Psychological Methods
2007, Vol. 12, No. 4, 451–466

1082-989X/07/$12.00 DOI: 10.1037/1082-989X.12.4.451

Twins and the Study of Rater (Dis)agreement

Meike Bartels and Dorret I. Boomsma
VU University

James J. Hudziak
University of Vermont College of Medicine

Toos C. E. M. van Beijsterveldt
VU University

Edwin J. C. G. van den Oord
Virginia Commonwealth University School of Medicine

Genetically informative data can be used to address fundamental questions concerning the measurement of behavior in children. The authors illustrate this with longitudinal multiple-rater data on internalizing problems in twins. Valid information on the behavior of a child is obtained for behavior that multiple raters agree upon and for rater-specific perception of the child's behavior. Rater-disagreement variance σ^2(rd) accounted for 35% of the individual differences in internalizing behavior. Up to 17% of this σ^2(rd) was accounted for by rater-specific additive genetic variance $\sigma^2(A_u)$. Thus, the disagreement should not be considered only to be bias/error but also as representing the unique feature of the relationships between that parent and the child. The longitudinal extension of this model helps to make a distinction between measurement error and the raters' unique perception of the child's behavior. For internalizing behavior, the results show large stability across time, which is accounted for by common additive genetic and common shared environmental factors. Rater-specific shared environmental factors show substantial influence on stability. This could mean that rater bias may be persistent and affect longitudinal studies.

Keywords: twin studies, rater bias, developmental psychology, structural equation modeling

Supplemental materials: http://dx.doi.org/10.1037/1082-989X.12.4.451

Genetically informative designs can be used to estimate the size of genetic and environmental effects on variation in behavior and other complex traits. It is less well known that these designs also have the potential to shed a unique light on fundamental measurement problems and could be an important addition to the traditional methodological arsenal for studying psychological data. This is illustrated in this article by studying sources of rater (dis)agreement in longitudinal data on internalizing problems in monozygotic (MZ) and dizygotic (DZ) twins rated at ages 3, 7, 10, and 12 years by their parents.

Rater (Dis)agreement

Self- and observer ratings of behavior are important sources of information in psychology and psychiatry. It is

Meike Bartels, Dorret I. Boomsma, and Toos C. E. M. van Beijsterveldt, Department of Biological Psychology, VU University, Amsterdam, the Netherlands; James J. Hudziak, Departments of Psychiatry and Medicine (Division of Human Genetics) and Center for Children, Youth and Families, University of Vermont College of Medicine; Edwin J. C. G. van den Oord, Virginia Institute of Psychiatric and Behavioral Genetics, Virginia Commonwealth University School of Medicine.

Edwin J. C. G. van den Oord is now at the Center for Biomarker Research and Personalized Medicine, Department of Pharmacy, Virginia Commonwealth University.

Financial support was given by the Netherlands Organization for Scientific Research (Grants NWO 575-25-012 and NWO/SPI 56-464-14192) and the National Institute of Mental Health (Grant RO1, MH58799-03). Meike Bartels is financially supported by the Netherlands Organization for Scientific Research (Grant VENI: 451-04-034). The Netherlands Organization for Scientific Research (Grant NWO: R 56-467) and the Stichting Simonsfonds (Grant SF053-iz) provided travel grants to facilitate collaboration with Edwin J. C. G. van den Oord at the Virginia Institute of Psychiatric and Behavioral Genetics, Virginia Commonwealth University School of Medicine. Special thanks to Conor Dolan and Reinoud Stoel for their helpful comments regarding the manuscript. We thank Rebecca Ortiz for the thorough editing of the manuscript.

Correspondence concerning this article should be addressed to Meike Bartels, Department of Biological Psychology, VU University, Van der Boechorststraat 1, 1081 BT, Amsterdam, the Netherlands. E-mail: m.bartels@psy.vu.nl

therefore important to understand and estimate the magnitude of potential sources of error in these ratings. Errors are usually distinguished into random and systematic components. Random errors can result from a variety of factors, such as misreading a question or fluctuations in the rater's psychological or emotional state. An important source of systematic rater errors is rater bias that occurs when raters consistently over- or underestimate behavioral scores (Judd, Smith, & Kidder, 1991).

Several models and methods have been proposed for studying rater bias, such as the weighted-average model (Kenny, 1991), the realistic accuracy model (Funder, 1995), or generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). Hoyt (2000) showed how the total variance of the scores assigned by a rater to a target on attribute P, $\sigma^2(P)$, can be decomposed as:

$$\sigma^2(P) = \sigma^2(t) + \sigma^2(r) + \sigma^2(d) + \sigma^2(\epsilon), \quad (1)$$

where $\sigma^2(t)$ is the target variance, $\sigma^2(r)$ is the rater variance, $\sigma^2(d)$ is the dyadic variance, and $\sigma^2(\epsilon)$ is the variance of the error term (residual variance). Target variance, $\sigma^2(t)$, is the variance of the deviations of the target's mean rating (averaging over all observers) from the grand mean. It reflects the score of the target on the trait of interest that is shared by all the raters. This variance is sometimes called universe variance in generalizability theory. Rater variance, $\sigma^2(r)$, is the variance of the deviations of a rater's mean rating (averaging over all targets) from the grand mean of all raters for that target. It reflects how a rater generally perceives targets on that trait or the tendency to be somewhat less/more critical than the average rater, that is, the unique view of a rater, or rater bias. Dyadic variance, $\sigma^2(d)$, is variance attributable to raters' unique perceptions of a specific target. To estimate this component, multiple ratings (e.g., items, forms of the rating scale, or occasions) should be available for each rater–target pair. A dyadic effect is present when the rater rates the target either higher or lower than one would predict on the basis of the rater bias and effect of the target. Finally, $\sigma^2(\epsilon)$ is the variance of the error term (residual effects).

The impact of rater bias can be substantial. Hoyt and Kerns (1999), for example, estimated that as much as 37% of observer ratings in psychological research may be attributed to rater bias. Bias will increase for attributes requiring rater inference (e.g., global ratings of achievement or personality traits) and decrease with the amount of training of the raters (Hoyt & Kerns, 1999). Rater bias reduces the reliability and validity with which a target construct is measured. However, its effect can be complex. For example, response styles when completing multiple questionnaire items will result in overestimates of the internal consistency as measured by Cronbach's alpha and underestimates of correlations between a scale and a criterion variable measured in another way.

The data that are collected determine which variance components can be estimated. As indicated, when only one observation is available for each rater–target pair, the dyadic variance, $\sigma^2(d)$, cannot be distinguished from the error variance, $\sigma^2(\epsilon)$, and the sum of these variances has to be estimated instead. Furthermore, the interpretation of the estimated variance components is based on several assumptions. For example, target variance, $\sigma^2(t)$, defined above, reflects the score of the target on the trait of interest as shared by all the raters. However, raters may observe targets in distinct situations or be exposed to distinct samples of the targets' behavior (e.g., teacher and parent ratings of children's behavior may differ as the teacher observes the child mainly at school and the parent observes the child mainly at home). If, as is usually done, the rater variance, $\sigma^2(r)$, is estimated as the variance of the deviations of a rater's mean rating (averaging over all targets) from the grand mean of all raters for that target, this means that the rater variance, $\sigma^2(r)$, includes not only rater effect but also target variance arising from the rater being exposed to unique samples of the target's attributes. Thus, the interpretation of the estimated variance components may be complex.

Multiple Raters and Multiple Targets

Genetically informative data can improve the interpretation of estimated variance components in rater-bias models. This is illustrated in this article by studying sources of rater (dis)agreement in longitudinal data on internalizing problems in MZ and DZ twins rated at ages 3, 7, 10, and 12 years by their parents.

Single Rater and Multiple Genetically Related Targets: The Classical Twin Design

When a study is expanded from a single target per rater to multiple genetically related targets per rater, which is, for example, the case in a twin study in which a mother rates the behavior of the two children of a twin pair, variance can be decomposed in genetic and environmental parts, and more insight in possible sources of variance can be obtained.

MZ twins derive from a single fertilized zygote and are (nearly always) genetically identical. Less than perfect MZ twin correlations ($r_{MZ} < 1$) therefore indicate environmental effects that are not shared between children growing up in the same family. Possible examples of such nonshared environmental influences are illnesses, accidents, and different peer groups. DZ twins develop from two zygotes and, like ordinary siblings, share on average 50% of their genes. A higher resemblance of MZ versus DZ twin pairs ($r_{MZ} > r_{DZ}$) typically reflects this higher genetic similarity and indicates genetic effects. The design also allows the estima-

tion of environmental influences common to or shared by twins growing up in the same family. Possible examples of factors that can make twin pairs from the same family more alike are socioeconomic status (SES) level, subculture, and style of parenting. These shared environmental influences are implied if the resemblance between twin pairs exceeds the resemblance expected on the basis of quantitative genetic theory.

