time. If MI holds, then it is straightforward to conclude that the same factor or trait is measured over time (Meredith & Horn, 2001). If, however, MI is expected to be violated, then the situation is more complicated. If loadings change over time in an expected direction (e.g., the loading of item "threatens people" on the underlying aggression factor increases as children's cognitive and verbal abilities increase), then one could argue that the violation of MI does not imply that different constructs are measured across time per se. In this case, it could be argued that the anticipated changes in factor loading structures reflect changes in the behavioral expression of the same underlying construct (see also arguments in Byrne, Shavelson, & Muthén, 1989).

[1] In practice, average scores are computed to account for missing data, but we use the terminology *sum score* to stay consistent with the literature.

To disentangle genetic and environmental contributions to childhood aggression in light of the problem of differential item relevance across age groups, we fit two different genetic simplex models that reflect different assumptions about these changes. The first model uses SS at each measurement occasion. When computing SS, all items have the same weight (usually 1), and these weights remain the same over time. Differences in relevance between items or changes in relevance over time are not acknowledged. In the genetic SS simplex, the SS variance is decomposed into genetic and environmental factors. Without measurement invariance, however, estimates of genetic and environmental stability and innovation extracted from SS are biased; a similar bias is observed in growth factor variances based on the standard twin growth model (Luningham, McArtor, Bartels, Boomsma, & Lubke, 2017). We include the SS model for reference because of the ubiquity of SS in empirical studies.

The second model is an item-level genetic simplex in which each item's contribution to the genetic and environmental factors is freely estimated at each observed time point rather than being fixed to one, as is the case in the SS model. This approach affords two advantages. First, by estimating the factor loadings onto the A, C, and E factors for each item, the genetic and environmental contributions can be evaluated for each questionnaire item separately. In the SS model, these loadings are constrained to be equal for all items (see Figure 2). Second, the free estimation of the loadings in the independent pathway (IP) model results in a more trustworthy estimation of genetic innovation and stability. To clarify this statement, consider the likely case that the measurement of aggression changes somewhat over time (i.e., the items have somewhat different relevance at different ages). In other words, measurement invariance is absent, and as a result, the A, C, and E factors at the different time points differ somewhat in their interpretation due to the changing factor loadings over time. For instance, the item "destroys things belonging to his/her family or others" might have a higher loading on the A factor at age 7 compared to age 12. In that case, the A factor at age 7 reflects to a larger extent the genetic component of "destroys" than at age 12. Importantly, the fact that the loadings in the IP model are freely estimated is expected to more accurately characterize the stability of genetic and environmental effects. This is because the estimated proportion of variance of A due to transmission (i.e., the estimate of stability) will only reflect what is truly shared between the freely estimated A factors across successive measurement occasions. Any contribution to the variance of A that is not transmitted from the previous time point, whether new genes or changes in measurement properties of the items, will be reflected in the estimates of innovation. Note that the same argument holds for C and E since the corresponding loadings in the IP model are also freely estimated.

To illustrate the genetic simplex for item-level data, consider the most extreme case in which an item is irrelevant as an indicator of aggression at age (time), $t - 1$, but is a very good indicator at age t. The variance of this item at age $t - 1$ is entirely due to error and is therefore not affected by genetic or environmental influences. Consequentially, we expect only innovation and zero stability at age t because there are no effects that could be transmitted from $t - 1$ to t. It is still possible, however, that the same genes or environmental drivers of aggression are active throughout childhood and that only the behavioral expression of aggression has changed. In other words, although the type of behavioral manifestation of aggression changes, the underlying genetic and environmental factors can be stable. When using a simplex model to directly decompose the variance of individual items at each measurement occasion, we expect at least some degree of genetic and environmental innovation due to changes in relevance of the items as indicators of aggression throughout childhood. By considering the genetic decomposition into A, C, and E for each item, the results of the item-level genetic simplex permit a more detailed look at the stability and change of genetic and environmental influences over time.

This investigation of the genetic and environmental stability and innovation of overt childhood aggression is carried out using longitudinal aggression data collected from a large population-based sample of Dutch twins, with age-appropriate checklists from the Achenbach System of Empirically Based Assessment (ASEBA) taxonomy being administered at six measurement occasions between ages 3 and 16 years.

Method

Participants

This study used repeated measures of aggression collected by the Young Netherlands Twin Register. Twins in the Young Netherlands Twin Register are registered shortly after birth by their parents; details concerning the sample and available data are described in van Beijsterveldt et al. (2013). The data analyzed in this study were collected between 1987 and 2011. The sample consisted of 42,827 individuals (21,565 girls and 21,262 boys from 21,608 families), with data collection occurring when the twins were approximately 3, 7, 10, 12, 14, and 16 years old.

The full dataset was partitioned into two independent subsets. A small exploratory set of 1,600 participants (1,009 girls and 591 boys from 1,347 families) was randomly selected from the pool of subjects who had complete data for at least five of the six measurement occasions. The exploratory set was used to investigate the factor structure of the aggression items as well as the longitudinal structure of aggression. These results were used to formulate expectations about the genetic simplex models that were fit using the confirmatory set. Using mutually exclusive data sets for the exploration of the longitudinal factor structure and the fitting of genetically informative models avoids capitalization on chance that can lead to spurious results (Lubke & Campbell, 2016). The exploratory sample size of 1,600 was selected with the goal of obtaining stable estimates from the initial exploratory models while preserving as many observations as possible for use in the confirmatory analyses.

The confirmatory set was comprised of the remaining 41,227 participants (20,556 girls and 20,671 boys from 21,355 families). Supplementary Table S1 (see the online supplementary materials) enumerates the sample size available at each measurement occasion for boys and girls separately; these counts are also divided into the number of individuals at each measurement occasion that had also provided data at a younger age (i.e., "n repeat") and those that were measured for the first time (i.e., "n new"). Note that the total sample size available in each age group (i.e., "n total") decreases across age groups because the NTR data collection is prospective, with data collected when twins reach a particular age.

Therefore, more data are available from twin pairs at younger ages. Detailed information concerning attrition such as specific response rates and temporary financial constraints limiting data collection are given in van Beijsterveldt et al. (2013). Importantly, Bartels, van Beijsterveldt, et al. (2007) found no differences in externalizing and internalizing behaviors between groups that participated at all ages, only at age 3, or had temporary nonparticipation.

Because of established sex differences in aggression (Bartels et al., 2003; Björkqvist, 1994; Crick, Casas, & Mosher, 1997; McEvoy, Estrem, Rodriguez, & Olson, 2003), all analyses were conducted for boys and girls separately (for explanations, please see the Measurement Noninvariance Across Sex section). The girls in the confirmatory set each belonged to one of 14,017 twin pairs (3,355 monozygotic [MZ], 10,562 dizygotic [DZ]), and the boys each belonged to one of 13,917 twin pairs (3,710 MZ, 10,307 DZ). Note that the combined number of male and female twin pairs exceeds the number of pairs in the complete data because we included individual twins from opposite sex twin pairs as DZ twins with missing data for their cotwins.

