Behav Genet (2008) 38:447–461
DOI 10.1007/s10519-008-9213-4



ORIGINAL RESEARCH

# Genetic and Environmental Influences on the Stability of Withdrawn Behavior in Children: A Longitudinal, Multi-informant Twin Study

**Rosa A. Hoekstra · Meike Bartels · James J. Hudziak · Toos C. E. M. Van Beijsterveldt · Dorret I. Boomsma**

Received: 19 February 2007 / Accepted: 27 May 2008 / Published online: 12 June 2008


**Abstract** We examined the contribution of genetic and environmental influences on the stability of withdrawn behavior (WB) in childhood using a longitudinal multiple rater twin design. Maternal and paternal ratings on the withdrawn subscale of the Child Behavior Checklist (CBCL) were obtained from 14,889 families when the twins were 3, 7, 10 and 12 years old. A longitudinal psychometric model was fitted to the data and the fit of transmission and common factor models were evaluated for each variance component. WB showed considerable stability throughout childhood, with correlation coefficients ranging from about .30 for the 9-year time interval to .65 for shorter time intervals. Individual differences in WB as observed by the mother and the father were found to be largely influenced by genetic effects at all four time points, in both boys (50–66%) and girls (38–64%). Shared environmental influences explained a small to modest proportion (0–24%) of the variance at all ages and were slightly more pronounced in girls. Non-shared environmental influences were of moderate importance to the variance and slightly increased with age, from 22–28% at age 3 to 35–41% at age 12 years. The stability of WB was largely explained by genetic effects, accounting for 74% of stability in boys and 65% in girls. Shared environmental effects explained 7% (boys) and 17% (girls) of the behavioral stability. Most shared environmental effects were common to both raters, suggesting little influence of rater bias in the assessment of WB. The shared environmental effects common to both raters were best described by a common factor model, indicating that these effects are stable and persistent throughout childhood. Non-shared environmental effects accounted for the remaining covariance over time.

**Keywords** Genetics · Twins · Childhood · Problem behavior · Longitudinal studies · Heritability

Edited by Hermine Maes.

R. A. Hoekstra · M. Bartels · T. C. E. M. Van Beijsterveldt · D. I. Boomsma
Department of Biological Psychology, VU University, Amsterdam, The Netherlands

R. A. Hoekstra (✉)
Autism Research Centre, Section of Developmental Psychiatry, University of Cambridge, Douglas House, 18b Trumpington Road, Cambridge CB2 8AH, UK
e-mail: rah58@medschl.cam.ac.uk

J. J. Hudziak
Department of Psychiatry, University of Vermont College of Medicine, Burlington, VT, USA

## Introduction

Children scoring high on withdrawn behavioral scales are characterized by shy, inhibited, introvert and withdrawn behavior (WB). WB correlates with symptoms of anxiety and depression (Verhulst et al. 1996), and WB in childhood has been shown to predict anxiety disorders and major depression in adolescence and adulthood (Goodwin et al. 2004). In a follow-up study spanning 14 years, Hofstra et al. (2000) found that parent reported WB was an important predictor for malfunctioning in adulthood. WB at the time of first measurement predicted both adult internalizing and externalizing problems 14 years later. Furthermore, inhibited 3-year-olds (children who are shy, fearful and easily upset) were more likely to meet diagnostic criteria for depression when they were 21 years old (Caspi et al. 1996). Children described as "shy" on

multiple time points showed increased incidence of anxiety problems in adolescence (Prior et al. 2000). The evidence that childhood WB is a predictor for anxiety and depression later in life is further supported by laboratory studies of behavioral inhibition. Behavioral inhibition (characterized by shy, inhibited behavior, and fear for novel situations) is present in about 10 to 15% of children (Kagan et al. 1988). Behaviorally inhibited children have higher rates of childhood anxiety disorders (Rosenbaum et al. 1993; Biederman et al. 2001) and are at increased risk of developing adolescent social phobia (Hayward et al. 1998).

Longitudinal studies indicate that behavioral problems show considerable stability. In a large population sample of Dutch children, a correlation of .48 for problem behaviors across an 8-year period was found (Verhulst et al. 1996). In a follow-up of this study the stability of general behavioral problems over 14 years was found to be .43 (Hofstra et al. 2000), whilst the continuity of withdrawn behavioral problems was .36. The continuity of problem behaviors advocates research into the underlying mechanisms influencing stability of behavioral traits. In the last decades a range of longitudinal studies have focused on childhood externalizing problem behaviors in general (e.g. Van der Valk et al. 2003a; Bartels et al. 2004b; Fergusson 1998; Haberstick et al. 2005) and more specific problem behaviors such as attention problems (e.g. Rietveld et al. 2004; Mannuzza et al. 2003), aggression (e.g. Campbell et al. 2006; Alink et al. 2006) or conduct disorder (e.g. Fergusson et al. 2005; Kim-Cohen et al. 2003). Likewise, substantial attention has been devoted to internalizing behaviors in general (e.g. Van der Valk et al. 2003a; Bartels et al. 2004b; Haberstick et al. 2005) and more narrowly defined problems such as anxiety disorder and depression (e.g. Fombonne et al. 2001; Tram and Cole 2006). However, surprisingly little research has focused on withdrawn behavioral problems.

A powerful way of unraveling the genetic and environmental effects on individual differences in the development of behavioral problems is the study of genetically related individuals. Both cross-sectional and longitudinal studies using the classical twin design have been conducted to assess heritability estimates for broad band internalizing and externalizing problem behaviors (Bartels et al. 2004b; Van der Valk et al. 2003a) as well as for specific syndrome scales such as aggression (Van Beijsterveldt et al. 2003; Haberstick et al. 2006), obsessive compulsive disorder (Hudziak et al. 2004; van Grootheest et al. 2007), juvenile bipolar disorder (Boomsma et al. 2006b), attention problems (Rietveld et al. 2004), and anxious/depression (Boomsma et al. 2005). However, no large scale longitudinal twin studies into WB have been reported.

Family studies into childhood WB have been scarce, but there are indications that familial factors play a role. Behavioral inhibition is more frequent in children whose parents have agoraphobia and panic disorder (Rosenbaum et al. 1988), and anxiety disorders are more frequent in the families of behaviorally inhibited children (Rosenbaum et al. 1991). Furthermore, a study in a large sample of 4-year-old twins reported a heritability of 76% (boys) and 66% (girls) for shyness/inhibition, as assessed with a 3-item questionnaire (Eley et al. 2003). A few twin studies examined the cross-sectional heritability of WB at various ages in childhood using the Child Behavior Checklist (CBCL; Achenbach 1991; Achenbach 1992). An early twin study in a relatively small sample of 2 to 3 year-old twins found no significant genetic effects on variance in WB (Schmitz et al. 1995). Contrary to these findings, Van den Oord et al. (1996) reported major genetic influences (74%) and no evidence for shared environmental influences on individual differences in WB in a sample of 1,358 3-year-old twin pairs. Eight years later, Derks et al. (2004b) analyzed data on WB of more than 9,000 3-year-old twin pairs, including the data used in Van den Oord's study and found moderate heritability (about 60% in boys; 45% in girls) and significant shared environmental effects. Two early twin studies examined the heritability of WB in middle childhood (sample sizes 181 and 203 pairs) and reported significant genetic effects (Edelbrock et al. 1995; Schmitz et al. 1995). On the other hand, a twin study from Taiwan including 279 12 to 16-year-old twin pairs (Kuo et al. 2004) found no significant genetic influences and major effects of shared and non-shared environment. One study compared WB data of biological and non-biological adopted siblings and found modest genetic influences at first assessment (age between 10 and 15 years), but no significant genetic effects 3 years later (Van der Valk et al. 1998). These family studies are all based on parental ratings of WB. Using teacher report data of WB in 5-year-old twins, Polderman et al. (2006) found moderate genetic (49%) and non-shared environmental (51%) effects.

Only two longitudinal studies (Van der Valk et al. 1998; Schmitz et al. 1995) have examined the genetic influences on the stability of childhood WB, and both failed to find significant genetic contributions to stability. However, in both studies the power to detect such effects was very low, due to limited sample size (Schmitz et al. 1995) or the design of the study (Van der Valk et al. 1998).

To summarize, the results of studies into the heritability of childhood WB have yielded varying results. Large scale studies into WB at later ages in childhood are lacking. Moreover, little is known about the genetic and environmental mechanisms underlying stability in WB. Unlike the WB syndrome scale of the CBCL, the broad band internalizing scale that it is part of, has received a fair amount of attention in the field of behavior genetics. Longitudinal studies in our Dutch twin sample suggested increasing

influence of the shared environment and decreasing genetic effects between the age of 3 and 12 years (Van der Valk et al. 2003a; Bartels et al. 2004b, 2007a). Exploring the relative influence of genes and environment on behavioral stability, Bartels et al. (2004b) found that both genes (average influence 43%) and shared environment (average influence 47%) had a major impact on the continuity of internalizing problems over time. A study using teacher ratings of internalizing behavior failed to detect significant effects of the shared environment and found that genetic effects were mainly responsible for the stability of the behavior (Haberstick et al. 2005). Apart from WB, the internalizing problems scale of the CBCL also encompasses the syndrome scales Somatic Complaints and Anxious/Depressed behavior. Focusing on the latter syndrome scale, a recent study in our twin sample showed decreasing genetic effects with age and increasing shared environmental effects (Boomsma et al. 2005). Furthermore, virtually no sex differences in the magnitude of the heritability estimates were observed, suggesting that the influences of genetic and environmental effects are similar across the sexes. Interestingly, the two largest studies into WB (Derks et al. 2004b) and shyness/inhibition (Eley et al. 2003), both conducted in early childhood, did detect significant sex differences in the heritability estimates, with the influence of genetic effects being larger in boys than in girls. These results suggest that, in contrast to the Anxious/Depressed scale, the genetic and environmental influences on WB may be different for boys and girls, thus warranting further studies into WB at later ages. The aims of the current study are twofold. Firstly, this project, which is a follow-up of the twin sample studied by Derks et al. (2004b), aims to examine the etiology of variation in WB at various time points across childhood. Secondly, using the longitudinal nature of this study, we aim to assess the genetic and environmental factors underlying the stability of childhood WB.