Figure 1 summarizes the standard path diagram used to represent two measured variables, or phenotypes in twins (squares). The phenotypes are influenced by the twins' genotypes (A), their shared environment (C), and their nonshared environment (E). These factors are unobserved (latent) and symbolized by circles. The latent genotype and the environmental factors usually are scaled to have unit variance. Their influence on the phenotype is given by path coefficients *a, c,* and *e*. The phenotypic variance, $\sigma^2(P)$, of a trait can, in absence of genotype–environment correlation and interaction, be decomposed into additive genetic, shared environmental, and nonshared environmental variance components, as in Equation 1:

$$\sigma^2(P) = a^2 + c^2 + e^2. \quad (2)$$

If latent factors have unit variance, a^2 is the estimate of the additive genetic variance, c^2 is the estimate of the shared environmental variance, and e^2 is the estimate of the nonshared environmental variance; δ is 1.0 for both MZ and DZ twins, representing the total overlap in shared environment; γ is 1.0 for MZ twins, representing their perfect genetic overlap; and γ is 0.5 for DZ twins because they share 50% of the segregating genes on average.

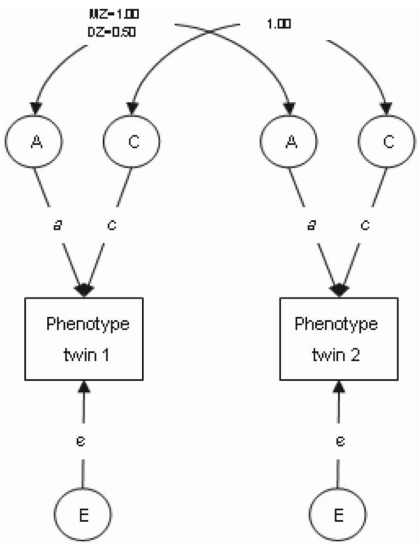

Figure 1. The univariate twin model: Squares represent measured variables, and circles represent latent, unobserved factors. A represents genotypes, C is common environment, and E is unique environment. Their influence on the phenotype is given by path coefficients *a, c,* and *e*. The correlation between the latent genetic factors is 1.00 for monozygotic (MZ) twins and 0.50 for dizygotic (DZ) twins. The correlation between the latent shared environmental factors is fixed to 1.00.

Multiple-Raters Twin Design

Hewitt, Silberg, Neale, Eaves, and Erickson (1992) proposed two models that combine data of multiple raters of twins' behavior. In general, the so-called psychometric model best describes twin data obtained from multiple raters (e.g., Bartels et al., 2003, 2004; van der Valk, van den Oord, Verhulst, & Boomsma, 2001, 2003).

The psychometric model for mother and father ratings of behavior in twins is shown in Figure 2. The latent (true), but not directly observable, phenotypes for Twin 1 and Twin 2 influence the ratings of the parents. The variance of these phenotypes can be decomposed into additive genetic (A), shared environmental (C), and nonshared environmental variance (E). The ratings of the parents are influenced by the genotype of the children that is expressed only in the presence of father or mother (A_m and A_f; from here on, a distinction between paternal and maternal variance is made is made by the subscript $_m$ or $_f$ to denote mother or father rating, respectively). Likewise, shared environmental effects (C_f and C_m) and nonshared environmental effects (E_f and E_m) contribute to the maternal and paternal ratings. The total variance in mother and father ratings thus can be written as

$$\sigma^2(MRT_1) = (a^2 + c^2 + e^2) + (a^2{}_m + c^2{}_m + e^2{}_m),$$

and

$$\sigma^2(FRT_1) = (a^2 + c^2 + e^2) + (a^2{}_f + c^2{}_f + e^2{}_f), \quad (3)$$

where MRT1 and FRT1 refer to the maternal and paternal ratings of Twin 1, respectively. The possibility of estimating and testing rater-specific component ($a^2{}_{m,f}$, $c^2{}_{m,f}$, and $e^2{}_{m,f}$) in genetically informative designs brings along a subtle, but important, conceptual shift. Rather than merely describing rater disagreement along the lines of rater variance, $\sigma^2(r)$; dyadic variance, $\sigma^2(d)$; and error variance, $\sigma^2(\epsilon)$, we can now begin to study mechanisms of rater disagreement along the lines of additive genetic effects, shared environmental effects, and nonshared environmental effects. An important improvement is the ability to test the significance and estimate the magnitude of rater-specific additive genetic variance ($a^2{}_m$). This component reflects real behavior of the child observed by a specific rater independent of any bias. Thus, valid information on the behavior of a target is obtained, and the use of multiple raters provides a more complete picture of the target. Note that c^2 represents real shared environmental influences, whereas $c^2{}_m$ and $c^2{}_f$ could

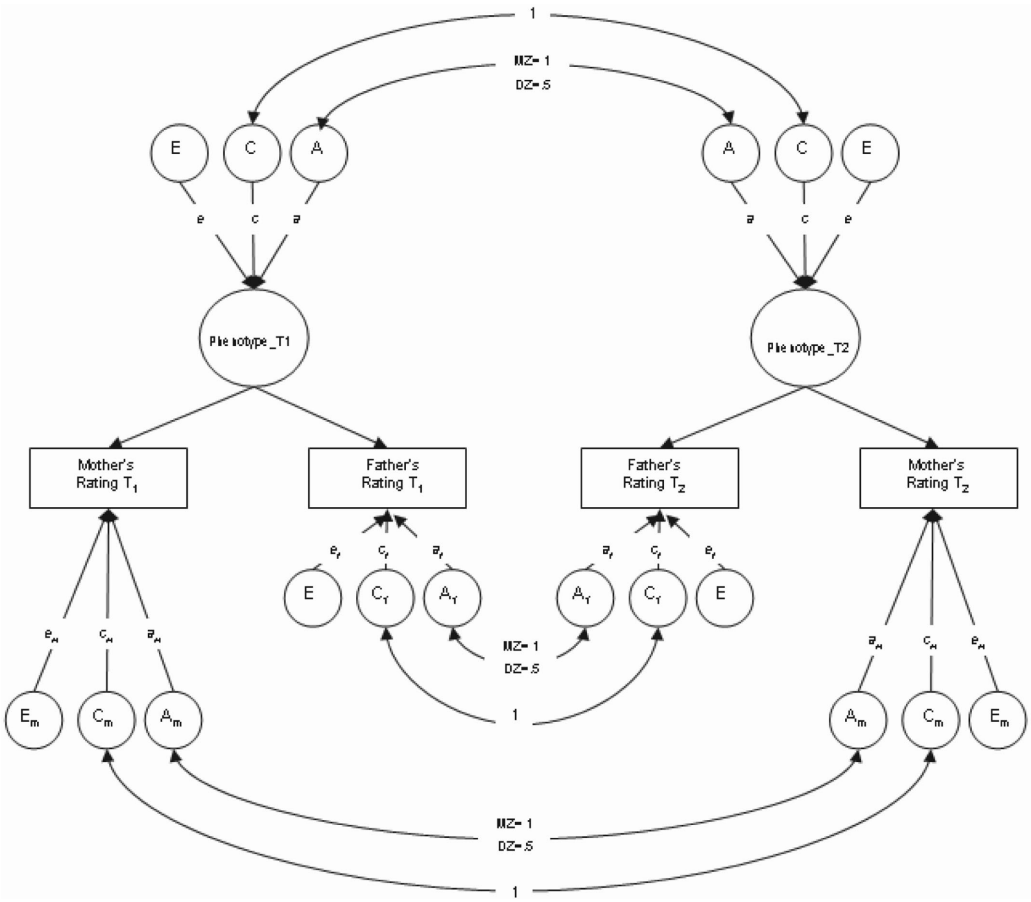

Figure 2. The multiple-rater psychometric model: Squares represent measured variables, and circles represent latent, unobserved factors. A represents genotypes, C is common environment, and E is unique environment. Their influence on the phenotype is given by path coefficients *a, c,* and *e.* Distinction between paternal and maternal variance components is made by subscript $_m$ or $_f$ to denote mother or father ratings. Phenotype_T1 and Phenotype_T2 represent the agreed-upon latent phenotype for the children of a twin pair. The correlation between the latent genetic factors is 1 for monozygotic (MZ) twins and .5 for dizygotic (DZ) twins. The correlation between the latent shared environmental factors is fixed to 1.

be rater bias. Furthermore, e^2 represents real idiosyncratic experience, unaffected by measurement error. The expected variance–covariance matrix for a twin pair rated by mother and father is given in the Appendix.

On the basis of Equation 3, we assume that raters agree and disagree. A parallel can be drawn between rater-agreement variance, $\sigma^2(\text{ra})$ (first part of Equation 3), and target variance, $\sigma^2(\text{t})$, in Equation 1. This variance can be decomposed into genetic, shared, and nonshared environmental effects.

$$\sigma^2(\text{ra}) = \sigma^2(\text{t}) = a^2 + c^2 + e^2. \quad (4)$$

This decomposition enables us to study the relative importance of genetic and environmental effects on common agreed-upon behavior free of bias and unreliability. So, one of the strengths of the proposed multiple-rater design is that a bias/error-free variance decomposition becomes available for the behavior under study.