Measures

Aggression was measured in this study throughout the course of childhood by using age-appropriate checklists from the ASEBA taxonomy. These scales are shown in Table 1. At age 3, mother-report data were collected by the Dutch version of the Preschool-Age Childhood Behavior Checklist (P-CBCL; Achenbach & Rescorla, 2000). The aggressive behavior subscale of the P-CBCL consists of 19 items rated with a 3-point Likert scale ranging from 0 (*not true*) to 2 (*very/often true of my child in the past 6 months*). Achenbach and Rescorla (2000) reported that the test–retest reliability of this subscale is 0.87.

When twins were around age 7, 10, and 12 years, mother-report data from the Dutch version of the School-Age Childhood Behavior Checklist (CBCL; Achenbach & Rescorla, 2001) were analyzed. The aggressive behavior subscale of the CBCL is comprised of 18 items rated with the same 3-point Likert scale as in the P-CBCL. Achenbach and Rescorla (2001) reported a test–retest reliability of 0.90 for this subscale. A total of eight of these items share similar content with the items comprising the P-CBCL based on face validity (see Table 1).

At age 14 and 16, self-report data were collected from twins and their siblings with the aggressive behavior subscale of the Dutch version of the Youth Self Report (YSR; Achenbach & Rescorla, 2001). The 17 items also had three ordered response categories, and the test–retest reliability of this scale is 0.88 (Achenbach & Rescorla, 2001). All of the items comprising this subscale share item content with the mother-rated CBCL at ages 7 through 12 based on face validity; the only difference in item wording between the scales is that items in the YSR are phrased in terms of the participant's own behavior to reflect the change to self-report assessment (see Table 1).

Analysis Plan

The analyses consist of several exploratory analyses and a main analysis, which is aimed at investigating the genetic and environmental contributions to childhood aggression with a focus on determining the extent to which the same genes are active in explaining aggression across childhood.

Table 1
Item Content From the P-CBCL, CBCL, and YSR

P-CBCL (Age 3)	CBCL (Age 7, 10, 12)	YSR (Age 14, 16)
Mother report	Mother report	Self-report
• **Destroys things belonging to his/her family or other children**	• Demands a lot of attention	• I am stubborn
• Disobedient	• **Destroys things belonging to his/her family or others**	• **I destroy things that belong to others**
• **Gets in many fights**	• Disobedient at home	• I disobey my parents
• **Physically attacks people**	• **Gets in many fights**	• **I get in many fights**
• Screams a lot	• **Physically attacks people**	• I have a hot temper
• Stubborn, sullen, or irritable	• Screams a lot	• **I physically attack people**
• Temper tantrums or hot temper	• Stubborn, sullen, or irritable	• I scream a lot
• Wants a lot of attention	• Temper tantrums or hot temper	• I try to get a lot of attention
• Angry moods	• Argues a lot	• I am louder than other kids
• Can't stand waiting; wants everything now	• Cruelty, bullying, or meanness to others	• I am mean to others
• Defiant	• **Destroys his/her own things**	• I am suspicious
• Demands must be met immediately	• Disobedient at school	• I argue a lot
• Doesn't seem to feel guilty after misbehaving	• Sudden changes in mood or feelings	• **I destroy my own things**
• Easily frustrated	• Sulks a lot	• I disobey at school
• **Hits others**	• Suspicious	• I tease others a lot
• **Hurts animals or people without meaning to**	• Teases a lot	• **I threaten to hurt people**
• Punishment doesn't change his/her behavior	• **Threatens people**	• My mood or feelings change suddenly
• Selfish or won't share	• Unusually loud	
• Uncooperative		

Note. Bold items measure the physical/overt aggression subscale. Note that the eight items in the top panel appear in all three scales with changes in phrasing to reflect the change in raters. The remaining 11 items comprising the Preschool Childhood Behavior Checklist do not appear in the other two scales, but all items in the Childhood Behavior Checklist except "sulks a lot" also appear in the Youth Self Report with slightly altered phrasing to reflect the shift to self-report assessment. Within each panel, items are sorted alphabetically.

Exploratory data analysis. An exploratory data analysis (EDA) was conducted in the exploratory subset of the data to guide modeling decisions in the main genetic analyses. The exploration consisted of (a) computing descriptive statistics, (b) fitting exploratory factor models, (c) fitting confirmatory factor models and assessing measurement invariance over time and across sex, and (d) computing autocorrelations to assess the longitudinal stability of the aggression data to confirms that simplex modeling is appropriate

Histograms of item responses, item means, variances, and correlations were obtained for each age group and for boys and girls separately. The dimensionality of the aggression scales at each measurement occasion was investigated by comparing exploratory factor analysis (EFA) models with up to 5 factors based on fit and interpretability. Tests of measurement invariance in confirmatory factor analyses (CFAs) assessed whether or not the items comprising a subscale measured the same latent construct across age, sex, and rater.

Measurement noninvariance across sex. Consistent with previous studies, we found substantial sex differences in the factor model relating aggression questionnaire items to underlying aggression factors in the exploratory data. It has been shown that when the sexes differ in their factor loadings, fitting a twin model that includes opposite sex twins leads to the confounding of measurement differences with differences in genetic and environmental contributions to the measured behavior (Lubke, Dolan, & Neale, 2004). To avoid bias in the estimated genetic and environmental effects, data from boys and girls were analyzed separately.

Main analysis: Estimation of genetic and environmental contributions to aggression. Analysis of twin data involves a multigroup design that leverages the expected genetic similarities between MZ twins, who have identical or nearly identical DNA, and DZ twins, who share 50% of their segregating genes on average. These expected genetic similarities for MZ and DZ twins are then used in a multigroup model to separate genetic and environmental contributions to the outcome variable(s). Using this approach, twin models can be used to decompose the variance of the outcome into portions due to additive genetic effects (A), shared environmental effects between twins (C), and the remainder (E), which is interpreted as the effect of the unshared environment between twins and error variance. A genetic simplex model (Boomsma & Molenaar, 1987) extends the classic twin model to longitudinal data by imposing a simplex structure on each of the three variance components (A, C, and E) at each measurement occasion. The resulting model further decomposes the variance attributable to A, C, and E into portions due to transmitted and innovated effects. That is, the genetic simplex model allows the variance of aggression at each observed age to be partitioned into portions due to stable genetic and environmental effects as well as genetic and environmental effects that change as children develop. Note that the genetic simplex model considers all variance components at the first measurement occasion to be fully innovated because there is no prior measurement that could transmit effects.

Transmitted genetic variance is interpreted as the variance at a given age (time = t) that is due to the same genes that were active in explaining aggression at the previous age (time = $t - 1$). Innovated genetic variance represents variance that is due to the activation of new genes that were not relevant in explaining aggression earlier in childhood. The sum of genetic innovation and genetic stability at a given measurement occasion equals the total genetic effect at that measurement, and dividing this effect by the total variance of the outcome yields an estimate of the total heritability of aggression at that measurement, which may vary across time due to changing genetic and environmental factors as children develop.