Ratings of both maternal and paternal reported WB were incorporated in the analyses. Several studies into childhood behavioral problems have shown that different informants can provide different information about children's behavior (Verhulst et al. 1996; Achenbach and Rescorla 2000; Achenbach and Rescorla 2001; Van der Ende and Verhulst 2005; Seiffge-Krenke and Kollmar 1998). Achenbach and Rescorla (2000; 2001) reported correlations between maternal and paternal ratings of WB of .69 for preschool children and of .57 for school-aged children. These correlations were based on data from a combined clinical and general population sample. In a general population only sample using the Dutch CBCL, the correlation between parental ratings of WB was found to be between .48 and .79 (Verhulst et al. 1996). The less-than-perfect correlation between parental ratings implies rater disagreement. Various studies have explored the sources of parental disagreement in ratings of problem behavior (Derks et al. 2004b; Bartels et al. 2003, 2004a; Van der Valk et al. 2003b) using different structural equation models. Generally, it was found that parental agreement and disagreement was best explained by a psychometric model. This model, developed by Hewitt et al. (1992), assumes that parents not only assess the exact same behavior of a child, but also rate an informant specific aspect of the child's behavior. This unique perception of the child's behavior can arise if the child behaves differently towards the different raters (e.g. the child is more withdrawn when it is with its mother than when it spends time with its father), or if the raters observe the child in different situations (e.g. the mother observes the child more often in the home environment, whilst the father often observes the child interacting with other children in the playground). Apart from these real differences in behavior, the unique perception of the child's behavior may also be influenced by rater bias and unreliability. Rater bias may arise if parents hold on to different normative standards, have specific response styles, or tend to stereotype the child's behavior. Unreliability may be an important source of rater disagreement if raters cannot give an accurate description of the behavior under study. This may be relevant to our analyses, as some studies suggest that parents may be relatively insensitive to the more covert internalizing problems of children (Seiffge-Krenke and Kollmar 1998; Ollendick and King 1994). A previous longitudinal multiple rater study in our sample found that rater disagreement variance accounted for 35% of the individual differences in internalizing problems (Bartels et al. 2007a).

In the present study, stability of WB was assessed in longitudinal CBCL data from a large sample of 3, 7, 10 and 12 year-old twin pairs. The sample included roughly equal numbers of boys and girls. Genetic and environmental effects on stability in childhood WB were examined for both sexes. As both mother and father ratings of the twin's behavior were incorporated in the analyses, we could partial out rater bias effects by distinguishing between variance that is shared between parents (i.e. perception of the child's behavior common to both raters) and variance that is specific to one rater and might include variance due to rater bias. To examine the mechanism underlying the stability of WB, various developmental models were fitted to this longitudinal data set.

## Method

### Participants

All participants were contacted via the Netherlands Twin Register (NTR), kept by the Department of Biological Psychology at the VU University in Amsterdam (Boomsma

et al. 2002, 2006a; Bartels et al. 2007b). From 1986 onwards the NTR has recruited families with multiples a few weeks or months after birth. Currently 40–50% of all multiple births are registered at the NTR. For the present study data from twins born in 1986–2001 were included. Parents of the twins were asked to fill out a questionnaire assessing the twin's behavior at age 3, 7, 10 and 12 years. The questionnaires were mailed within 3 months of the twin's 3rd, 7th, 10th and 12th birthdays. Two to three months after this mailing reminders were sent to the non-responders. If finances permitted, persistent non-responders were contacted by phone. This procedure yielded a response rate between 61% and 73% (Bartels et al. 2007b). Non-responders also include twin families who moved to an unknown address. From the original sample 281 families were excluded because either one or both of the children had a disease or handicap that interfered with daily functioning. The total sample consisted of 14,889 twin families. Ratings from both parents were available for 8,479 families when the twins were 3 years old, 6,414 at age 7, 4,133 at age 10, and 2,900 at age 12. Complete data from both parents at all time points were available for 1,160 families. Maternal ratings were available for 14,735 families, of which 13,095 participated at age 3, 8,855 at age 7, 5,863 at age 10, and 3,958 at age 12. Maternal data at all ages were available for 2,797 families. Paternal ratings were available for 11,499 families, of which 8,794 families participated when the twins were 3 years old, 6,522 at age 7, 4,237 at age 10, 2,974 at age 12. Complete father data on all four time points were available for 1,290 families. This study is part of an ongoing project; the children born in later birth cohorts have not reached the age of 7, 10 or 12 years yet, which explains the decreasing numbers of participating families at the later ages.

To examine effects of sample attrition, we compared WB scores at age 3 of families who continued to participate at all other time points (when the twins were 7, 10, and 12 years of age), with families who only responded twice, once, or zero times at the subsequent time points. For the father ratings there were no mean differences between these groups, neither for boys nor for girls. For the mother ratings a significant effect of attrition was observed for both boys and girls. Mothers who continued to participate reported lower WB scores when their twins were 3 years old than mothers who did not participate at one or more of the later measurement occasions However, these effects were small (effect size $r = .07$ in both boys and girls). No systematic differences in variance were observed between complete versus partial responders.

Of all participating twin pairs, 2,310 were monozygotic males (MZM), 2,591 were dizygotic males (DZM), 2,619 monozygotic females (MZF), 2,339 dizygotic females (DZF), 2,566 opposite sex twins with a male firstborn (DOSMF), and 2,464 opposite sex twins with a female firstborn (DOSFM). For 1,380 same-sex twin pairs zygosity was based on DNA polymorphisms ($n = 1{,}039$) or blood group ($n = 341$; Van Dijk et al. 1996). For the remaining same sex twin pairs ($n = 8{,}479$) zygosity was determined by discriminant analysis using longitudinally collected questionnaire items. This method has proven to be of sufficient reliability: Rietveld et al. (2000) reported that agreement between this method and zygosity determination by blood/DNA polymorphisms was 93%.

## Measures

Mother and father ratings of WB problems were obtained from the withdrawn/depressed syndrome scale of the CBCL/2-3 at age 3. The CBCL/2-3 (Achenbach 1992) has been translated and validated for the Dutch population (Koot et al. 1997). The withdrawn/depressed scale in the CBCL/2-3 consists of 10 items (see Table 1). The parents were asked to rate the behavior of the child on a 3-point scale based on the occurrence of the behavior in the past 2 months. They were asked to rate the behavior as 0 if the problem item was not true; 1 if the item was somewhat or sometimes true; and 2 if it was very true or often true.

At age 7, 10 and 12 WB was assessed using ratings of the WB syndrome scale of the CBCL/4-18 (Achenbach 1991; Verhulst et al. 1996). The syndrome scale withdrawn encompasses nine items, partially overlapping with the items in the CBCL/2-3 (see Table 1). This time the parents were asked to rate the behavior of the child in the preceding 6 months, on a 3-point scale identical to the scale

**Table 1** Overview of the items included in the withdrawn behavior scale of the CBCL

| Items CBCL/2-3 | Items CBCL/4-18 |
|---|---|
| Acts too young for age | Would rather be alone than with others |
| Doesn't answer when people talk to him/her | Refuses to talk |
| Doesn't know how to have fun; acts like a little adult | Secretive, keeps things to self |
| Looks unhappy without a good reason | Too shy or timid |
| Seems unresponsive to affection | Stares blankly |
| Shows little affection toward people | Sulks a lot |
| Shows little interest in things around him/her | Underactive, slow moving, or lacks energy |
| Stares into space or seems preoccupied | Unhappy, sad, or depressed |
| Strange behavior | Withdrawn, doesn't get involved with others |
| Unhappy, sad or depressed | |

used in the CBCL/2-3. For all ages, if more than two items on the WB scale were missing, the data were regarded incomplete and excluded from the analyses.

## Data analyses

Descriptive statistics for WB at age 3, 7, 10 and 12 were calculated using SPSS 13, correlations and cross-correlations were estimated using the software package Mx (Neale et al. 2006). To assess stability of WB over time, phenotypic correlations between the time points were estimated for boys and girls separately. Twin correlations at each age and cross-twin-cross-age correlations were estimated for each zygosity group separately. These correlations give a first impression of the contribution of genetic and environmental effects on the variance of WB at each age, and on the stability of WB over time. Within-person inter-parent correlations were inspected to examine parental agreement on WB. The cross-rater cross-twin correlations (e.g. the correlation between the mother rating of the oldest of the twins with the father rating of the youngest of the twins) were estimated to gain insight in the contribution of genes and environment on the phenotypic variance that both raters agree on.