Furthermore, raters may disagree, and several factors potentially underlying disagreement can be found. Rater-disagreement variance, $\sigma^2(\text{rd})$, is presented in Equation 5. This rater-disagreement variance can now be decomposed into variance related to true behavior of the child and bias and/or error. For example, in the case of maternal rating, we have

$$\sigma^2(\text{rd}) = (a^2{}_m + mc^2{}_m + ne^2{}_m) + (1 - m)c^2{}_m + (1 - n)e^2{}_m, \quad (5)$$

where the first part within parentheses represents the part of

the rater-disagreement variance that reflects real behavior of the target. This part will be nonzero if mothers observe targets in distinct situations or are exposed to distinct samples of the targets' behavior. This real behavior, uniquely assessed by a certain rater, can be influenced by additive genetic, shared environmental, and nonshared environmental factors. So, constants m and n represent the part of the rater-specific shared and nonshared environmental variance that reflects real behavior; the rater-specific additive genetic variance (a^2_m) also reflects real behavior of the child observed by a specific rater independent of any bias. Furthermore, disagreement between raters arises due to bias and error, and the complements $(1 - m)$ and $(1 - n)$ represent this part that is either bias or error. In a nongenetically informative sample, it is not possible to make this distinction, and all variance not shared by raters will be included in the estimate of variance due to rater bias or error.

Sources of bias can be distinguished. According to the definitions of Hoyt (2000), rater variance, $\sigma^2(r)$, reflects how a rater perceives targets on that trait or the tendency to be somewhat less/more critical than the average rater, and dyadic variance, $\sigma^2(d)$, is variance attributable to raters' unique perception of a specific target. Rater variance, $\sigma^2(r)$, is independent of the zygosity of the twin pair but will make two individuals of a twin pair more alike. In a multiple-rater design with genetically related targets, rater variance thus will show up as rater-specific shared environment (c^2_m). However, not all rater-specific shared environment will be due to leniency error, and this distinction is made by constant m in Equation 5. Furthermore, rater disagreement can be due to rater-specific nonshared environmental effects (e^2_m). This can be either nonshared environmental effects on the real behavior of the child (ne^2_m in Equation 5) or variance not shared by targets and not related to the real behavior of the child ($[1 - n]e^2_m$ in Equation 5). As the basic univariate multiple-rater twin design is essentially the scenario where one observation is available for each rater–target pair, it has the same limitation that the dyadic variance, $\sigma^2(d)$, cannot be distinguished from the error variance, $\sigma^2(\epsilon)$. So, in Equation 5,

$$(1 - n)\, e^2_m = \sigma^2(d) + \sigma^2(\epsilon).$$

Practical implications exist in being able to discriminate between sources of variance for rater agreement and rater disagreement. If each rater contributes valid information from his or her own unique perspective, focusing solely on the shared variance gives an incomplete picture of the target. Instead of correcting for rater bias, the rater-specific variance may contain useful information. There is an analogy with the well-known reliability–validity trade-off in observer-rating studies. Relying on a limited behavioral sample with complete overlap among raters enhances interrater agreement but reduces validity.

Although the interpretation of the estimated variance components improves, assumptions still need to be made. When assumptions are not met, the interpretation of the estimated variance components becomes more complex. More specifically, constants n and m cannot be estimated in the basic univariate multiple-rater twin design; we can only estimate the sum: c^2_m and e^2_m. Thus, c^2_m will reflect shared environmental effects on real behavior of the child observed only by mothers plus maternal rater bias effects. In addition, e^2_m will reflect nonshared environmental effects on real behavior of the child observed only by mothers plus error in maternal ratings. To improve this interpretation, the basic univariate design can be extended to multivariate or longitudinal designs (e.g., parents rate different types of behavior or parents rate the behavior of their child at different ages). Because random errors of measurement, captured in $\sigma^2(\epsilon)$, are completely trait or age specific, these will not contribute to the correlations across time. This extension can help to make a finer distinction between dyadic variance, $\sigma^2(d)$, and measurement error, $\sigma^2(\epsilon)$.

Assumptions in Our Model

The interpretation of the biometric variance components explained above holds only under certain assumptions. An assumption of the biometric decomposition is that genetic effects are additive. Behavior is influenced by genetic information at multiple loci on the chromosomes where each locus consists of two or more variants, called alleles. The additive genetic values are simply the sum of the effects of the different alleles at each locus, as well as the sum of the effects across all causal loci. This assumption will be violated if there are interactions between alleles at the same locus (dominance) or interactions between alleles at different loci (epitasis). The consequence of nonadditive genetic effects is an increase in the difference between MZ and DZ correlations. Failure to account for these nonadditive genetic effects will overestimate the total genetic variance and underestimate environmental effects. With only twin data, the full model estimating the nonadditive genetic plus shared environmental variance components is not identified and cannot be estimated. In practice, the shared environmental component is typically estimated by comparing a full model with shared environment with a restricted model with the contribution of shared environment fixed at zero. In cases in which the observed twin correlations are inconsistent with a model assuming only additive genetic effects (e.g., when the DZ twin correlation is less than half the MZ twin correlation), a full model with additive genetic, dominant genetic, and nonshared environment is compared with a model with additive genetic and nonshared environmental effects.

A second assumption concerns the absence of assortative mating between spouses. The presence of assortative mating

may result in genetic similarity between spouses. Positive assortative mating increases the resemblance between DZ twins. MZ twins, however, are already at the point of maximum genetic resemblance, and the correlation between their phenotypes is unaffected by assortative mating (Plomin, DeFries, McClearn, & McGuffin, 2000). As a result, the genetic effects of assortative mating will artificially inflate estimates of the shared environmental influences. This means, in turn, that estimates of the genetic component based primarily on the difference between MZ correlations and DZ correlations will tend to be biased downward. The resolution of the mechanisms of assortment relies on studies that include the spouses or parents of twins (see, e.g., Heath & Eaves, 1985; van Leeuwen, van den Berg, & Boomsma, in press).

Furthermore, absence of gene–environment correlation (*r*GE) and gene–environment interaction (G × E) is assumed. Genotype–environment correlation refers to genetic effects on individual differences in liability to exposure to particular environmental circumstances, that is, it reflects a nonrandom distribution of environments among genotypes. *r*GE adds to the phenotypic variance for a trait, but it is difficult to detect the overall extent to which phenotypic variance is due to the correlation between genetic and environmental effects (Plomin, DeFries, & Loehlin, 1977). Genotype–environment interaction refers to the genetic control of sensitivity or susceptibility to differences in the environment. In other words, different genotypes respond differently to the same environment (Boomsma & Martin, 2002; Eaves, 1984; Falconer & Mackay, 1996; Mather & Jinks, 1977). The contribution of G × E to the overall population variance is typically smaller than the main effects of genotype and environment even in controlled experiments using extreme environments. The interaction between genetic effects and nonshared environment, G × E, will contribute to the total variance but not to the resemblance of twin pairs. In other words, this interaction term will be confounded with nonshared environmental effects (Eaves, Last, Martin, & Jinks, 1977). If, however, G × E represents an interaction between genes and shared environmental influences, G × C, models assuming its absence will result in overestimation of the effect of genes on the phenotype.

A next important assumption is that processes underlying the resemblance between twin pairs are similar in MZ and DZ twins. This assumption is important for the target and rater components. At the target level, this is known as the equal environment assumption, stating that the influence of the environments on MZ and DZ twins is similar. This assumption could, for instance, be violated if MZ twins are treated more alike by other people than DZ twins and if this treatment influences the trait under study. A larger resemblance of MZ twins then would have partly environmental causes. Failure to take this into account would result in an overestimate of the genetic variance, $\sigma^2(A)$, and an underestimate of shared environmental variance, $\sigma^2(C)$. The study of the equal environment assumption is complicated by the fact that individuals may actively shape their environments (Plomin, 1995). For example, MZ twins may spend more time with the same peers than DZ twins because they select similar friends. Instead of a more similar environment inflating the resemblance between MZ twins, part of the more similar environment then reflects the higher genetic similarity of MZ twins and represents true genetic variance. At the rater level, there may be an expectation that MZ twins are more similar than DZ twins. This could inflate the parental ratings of their twins' resemblance and overestimate genetic effects in the rater-specific component $\sigma^2(A_u)$. By comparing twin correlations in correctly classified and misclassified twins, this assumption can be studied. If twin correlations differ for correctly (e.g., as validated by genotyping multiple genetic markers) and incorrectly classified twins, this may reflect a rater effect. Cohen, Dibble, and Grawe (1977) and Scarr and Carter-Salzman (1979) found no differences between a group of misclassified MZ twins and correctly classified MZ twins for mother and father ratings of personality characteristics and extraversion/self esteem, respectively. In contrast, Goodman and Stevenson (1989) found, for a hyperactivity scale, somewhat lower twin correlations for mother, father, and teacher ratings of MZ pairs mistakenly thought by their parents to be DZ.