Two different genetic simplex models were fit in the confirmatory dataset, and boys and girls were analyzed separately based on

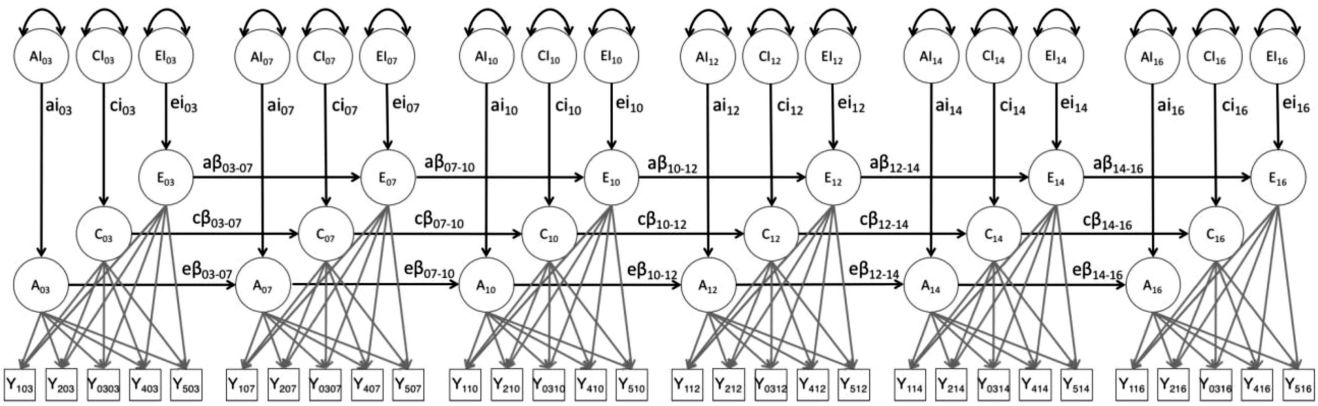

Figure 1. The independent pathway genetic simplex model for one twin. The variance of each aggression item is separately decomposed into additive genetic (A), shared environmental (C), and nonshared environmental (E) components through the loadings of the items onto the age-specific A, C, and E factors. The variances of the A, C, and E factors are decomposed into a component transmitted from the previous measurement and an innovated portion (AI, CI, EI). The paths labeled with β correspond to transmissions from the previous measurement occasion, and the paths labeled with an *i* correspond to innovated effects, passed on from the latent "innovation factors" (i.e., AI, CI, and EI at each age). The model for the other twin is the same, and the additive genetic and shared environmental innovation terms are correlated between the twins. AI is correlated 1.0/0.5 for MZ/DZ twins respectively, and CI is correlated 1.0 for all twins. All unlabeled single-headed arrows are freely estimated, and the variance of each innovation term is fixed to 1 to identify the model.

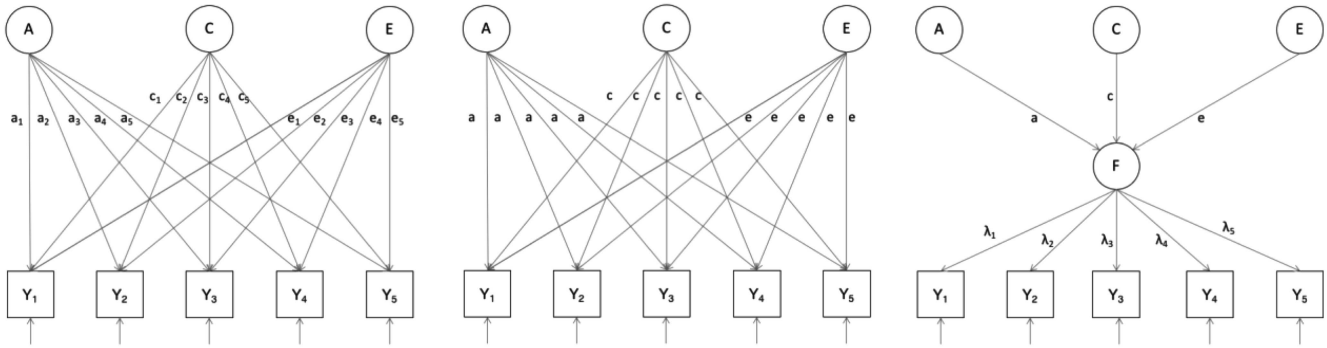

Figure 2. Loadings of each item onto each variance component at a single measurement occasion for one twin. All loadings are freely estimated in the independent pathway (IP) model (left), accounting for the possibility of items being affected differentially by the genetic and environmental effects. The sum score (SS) model (center) assumes that genetics and the environment affect all items in the same manner, so all items are constrained to have the same a, c, and e loadings onto the three variance components. Note that the a, c, e loadings are freely estimated at each measurement occasion. For comparison, we have also depicted the common pathway model (right), which reflects the assumption that the proportions of A, C, and E, respectively, are the same for each item since the A, C, and E factors are transmitted through the common factor instead of decomposing each individual item separately. However, when fitted to the aggression data the common pathway model did not converge properly. Because the results of the IP model provide evidence that the loadings on A, C, and E are indeed item specific, the common pathway model was not further investigated.

an absence of measurement invariance found in previous studies and confirmed in our exploratory analyses. Both simplex models used item-level aggression data and are depicted in Figure 1. The critical constraints that differentiate the two models are illustrated in Figure 2. The more complex model allows each item to load directly onto the A, C, and E factors at each age. All loadings onto each factor (gray paths in Figure 1) are freely estimated at each age. Therefore, this approach yields item-specific estimates of genetic and environmental effects that allow for the possibility of changes in item relevance as children grow up. Twin models that use item-level variance decompositions to study multivariate outcome data are commonly referred to as the IP models to reflect the independent loadings of each item onto each variance component (Kendler, Heath, Martin, & Eaves, 1987; Martin & Eaves, 1977; McArdle & Goldsmith, 1990). We therefore refer to the more complex genetic simplex model as the IP model.

The more constrained model mimics the assumptions of fitting a genetic simplex model to a SS outcome. Instead of carrying out a two-step procedure (first computing SS followed by fitting a genetic simplex to the SS), we fit a constrained genetic simplex to the item-level data that reflects the implicit assumptions involved when modeling SS. This item-level approach to the SS model allows the fit of the IP and SS model to be compared because the two models use the exact same data, which would not be the case if aggregate scores were computed a priori and modeled directly. Importantly, the SS model yields estimates of the genetic and environmental effects on the common content of all items.

For the model depicted in Figure 1 to reflect the assumptions invoked when modeling aggregate scores, the loadings onto the A, C, and E factors were fixed to be equal for all items in this model (see Figure 2). For example, at age 10, all of the item loadings onto A_{10} were fixed to be equal (referred to as a_{10}), all of the item loadings onto C_{10} were fixed to be equal (c_{10}), and all of the loadings onto E_{10} were fixed to be equal (e_{10}). Importantly, however, a_{10}, c_{10}, and e_{10} were allowed to differ from one another to account for the different effects of genes and the environment at each age, and these coefficients were free to vary across age to account for changes in these effects as children develop. These constraints correspond to computing a SS prior to fitting a genetic simplex model, that is, taking an unweighted sum of the item-scores and subjecting the resulting aggregate to the genetic simplex model (illustrated in Figure 3).