## Genetic modeling

Since monozygotic (MZ) twins are (nearly) genetically identical, while dizygotic (DZ) twins on average only share 50% of their segregating genes, genetic model fitting of twin data allows for separation of the observed phenotypic variance into variance due to additive genetic factors (A), shared environmental (C) and non-shared environmental (E) factors. To incorporate the WB ratings from both parents simultaneously, a psychometric model was used (see Fig. 1). The psychometric model enables a distinction

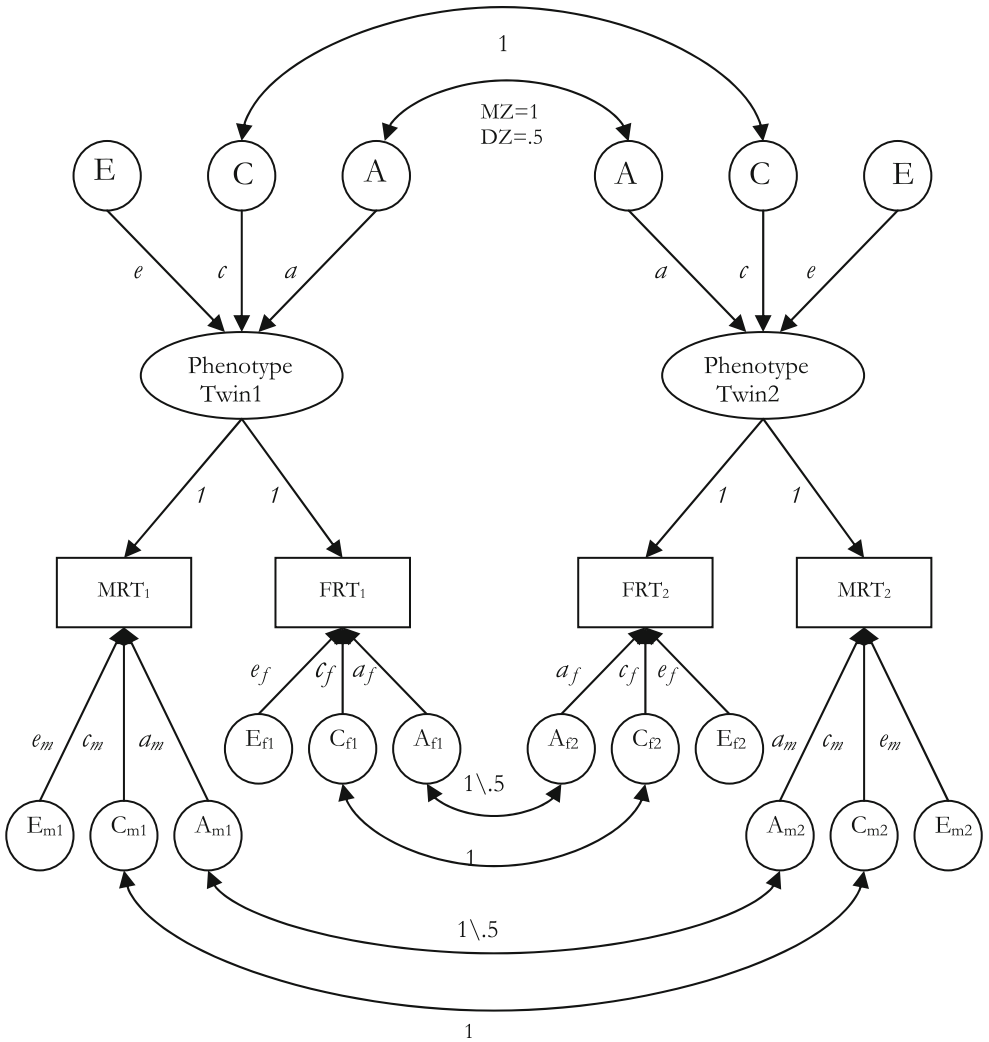


**Fig. 1** Psychometric model for multiple raters. A = Additive genetic effects; C = Shared environmental effects; E = Non-shared environmental effects; $MRT_{1/2}/FRT_{1/2}$ = Mother/Father rating twin 1/2

between the variance that is shared by both raters and is independent of rater bias and unreliability. This shared variance is also called common phenotypic variance or reliable trait variance. The variance that is rater specific is variance in the child's behavior that is uniquely perceived by one of the parents, or is associated with parental characteristics. This is also called unique or rater specific variance. In the psychometric twin model, both the common and the unique phenotypic variance are decomposed into components due to genetic, shared environmental and non-shared environmental influences. Significant genetic effects on the rater specific variance indicate that these unique perceptions of the child's behavior are real, in the sense that they reflect variance associated with heritable behavior rather than systematic or unsystematic error, as error and unreliability do not cause systematic effects and cannot mimic genetic influences. Shared environmental effects on the unique phenotype may be confounded by rater bias, as possible influences of rater bias will act independently of the zygosity of the twins. Unique non-shared environmental influences may be confounded by measurement error or unreliability.

To examine the stability of WB throughout childhood, the psychometric model was extended to incorporate data on all four time points (see Bartels et al. (2007a) for a comprehensive description of the longitudinal application of the psychometric model). To gain insight in what factors are important for the continuity of WB, all possible genetic and environmental influences were specified using Cholesky decompositions for both the common and rater specific phenotype. A path diagram depicting the genetic influences in this model is shown in Fig. 2; the same structures were applied to capture the shared environmental and non-shared environmental influences. This model (a saturated psychometric model) is descriptive rather than driven by any specific developmental hypothesis. However, it is useful to gain a first insight in what factors are important for the stability of WB and serves as a reference to evaluate the fit of more specific developmental models. If the more constrained developmental model fits the data significantly worse than the fully parameterized model, this indicates that the predicted developmental mechanism is inconsistent with the data, and the hypothesized model should be rejected.

Two types of developmental models were fitted to the data. Both models were fitted for A, C, and E influences separately (leaving the other influences specified using Chlolesky decompositions), and evaluated against the fit of the saturated model. Firstly, the transmission, or simplex model (Boomsma and Molenaar 1987) was tested. In this model, covariances between the four times of measurement are accounted for by genetic and/or environmental

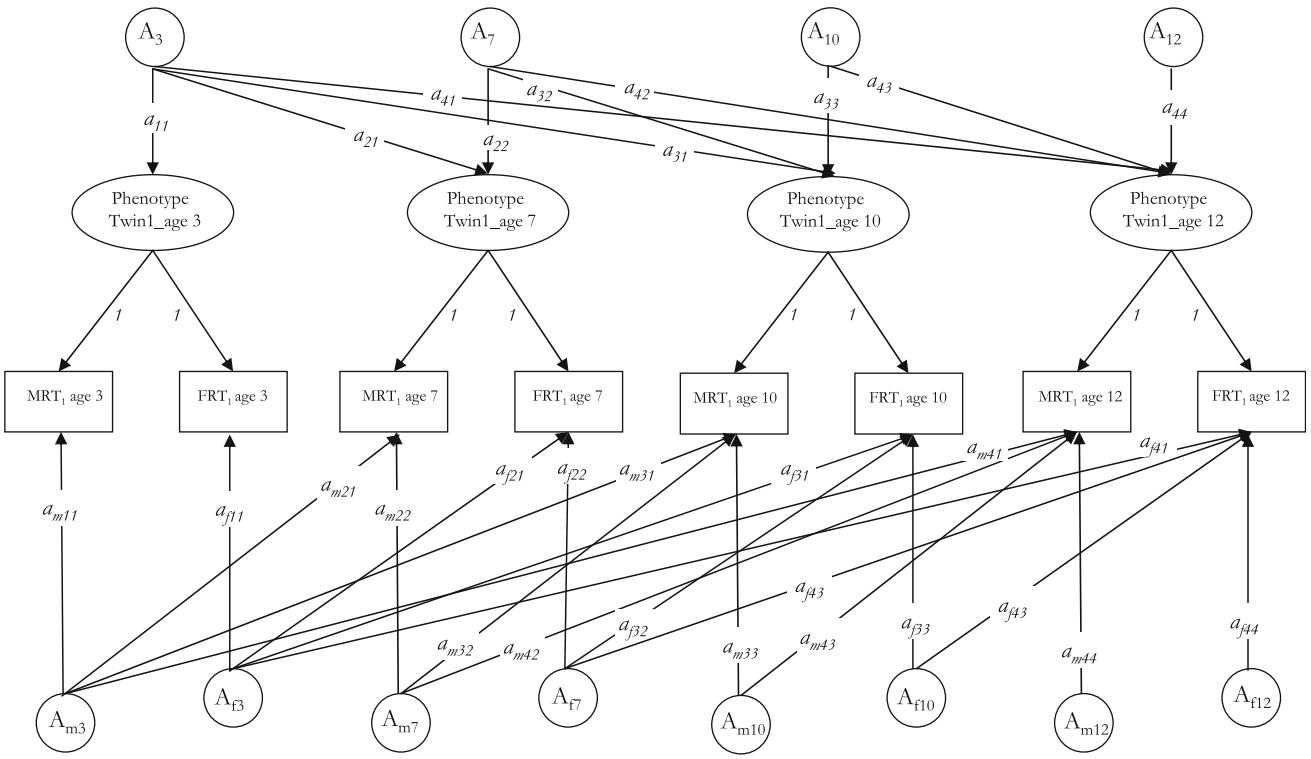


**Fig. 2** Longitudinal psychometric model for multiple raters, shown for one member of the twin pair and for additive genetic influences (A) only. $MRT_1/FRT_1$ = Mother/Father rating twin 1

influences that are carried over to subsequent time points. Apart from the influences from prior time points, an innovation term unique to each measurement occasion can affect the variance. The total variance at each time point is the sum of the innovation effect and the age-to-age carry-over effect. The second developmental model that was tested was the common factor model (Martin and Eaves 1977). In the common factor model, one underlying factor is specified, implying a continuous influence throughout the different time points. To also account for age specific variance, age specific influences are added to the model. If the common factor model fitted the data well, the significance of the age specific influences was tested subsequently, by dropping these effects from the model. Likewise, the fit of a model with only age specific influences (without an underlying common factor) was tested. Since the non-shared environmental component also includes measurement error, the age specific non-shared environmental influences on the unique phenotype were per definition specified in the model.

The model fit was evaluated using likelihood ratio tests and Akaike's information criterion (AIC = $\chi^2 - 2\Delta df$). The best fitting most parsimonious model was used to obtain estimates of genetic, shared environmental and non-shared environmental effects on the variances and covariances of WB. Genetic modeling was performed in Mx (Neale et al. 2006). In order to use all available data, including information of incomplete longitudinal data or data of which one of the parental ratings is missing, analyses were performed on the raw data.