Furthermore, zygosity-specific or zygosity-independent rater bias is assumed not to be present. When parents are asked to assess their children's phenotype, they may compare the twins' behavior. The behavior of one twin then becomes the standard against which the behavior of the cotwin is rated. Parents may stress either the similarities or the differences between the children, resulting in an apparent cooperation or competition effect. This so-called rater contrast may be hard to distinguish from sibling interaction (Carey, 1986; Eaves et al., 2000; Neale & Stevenson, 1989; Saudino & Eaton, 1991; Simonoff et al., 1998). Both, however, will result in a difference in trait variance of MZ and DZ twins. Evidence for rater contrast/sibling interaction is found in some studies (for a review, see Garcia, Shaw, Winslow, & Yaggi, 2000).

Finally, it is assumed that rater bias is uncorrelated for the distinct raters. On the basis of this assumption, rater bias is part of the rater-disagreement variance (see Equation 5), resulting in bias-free estimates of additive genetic, shared environmental, and nonshared environmental influences on the target variance. If, however, this assumption is violated and bias is shared between raters, this bias will be zygosity independent and thus end up being part of the shared environmental variance of the behavior on which raters agree (C_{common}). Correlated bias could arise if, for example, parents share views on normative standards.

An Application to Internalizing Problems

We use structural equation modeling for longitudinal and genetically informative data obtained from multiple raters to illustrate our model and its results. For the analyses with longitudinal data, we use a Cholesky or triangular decomposition (see Figure 3). The Cholesky decomposition is descriptive and not driven by a specific developmental hypothesis. It decomposes a covariance matrix into genetic and nongenetic covariance matrices and is a first approach to obtaining genetic and environmental correlations across time in longitudinal data sets. Combining the psychometric rater model and the Cholesky decomposition gives a path diagram as depicted in Figure 4, considering one member of a twin pair and the additive genetic part of the model solely, and is explained in detail in the Appendix.

Subjects and Measures

Longitudinal survey data were collected in a large sample of Dutch twin pairs. All participants are registered by the Netherlands Twin Registry (NTR), kept by the Department of Biological Psychology at the VU University in Amsterdam, the Netherlands (Bartels et al., 2007; Boomsma et al., 2006). For this study, data from twins from birth cohorts 1986–1993 were used. Behavioral checklists were collected longitudinally at ages 3, 7, 10, and 12 years. Mother and father ratings were collected by age-appropriate Child Behavior Checklists (CBCL/2-3, Achenbach, 1992; CBCL/4-18, Achenbach, 1991). The CBCL is a standardized questionnaire for parents to report on the frequency of problem behavior as exhibited by the child during the last 6 months. Two broadband groupings, called internalizing (INT) and externalizing behavior, can be formed. In this article, we analyze INT, which reflects withdrawn behavior and anxious/depressed behavior. As in every longitudinal project, changes due to change of measurement instrument cannot be distinguished from biological changes. However, in our study, we used age-appropriate questionnaires from the same taxonomy (the Achenbach System of Empirically Based Assessment) and the broadband scale INT, which minimizes changes due to instrument change. Details on the CBCL and the construction of the INT scale can be found elsewhere (Achenbach, 1991, 1992; Koot, van den Oord, Verhulst, & Boomsma, 1997). Mother and father ratings were available for 3,207 twin pairs at age 3 years, for 3,859 twin pairs at age 7 years, for 2,196 twin pairs at age 10 years, and for 1,105 twin pairs at age 12 years. Because of funding constraints, the CBCL was sent only to the mother of 3-year-old twins born between May and November 1989, resulting in lower numbers of both mother and father ratings at age 3 compared with age 7. For 2,395 twin pairs, only maternal ratings were available at age 3. Furthermore, only maternal ratings were available for 1,256 twin pairs at age 7, for 760 twin pairs at age 10, and for 376 twin pairs at age 12. For a small number of twin pairs, only father ratings were available (182 pairs at age 3, 136 pairs at age 7, 62 pairs at age 10, and 50 pairs at age 12).

In this longitudinal sample, no differences in mean levels of INT were found when comparing twins who participated at all ages, twins who dropped out after age 3, and twins who dropped out after age 3 but returned to the study at later ages (Bartels et al., 2007). However, a significant association between (non)response and level of SES was found (Derks, 2006). At ages 7 and 10, the level of SES was higher in families who returned the questionnaire than in families who did not return the questionnaire. In contrast, no difference was found for the level of SES at age 3. Although significant differences in the level of problem behaviors were found between responders and nonresponders, the effect sizes were all near zero. This implies that the differences between responders and nonresponders are statisti-

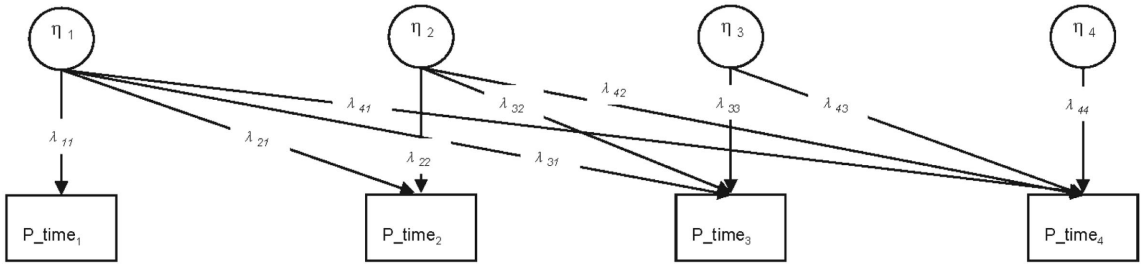

Figure 3. The Cholesky decomposition: Squares represent measured variables. In this figure, the observed variables P_time_1–P_time_4 represent behavior assessed at the four different time points (in this study, ages 3, 7, 10, and 12 years). Circles represent latent, unobserved factors (η_1–η_4), which could be replaced by genetic and/or environmental factors. Their influence on the phenotype is given by path coefficients λ_{11}–λ_{44}. λ_{11} represents the influence of the first latent factor on the observed variable at the first measurement occasion, λ_{21} represents the influence of the first latent factor on the observed variable at the second measurement occasion, and so on.