Importantly, in both the IP and SS models, loadings on the A, C, and E factors are permitted to be different over time, which implies that the A, C, and E factors need to be interpreted as "age-appropriately measured genetic and environmental effects" (McArdle et al., 2009).

Expectations regarding the IP and SS models. Consider an item that has very low relevance at the first measured time point. When computing a SS, this item has a weight of 1 at all timepoints, including the first, and therefore the item variance adds noise to the variability of the SS. Consequently, when decomposing the SS into genetic and environmental and error variance, the error variance E will be larger than its counterpart in the IP model, where that item contributes very little to the E component because its loadings are freely estimated rather than being fixed to 1. The relative size of the A, C, and E components that are estimated at each time point are therefore expected to be biased in the SS model, and more accurately estimated in the IP model. The amount of bias is likely to differ across time and will affect the estimates of genetic and environmental stability and innovation in the SS model. Note that not only the relevance of items as indicators of aggression can change over time, but also their age-specific decomposition into genetic and environmental effects as quantified by the loadings on the A, C, and E factors. For instance, a specific aspect of aggression might be under more environmental control during early childhood, resulting in less individual differences due to environment. Again, estimating the loadings freely in the IP model permits

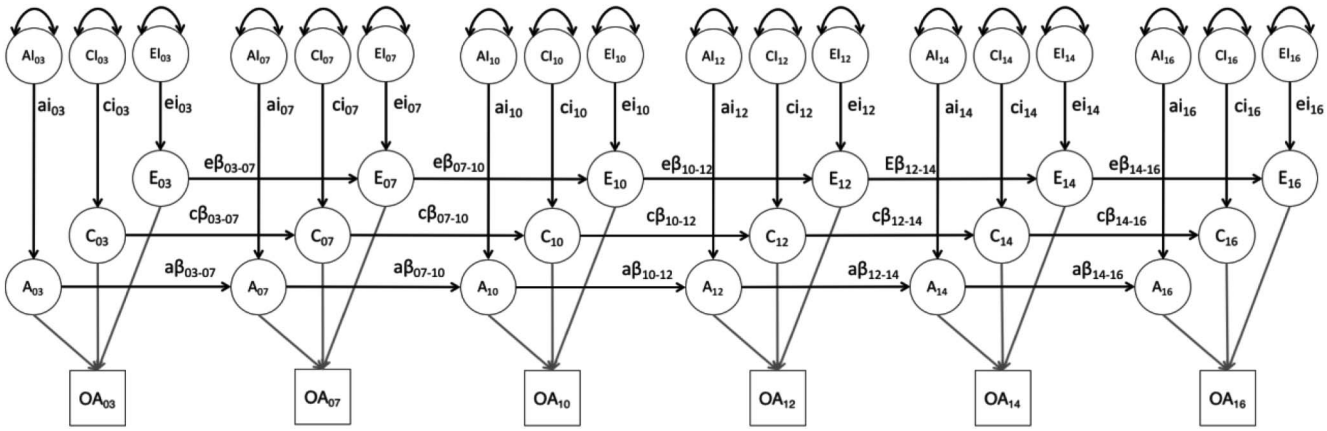

Figure 3. Illustration of how the sum score (SS) model based on Figure 1 and the right panel of Figure 2 should be interpreted. The SS model fit to the item-level data is conceptually equivalent to first computing an aggregate overt aggression (OA) score at each measurement occasion by summing the item-level data and then subjecting these OA variables to the genetic simplex model. This approach, depicted here, has the disadvantage that its resulting fit statistics cannot be directly compared to the IP model because it utilizes different data (aggregate scores rather than item-level data).

such influences to be reflected in the corresponding model parameters, and therefore also in the estimates of innovation and transmission. In sum, estimates from the IP model should be more trustworthy than those based on the SS model.

All models were fit using Mplus version 7.2 using the WLSMV estimator for categorical outcomes. The models were fit using full information maximum likelihood to account for missingness.

Results

EDA

Descriptive statistics. All items in the P-CBCL, CBCL and the YSR were positively skewed. The skew was more pronounced in the CBCL and YSR than it was in the P-CBCL. Boys had slightly higher scores on average, and there was a slightly stronger skewness for girls. Total scores were also highly positively skewed (see Supplementary Figure S1 in the online supplementary material). Items were treated as categorical for all subsequent analyses, and were dichotomized into "not true" (original score of 0) and "at least somewhat true" (original score of 1 or 2) to avoid numerical problems caused by the infrequent endorsement of "very true" for most items.

Within the P-CBCL and CBCL, most items were moderate-to-highly correlated, with some low correlations associated with items that had particularly low endorsement rates. In the YSR, correlations were lower in general but displayed the same trends.

Factor analyses. A three-factor solution had the best fit at each age and produced the most interpretable results across sex and age. The three factors can be interpreted as "physical/overt aggression," "non-physical aggression," and "mood instability." All subsequent genetic analyses focused on overt aggression (OA), as this is the more severe and likely the more personally and societally problematic component of aggressive behavior. Five items at each measurement occasion had high loadings (0.60–0.95) onto the OA factor, and these items were characterized by destruction of property and violence toward others. See Table 1 for specific items comprising OA. In the exploratory data set the reliability of the OA items as measured by Cronbach's alpha for ages 3, 7, 10, 12, 14, and 16 ranged between 0.812 and 0.892 for boys and between 0.793 and 0.873 for girls.

As expected, the results of the CFAs fitted to the OA items showed that factor loadings and item thresholds differed substantially across age, sex, and rater, thus indicating measurement noninvariance for sex, rater, and across time.

Longitudinal modeling. The autocorrelation matrix of OA total scores (treated as an ordered categorical variable with six categories due to skewness) suggested that OA is moderately stable as children grow up. The correlation between ages assessed with different raters (change from mother to self-ratings) was smaller than the correlation between comparably spaced measurements assessed using the same rater and instrument (see Supplementary Figure S2 in the online supplemental material).

Genetic Analyses

Model fit. Although the SS model is nested under the IP model, we were not able to use a chi-squared test to compare the model fit because the effect of shared environment on two items had to be fixed to zero in the IP model to achieve proper convergence. These constraints compromised proper nesting. Inspection of the fit statistics showed that both models fit the data at least reasonably well, especially based on the RMSEA. The IP model provided a better fit than the SS model; the CFI was 0.965 versus 0.876 (males) and 0.959 versus 0.873 (females), and the RMSEA was 0.012 versus 0.022 (males) and 0.009 versus 0.016 (females) for the IP and SS models, respectively. In light of the expectation that the aggression items would change in relevance over time, the better fit of the IP model was unsurprising due to its free estimation of loadings onto the A, C, and E factors at each measurement occasion. Furthermore, items differed with respect to their heritability estimates (see below); this also contributes to the better fit of

the IP model relative to the SS model, which reflects an assumption of uniform item heritability at each measurement occasion via its item loading constraints.