## Results

Table 2 shows the means and standard deviations for maternal and paternal rated WB, as assessed with the WB syndrome scale of the CBCL. Problem scores are shown for boys and girls from the twin sample, and from a community sample (Verhulst et al. 1996; Van den Oord et al. 1995). Mother ratings of WB were significantly higher than father ratings at all ages and in both boys and girls (all tests significant at $P < .001$ level). Withdrawn behavioral problems were significantly higher in boys than in girls at age 3 ($P < .001$ for both raters), age 10 (mother rating $P = .04$; father rating $P = .01$) and age 12 years (mother rating $P = .05$; father rating $P < .001$), but not at age 7 years (mother rating $P = .24$; father rating $P = .72$). As the power to detect mean differences was high in our

**Table 2** Sample sizes, means and standard deviations (SD) for mother and father ratings of withdrawn behavior at age 3, 7, 10 and 12 for boys and girls separately

| | Mother ratings | | | | Father ratings | | | |
|---|---|---|---|---|---|---|---|---|
| | *N* | Mean (SD) | Skewness (SE) | Kurtosis (SE) | *N* | Mean (SD) | Skewness (SE) | Kurtosis (SE) |
| Age 3 | | | | | | | | |
| Boys | 12,940 | 1.23 (1.61) | 2.45 (.02) | 9.22 (.04) | 8,686 | 1.12 (1.49) | 2.28 (.03) | 7.50 (.05) |
| Girls | 13,078 | 1.08 (1.42) | 2.38 (.02) | 8.75 (.04) | 8,761 | 1.00 (1.32) | 2.13 (.03) | 6.36 (.05) |
| Community boys | 214 | 1.26 (2.09) | | | | | | |
| Community girls | 199 | 1.02 (1.54) | | | | | | |
| Age 7 | | | | | | | | |
| Boys | 8,698 | 1.61 (1.77) | 2.02 (.03) | 6.77 (.05) | 6,442 | 1.34 (1.62) | 2.00 (.03) | 6.1 (.06) |
| Girls | 8,963 | 1.68 (1.72) | 1.70 (.03) | 4.35 (.05) | 6,578 | 1.35 (1.56) | 1.94 (.03) | 6.17 (.06) |
| Community boys | 579 | 1.61 (1.75) | | | | | | |
| Community girls | 593 | 1.79 (1.95) | | | | | | |
| Age 10 | | | | | | | | |
| Boys | 5,644 | 1.65 (1.92) | 1.92 (.03) | 5.07 (.07) | 4,116 | 1.32 (1.69) | 2.05 (.04) | 5.91 (.08) |
| Girls | 6,043 | 1.60 (1.83) | 2.07 (.03) | 6.65 (.06) | 4,347 | 1.27 (1.63) | 2.17 (.04) | 7.03 (.07) |
| Community boys | 579 | 1.61 (1.75) | | | | | | |
| Community girls | 593 | 1.79 (1.95) | | | | | | |
| Age 12 | | | | | | | | |
| Boys | 3,836 | 1.47 (1.92) | 2.21 (.04) | 7.19 (.08) | 2,863 | 1.22 (1.73) | 2.24 (.05) | 6.78 (.09) |
| Girls | 4,055 | 1.42 (1.74) | 2.00 (.04) | 6.55 (.08) | 3,068 | 1.15 (1.58) | 2.12 (.04) | 6.39 (.09) |
| Community boys | 440 | 2.06 (2.41) | | | | | | |
| Community girls | 456 | 2.10 (2.32) | | | | | | |

*Note*: Data from the Dutch community sample are derived from (Verhulst et al. 1996) and (Van den Oord et al. 1995). At all ages, the majority (≥95%) of the children from the community sample were rated by their mothers

study, these mean differences seem only of practical importance at age 3. In all subsequent analyses, the means were specified per sex, zygosity, birth order and age. The ratings of WB in our twin sample were similar to the scores in the community sample at age 3, 7, and 10 years. At age 12, both the mean scores and the variances were higher in the community sample.

Although skewness was observed, we used the untransformed scores. A simulation study by Derks et al. (2004a) showed that a square root transformation of the data (the most commonly used transformation when data are censored) does not remove bias induced by non-normality of the data.

**Table 3** Phenotypic correlations across age for mother and father ratings of withdrawn behavior in boys (above diagonal) and girls (below diagonal). The last 2 columns give cross-rater correlations ($r_{m-f}$) for boys and girls as a function of age

| Age | Mother ratings | | | | Father ratings | | | | $r_{m-f}$ | $r_{m-f}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| | 3 | 7 | 10 | 12 | 3 | 7 | 10 | 12 | Boys | Girls |
| 3 | – | .33 | .30 | .29 | – | .27 | .23 | .26 | .55 | .51 |
| 7 | .32 | – | .59 | .51 | .27 | – | .54 | .47 | .58 | .55 |
| 10 | .27 | .56 | – | .65 | .27 | .55 | – | .60 | .58 | .57 |
| 12 | .27 | .51 | .59 | – | .23 | .44 | .56 | – | .62 | .59 |

The phenotypic correlations across age are given in Table 3 for both the mother and the father ratings, separately for boys and girls. These correlations give an indication of the stability of WB over time. Correlations were around .30 between age 3 and later ages, and increased to .44–.65 between age 7, 10, and 12 years. This pattern was similar in both parental ratings and was observed in both boys and girls. The last two columns of Table 3 show the level of agreement between maternal and paternal ratings. Within-person-cross-rater correlations were similar across sex and age and varied between .51 and .62.

The twin correlations and cross-twin-cross-age correlations are presented in Table 4 for both the mother and the father ratings. Inspection of the MZ and DZ twin correlation (on the diagonal) at the four time points gives a first impression of what factors influence variance in WB. At all ages, MZ correlations were higher than DZ correlations in both sexes, indicating that genetic factors play a role. Twin correlations in opposite sex twins were similar to the correlations in DZ same sex twins, suggesting that there are no qualitative sex differences in genes or shared environment influencing WB. This was confirmed in the genetic model. When the genetic correlation in opposite sex twins was estimated freely (within the biologically plausible range of 0–.5) it was estimated at .5, identical to the genetic correlation in same sex twins. Apart from age 7, the MZ

**Table 4** Twin correlations, cross-twin-cross-age correlations and cross-twin-cross-rater correlations (within and across age) for mother and father ratings of withdrawn behavioral problems per zygosity group (MZM, DZM, DOSMF above diagonals; MZF, DZF, DOSFM below diagonals)

| Age | Mother ratings | | | | Father ratings | | | | Cross-rater (Father/Mother ratings) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 3 | 7 | 10 | 12 | 3 | 7 | 10 | 12 | 3 | 7 | 10 | 12 |
| MZF/MZM | | | | | | | | | | | | |
| 3 | .72/.71[a] | .23 | .21 | .23 | .73/.78[a] | .22 | .20 | .17 | 41/.39[a] | .19 | .18 | .17 |
| 7 | .29 | .65/.60[a] | .42 | .37 | .29 | .68/.64[a] | .38 | .35 | .26 | .40/.38[a] | .27 | .27 |
| 10 | .22 | .44 | .63/.61[a] | .44 | .26 | .43 | .63/.60[a] | .39 | .22 | .31 | .36/.33[a] | .35 |
| 12 | .27 | .44 | .47 | .67/.63[a] | .23 | .39 | .36 | .56/.63[a] | .26 | .30 | .30 | .39/.38[a] |
| DZF/DZM | | | | | | | | | | | | |
| 3 | .47/.40[b] | .18 | .20 | .20 | .50/.48[b] | .14 | .14 | .19 | .27/.24[b] | .11 | .12 | .14 |
| 7 | .20 | .32/.27[b] | .23 | .23 | .21 | .40/.29[b] | .18 | .18 | .18 | .18/.12[b] | .16 | .14 |
| 10 | .23 | .22 | .36/.38[b] | .29 | .24 | .28 | .44/.33[b] | .27 | .20 | .20 | 24/.22[b] | .19 |
| 12 | .21 | .24 | .28 | .37.38[b] | .24 | .23 | .27 | .41/.40[b] | .15 | .16 | .21 | .20/.26[b] |
| DOSFM/DOSMF | | | | | | | | | | | | |
| 3 | .40/.45[c] | .16 | .16 | .16 | .51/.50[c] | .16 | .15 | .18 | .26/.26[c] | .12 | .13 | .13 |
| 7 | .17 | .37/.33[c] | .22 | .18 | .12 | .39/.34[c] | .22 | .18 | .10 | .17/.14[c] | .12 | .06 |
| 10 | .20 | .25 | .36/.35[c] | .22 | .18 | .28 | .46/.32[c] | .26 | .18 | .15 | .23/.18[c] | .13 |
| 12 | .16 | .23 | .23 | .34/.41[c] | .14 | .19 | .30 | .31/.45[c] | .08 | .11 | .14 | .17/.24[c] |

*Note*: MZM = monozygotic males; DZM = dizygotic males; MZF = monozygotic females; DZF = dizygotic females; DOSMF = dizygotic opposite sex, male firstborn; DOSFM = dizygotic opposite sex, female firstborn. [a]First figure correlation MZF, second figure correlation MZM; [b]correlation DZF/DZM; [c]correlation DOSFM/DOSMF

correlations were not twice as high as the DZ correlations, suggesting that shared environmental factors also play a role. Inspection of the MZ and DZ cross-twin-cross-age correlations (off-diagonal in Table 4) can provide insight in what factors are important for the stability of WB over time. As compared to the DZ cross correlations, MZ cross correlations were slightly higher between age 3 and subsequent ages, and considerably higher between age 7 and later ages, indicating genetic effects on stability. However, MZ cross correlations were not twice as high, particularly between age 3 and later ages, suggesting that shared environmental effects on stability are also important.