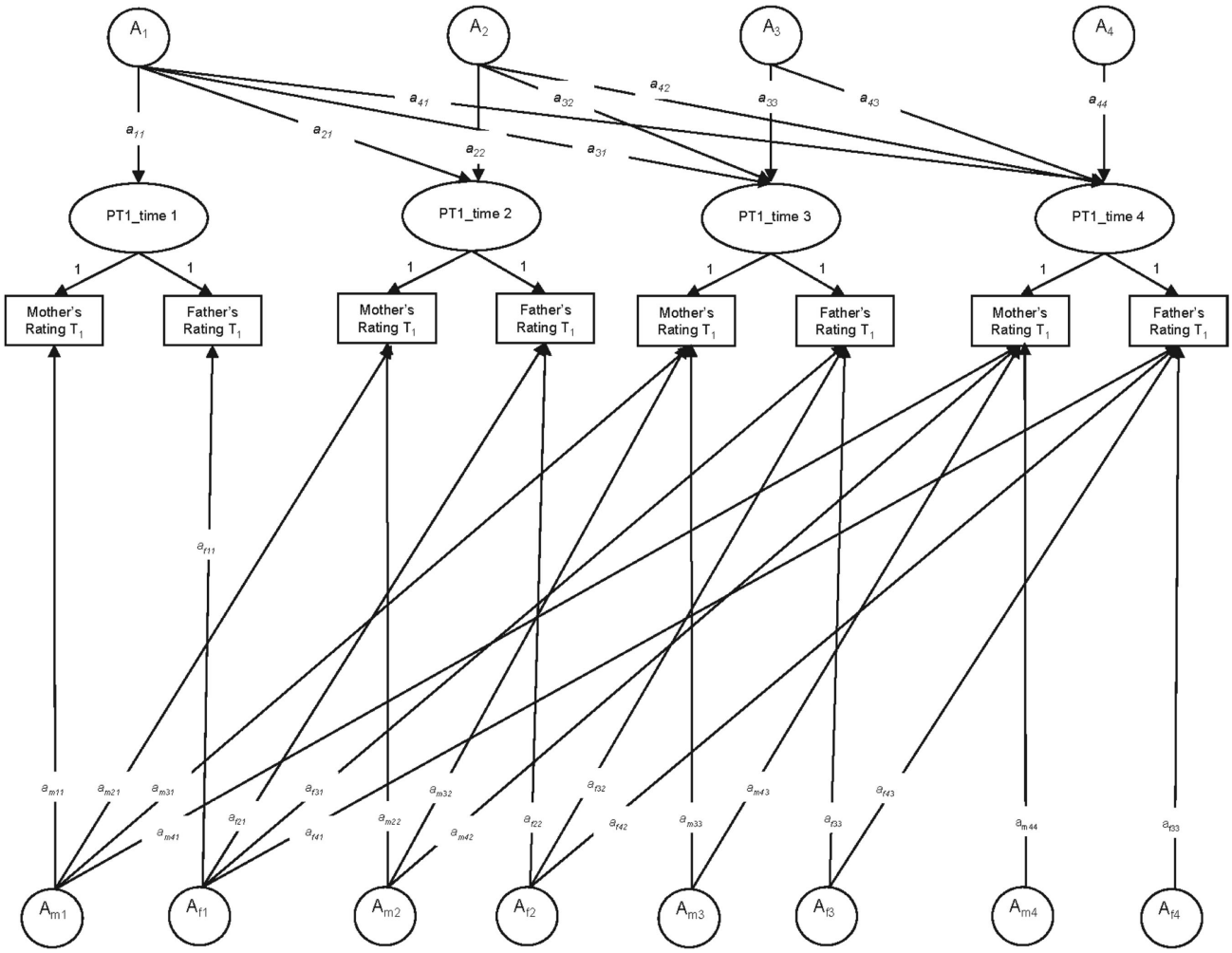

Figure 4. The combined multiple-rater longitudinal model for one member of a twin pair, with a Cholesky decomposition for the additive genetic influences on the reliable trait variance and the parental unique additive genetic influences. All other variance components can be expressed in this way but are left out the figure for sake of simplicity. Squares represent measured/observed variables. In this figure, the observed variables represent mother and father ratings for the older of the twin pair (T_1) at the first measurement occasion (time 1) to the fourth measurement occasion (time 4). PT1_time 1 represents the overlap in parental ratings (P) for the older of the twin pair (T1) at the first measurement occasion (time 1; age 3 years), PT_1_time 2 represents the overlap in parental ratings for the older of the twin pair (T1) at the second measurement occasion (time 2; age 7), and so on. A_1 to A_4 represent genotypes at measurement occasions 1 to 4. Their influence on the phenotype is given by path coefficients a_{11}–a_{44}. a_{11} represents the influence of the first latent factor on the observed variable at the first measurement occasion, a_{21} represents the influence of the first latent factor on the observed variable at the second measurement occasion, and so on. Distinction between paternal and maternal variance components is made by subscript $_m$ or $_f$ to denote mother or father ratings.

cally significant, probably as a result of the large sample sizes, but are not practically significant.

Zygosity was determined for 1,249 same-sex twin pairs based on DNA or blood group polymorphisms. For all other same-sex twin pairs, zygosity was determined by discriminant analysis, using questionnaire items at each age separately. Agreement between zygosity assignment by the replies to the questions and zygosity assignment by DNA markers/blood typing is around 93% (for details, see Rietveld et al., 2000).

Genetic Modeling

Full-information maximum-likelihood analysis of raw data (so that all available data are used; i.e., also twin pairs

with only maternal ratings) was used to obtain parameter estimates. The procedure follows the theory described by Lange, Westlake, and Spence (1976). The package Mx (Neale, Boker, Xie, & Maes, 2003) was used to estimate genetic and environmental variance components (to download Mx software, go to http://www.vcu.edu/mx; to obtain Mx scripts, see the Mx-scripts library at http://www.psy-.vu.nl/mxbib; for the Mx script used in this article, see the doi.org Web site URL at the head of the article). A Cholesky decomposition was specified for all common and rater-specific genetic and environmental components. It is possible that the processes underlying parents' judgments of INTs may differ for boys and girls. Furthermore, the genetic architecture of the agreed-upon phenotype may differ for boys and girls. Consequently, analyses were based on a five-group design so as to be able to detect sex differences in the variance components. In this design, the five groups are MZ males, DZ males, MZ females, DZ females, and twins of opposite sex. The significance of each variance component (rater-specific and common) was tested by constraining them at zero in a submodel, which was compared with the full model in which the component was freely estimated. For instance, it was investigated whether disagreement between parents were at least partly due to rater bias or whether rater-specific views were involved. To make this distinction, the significance of the rater-specific additive genetic effects (A_m or A_f) were tested. If these rater-specific genetic effects were significant, systematic effects must be present, which would not be expected when differences in parental ratings were caused only by rater bias and unreliability, which are independent of zygosity of the children. The significance of other influences, for example, genetic and shared environmental factors on the agreed-upon part of the behavior, was also tested. The only factor that was never dropped from the model was the rater-specific nonshared environmental factor ($E_{m,f}$) because measurement errors are included in this factor. Sex differences in variance components were tested by constraining these to be equal for boys and girls.

Submodels were compared by hierarchical chi-square tests. The chi-square statistic is computed by subtracting $-2LL$ (log likelihood) for a reduced model from that for the full model: $\chi^2 = -2LL_0 - (-2LL_1)$. Given that the full model is correct, this statistic is chi-square-distributed with degrees of freedom (*df*s) equal to the difference in the number of parameters estimated in the two models ($\Delta df = df_0 - df_1$).

Results

Genetic Model Fitting: Significance of the Distinct Variance Components

We first tested the significance of each variance component. The saturated model was taken as a reference for evaluating changes in chi-square and associated degrees of freedom of more parsimonious models. All variance components were significant, as indicated by the poorer fit of the reduced models: The chi-square increased dramatically after constraining variance components at zero (for all variance components, $p < .001$). The total observed variance can be decomposed into rater-agreement variance, $\sigma^2(ra)$, which is accounted for by significant additive genetic, shared environmental, and nonshared environmental effects, and into rater-disagreement variance, $\sigma^2(rd)$.

Significance of the rater-specific additive genetic effects (a^2_m and a^2_f) indicated that rater disagreement was not solely due to bias or error but that each parent provided specific and reliable information on the behavior of his or her child. Furthermore, significant rater-specific shared and nonshared environmental influences were present, partly representing bias and/or error. Finally, sex differences in the magnitude of the variance components were found, $\chi^2(90) = 131.044$, $p < .05$.

Variance and Covariance Decomposition

The percentages of the total age-specific variance (bolded cells) and the total between-age covariances (off diagonal) decomposed into common (rater-agreement) and rater-specific (rater-disagreement) additive genetic, shared environmental, and nonshared environmental factors for boys (below diagonal) and girls (above diagonal) based on the best fitting model are presented in Table 1. Common factors ($A_{common} + C_{common} + E_{common}$) are more important than rater-specific factors ($A_{unique} + C_{unique} + E_{unique}$). The ratio of the common factor variance to the total variance could be treated as an index of interrater reliability. For example, 70% of the total variance in INT at age 3 based on maternal ratings is agreed upon by both raters. The remaining variance (about 30%) is rater-disagreement variance. Common additive genetic factors are most important in explaining individual differences in INT at ages 3, 7, 10, and 12. However, a decrease in common additive genetic influences is found over the years. A complementary increase in common shared environmental influences is found.

Rater-specific variance components are significant as well. Between 1% to 17% of the total variance in INT (the observed behavior) is accounted for by rater-unique views of mothers and fathers (A_m, A_f). Differences in the magnitude of variance components is based on the rater (mother vs. father), age of the target (age 3, 7, 10, or 12), and gender of the target (boy or girl). Asking both mothers and fathers to rate problem behavior in children does add additional information on the child's behavior. The use of this design shows that 1% to 17% of rater-disagreement variance is reflecting real behavior. In a nongenetically informative sample, this part of the rater-disagreement variance would be labeled as bias or error.

Table 1
Proportions of the Total Genetic and Environmental Variances (Diagonal; Bolded Cells) and Covariances (Off Diagonal) for Internalizing Problem Behavior Based on the Best Fitting Model for Boys (Below Diagonal) and Girls (Above Diagonal)

	Internalizing							
	Mother				Father			
	3	7	10	12	3	7	10	12
A_{common}								
3	**.51/.45**	.43	.43	.41	**.53/.46**	.47	.48	.44
7	.56	**.36/.29**	.35	.27	.71	**.38/.31**	.42	.32
10	.50	.44	**.28/.24**	.28	.70	.53	**.32/.27**	.34
12	.49	.39	.35	**.28/.22**	.66	.44	.39	**.30/.23**
C_{common}								
3	**.04/.09**	.31	.33	.43	**.04/.09**	.34	.37	.46
7	.17	**.11/.16**	.22	.33	.21	**.11/.17**	.27	.39
10	.23	.20	**.18/.19**	.27	.33	.23	**.21/.22**	.33
12	.26	.26	.25	**.18/.23**	.36	.29	.28	**.20/.25**
E_{common}								
3	**.15/.13**	.03	.01	−.03	**.15/.14**	.04	.01	−.03
7	.03	**.15/.16**	.12	.11	.04	**.16/.16**	.15	.13
10	.00	.10	**.16/.14**	.15	.00	.12	**.18/.16**	.18
12	−.01	.12	.16	**.14/.14**	−.01	.13	.18	**.16/.15**
A_{unique}								
3	**.09/.14**	.04	.00	−.12	**.11/.01**	−.02	−.03	−.04
7	.03	**.13/.11**	.06	.00	.06	**.10/.11**	−.06	−.02
10	.04	.13	**.12/.10**	.06	−.07	−.04	**.05/.07**	.02
12	−.11	.12	.10	**.14/.06**	−.07	−.02	.01	**.17/.13**
C_{unique}								
3	**.08/.04**	.17	.22	.26	**.05/.19**	.14	.22	.12
7	.20	**.13/.16**	.20	.19	.01	**.12/.15**	.22	.17
10	.27	.13	**.12/.17**	.16	.12	.12	**.13/.19**	.16
12	.31	.10	.10	**.16/.21**	.13	.13	.12	**.07/.13**
E_{unique}								
3	**.13/.15**	.02	.01	.05	**.12/.11**	.03	−.05	.05
7	.01	**.12/.12**	.05	.10	−.03	**.13/.10**	.00	.01
10	−.04	.00	**.14/.16**	.08	−.08	.04	**.11/.09**	−.03
12	.06	.01	.04	**.10/.14**	−.07	.03	.02	**.10/.11**