Genetic stability. The results provide evidence of substantial genetic stability when considering adjacent measurement occasions in which aggression was assessed by the same scale and rater. The IP model results showed that 62–100% of the total genetic variance of OA was attributable to the same genes that were active at the previous measurement occasion. Changes in item content (P-CBCL to CBCL) or rater (CBCL to YSR), however, resulted in substantially lower genetic stability. The results of the SS model were similar, but suggested somewhat higher genetic stability than the IP model in mother-report data. Table 2 shows the proportion of the total genetic variance at each measurement occasion that could be attributed to the same genes that were also active in explaining aggression at the prior measurement.

Figure 4 illustrates how these variance components were computed based on point estimates resulting from fitting the genetic simplex models. To demonstrate how the variance components attributable to genetic transmission were derived from the fitted model, this illustration focuses on a single measurement occasion (age 10) from the IP model that was fit to the sample of girls. Its caption also describes the derivation of the genetic innovation terms (see "total heritability" below), and these principles also apply to the estimated environmental variance components, which are described in the remainder of this subsection. Importantly, the principles illustrated in Figure 4 apply to the SS model as well because the SS model is obtained by applying constraints to the IP model (see Figure 2).

Recall that with the change in item content between ages 3 and 7, some items (e.g., "threatens people") were added to the aggression questionnaire at age 7. Importantly, the transmitted additive genetic effects of these items can still be estimated in the IP model despite the fact that they were not used to measure aggression at the previous measurement occasion. This is due to the fact that transmitted and innovated effects are estimated at the factor level. Consider additive genetic innovation at age 3 and transmission from age 3 to age 7. As can be seen in Figure 1, the additive genetic factor A_{03} represents what items measured at age 3 have genetically in common. Each item's loading onto A_{03} quantifies the item specific additive genetic effects (e.g., loadings of Y_{103}–Y_{503} onto A_{03}). Genetic innovation at age 3 is the variance of A_{03} due to innovation, denoted as AI_{03}, and $e\beta_{03-07}$ is the transmission of additive genetic effects from age 3 to age 7 (see respective arrows from AI_{03} to A_{03} and from A_{03} to A_{07} in Figure 1). These variance components can be estimated in the IP model even if different items (or raters) are used at the different ages. For instance, the item "threatens people" is not included in the model at age 3 (i.e., no loading is estimated since there are no data available), but it is included in the model at age 7. The same principle of estimating innovation and transmission effects at the factor level while simultaneously quantifying item specific contributions to the factors through the loadings also applies to the environmental effects C and E (see Figure 1).

Genetic innovations. A change of scale or rater resulted in high estimates of genetic innovation effects, comprising roughly 65–75% of the total genetic effects in both models (see Table 3). Note that the higher estimates of genetic innovation should most likely be attributed to the change in measurement, given the fact that when considering adjacent measurement occasions in which the same sets of items were evaluated by the same rater, the genetic variance due to innovations was generally low in both the IP and SS models, comprising 0–38% of the total genetic variance. Although the results were similar for both models in general, the IP model showed slightly higher innovation variance in the mother-report measurements, and the SS model resulted in slightly higher innovation variance in the self-report measurements.

Total heritability. Results from the IP model indicate that items measuring OA were differentially heritable in the mother-report data. For boys and girls alike, mother-report items relating to destruction of property were more highly heritable (>80%) than items related to physical violence against others or threatening others (~25–40%). The differences between item-level heritability were smaller in the self-report YSR, with typical estimates

Table 2
Transmitted Additive Genetic Effects From One Measurement Occasion to the Next

Variable	Age 3 to 7	Age 7 to 10	Age 10 to 12	Age 12 to 14	Age 14 to 16
Males					
Sum score	*.206 (.017)*	*.498 (.027)*	*.560 (.027)*	*.113 (.021)*	*.434 (.065)*
Threatens people	.077 (.014)	.222 (.035)	.274 (.043)	.126 (.028)	.491 (.081)
Destroys his or her things	.281 (.018)	.582 (.026)	.639 (.032)	.105 (.021)	.301 (.063)
Destroys others things	.291 (.018)	.556 (.027)	.663 (.032)	.156 (.032)	.379 (.067)
Gets in fights	.088 (.008)	.217 (.020)	.287 (.029)	.172 (.034)	.469 (.074)
Physically attacks people	.082 (.009)	.187 (.021)	.300 (.034)	.145 (.029)	.536 (.088)
Females					
Sum score	*.290 (.027)*	*.447 (.034)*	*.496 (.036)*	*.208 (.032)*	*.579 (.084)*
Threatens people	.044 (.012)	.162 (.043)	.201 (.039)	.250 (.047)	.688 (.130)
Destroys his or her things	.306 (.024)	.541 (.038)	.563 (.045)	.226 (.039)	.522 (.083)
Destroys others things	.294 (.023)	.512 (.036)	.616 (.045)	.269 (.043)	.434 (.088)
Gets in fights	.109 (.013)	.194 (.026)	.225 (.038)	.214 (.042)	.774 (.110)
Physically attacks people	.070 (.012)	.174 (.030)	.231 (.047)	.275 (.049)	.641 (.114)

Note. Reported are the estimated proportions of total variance due to transmitted additive genetic effects. Standard errors are shown in parentheses. Note that the scale changes from age 3 to age 7, resulting in lower transmitted effects. Results for the sum score models are presented in italics to differentiate them from the item-level results of the independent pathway model.