The far right panel of Table 4 displays the cross-twin-cross-rater correlations within age (diagonal) and across age (off-diagonal). These correlations yield a first impression on the importance of genes and environment on the common phenotypic variance, and thus the reliable trait variance. At all ages, the MZ cross-rater correlations were larger than the DZ cross-rater correlations, indicating genetic effects on the common phenotype. Apart from the correlations at age 7, the DZ correlations were higher than would be expected based on genetic influences alone, therefore shared environmental influences also seem to influence the common phenotypic variance. For all ages the cross-twin-cross-rater correlations were lower than the cross-twin-within-rater correlations. These differences indicate parental disagreement, and reveal the part of the total variance that is uniquely observed by a specific rater. Similar to the pattern of the within age twin correlations, the cross-twin-cross-rater correlations across age were higher in MZ than in DZ twins, indicating genetic effects. These correlations were less than twice as high in MZ twins compared to DZ twins, especially in girls, indicating that shared environmental influences also play a role in the stability of the common phenotype. The cross-twin-cross-rater-cross-age correlations were similar to the cross-twin-within-rater-cross-age correlations between age 3 and later ages. This pattern indicates that practically all the stability of WB was perceived by both raters. In later phases of childhood, the cross-twin-within-rater-cross-age correlations were larger than the cross-twin-cross-rater-cross-age correlations, indicating that both the common and the unique phenotype show considerable continuity over time.

Table 5 displays the results of the model fitting procedure. The relative importance of the different variance components was found to be significantly different in boys

**Table 5** Model fitting results for multivariate longitudinal analyses of withdrawn behavior

| Model | −2LL | df | vs. | $\chi 2$ | *P* | AIC |
|---|---|---|---|---|---|---|
| 1. ACE Saturated | 371758.639 | 107,842 | | | | |
| 2. ACE Saturated no sex diff | 372277.660 | 107,932 | 1 | 519.021 | <.001 | 339.021 |
| 3. ACE drop $A_c$ | 372472.376 | 107,862 | 1 | 713.737 | <.001 | 673.737 |
| 4. ACE drop $A_{um}$ | 372316.332 | 107,862 | 1 | 557.693 | <.001 | 517.693 |
| 5. ACE drop $A_{uf}$ | 372097.704 | 107,862 | 1 | 339.065 | <.001 | 299.065 |
| 6. ACE drop $C_c$ | 371797.927 | 107,862 | 1 | 39.158 | .006 | −.842 |
| **7. ACE drop $C_{um}$** | **371783.772** | **107,862** | **1** | **25.133** | **.196** | **−14.867** |
| 8. ACE drop $C_{uf}$ | 371842.211 | 107,862 | 1 | 83.572 | <.001 | 43.572 |
| 9. ACE drop $E_c$ | 373023.669 | 107,862 | 1 | 1265.03 | <.001 | 1225.03 |
| 10. $A_c$ transmission | 372338.480 | 107,848 | 1 | 579.841 | <.001 | 567.841 |
| 11. $A_c$ common + age specific | 371803.511 | 107,846 | 1 | 44.872 | <.001 | 36.872 |
| 12. $A_c$ age specific | 372471.108 | 107,854 | 1 | 712.469 | <.001 | 688.469 |
| 13. $C_c$ transmission | 371800.769 | 107,848 | 1 | 42.130 | <.001 | 30.130 |
| 14. $C_c$ common + age specific | 371760.347 | 107,846 | 1 | 1.708 | .789 | −6.292 |
| **15. $C_c$ common** | **371758.769** | **107,854** | **1** | **.130** | **.999** | **−23.870** |
| 16. $C_c$ age specific | 371801.204 | 107,854 | 1 | 42.565 | <.001 | 18.565 |
| 17. $E_c$ transmission | 372162.778 | 107,848 | 1 | 404.139 | <.001 | 392.139 |
| 18. $E_c$ common + age specific | 372067.131 | 107,846 | 1 | 308.492 | <.001 | 300.492 |
| 19. $E_c$ age specific | 372924.896 | 107,854 | 1 | 1166.257 | <.001 | 1142.257 |
| **20. No $C_{um}$; $C_c$ common** | **371789.382** | **107,874** | **1** | **30.743** | **.530** | **−33.266** |

*Note*: −2LL = minus 2 log likelihood; vs. = compared to model; AIC = Akaike's Information Criterion; ACE Saturated no sex diff = saturated psychometric ACE model without sex differences in the variance components. $A_c$ = Additive genetic influences on the common phenotype; $C_c$ = Shared environmental influences on the common phenotype; $E_c$ = Non-shared environmental influences on the common phenotype; $A_u$ = Additive genetic influences on the unique phenotype as rated by the mother ($A_{um}$) or the father ($A_{uf}$); $C_u$ = Shared environmental influences on the unique phenotype rated by the mother/father ($C_{um}$/$C_{uf}$); $E_u$ = Non-shared environmental influences on the unique phenotype rated by the mother/father ($E_{um}$/$E_{uf}$); best fitting submodels and final model are expressed in bold

and girls ($\chi^2(90) = 579.841$, $P < .001$; model 2), indicating sex differences in heritability. The significance of all genetic and environmental components was tested by examining the deterioration of the model fit after each component was dropped from the saturated psychometric model. Shared environmental influences on the phenotype unique to the mother could be omitted from the model without a significant reduction in fit ($\chi^2(20) = 25.133$, $P = .196$; model 7). All other variance components were significant, for both the common phenotype (models 3, 6, and 9) and the unique phenotype (models 4, 5, and 8). Next, both transmission and common factor models were fitted to the different variance components. The saturated psychometric model indicated that the contribution of the rater specific variance to the stability of WB over time was negligible. Therefore, the developmental models were only tested for the variance common to both raters. When the developmental models were fitted to the common phenotype, the Cholesky decomposition was maintained for the rater specific factors. Both the genetic and the non-shared environmental influences appeared to be too complex to be captured by one of the developmental models; fitting these models to the common genetic and non-shared environmental influences resulted in a significant deterioration of the model fit (models 10–12 and 17–19). The shared environmental influences however, could be described by a common factor model without age specific influences ($\chi^2(12) = .130$, $P = .999$; model 15). A good fit of this developmental structure implies that there is a continuous influence of one underlying factor (in this case: the environment shared by the twins) over time. The best fitting most parsimonious model was a longitudinal psychometric model without shared environmental influences on the mother specific phenotype and including a common factor structure capturing the shared environmental influences on the common phenotype (model 20).

**Table 6** Relative contributions of genetic (A), shared (C) and non-shared (E) environmental influences to the variance (diagonal) and covariances (off-diagonal) of withdrawn behavior for the common phenotype and the unique (rater specific) phenotype for boys (above diagonal) and girls (below diagonal)

| | Age | Mother ratings | | | | Father ratings | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 3 | 7 | 10 | 12 | 3 | 7 | 10 | 12 |
| $A_c$ | 3 | .26/.32[a] | .85 | .69 | .77 | .29/.34[a] | .88 | .69 | .76 |
| | 7 | .69 | .40/.44[a] | .46 | .73 | .75 | .39/.42[a] | .54 | .62 |
| | 10 | .60 | .36 | .16/.20[a] | .32 | .60 | .41 | .19/.24[a] | .40 |
| | 12 | .58 | .66 | .24 | .22/.25[a] | .60 | .59 | .29 | .22/.26[a] |
| $C_c$ | 3 | .11/.06[a] | .01 | .12 | .08 | .12/.06[a] | .01 | .12 | .07 |
| | 7 | .09 | .01/.00[a] | .01 | .00 | .10 | .01/.00[a] | .01 | .00 |
| | 10 | .24 | .05 | .06/.04[a] | .03 | .24 | .05 | .07/.04[a] | .04 |
| | 12 | .26 | .05 | .11 | .06/.01[a] | .27 | .05 | .13 | .06/.01[a] |
| $E_c$ | 3 | .08/.09[a] | .11 | .19 | .10 | .08/.10[a] | .11 | .19 | .10 |
| | 7 | .09 | .12/.13[a] | .06 | .26 | .10 | .12/.13[a] | .07 | .22 |
| | 10 | .16 | .04 | .04/.05[a] | .04 | .16 | .05 | .05/.06[a] | .05 |
| | 12 | .12 | .25 | .05 | .11/.11[a] | .13 | .23 | .06 | .11/.11[a] |
| $A_u$ | 3 | .38/.34[a] | .02 | .00 | .03 | .21/.26[a] | .00 | .00 | .03 |
| | 7 | .13 | .24/.22[a] | .34 | .00 | .00 | .22/.24[a] | .20 | .10 |
| | 10 | .00 | .39 | .41/.40[a] | .39 | .00 | .31 | .30/.26[a] | .07 |
| | 12 | .04 | .03 | .44 | .34/.34[a] | .00 | .00 | .10 | .16/.24[a] |
| $C_u$ | 3 | .00/.00[a] | .00 | .00 | .00 | .12/.12[a] | .00 | .00 | .03 |
| | 7 | .00 | .00/.00[a] | .00 | .00 | .05 | .09/.05[a] | .07 | .05 |
| | 10 | .00 | .00 | .00/.00[a] | .00 | .00 | .04 | .07/.09[a] | .19 |
| | 12 | .00 | .00 | .00 | .00/.00[a] | .00 | .14 | .15 | .15/.14[a] |
| $E_u$ | 3 | .17/.19[a] | .01 | .00 | .02 | .18/.12[a] | .00 | .00 | .00 |
| | 7 | .00 | .23/.21[a] | .13 | .00 | .00 | .17/.16[a] | .12 | .00 |
| | 10 | .00 | .16 | .33/.31[a] | .22 | .00 | .14 | .32/.31[a] | .25 |
| | 12 | .00 | .00 | .16 | .27/.29[a] | .00 | .00 | .27 | .30/.24[a] |

*Note*: $A_c$ = Additive genetic influences on the common phenotype; $C_c$ = Shared environmental influences on the common phenotype; $E_c$ = Non-shared environmental influences on the common phenotype; $A_u$ = Additive genetic influences on the unique phenotype; $C_u$ = Shared environmental influences on the unique phenotype; $E_u$ = Non-shared environmental influences on the unique phenotype. [a]First figure is the relative contribution for girls, second figure for boys

Table 6 presents the relative contributions of genetic, shared and non-shared environmental influences to the common phenotype ($A_c$, $C_c$, and $E_c$), and to the phenotype unique to either the mother or the father ratings ($A_u$, $C_u$, $E_u$) in the best fitting model. Together these influences make up the total (common + unique) variances of WB at each age (diagonal) and the total covariance of WB over time (off-diagonal). The heritability of paternal and maternal rated behavior ranged between 50 and 66% in both sexes at age 3 and age 7, and decreased slightly to 38–59% at age 12 years. Shared environmental effects were of modest importance for the variance in both sexes and at all ages but were slightly more pronounced in girls. Non-shared environmental influences increased somewhat with age, and explained 22–28% of the variance at age 3 to 35–41% at age 12 years. This increase is mainly reflected in the rater specific non-shared environmental factors. At age 3 and age 7, the common and the unique phenotypic variance contributed equally to the total variance. At later ages in childhood, substantial extra information was added by the specific raters, especially by the mothers. A large proportion of the rater specific variance was due to genetic influences, and therefore reflects variance associated with heritable behavior rather than systematic or unsystematic error. No significant mother specific shared environmental influences were found, about half of the shared environmental variance in father-rated WB was rater specific. These effects may be real father specific shared environmental effects, but may also be due to rater bias.