Note. In boldfaced cells, the first number is the estimate for boys, and the second number is the estimate for girls. Distinct estimates for the common variance component for mothers and fathers arise due to standardization based on the total observed variance for each rater. A_{common} = additive genetic influence on common agreed-upon variance; C_{common} = shared environmental influence on the common agreed-upon variance; E_{common} = nonshared environmental influence on the common agreed-upon variance; A_{unique} = parental unique genetic influences; C_{unique} = parental unique shared environmental variance; E_{unique} = parental unique nonshared environmental variance.

A salient finding is the significant and rather high influence of rater-specific shared environmental factors (C_m, C_f). This factor can represent two components. First, it can represent real shared environmental influences, uniquely assessed by one of the parents. Second, it can represent rater bias. Rater-specific nonshared environmental influences (E_m, E_f) account for 9% to 16% of the variance at the distinct ages. Measurement error and real nonshared environmental influences are captured in these estimates.

More important in Table 1 are the influences of common and rater-specific genetic and environmental factors on the covariances (off diagonal), representing genetic and environmental influences on the stability of INT throughout childhood. Stability in the behavior similarly assessed by both parents is explained by common additive genetic influences (51% of the covariance on average for boys and 39% for girls) and common shared environmental influences (26% of the covariance on average for boys and 34% for girls). Common nonshared environmental influences seem to be less important for stability in problem behavior, as indicated by very low influences on the covariances. Rater-specific influences are generally less important for stability in INT over the years. The one exception is the mother-specific and the father-specific shared environmental influences. About 19% of the covariance in INT based on the mother ratings is accounted for by these rater-specific shared environmental influences. Therefore, rater bias possibly accounts for a significant part of the stability, although

its account cannot be fully distinguished from valid variance in this model. It can further be observed that the paternal unique additive genetic component (A_f) does not add information on stability of behavior. This has been found for other phenotypes as well, for example, obsessive-compulsive behavior (OCB; van Grootheest et al., 2007), which indicates that it is not likely to be a chance finding. Father ratings do not appear to be the best source of information for studying stability of behavior. Note that the A_f component, being age specific, is not inconsistent with the interpretation of unique perceptions of inherited characteristics. Rather, it indicates that this component is not stable over time and that, apparently, the father perceives somewhat different aspects of the child's behavior at each age. In contrast to the findings for OCB (van Grootheest et al., 2007), maternal unique views (A_m) account for a small to modest percentage of the total covariance. These results indicate that for studying the stability of INT, most of the information comes from the part of the covariance on which both parents agree.

Discussion

In this article, we have illustrated how genetically informative data can be used to address fundamental questions concerning the assessment of behavior and behavior problems in young children. For this purpose, we analyzed longitudinal data on internalizing problem behavior in children as assessed by both parents and collected in a large sample of Dutch twin pairs. The extension of the multiple-rater model to a longitudinal model allowed the decomposition of the longitudinal variance–covariance matrix into components due to common additive genetic, shared environmental, and nonshared environmental influences, as well as components due to rater-specific additive genetic, shared environmental, and nonshared environmental influences. Conditional on the assumptions discussed in the introduction, the common components can be interpreted as reflecting behavior similarly assessed by both parents, that is, the target variance, $\sigma^2(t)$. The rater-specific components reflect disagreement in behavioral assessment by mothers and fathers and may include rater effects, dyadic effects, and residual error effects.

The significant influences of additive genetic factors on the target variance (common agreed-upon behavior of the child) indicate the child's innate vulnerability to childhood psychopathology. The significant influences of common nonshared environmental influences indicate the importance of pure idiosyncratic experiences. The significant influence of common shared environmental factors indicate that environmental factors that are overlapping for the two children of a twin pair, for example, family environment or neighborhood, are of importance as well.

In our example, rater-disagreement variance, $\sigma^2(rd)$, that is, $A_u + C_u + E_u$, accounted on average for 35% of the individual differences in INT, which is in line with the 37% mentioned by Hoyt and Kerns (1999). One percent to 17% (*Mdn* = 10%) of this rater-disagreement variance was accounted for by rater-specific additive genetic variance, $\sigma^2(A_u)$. This suggests that parents assess reliable unique aspects of their child's behavior. Thus, the lack of agreement not only should be considered to be bias but also partly represents the unique feature of the relationships between that parent and his or her child. The remaining part of the rater-disagreement variance consisted on average (across the multiple measurements) of 11% (*Mdn* = 12%) shared environmental variance, $\sigma^2(C_u)$, and 12% (*Mdn* = 12%) nonshared environmental variance, $\sigma^2(E_u)$. With this design, it is not possible to distinguish rater-specific shared environmental variance from rater bias, and we cannot distinguish rater-specific nonshared environmental effects from random measurement error plus dyadic effects. However, the rater-specific shared environment components, $\sigma^2(C_u)$,were about equal to the common shared environmental effects, $\sigma^2(C)$, while the rater-specific additive genetic components, $\sigma^2(A_u)$, were relatively small compared with the common additive genetic effects, $\sigma^2(A)$. An explanation of the different balance for C compared with A could be that a part of $\sigma^2(C_u)$ is rater bias. However, this finding should be interpreted with care because no real distinction between rater bias and real rater-specific shared environmental effect can be made with the proposed design.

Longitudinal data can shed some further light on the nature of rater disagreement. The finding that the contributions of both A_m and A_f to cross-time covariances are near zero and often negative indicates that the rater-specific variance reflects real but only age-specific behavior of the child. Stability of INT is reflected by the influences of common additive genetic effects to the covariance. Furthermore, the finding that rater-specific shared environment component $\sigma^2(C_u)$ contributes at least moderately to stability over time could mean that rater bias may be persistent and affect longitudinal studies. Finally, the absence of stability in the rater-specific nonshared environmental influences (especially for father raters) suggests the presence of random error variance that would not be expected to be stable across time. However, nonshared environmental influences on the target covariance (the covariance of the common agreed-upon behavior) over time were also small, so the later finding should be interpreted with care as these estimates could also reflect real rater-specific nonshared environmental influences.

Our results suggest that two strategies are possible in studying childhood psychopathology given data obtained from raters: (a) Only focus on the behavior on which raters agree, which could be incomplete due to the absence of rater-specific additive genetic variance, representing the unique view of a rater on the target behavior. However, a bias-free estimate of genetic and environmental influences

on the target variance is obtained. (b) Focus on behavior on which raters both agree and disagree, which gives a more complete picture of the child's behavior. However, this strategy has the disadvantages of a possible confounding with leniency/severity effects and measurement error.

We have discussed the unique contributions of the classical twin design when it is combined with information from multiple raters. The model may be applied to other genetically informative designs, such as the full family design, the step/half-sibling design, or the adoption design. The full family design and the adoption design, however, lack information for studying rater (dis)agreement in the detailed sense we have illustrated for the twin design. For instance, within the family design, one can collect data from multiple raters on multiple offspring, but the family design does not permit one to disentangle genetic and shared environmental factors as the underlying source of familial aggregation. For studies on rater disagreement, this distinction is essential to distinguish rater-specific (informative) views, represented by rater-specific additive genetic effects, from (noninformative) rater bias, represented by rater-specific shared environmental effects. Furthermore, in the adoption design, the two (or more) raters do not rate genetically related subjects but, rather, genetically unrelated siblings. Information from biological parents is often lacking in this design, so effects of shared environment can be estimated, whereas information to estimate heritability and rater-specific additive genetic effects in the case of our model is lacking. A design with additional groups of different levels of genetic relatedness, such as, full siblings, half-siblings, and step-siblings, could be as informative as the proposed classical twin design but will be less powerful.