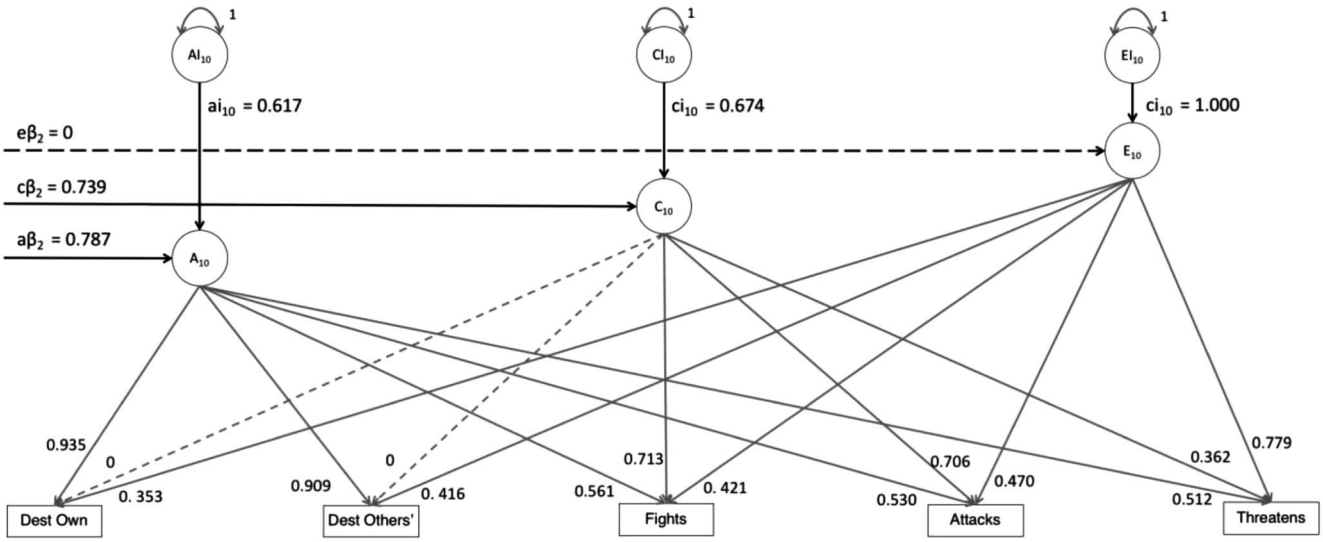

Figure 4. Estimated independent pathway (IP)model for females, zoomed in to illustrate the estimated structure at age 10. All latent and observed variables are normalized to have variance equal to 1. Dashed lines indicate paths that were fixed to zero. For example, the estimated variance of "Gets in fights" that is attributable to transmitted genetic effects can be found in the following way. First find the total genetic variance by squaring the corresponding loading onto A_{10} ($0.561^2 = 0.315$), and then multiply it by the proportion of variance in A_{10} that can be explained by A_7, which is accomplished by squaring the estimated genetic transmission coefficient ($0.787^2 = 0.620$). This leads to the value $0.561^2 * 0.787^2 = 0.194$ reported in Table 2. Similarly, the innovated genetic variance is $0.561^2 * 0.617^2 = 0.120$, reported in Table 3. Note that when the constraints of the sum score (SS) model are applied, the measurement properties of all five items are constrained to be equal within each measurement occasion. All of their estimated variance components will therefore be equal and interpreted as the variance components of the SS.

ranging from 40–65% for all items. The heritability estimates of the SS in the SS model ranged between 45% and 65%. The estimates of total heritability for both models can be found in Supplementary Table S2 (see the online supplementary materials).

Shared environmental effects. The common environment shared between twins was a substantial source of variation in the data collected using the P-CBCL and CBCL, but this was not the case for the self-report YSR data collected at 14 and 16 years old.

Table 3
Innovative Additive Genetic Effects Specific to Each Measurement Occasion

Variable	Age 3	Age 7	Age 10	Age 12	Age 14	Age 16
Males						
Sum score	*.504 (.023)*	*.361 (.024)*	*.147 (.024)*	*.115 (.054)*	*.400 (.025)*	*.144 (.062)*
Hits others	.341 (.014)					
Hurts others accidentally	.048 (.007)					
Threatens people		.165 (.028)	.122 (.020)	.101 (.019)	.330 (.041)	.154 (.070)
Destroys his or her things		.599 (.019)	.320 (.025)	.237 (.028)	.275 (.036)	.094 (.042)
Destroys others things	.899 (.007)	.619 (.019)	.306 (.024)	.246 (.029)	.409 (.041)	.119 (.056)
Gets in fights	.575 (.014)	.188 (.013)	.119 (.012)	.107 (.014)	.452 (.043)	.147 (.069)
Physically attacks people	.328 (.018)	.175 (.016)	.103 (.012)	.111 (.015)	.381 (.042)	.168 (.077)
Females						
Sum score	*.433 (.025)*	*.316 (.032)*	*.166 (.030)*	*.157 (.062)*	*.343 (.033)*	*.042 (.086)*
Hits others	.518 (.015)					
Hurts others accidentally	.073 (.010)					
Threatens people		.088 (.023)	.100 (.027)	.090 (.021)	.306 (.047)	.000 (.103)
Destroys his or her things		.615 (.024)	.334 (.033)	.252 (.041)	.278 (.040)	.000 (.078)
Destroys others things	.893 (.007)	.591 (.024)	.315 (.031)	.276 (.044)	.330 (.051)	.000 (.065)
Gets in fights	.242 (.015)	.220 (.020)	.120 (.018)	.101 (.021)	.263 (.043)	.000 (.116)
Physically attacks people	.445 (.023)	.140 (.019)	.107 (.018)	.103 (.024)	.338 (.049)	.000 (.096)

Note. Standard errors appear in parentheses following their corresponding variance components. Results for the sum score models are presented in italics to differentiate them from the item-level results of the independent pathway model. Blank cells correspond to items that were not measured at a given measurement occasion. Note that genetic variance at the first measurement occasion is entirely innovated by definition, therefore column 1 shows estimates of the proportion the total variance that is due to all additive effects. Scale or rater changes from age 3 to 7 and from age 12 to 14.

In the self-report measures, the effect of the shared environment was near zero and statistically insignificant.

Transmitted shared environmental effects. The transmitted effects of shared environmental factors were generally close to zero or very low, and several effects had to be fixed to zero to ensure model convergence. Notably, only the items "threatens people," "gets in fights, and "physically attacks people" had nonzero transmitted shared environmental effects. The details are shown in Supplementary Table S3 (see the online supplementary materials).

Innovative shared environmental effects. Up to age 12, and with the exception of the two items tapping into destruction of property, which had zero shared environmental variability, innovated effects of the shared environment were larger than the transmitted effects. This implies that when relatively young children have their aggressive behaviors rated by their mothers, the environmental effects change over time, but these effects are shared between twins within a family. When aggressive behavior is assessed via self-report at age 14 and 16, the innovated shared environmental effects are zero. These results are shown in Supplementary Table S3 (see the online supplementary materials).

Total shared environmental effects. Supplementary Table 4 (see Supplementary Materials) summarizes the total shared environmental effects, which sum the transmitted and innovative effects. The shared environmental effects on self-rated behaviors are zero throughout. The SS model suggests that in 15–21% of the variance in mother-report aggression is due to shared environmental effects. The IP model provides more detail. Whereas the items measuring destruction are not affected by the shared environment, behaviors such as threatening, fighting, and physically attacking others can be substantially explained by shared environmental factors. In particular, it was found to explain between 44% and 58% of the total variance in the items assessing attacking and fighting behaviors.

Nonshared environmental effects. The E factor in ACE models captures nonshared environment as well as error variance. No evidence for the transmission of E was found, so these effects were constrained to be unique to each age (i.e., fully innovated). Supplementary Table S5 (see the online supplementary materials) lists the proportion variance in the SS (SS model) and each OA item (IP model) attributable to measurement error and the impact of nonshared environmental factors at each age for each sex. These estimates reflect the proportion of the total variance of the items or the SS that is not explained by additive genetic effects or shared environmental effects.