The off-diagonals of Table 6 show the influences of common and rater specific genetic and environmental influences on the covariances. Common genetic influences were most important for explaining continuity of WB, and accounted on average for 53% ([69% + 60% + 58% + 36% + 66%+ 24% + 75% + 60% + 60% + 41% + 59% + 29%=]/12 = 53%) of the total covariance in girls and 64% of the covariance in boys. Rater specific genetic influences were mainly of importance for explaining stability between short time intervals (e.g. between age 3 and 7, 7 and 10, or 10 and 12 years) and accounted for 0–44% of the stability. Shared environmental influences common to both raters were important for the stability of WB in girls, and explained on average 14% of the behavioral continuity. In boys, these effects only explained about 4% of the covariance. Shared environmental influences unique to the father explained 6% of the stability in both sexes; these effects might reflect rater bias. Shared environmental influences on the mother specific phenotype were not significant. All in all, most of the shared environmental influences to stability were common to both raters. As this covariance reflects behavior that was perceived by both raters, these effects cannot be due to rater bias. The bulk of the non-shared environmental influences on the continuity of WB was perceived by both raters and can thus not be attributed to systematic or unsystematic error.

**Table 7** Genetic, shared and non-shared environmental correlations across time for boys (above diagonal) and girls (below diagonal)

| Age | 3 | 7 | 10 | 12 | 3 | 7 | 10 | 12 | 3 | 7 | 10 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $A_c$ | | | | $A_{um}$ | | | | $A_{uf}$ | | | |
| 3 | – | .85 | 1.00 | .92 | – | .02 | .00 | .03 | – | .00 | .00 | .05 |
| 7 | .75 | – | .85 | .99 | .15 | – | .61 | .00 | .00 | – | .39 | .21 |
| 10 | 1.00 | .75 | – | .92 | .00 | .64 | – | .66 | .00 | .59 | – | .15 |
| 12 | .75 | 1.00 | .75 | – | .04 | .04 | .68 | – | .00 | .00 | .23 | – |
| | $C_c$ | | | | $C_{um}$ | | | | $C_{uf}$ | | | |
| 3 | – | 1.00 | 1.00 | 1.00 | – | .00 | .00 | .00 | – | .00 | .00 | .08 |
| 7 | 1.00 | – | 1.00 | 1.00 | .00 | – | .00 | .00 | .17 | – | .51 | .36 |
| 10 | 1.00 | 1.00 | – | 1.00 | .00 | .00 | – | .00 | .00 | .24 | – | .96 |
| 12 | 1.00 | 1.00 | 1.00 | – | .00 | .00 | .00 | – | .00 | .59 | .79 | – |
| | $E_c$ | | | | $E_{um}$ | | | | $E_{uf}$ | | | |
| 3 | – | .38 | 1.00 | .36 | – | .02 | .00 | .04 | – | .00 | .00 | .00 |
| 7 | .34 | – | .38 | .99 | .00 | – | .29 | .00 | .00 | – | .26 | .00 |
| 10 | 1.00 | .34 | – | .36 | .00 | .30 | – | .47 | .00 | .29 | – | .52 |
| 12 | .42 | .99 | .42 | – | .00 | .00 | .31 | – | .00 | .00 | .46 | – |

*Note*: $A_c$ = Additive genetic influences on the common phenotype; $C_c$ = Shared environmental influences on the common phenotype; $E_c$ = Non-shared environmental influences on the common phenotype; $A_u$ = Additive genetic influences on the unique phenotype as rated by the mother ($A_{um}$) or the father ($A_{uf}$); $C_u$ = Shared environmental influences on the unique phenotype rated by the mother/father ($C_{um}$/$C_{uf}$); $E_u$ = Non-shared environmental influences on the unique phenotype rated by the mother/father ($E_{um}$/$E_{uf}$)

Lastly, the correlations between the genetic influences over time and the correlations between shared and non-shared environmental effects across development are displayed in Table 7. The genetic correlation of the common phenotype remained high across time, indicating that roughly the same genes influence the stability of the reliable WB phenotype between age 3 and age 12 years. The shared environmental effects on the stability of the common phenotype could be described by a common factor model and thus reflect one continuous influence over time. The non-shared environmental correlations and the rater specific genetic and shared environmental correlations over time were lower, indicating that these effects are more variable over time.

## Discussion

We studied the etiology of WB in a large longitudinal sample of 3, 7, 10, and 12-year-old twins and explored the genetic and environmental influences on stability of WB across childhood. Both maternal and paternal ratings of their children's behavior were analyzed in order to identify the part of the phenotype that both raters agree on. This

way the variance and covariance in behavioral measurements could be distinguished from the effects of rater bias.

Individual differences in WB were largely influenced by genetic effects at all ages, in both boys (heritability estimates 50–66%) and girls (heritability estimates 38–64%). Shared environmental influences explained a small to modest proportion (0–24%) of the variance at all ages in both sexes, but were slightly more pronounced in girls. Non-shared environmental influences increased slightly with age and accounted for 22–28% of the variance at age 3 to 35–41% at age 12 years. This study is the first large scale study examining genetic and environmental influences on WB at multiple time points in childhood which is adequately powered to test for shared environmental effects. Results from previous family studies into childhood WB provided varying results. Two earlier twin studies in 3-year-olds from the NTR (Van den Oord et al. 1996; Derks et al. 2004b) found significant genetic effects, varying in magnitude from 45 to 74%. The most recent study with the largest sample size (Derks et al. 2004b) also found significant shared environmental effects that were more pronounced in girls than in boys. Two studies in middle to late childhood reported heritabilities of 40% (Schmitz et al. 1995) and 53% (Edelbrock et al. 1995). Two other family studies in middle to late childhood found no or little genetic effects (Kuo et al. 2004; Van der Valk et al. 1998; Schmitz et al. 1995). In the Taiwanese twin study (Kuo et al. 2004) the 95% confidence interval for additive genetic effects, however, was between 0% and 55%, indicating that the proportion of the variance explained by additive genetic influences may have been as large as 55%. On the other hand, the different results in the Taiwanese study compared to ours may also be due to cultural differences between the populations. Previous studies have suggested cultural effects on WB (Murad et al. 2003; Crijnen et al. 1999).

The longitudinal nature of the current study allowed for examination of the stability of WB throughout childhood. WB showed considerable continuity over time, with stability coefficients ranging from .23–.29 between age 3 and 12 years to .56–.65 for the shortest time interval. These correlations across age were similar to the correlations reported in other studies of childhood WB. In a small longitudinal twin sample, Schmitz et al. (1995) found a correlation of .33 between CBCL/2-3 scores and CBCL/4-18 scores of WB. In a large general population sample from the Netherlands (Verhulst et al. 1996), an 8-year stability coefficient of .36 was reported for the CBCL/4-18 WB scale. Smaller time intervals gave higher stability coefficients (.47 for a 6-year interval; .46 for a 4 year interval and .60 for a 2-year interval). Studying the development of behavioral problems over a broad time span necessitates the use of different instruments at different ages. At ages 7, 10 and 12 years the CBCL/4-18 was used, whilst the age-adjusted CBCL/2-3 was used when the twins were 3 years old. The use of different instruments could effect the longitudinal correlations in our study. However, the CBCL/2-3 and CBCL/4-18 are developed in parallel, and our 9-year stability coefficient (.23–.29) using the two different instruments is not much lower than the 8-year stability coefficient (.36) using the same instrument reported by Verhulst et al. (1996). Therefore we feel that it is likely that our longitudinal measures capture true development. Since both maternal and paternal ratings of WB were included in this study, we could also examine the extent to which the raters agree on the behavior of the child at the various ages. The cross rater correlations varied between .51 and .62 in both sexes and at all ages, indicating that rater agreement was similar in boys and girls and throughout childhood.

The stability in WB problems was largely accounted for by genetic effects, both in boys (on average 74%) and in girls (on average 65%). In girls, the shared environment was of moderate importance for continuity of WB, these influences explained 17% of the stability over time. In boys, these effects were less important, explaining about 7% of the stability. Interestingly, most of the shared environmental influences on stability were common to both raters, indicating that these effects are not due to rater bias. Non-shared environmental effects explained 18–19% of the stability over time in both sexes. Most of these effects were common to both raters.

In a longitudinal study into childhood internalizing behavior conducted in the same sample as the current report, Bartels et al. (2004b) found that stability in internalizing problems was accounted for by both genetic and shared environmental effects, and that these effects were roughly of the same importance for stability (43 vs. 47%). A multiple rater analysis including both maternal and paternal ratings (Bartels et al. 2007a) indicated similar influence of genes and shared environment on stability of internalizing problems. We found that genetic effects largely explained stability in WB (65% in girls; 74% in boys), whilst shared environmental influences were only of modest importance (7% in boys, 17% in girls). Our study suggests that, unlike the influences on general internalizing problems, shared environmental effects on stability in WB are only modest. Moreover, in line with findings in early childhood (Derks et al. 2004b; Eley et al. 2003), we found that genetic influences were stronger in boys than in girls. This is in contrast to studies into Anxious/Depressed behavior in which no meaningful sex differences in heritability estimates were found (Derks et al. 2004b; Boomsma et al. 2005).