Rater Contrast

To use the proposed model, several assumptions have been presented in the introduction. Although the analysis may offer a more refined picture of the nature of rater differences or developmental processes, reality is likely to be more complex. One example involves a mix of phenomena that have been referred to in the literature as contrast effects, sibling interaction, or the equal environment assumption. For example, the current model assumes that the INT of one twin does not directly affect the other twin's INT. For common childhood psychopathology, claims of cooperation and competition effects have been made (for a review, see Garcia, Shaw, Winslow, & Yaggi, 2000). Sibling interactions may be very hard to distinguish from certain types of rater effects (Eaves et al., 2000; Neale & Stevenson, 1989; Simonoff et al., 1998). When parents are asked to assess their children's phenotype, they may compare the twins' behavior. The behavior of one twin could then become the standard against which the behavior of the co-twin is rated. Parents may stress either the similarities or the differences between the children, resulting in an apparent cooperation or competition effect. Furthermore, genetic nonadditivity, such as allelic interaction effects on the same or different loci, may also produce patterns that resemble those of sibling interaction and rater effects. More specifically, these effects will result in a difference in trait variance of MZ and DZ twins, and with large sample sizes, the variance differences between zygosities provide a good test. Variances and twin correlations for the data used in the current article do not give any indication of the presence of contrast effects or genetic dominance, so we have not taken these effects into account.

We have further assumed that processes are similar for MZ versus DZ twins. Because, in reality, this assumption might not hold and processes might differ for MZ and DZ twins, this could affect the validity of the results. For instance, it has been argued that the similarity of MZ twins might be inflated because they grow up in more similar environments than DZ twins. However, individuals may actively shape their environments so that part of the more similar environment may actually be a reflection of the higher genetic similarity of MZ twins. Also, it has been suggested that raters may expect DZ twins to be dissimilar to a certain extent. Some support for possible zygosity-dependent rater-contrast effects comes from the finding that when more behavioral measures of temperament are used (e.g., actometer readings or behavioral observations), DZ correlations may be higher (Saudino & Eaton, 1991). Borkenau, Riemann, Angleitner, and Spinath (2001) found, for instance, that video-based personality ratings yielded estimates of shared environmental influences about .15 higher than those obtained using self-reports and peer reports. These authors stated that the main source of this difference was the relatively high DZ correlation in video-based personality ratings. However, no tests were performed to examine whether these results were significant nor were other possible explanations evaluated systematically. Furthermore, it is not clear whether behavioral observations assess the same behavior as, for instance, ratings by parents, who may observe their children in more diverse and different situations. In fact, what may appear to be a limitation of the twin method could merely be a reflection of the true complexity of psychological processes. In a sense, this underscores the point made in this article that there is a need to extend the methodological arsenal. Collecting information from multiple raters in large genetically informative samples may be important in this respect. It has the potential to shed light on some of the these issues by systematically evaluating many of the processes that may underlie differences between raters.

Conclusions

Genetically informative designs offer some unique possibilities. This has been illustrated in this article by focusing on rater agreement and disagreement using parental ratings.

The same model could be applied to parent and teacher data. Using teacher data will probably increase the rater-specific components as, by definition, the information available to a teacher and parent will overlap less than the information available to a mother and father. Furthermore, expansion can be in the direction of examples for the study of environmental (Collins, Maccoby, Steinberg, Hetherington, & Bornstein, 2000) and developmental mechanisms (van den Oord & Rowe, 1997). Clearly, genetically informative designs will not solve all problems, and there are inherent limitations as to the complexity of the phenomena that can be modeled. These designs will, however, provide a unique piece of information that cannot be obtained using nongenetically informative samples.

References

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

Achenbach, T. M. (1992). *Manual for the Child Behavior Checklist/2–3 and 1992 profile.* Burlington: University of Vermont, Department of Psychiatry.

Bartels, M., Boomsma, D. I., Rietveld, M. J. H., van Beijsterveldt, C. E. M., Hudziak, J. J., & van den Oord, E. J. C. G. (2004). Disentangling genetic, environmental, and rater effects on internalizing and externalizing problem behavior in 10-year-old twins. *Twin Research, 7,* 162–175.

Bartels, M., Hudziak, J. J., Boomsma, D. I., Rietveld, M. J. H., van Beijsterveldt, C. E. M., & van den Oord, E. J. C. G. (2003). A study of parent ratings of internalizing and externalizing behavior in 12-year-old twins. *Journal of the American Academy of Child and Adolescent Psychiatry, 42,* 1351–1359.

Bartels, M., van Beijsterveldt, C. E. M., Derks, E. M., Stroet, T. M., Polderman, J. C., Hudziak, J. J., et al. (2007). Young Netherlands Twin Register (Y-NTR): A longitudinal multiple informant study of problem behavior. *Twin Research and Human Genetics, 10,* 3–11.

Boomsma, D. I., de Geus, E. J. C., Vink, J. M., Stubbe, J. H., Distel, M. A., Hottenga, J. J., et al. (2006). Netherlands Twin Register: From twins to twin families. *Twin Research and Human Genetics, 9,* 849–857.

Boomsma, D. I., & Martin, N. G. (2002). Gene–environment interactions. In H. D. D'haenen, J. A. den Boer, & P. Willner (Eds.), *Biological psychiatry* (pp. 181–187). Chichester, England: Wiley.

Borkenau, P., Riemann, R., Angleitner, A., & Spinath, F. M. (2001). Genetic and environmental influences on observed personality evidence from the German Observational Study of Adult Twins. *Journal of Personality and Social Psychology, 80,* 655–668.

Carey, G. (1986). Sibling imitation and contrast effects. *Behavior Genetics, 16,* 319–341.

Cohen, D. J., Dibble, E., & Grawe, J. M. (1977). Fathers' and mothers' perceptions of children's personality. *Archives of General Psychiatry, 34,* 480–487.

Collins, A. W., Maccoby, E. E., Steinberg, L., Hetherington, M., & Bornstein, M. H. (2000). Contemporary research on parenting: The case for nature and nurture. *American Psychologist, 3,* 218–233.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Derks, E. M. (2006). *Assessment and genetic aetiology of attention problems, hyperactivity, and related disorders.* Unpublished master's thesis, VU University, Amsterdam, the Netherlands.

Eaves, L. J. (1984). The resolution of Genotype × Environment interaction in segregation analysis of nuclear families. *Genetic Epidemiology, 1,* 215–228.

Eaves, L. J., Last, K. A., Martin, N. G., & Jinks, J. L. (1977). A progressive approach to non-additivity and genotype–environmental covariance in the analysis of human differences. *British Journal of Mathematical and Statistical Psychology, 30,* 1–42.

Eaves, L. J., Rutter, M., Silberg, J. L., Shillady, L., Maes, H., & Pickles, A. (2000). Genetic and environmental causes of covariation in interview assessments of disruptive behavior in child and adolescent twins. *Behavior Genetics, 30,* 321–334.

Falconer, D. S., & Mackay, T. F. C. (1996). *Introduction to quantitative genetics* (4th ed.). Harlow, England: Longman.

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review, 102,* 652–670.

Garcia, M. M., Shaw, D. S., Winslow, E. B., & Yaggi, K. E. (2000). Destructive sibling conflict and the development of conduct problems in young boys. *Developmental Psychology, 36,* 44–53.

Goodman, R., & Stevenson, J. (1989). A twin study of hyperactivity: II. The aetiological role of genes, family relationships and perinatal adversity. *Journal of Child Psychology and Psychiatry, 30,* 691–709.

Heath, A. C., & Eaves, L. J. (1985). Resolving the effects of phenotype and social background on mate selection. *Behavior Genetics, 15,* 15–30.

Hewitt, J. K., Silberg, J. L., Neale, M. C., Eaves, L. J., & Erickson, M. (1992). The analysis of parental ratings of children's behavior using LISREL. *Behavior Genetics, 22,* 293–317.

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

Hoyt, W. T., & Kerns, M.-D. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods, 4,* 403–424.

Judd, C. M., Smith, E. R., & Kidder, L. H. (1991). *Research methods in social relations* (6th ed.). Fort Worth, TX: Holt, Rinehart & Winston.

Kenny, D. A. (1991). A general model of consensus and accuracy in interpersonal perception. *Psychological Review, 98,* 155–163.

Koot, H. M., van den Oord, E. J. C. G., Verhulst, F. C., & Boomsma, D. I. (1997). Behavioral and emotional problems in

young preschoolers: Cross-cultural testing of the validity of the Child Behavior Checklist/2–3. *Journal of Abnormal Child Psychology, 25,* 183–196.

Lange, K., Westlake, J., & Spence, M. A. (1976). Extension to pedigree analysis: III. Variance components by the scoring method. *Annals of Human Genetics, 48,* 47–59.

Mather, K., & Jinks, J. L. (1977). *Introduction to biometrical genetics.* London: Chapman & Hall.

Neale, M. C., Boker, S. M., Xie, G., & Maes, H. H. (2003). *Mx: Statistical Modeling* (6th ed.). Richmond: Virginia Commonwealth University, Department of Psychiatry.

Neale, M. C., & Stevenson, J. (1989). Rater bias in the EASI temperament scales: A twin study. *Journal of Personality and Social Psychology, 56,* 446–455.

Plomin, R. (1995). Genetics and children's experiences in the family. *Journal Of Child Psychology and Psychiatry, 36,* 33–68.

Plomin, R., DeFries, J. C., & Loehlin, J. C. (1977). Genotype–environment interaction and correlation in the analysis of human behavior. *Psychological Bulletin, 84,* 309–322.