Discussion

The current study shows that the overt and more physical aspects of childhood aggression (OA) are highly heritable between ages 3 and 16. Additive genetic effects explain between 50% and 68% of the variance of a total score computed from five CBCL items, with heritability estimates for individual items measuring destruction of property reaching 90%. Importantly, our study showed that OA is largely affected by the same genes throughout childhood. This genetic stability was evident at the level of the SS as well as at the level of the individual items, and it is in line with a recent paper that compares this same NTR cohort to a large longitudinal cohort from the United Kingdom (Porsch et al., 2016). Environmental effects shared by children from the same family typically account for 20% of the variance of OA early in childhood, but they are negligible in adolescence. Nonshared environmental effects on OA are comparatively smaller in childhood (20% variance explained) than in adolescence (40%).

Our comparison of a model representing the common practice of analyzing total or SS (SS model) and a model that provided estimates of genetic and environmental effects for each questionnaire item (IP model) provided detailed insight into the genetic architecture of OA during childhood and early adolescence. CBCL items measuring the destruction of property are highly heritable throughout childhood, whereas fighting behaviors are much less heritable and more influenced by shared environment. This pattern is reversed during early adolescence, with increased additive genetic effects for fighting behaviors and somewhat decreased heritability of destructive behaviors. One of the factors contributing to these results might be the slightly higher reliability of property items when rated by the mother, however, this interesting difference clearly deserves additional research.

Apart from differences in heritability across the different expressions of aggression, we also observed a notable decrease in the autocorrelation of the observed SS (shown in Supplementary Figure 2) when the scale changed between ages 3 and 7, and when the rater changed from mother to self between ages 12 and 14. These changes were also evident in the subsequent decomposition of the variance into genetic and environmental variance components. While it is likely that shared environmental influences decrease when twins spend less time together, it is also known that mothers can have the tendency to rate their children more similarly than the children would rate themselves, which can result especially in higher shared environmental effects than can be found based on self-report data (Bartels, Boomsma, Hudziak, van Beijsterveldt, & van den Oord, 2007). Therefore, some caution is warranted when interpreting the observed structural changes between ages 3 and 7 and between ages 12 and 14.

The genetic stability in the SS model was estimated to be higher during childhood compared to estimates based on the IP model. As outlined in the method section, the decomposition of the total variance of the SS at each time point into A, C, and E components is biased because all loadings are fixed to 1, independent of item relevance. An item with a true loading of zero therefore incorrectly inflates the E component, and this implies bias in A and C because the three components must sum to 100% of the total variance. Because the amount of bias is unlikely to be equal at all time points, it also induces bias in the estimates of genetic stability and innovation. For this reason, the estimates of the IP model are likely to be more trustworthy.

The results of our analyses have several implications for gene-finding studies, and more generally, they facilitate a better understanding of childhood aggression. Gene-finding studies necessitate extremely large sample sizes, which are commonly obtained by combining data sets from different individual studies in a meta-analysis. The finding that genetic effects are largely stable across childhood even though the behavioral expression of aggression changes is important because it suggests that data from different age groups (e.g., 7–12 and 12–14) can be analyzed jointly without losing too much power due to age-related heterogeneity. Although this study provides support for combining data collected from different age groups, combining data from different raters is not

advisable due to the finding that a change of rater induced a substantial increase in the estimate of genetic innovation and, simultaneously, a decrease in genetic transmission. High transmission effects provide evidence that the same genes continue to affect the behavior, whereas high innovation indicates either that new genes come into play, or that there is substantial phenotypic heterogeneity, which also reduces power in gene finding studies (Laurin et al., 2015; Levinson et al., 2014; Lubke et al., 2014). Because aggression is most often rated by the mother in early childhood and assessed via self-report in adolescence, this suggests that gene finding studies should focus on childhood and adolescent aggression separately.

Furthermore, our analyses provided important information regarding differences in aggressive behaviors between girls and boys. Although the exploratory analyses supported previous findings of substantial violations of measurement invariance across sex, the decomposition of the total variance into genetic and environmental effects had largely the same patterns for girls and for boys. It is therefore plausible that questionnaire data collected from boys and girls could be meaningfully combined in the context of gene-finding studies, but more research is necessary to formulate a confident recommendation about how to account for sex differences in the measurement of aggression.

As discussed and illustrated here, the behavioral manifestation of complex traits like aggression can change over the course of development from early childhood into adolescence. Measurement invariance will not hold in these scenarios, complicating the interpretation of many common longitudinal models that are based on the assumption of a consistent measurement model across ages. To appropriately study the development of such traits, it is important to understand these changes in measurement across age and/or rater. The utilization of item-level data, illustrated here with the IP model, can serve to assess whether or not the measurement properties of the scales used to assess these traits change in a manner that reflects theoretical expectations. Understanding these changes is critical to the development of a model for the phenomenon of interest, and they should be taken into consideration when drawing inference and interpreting results. In the case of the genetic and environmental contributions to the development of childhood aggression, item-level analyses seemed to corroborate the overall story told by the models fit to SS outcomes while also providing more details to the story, thereby facilitating a more confident interpretation of the results than would be warranted based on the SS analyses alone.

References

Achenbach, T. M., & Rescorla, L. A. (2000). *ASEBA preschool forms & profiles*. Burlington, VT: Research Center for Children, Youth and Families, University of Vermont.

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

Althoff, R. R., Kuny-Slock, A. V., Verhulst, F. C., Hudziak, J. J., & van der Ende, J. (2014). Classes of oppositional-defiant behavior: Concurrent and predictive validity. *Journal of Child Psychology and Psychiatry, 55,* 1162–1171. http://dx.doi.org/10.1111/jcpp.12233

Bartels, M., Boomsma, D. I., Hudziak, J. J., van Beijsterveldt, T. C., & van den Oord, E. J. C. G. (2007). Twins and the study of rater (dis)agreement. *Psychological Methods, 12,* 451–466. http://dx.doi.org/10.1037/1082-989X.12.4.451

Bartels, M., Hudziak, J. J., van den Oord, E. J., van Beijsterveldt, C. E., Rietveld, M. J., & Boomsma, D. I. (2003). Co-occurrence of aggressive behavior and rule-breaking behavior at age 12: Multi-rater analyses. *Behavior Genetics, 33,* 607–621. http://dx.doi.org/10.1023/A:1025787019702

Bartels, M., van Beijsterveldt, C. E. M., Derks, E. M., Stroet, T. M., Polderman, T. J. C., Hudziak, J. J., & Boomsma, D. I. (2007). Young Netherlands Twin Register (Y-NTR): A longitudinal multiple informant study of problem behavior. *Twin Research and Human Genetics, 10,* 3–11. http://dx.doi.org/10.1375/twin.10.1.3

Björkqvist, K. (1994). Sex differences in physical, verbal, and indirect aggression: A review of recent research. *Sex Roles, 30*(3–4):177–188. http://dx.doi.org/10.1007/BF01420988

Boomsma, D. I., & Molenaar, P. C. (1987). The genetic analysis of repeated measures. I. Simplex models. *Behavior Genetics, 17,* 111–123. http://dx.doi.org/10.1007/BF01065991

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105,* 456–466. http://dx.doi.org/10.1037/0033-2909.105.3.456

Connor, D. F. (2004). *Aggression and antisocial behavior in children and adolescents: Research and treatment*. New York, NY: Guilford Press.