The correlation between the genetic influences on the common phenotype over the course of development approaches unity. As the common phenotype represents the

behavior that both raters agree upon, and can thus be considered as a reliable phenotype, the high genetic correlations suggest that the same genes influence the continuity of WB over time. The shared environmental influences on the common phenotype could be modeled as a common factor, indicating that a stable persistent shared environmental influence is important for behavioral stability. Non-shared environmental correlations on the other hand vary over time in both boys and girls, suggesting that these effects are less persistent. Some studies suggest that parental behavior may moderate the stability of behavioral inhibition and shyness in young children. Inappropriate affectionate parenting (Park et al. 1997) or maternal overcontrol (Rubin et al. 2002) could increase stability of these behaviors. A study by Hastings and Rubin (1999) suggested that over-protective mothers may particularly show this behavior toward shy daughters. Such parenting practices (if they are employed toward both twins in the family) could explain why the shared environment has a stronger influence on WB in girls than in boys. If the parenting is child specific, this influence would be reflected in the non-shared environmental component. Other non-shared environmental influences, such as traumatic experiences, or the consequences of an accident or illness could also account for stability in WB.

The extent to which a child displays WB might be influenced by the composition of the family in which the child is raised. Our study focused on variation in WB in twins. It may be that twins show less WB compared to singletons, because they are raised with a sibling of the exact same age. On the other hand, if twins mainly interact with each other in childhood, twins may be more inhibited than non-twin siblings or singletons in interaction with others than their co-twin. Comparing the mean scores of our twin sample with the scores in a Dutch community sample showed that withdrawn scores are similar at age 3, 7 and 10 years. At age 12, however, mean problem scores were higher in the community sample than in the twin sample. In a large twin-singleton comparison study (Pulkkinen et al. 2003), 12-year-old twins were reported to be more socially adaptable than non-twins, but no twin-singleton differences were found for social anxiety. A twin-singleton comparison of both maternal CBCL withdrawn ratings and laboratory assessment of inhibition in 5-year-olds yielded inconsistent results (DiLalla and Caraway 2004). According to laboratory ratings, twins were more inhibited than non-twins, whilst maternal ratings showed the opposite. These studies yield no explanation for the low mean withdrawn scores observed in our twin sample at age 12. To explore possible twin-singleton differences, future studies are needed.

The present study assessed withdrawn behavioral problems using the WB syndrome scale of the CBCL. The CBCL is a clinical rating scale; since the majority of the twins included in our study are typically developing children, mean problem scores were low, and the score distribution was significantly skewed. A data simulation study (Derks et al. 2004a) showed that skewness could lead to overestimation of non-shared environmental effects and underestimation of shared environmental effects. The same study indicated that a square root data transformation did not attenuate the observed bias. Inspection of the level of censoring in our data showed that respectively 41/44% (3-year-olds maternal/paternal ratings); 28/36% (7-year-olds); 32/40% (10-year-olds) and 39/46% (12-year-olds) of the children scored at base level. Data simulation showed that a similar degree of censoring (39%) resulted in an underestimation of the shared environmental influences of 8%, and an overestimation of the non-shared environment by 10% (Derks et al. 2004a). These findings should be kept in mind when interpreting the results of our present study. Future studies could avoid the problem of censoring by using a questionnaire assessing continuously distributed behaviors.

The current study highlights the importance of genetic effects on stability in WB. Unlike general internalizing problems, shared environmental influences do only have a modest effect on the stability of WB throughout childhood. Results from longitudinal (Bongers et al. 2003) and cross-sectional studies (Achenbach 1991; Verhulst et al. 1996) indicate an increase in WB from childhood to adolescence. Future longitudinal research should extend the current study and investigate stability of WB into adolescence. Since childhood WB has been shown to be a predictor for anxiety disorders and depression later in life (Goodwin et al. 2004), insight into the developmental mechanisms underlying stability of WB into adolescence would be highly desirable. This is particularly true given the differences between the Anxious/Depressed and WB scales of the broad band internalizing scale. It appears from our studies that WB has a different genetic architecture and longitudinal course than Anxious/Depressed. While Anxious/Depressed has received a great deal of research attention, there has been relatively little investigation into the long term sequelae of WB. Our research provides evidence for the need for further work on this phenotype.

**Acknowledgements** Financial support was given by The Netherlands Organization for Scientific Research (NWO 575-25-006), (NWO 400-05-717) & (NWO/SPI 56-464-14192) and by the National Institute of mental health (NIMH, RO1 MH58799-03. Dr Bartels is financially supported by NWO (VENI 451-04-034), Dr Hoekstra is supported by NWO (Rubicon). Statistical analyses were carried out on the Genetic Cluster Computer (http://www.geneticcluster.org) which is financially supported by (NWO 480-05-003). We are indebted to all the participating twin families.

## References


Achenbach TM (1991) Manual for the child behavior checklist/4-18 and 1991 profile. University of Vermont, Department of Psychiatry, Burlington

Achenbach TM (1992) Manual for the child behavior checklist/2-3 and 1992 profile. University of Vermont, Department of Psychiatry, Burlington

Achenbach TM, Rescorla LA (2000) Manual for the ASEBA preschool forms & profiles. University of Vermont, Research Center for Children, Youth & Families, Burlington, VT

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

Alink LR, Mesman J, Van Zeijl J, Stolk MN, Juffer F, Koot HM, Bakermans-Kranenburg MJ, Van IJzerdoorn MH (2006) The early childhood aggression curve: development of physical aggression in 10- to 50-month-old children. Child Dev 77:954–966

Bartels M, Hudziak JJ, Boomsma DI, Rietveld MJ, Van Beijsterveldt CEM, Van den Oord EJCG (2003) A study of parent ratings of internalizing and externalizing problem behavior in 12-year-old twins. J Am Acad Child Adolesc Psychiatry 42:1351–1359

Bartels M, Boomsma DI, Hudziak JJ, Rietveld MJ, Van Beijsterveldt CEM, Van den Oord EJCG (2004a) Disentangling genetic, environmental, and rater effects on internalizing and externalizing problem behavior in 10-year-old twins. Twin Res 7:162–175

Bartels M, Van den Oord EJCG, Hudziak JJ, Rietveld MJ, Van Beijsterveldt CEM, Boomsma DI (2004b) Genetic and environmental mechanisms underlying stability and change in problem behaviors at ages 3, 7, 10, and 12. Dev Psychol 40:852–867

Bartels M, Boomsma DI, Hudziak JJ, Van Beijsterveldt CEM, Van den Oord EJCG (2007a) Twins and the study of rater (dis)agreement. Psychol Methods 12:451–466

Bartels M, Van Beijsterveldt CEM, Derks EM, Stroet TM, Polderman TJC, Hudziak JJ, Boomsma DI (2007b) Young Netherlands Twin Register (Y-NTR): a longitudinal multiple informant study of problem behavior. Twin Res Hum Genet 10:3–11

Biederman J, Hirshfeld-Becker DR, Rosenbaum JF, Herot C, Friedman D, Snidman N, Kagan J, Faraone SV (2001) Further evidence of association between behavioral inhibition and social anxiety in children. Am J Psychiatry 158:1673–1679

Bongers IL, Koot HM, Van der Ende J, Verhulst FC (2003) The normative development of child and adolescent problem behavior. J Abnorm Psychol 112:179–192

Boomsma DI, Molenaar PC (1987) The genetic analysis of repeated measures. I. Simplex models. Behav Genet 17:111–123

Boomsma DI, Vink JM, Van Beijsterveldt CEM, De Geus EJC, Beem AL, Mulder EJ, Derks EM, Riese H, Willemsen AHM, Bartels M, Van den Berg M, Kupper NH, Polderman TJC, Posthuma D, Rietveld MJH, Stubbe JH, Knol LI, Stroet TM, Van Baal GCM (2002) Netherlands Twin Register: a focus on longitudinal research. Twin Res 5:401–406

Boomsma DI, Van Beijsterveldt CEM, Hudziak JJ (2005) Genetic and environmental influences on anxious/depression during childhood: a study from the Netherlands Twin Register. Genes Brain Behav 4:466–481

Boomsma DI, De Geus EJC, Vink JM, Stubbe JH, Distel MA, Hottenga JJ, Posthuma D, Van Beijsterveldt CEM, Hudziak JJ, Bartels M, Willemsen AHM (2006a) Netherlands Twin Register: from twins to twin families. Twin Res Hum Genet 9:849–857

Boomsma DI, Rebollo I, Derks EM, Van Beijsterveldt CEM, Althoff RR, Rettew DC, Hudziak JJ (2006b) Longitudinal stability of the CBCL-juvenile bipolar disorder phenotype: a study in Dutch twins. Biol Psychiatry 60:912–920

Campbell SB, Spieker S, Burchinal M, Poe MD (2006) Trajectories of aggression from toddlerhood to age 9 predict academic and social functioning through age 12. J Child Psychol Psychiatry 47:791–800

Caspi A, Moffitt TE, Newman DL, Silva PA (1996) Behavioral observations at age 3 years predict adult psychiatric disorders. Longitudinal evidence from a birth cohort. Arch Gen Psychiatry 53:1033–1039

Crijnen AA, Achenbach TM, Verhulst FC (1999) Problems reported by parents of children in multiple cultures: the Child Behavior Checklist syndrome constructs. Am J Psychiatry 156:569–574

Derks EM, Dolan CV, Boomsma DI (2004a) Effects of censoring on parameter estimates and power in genetic modeling. Twin Res 7:659–669