Plomin, R., DeFries, J. C., McClearn, G. E., & McGuffin, P. (2000). *Behavioral genetics* (4th ed.). New York: Worth Publishers.

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

Saudino, K. J., & Eaton, W. O. (1991). Infant temperament and genetics: An objective twin study of motor activity level. *Child Development, 62,* 1167–1174.

Scarr, J. C., & Carter-Salzman, L. (1979). Twin method: Defense of a critical assumption. *Behavior Genetics, 9,* 527–542.

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

van den Oord, E. J. C. G., & Rowe, D. C. (1997). Continuity and change in children's social maladjustment: A developmental behavior genetic study. *Developmental Psychology, 33,* 319–332.

van der Valk, J. C., van den Oord, E. J. C. G., Verhulst, F. C., & Boomsma, D. I. (2001). Using parental ratings to study the etiology of 3-year-old twins' problem behaviors: Different views or rater bias? *Journal of Child Psychology and Psychiatry, 42,* 921–931.

van der Valk, J. C., van den Oord, E. J. C. G., Verhulst, F. C., & Boomsma, D. I. (2003). Using common and unique parental views to study the etiology of 7-year-old twins' internalizing and externalizing problems. *Behavior Genetics, 33, 409*–420.

van Grootheest, D. S., Bartels, M., Cath, D. C., Beekman, A. T., Hudziak, J. J., & Boomsma, D. I. (2007). Genetic and environmental contributions underlying stability in childhood obsessive-compulsive behavior. *Biological Psychiatry, 61,* 308–315.

van Leeuwen, M., van den Berg, S. M., & Boomsma, D. I. (in press). A twin-family study of general IQ. *Learning and Individual Differences.*

Appendix

The Expected Variance–Covariance Matrix for a Twin Pair Rated by Mother and Father

To derive the expected covariances between ratings and between twins, we first focus on the ratings of the mother for Twin 1 (MRT_1) and of the father for Twin 1 (FRT_1). On the basis of the rules of path analysis (see Figure 2 in the main text), we write the model as a matrix equation:

$$\begin{bmatrix} MRT_1 \\ FRT_1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \times [[a] \times [A] + [c] \times [C] + [e] \times [E]] + \begin{bmatrix} a_m & 0 \\ 0 & a_f \end{bmatrix} \times \begin{bmatrix} A_m \\ A_f \end{bmatrix} + \begin{bmatrix} c_m & 0 \\ 0 & c_f \end{bmatrix} \times \begin{bmatrix} C_m \\ C_f \end{bmatrix} + \begin{bmatrix} e_m & 0 \\ 0 & e_f \end{bmatrix} \times \begin{bmatrix} E_m \\ E_f \end{bmatrix}, \quad \text{(A1)}$$

where A, C, and E represent the additive genetic, shared environmental, and nonshared environmental latent factors, respectively, and where their influence on the phenotype is given by path coefficients *a, c,* and *e.* Furthermore, a distinction between paternal and maternal factors is made by the subscript $_m$ or $_f$ to denote mother or father rating. The first part of this matrix equation represents influences of A, C, and E on the reliable trait variance (behavior similarly observed by both mother and father). The second part of the matrix equation represents rater disagreement. All genetic and environmental factors are uncorrelated.

To extend the model to a twin pair, we define the (4 × 1) data vector with parental ratings as $\mathbf{y'} = [MRT_1, FRT_1, MRT_2, FRT_2]$. By taking the expectation, we obtain the (4 × 4) expected covariance matrix for the full model in Figure 2:

$$\Sigma_Y = \mathbf{L} \times \begin{bmatrix} \Sigma_A + \Sigma_C + \Sigma_E | r_g \otimes \Sigma_A + \Sigma_C \\ r_g \otimes \Sigma_A + \Sigma_C | \Sigma_A + \Sigma_C + \Sigma_E \end{bmatrix} \times \mathbf{L}' + \begin{bmatrix} \mathbf{G + S + F} | r_g \otimes \mathbf{G + S} \\ r_g \otimes \mathbf{G + S} | \mathbf{G + S + F} \end{bmatrix}, \quad \text{(A2)}$$

where $(\Sigma_A + \Sigma_C + \Sigma_E)$ represents the within-person variance structure for one twin and $(r_g \otimes \Sigma_A + \Sigma_C)$ the between-twins covariance, representing the twin variance–covariance matrix for the reliable trait variance (see also the upper part of Figure 2 in the main text). Correlation r_g can be derived from quantitative genetic theory (Falconer & Mackay, 1996) and equals 1.0 for monozygotic twins and 0.5 for dizygotic twins. The $\otimes$ is the Kronecker product. The twin variance–covariance matrix for the reliable trait variance is multiplied by matrix **L**. **L** is a (4 × 2) matrix with loadings of the latent phenotypes (PT_1 and PT_2) on the parental ratings. This factor-loading matrix is of the general form $\mathbf{L} = \mathbf{I}_t \otimes (\mathbf{I}_s \otimes \mathbf{d})$. $\mathbf{I}_t$ is a $n_t \times n_t$ identity matrix where n_t is the number of measurement occasions. $\mathbf{I}_s$ is a 2 × 2 identity matrix determined by the fact there are two children in a twin pair, and **d** is an $n_r \times 1$ vector determined by the number of raters (n_r). For model identification, the elements in matrix **L** are fixed to one. **G**, **S**, and **F** (2 × 2) are:

$$\mathbf{G} = \begin{bmatrix} \Sigma_{A_m} & 0 \\ 0 & \Sigma_{A_f} \end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix} \Sigma_{C_m} & 0 \\ 0 & \Sigma_{C_f} \end{bmatrix}, \quad \text{and} \quad \mathbf{F} = \begin{bmatrix} \Sigma_{E_m} & 0 \\ 0 & \Sigma_{E_f} \end{bmatrix},$$

where Σ_{A_m} and Σ_{A_f} represent the additive genetic variance based on mother and father ratings, Σ_{C_m} and Σ_{C_f} represent the within-person shared environmental variance based on mother and father ratings, and Σ_{Em} and Σ_{E_f} represent the within-person nonshared environmental variance based on mother and father ratings (see also the lower part of Figure 2 in the main text). Combining the distinct components of both raters, who each rate two children of a twin pair, results in a 4 × 4 covariance matrix for the rater-specific factors.

Implementation of the Longitudinal Models in the Psychometric Model

In the case of multiple measurement occasions, we write the data vector as $\mathbf{y'} = [MRT_{1(t=1)}, \ldots, MRT_{1(t=n_t)}, FRT_{1(t=1)}, \ldots, FRT_{1(t=n_t)}, MRT_{2(t=1)}, \ldots, MRT_{2(t=n_t)}, FRT_{2(t=2)}, \ldots, FRT_{2(t=n_t)}]$ where t indexes the measurement occasion and n_t the maximum number of measurement occasions. The expectation Σ_Y is again given by Equation A2. However, matrices Σ_A, Σ_C, and Σ_E may now be replaced by the matrix Σ in Equation A3. This imposes a saturated (Cholesky) structure on the covariances among the genetic and environmental factors at the measurement occasions that are common to both parents. Matrices **G**, **S**, and **F** in Equation A2 represent the covariances among time points that involve factors unique for each rater. These unique covariances can also be modeled using a saturated structure. In this case of multiple measurement occasions, **G**, **S**, and **F** have the following block diagonal structure:

$$\begin{bmatrix} \Sigma_m | \mathbf{Z} \\ \mathbf{Z} | \Sigma_f \end{bmatrix},$$

(*Appendix continues*)

in which $\mathbf{Z}$ is a $n_t \times n_t$ full matrix with zeroes. Matrix Σ_m is for mother and Σ_f for father to indicate that parameter estimates may differ for both parents.

The Cholesky decomposition model in Figure 3 in the main text is an unconstrained model for the (co)variances among measurement occasions. It implies the covariance structure

$$\Sigma = \mathbf{X} \times \mathbf{X}' \quad \text{(A3)}$$

where $'$ indicates transposition. In matrix terminology, matrix $\mathbf{X}$ is an $n_t \times n_t$ lower triangular matrix with n_t equal to the number of measurement occasions. For instance and illustrative for our application, for $n_t = 4$ matrixes, $\mathbf{X}$ would be

$$\mathbf{X} = \begin{bmatrix} \lambda_{11} & 0 & 0 & 0 \\ \lambda_{21} & \lambda_{22} & 0 & 0 \\ \lambda_{31} & \lambda_{32} & \lambda_{33} & 0 \\ \lambda_{41} & \lambda_{42} & \lambda_{43} & \lambda_{44} \end{bmatrix}.$$

In this matrix, λ_{11} represents the influence of the first latent factor (η_1) on the first measurement occasion, while λ_{32} represents the influence of the second latent factor (η_2) on the third measurement occasion (the matrix elements correspond to the path coefficients in Figure 3 in the main text).

Received May 4, 2005
Revision received August 9, 2007
Accepted September 27, 2007 ■