Crick, N. R., Casas, J. F., & Mosher, M. (1997). Relational and overt aggression in preschool. *Developmental Psychology, 33,* 579–588. http://dx.doi.org/10.1037/0012-1649.33.4.579

Hannigan, L. J., Walaker, N., Waszczuk, M. A., McAdams, T. A., & Eley, T. C. (2017). Aetiological influences on stability and change in emotional and behavioural problems across development: A systematic review. *Psychopathology Review, 4,* 52–108. Advance online publication.

Huesmann, L. R., Eron, L. D., Lefkowitz, M. M., & Walder, L. O. (1984). Stability of aggression over time and generations. *Developmental Psychology, 20,* 1120–1134. http://dx.doi.org/10.1037/0012-1649.20.6.1120

Kendler, K. S., Heath, A. C., Martin, N. G., & Eaves, L. J. (1987). Symptoms of anxiety and symptoms of depression. Same genes, different environments? *Archives of General Psychiatry, 44,* 451–457. http://dx.doi.org/10.1001/archpsyc.1987.01800170073010

Lacourse, E., Boivin, M., Brendgen, M., Petitclerc, A., Girard, A., Vitaro, F., . . . Tremblay, R. E. (2014). A longitudinal twin study of physical aggression during early childhood: Evidence for a developmentally dynamic genome. *Psychological Medicine, 44,* 2617–2627. http://dx.doi.org/10.1017/S0033291713003218

Laurin, C. A., Hottenga, J. J., Willemsen, G., Boomsma, D. I., & Lubke, G. H. (2015). Genetic analyses benefit from using less heterogeneous phenotypes: An illustration with the hospital anxiety and depression scale (HADS). *Genetic Epidemiology, 39,* 317–324. http://dx.doi.org/10.1002/gepi.21897

Levinson, D. F., Mostafavi, S., Milaneschi, Y., Rivera, M., Ripke, S., Wray, N. R., & Sullivan, P. F. (2014). Genetic studies of major depressive disorder: Why are there no genome-wide association study findings, and what can we do about it? *Biological Psychiatry, 76,* 510–512. http://dx.doi.org/10.1016/j.biopsych.2014.07.029

Lewis, G. J., & Plomin, R. (2015). Heritable influences on behavioural problems from early childhood to mid-adolescence: Evidence for genetic stability and innovation. *Psychological Medicine, 45,* 2171–2179. http://dx.doi.org/10.1017/S0033291715000173

Ligthart, L., Bartels, M., Hoekstra, R. A., Hudziak, J. J., & Boomsma, D. I. (2005). Genetic contributions to subtypes of aggression. *Twin Research and Human Genetics, 8,* 483–491. http://dx.doi.org/10.1375/twin.8.5.483

Loeber, R., & Hay, D. (1997). Key issues in the development of aggression and violence from childhood to early adulthood. *Annual Review of Psychology, 48,* 371–410. http://dx.doi.org/10.1146/annurev.psych.48.1.371

Lubke, G. H., & Campbell, I. (2016). Inference based on the best-fitting model can contribute to the replication crisis: Assessing model selection uncertainty using a bootstrap approach. *Structural Equation Modeling, 23,* 479–490. http://dx.doi.org/10.1080/10705511.2016.1141355

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,* 292–298. http://dx.doi.org/10.1375/136905204774200578

Lubke, G. H., Laurin, C., Amin, N., Hottenga, J. J., Willemsen, G., van Grootheest, G., . . . Boomsma, D. I. (2014). Genome-wide analyses of borderline personality features. *Molecular Psychiatry, 19,* 923–929. http://dx.doi.org/10.1038/mp.2013.109

Luningham, J. M., McArtor, D. B., Bartels, M., Boomsma, D. I., & Lubke, G. H. (2017). Sum scores in twin growth curve models: Practicality vs. bias. *Behavior Genetics*, *47,* 516–536. http://dx.doi.org/10.1007/s10519-017-9864-0

Martin, N. G., & Eaves, L. J. (1977). The genetical analysis of covariance structure. *Heredity, 38,* 79–95. http://dx.doi.org/10.1038/hdy.1977.9

McArdle, J. J., & Goldsmith, H. H. (1990). Alternative common factor models for multivariate biometric analyses. *Behavior Genetics, 20,* 569–608. http://dx.doi.org/10.1007/BF01065873

McArdle, J. J., Grimm, K. J., Hamagami, F., Bowles, R. P., & Meredith, W. (2009). Modeling life-span growth curves of cognition using longitudinal data with multiple samples and changing scales of measurement. *Psychological Methods, 14,* 126–149. http://dx.doi.org/10.1037/a0015857

McEvoy, M. A., Estrem, T. L., Rodriguez, M. C., & Olson, M. L. (2003). Assessing relational and physical aggression among preschool children: Intermethod agreement. *Topics in Early Childhood Special Education, 23,* 51–61. http://dx.doi.org/10.1177/02711214030230020101

Meredith, W., & Horn, J. (2001). The role of factorial invariance in modeling growth and change. In A. Sayer & L. Collins (Eds.), *New methods for the analysis of change* (pp. 203–240). Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/10409-007

Niv, S., Tuvblad, C., Raine, A., & Baker, L. A. (2013). Aggression and Rule-breaking: Heritability and stability of antisocial behavior problems in childhood and adolescence. *Journal of Criminal Justice, 41,* 285–291. http://dx.doi.org/10.1016/j.jcrimjus.2013.06.014

Pingault, J. B., Rijsdijk, F., Zheng, Y., Plomin, R., & Viding, E. (2015). Developmentally dynamic genome: Evidence of genetic influences on increases and decreases in conduct problems from early childhood to adolescence. *Scientific Reports, 5,* 10053. http://dx.doi.org/10.1038/srep10053

Porsch, R. M., Middeldorp, C. M., Cherny, S. S., Krapohl, E., van Beijsterveldt, C. E., Loukola, A., . . . Bartels, M. (2016). Longitudinal heritability of childhood aggression. *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: The Official Publication of the International Society of Psychiatric Genetics*, *171,* 697–707. http://dx.doi.org/10.1002/ajmg.b.32420

Posthuma, D., & Boomsma, D. I. (2000). A note on the statistical power in extended twin designs. *Behavior Genetics, 30,* 147–158. http://dx.doi.org/10.1023/A:1001959306025

Tuvblad, C., Raine, A., Zheng, M., & Baker, L. A. (2009). Genetic and environmental stability differs in reactive and proactive aggression. *Aggressive Behavior, 35,* 437–452. http://dx.doi.org/10.1002/ab.20319

van Beijsterveldt, C. E. M., Bartels, M., Hudziak, J. J., & Boomsma, D. I. (2003). Causes of stability of aggression from early childhood to adolescence: A longitudinal genetic analysis in Dutch twins. *Behavior Genetics, 33,* 591–605. http://dx.doi.org/10.1023/A:1025735002864

van Beijsterveldt, C. E., 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. http://dx.doi.org/10.1017/thg.2012.118

Received August 29, 2016
Revision received June 13, 2017
Accepted June 23, 2017 ■