Derks EM, Hudziak JJ, Van Beijsterveldt CEM, Dolan CV, Boomsma DI (2004b) A study of genetic and environmental influences on maternal and paternal CBCL syndrome scores in a large sample of 3-year-old Dutch twins. Behav Genet 34:571–583

DiLalla LF, Caraway RA (2004) Behavioral inhibition as a function of relationship in preschool twins and siblings. Twin Res 7:449–455

Edelbrock C, Rende R, Plomin R, Thompson LA (1995) A twin study of competence and problem behavior in childhood and early adolescence. J Child Psychol Psychiatry 36:775–785

Eley TC, Bolton D, O'Connor TG, Perrin S, Smith P, Plomin R (2003) A twin study of anxiety-related behaviours in pre-school children. J Child Psychol Psychiatry 44:945–960

Fergusson DM (1998) Stability and change in externalising behaviours. Eur Arch Psychiatry Clin Neurosci 248:4–13

Fergusson DM, Horwood LJ, Ridder EM (2005) Show me the child at seven: the consequences of conduct problems in childhood for psychosocial functioning in adulthood. J Child Psychol Psychiatry 46:837–849

Fombonne E, Wostear G, Cooper V, Harrington R, Rutter M (2001) The Maudsley long-term follow-up of child and adolescent depression. 1. Psychiatric outcomes in adulthood. Br J Psychiatry 179:210–217

Goodwin RD, Fergusson DM, Horwood LJ (2004) Early anxious/withdrawn behaviours predict later internalising disorders. J Child Psychol Psychiatry 45:874–883

Haberstick BC, Schmitz S, Young SE, Hewitt JK (2005) Contributions of genes and environments to stability and change in externalizing and internalizing problems during elementary and middle school. Behav Genet 35:381–396

Haberstick BC, Schmitz S, Young SE, Hewitt JK (2006) Genes and developmental stability of aggressive behavior problems at home and school in a community sample of twins aged 7–12. Behav Genet 36:809–819

Hastings PD, Rubin KH (1999) Predicting mothers' beliefs about preschool-aged children's social behavior: evidence for maternal attitudes moderating child effects. Child Dev 70:722–741

Hayward C, Killen JD, Kraemer HC, Taylor CB (1998) Linking self-reported childhood behavioral inhibition to adolescent social phobia. J Am Acad Child Adolesc Psychiatry 37:1308–1316

Hewitt JK, Silberg JL, Neale MC, Eaves LJ, Erickson M (1992) The analysis of parental ratings of children's behavior using LISREL. Behav Genet 22:293–317

Hofstra MB, Van der Ende J, Verhulst FC (2000) Continuity and change of psychopathology from childhood into adulthood: a 14-year follow-up study. J Am Acad Child Adolesc Psychiatry 39:850–858

Hudziak JJ, Van Beijsterveldt CEM, Althoff RR, Stanger C, Rettew DC, Nelson EC, Todd RD, Bartels M, Boomsma DI (2004) Genetic and environmental contributions to the Child Behavior Checklist Obsessive-Compulsive Scale: a cross-cultural twin study. Arch Gen Psychiatry 61:608–616

Kagan J, Reznick JS, Snidman N (1988) Biological bases of childhood shyness. Science 240:167–171

Kim-Cohen J, Caspi A, Moffitt TE, Harrington H, Milne BJ, Poulton R (2003) Prior juvenile diagnoses in adults with mental disorder: developmental follow-back of a prospective-longitudinal cohort. Arch Gen Psychiatry 60:709–717

Koot HM, Van den Oord EJCG, Verhulst FC, Boomsma DI (1997) Behavioral and emotional problems in young preschoolers: cross-cultural testing of the validity of the Child Behavior Checklist/2-3. J Abnorm Child Psychol 25:183–196

Kuo PH, Lin CC, Yang HJ, Soong WT, Chen WJ (2004) A twin study of competence and behavioral/emotional problems among adolescents in taiwan. Behav Genet 34:63–74

Mannuzza S, Klein RG, Moulton JLIII (2003) Persistence of attention-deficit/hyperactivity disorder into adulthood: what have we learned from the prospective follow-up studies? J Atten Disord 7:93–100

Martin NG, Eaves LJ (1977) The genetical analysis of covariance structure. Heredity 38:79–95

Murad SD, Joung IM, Van Lenthe FJ, Bengi-Arslan L, Crijnen AA (2003) Predictors of self-reported problem behaviours in Turkish immigrant and Dutch adolescents in the Netherlands. J Child Psychol Psychiatry 44:412–423

Neale MC, Boker SM, Xie G, Maes HH (2006) Mx: statistical modeling, 7th edn. VCU, Department of Psychiatry, Richmond, VA 23298

Ollendick TH, King NJ (1994) Diagnosis, assessment, and treatment of internalizing problems in children: the role of longitudinal data. J Consult Clin Psychol 62:918–927

Park SY, Belsky J, Putnam S, Crnic K (1997) Infant emotionality, parenting, and 3-year inhibition: exploring stability and lawful discontinuity in a male sample. Dev Psychol 33:218–227

Polderman TJC, Posthuma D, De Sonneville LM, Verhulst FC, Boomsma DI (2006) Genetic analyses of teacher ratings of problem behavior in 5-year-old twins. Twin Res Hum Genet 9:122–130

Prior M, Smart D, Sanson A, Oberklaid F (2000) Does shy-inhibited temperament in childhood lead to anxiety problems in adolescence? J Am Acad Child Adolesc Psychiatry 39:461–468

Pulkkinen L, Vaalamo I, Hietala R, Kaprio J, Rose RJ (2003) Peer reports of adaptive behavior in twins and singletons: is twinship a risk or an advantage? Twin Res 6:106–118

Rietveld MJH, Van der Valk JC, Bongers IL, Stroet TM, Slagboom PE, Boomsma DI (2000) Zygosity diagnosis in young twins by parental report. Twin Res 3:134–141

Rietveld MJH, Hudziak JJ, Bartels M, Van Beijsterveldt CEM, Boomsma DI (2004) Heritability of attention problems in children: longitudinal results from a study of twins, age 3 to 12. J Child Psychol Psychiatry 45:577–588

Rosenbaum JF, Biederman J, Gersten M, Hirshfeld DR, Meminger SR, Herman JB, Kagan J, Reznick JS, Snidman N (1988) Behavioral inhibition in children of parents with panic disorder and agoraphobia. A controlled study. Arch Gen Psychiatry 45:463–470

Rosenbaum JF, Biederman J, Hirshfeld DR, Bolduc EA, Faraone SV, Kagan J, Snidman N, Reznick JS (1991) Further evidence of an association between behavioral inhibition and anxiety disorders: results from a family study of children from a non-clinical sample. J Psychiatr Res 25:49–65

Rosenbaum JF, Biederman J, Bolduc-Murphy EA, Faraone SV, Chaloff J, Hirshfeld DR, Kagan J (1993) Behavioral inhibition in childhood: a risk factor for anxiety disorders. Harv Rev Psychiatry 1:2–16

Rubin KH, Burgess KB, Hastings PD (2002) Stability and social-behavioral consequences of toddlers' inhibited temperament and parenting behaviors. Child Dev 73:483–495

Schmitz S, Fulker DW, Mrazek DA (1995) Problem behavior in early and middle childhood: an initial behavior genetic analysis. J Child Psychol Psychiatry 36:1443–1458

Seiffge-Krenke I, Kollmar F (1998) Discrepancies between mothers' and fathers' perceptions of sons' and daughters' problem behaviour: a longitudinal analysis of parent-adolescent agreement on internalising and externalising problem behaviour. J Child Psychol Psychiatry 39:687–697

Tram JM, Cole DA (2006) A multimethod examination of the stability of depressive symptoms in childhood and adolescence. J Abnorm Psychol 115:674–686

Van Beijsterveldt CEM, Bartels M, Hudziak JJ, Boomsma DI (2003) Causes of stability of aggression from early childhood to adolescence: a longitudinal genetic analysis in Dutch twins. Behav Genet 33:591–605

Van den Oord EJCG, Koot HM, Boomsma DI, Verhulst FC, Orlebeke JF (1995) A twin-singleton comparison of problem behaviour in 2–3-year-olds. J Child Psychol Psychiatry 36:449–458

Van den Oord EJCG, Verhulst FC, Boomsma DI (1996) A genetic study of maternal and paternal ratings of problem behaviors in 3-year-old twins. J Abnorm Psychol 105:349–357

Van der Ende J, Verhulst FC (2005) Informant, gender and age differences in ratings of adolescent problem behaviour. Eur Child Adolesc Psychiatry 14:117–126

Van der Valk JC, Verhulst FC, Neale MC, Boomsma DI (1998) Longitudinal genetic analysis of problem behaviors in biologically related and unrelated adoptees. Behav Genet 28:365–380

Van der Valk JC, Van den Oord EJCG, Verhulst FC, Boomsma DI (2003a) Genetic and environmental contributions to stability and change in children's internalizing and externalizing problems. J Am Acad Child Adolesc Psychiatry 42:1212–1220

Van der Valk JC, Van den Oord EJCG, Verhulst FC, Boomsma DI (2003b) Using shared and unique parental views to study the etiology of 7-year-old twins' internalizing and externalizing problems. Behav Genet 33:409–420

Van Dijk BA, Boomsma DI, de Man AJ (1996) Blood group chimerism in human multiple births is not rare. Am J Med Genet 61:264–268

Van Grootheest DS, Bartels M, Cath DC, Beekman AT, Hudziak JJ, Boomsma DI (2007) Genetic and environmental contributions underlying stability in childhood obsessive-compulsive behavior. Biol Psychiatry 61:308–315

Verhulst FC, Van der Ende J, Koot HM (1996) Handleiding voor de CBCL/4-18 [Dutch manual for the CBCL/4-18]. Academic Medical Centre Rotterdam/Erasmus University, Sophia Children's Hospital, Department of Child Psychiatry, Rotterdam, The Netherlands