Assessment issues in behavior-genetic research on personality

PETER BORKENAU
Martin-Luther-Universität, Halle, Germany
RAINER RIEMANN
Friedrich-Shiller Universität, Jena, Germany
ALOIS ANGLEITNER
Universität Bielefeld, Bielefeld, Germany
FRANK M. SPINATH
Universität Bielefeld, Bielefeld, Germany

ABSTRACT Behavior-genetic research on the sources of individual differences in personality relies on self-report data almost exclusively. As indices of inter-judge agreement yield the most adequate reliability estimates in behavior-genetic research on personality, and inter-judge agreement can not be estimated from self-report data alone, behavior-genetic self-report studies do not allow for adequate reliability estimates. Therefore, the reliability problem is usually ignored and the total variance is treated as true-score variance, resulting in underestimates of genetic and shared environmental and in overestimates of nonshared environmental influence. This problem may be overcome by using descriptions of the target persons by (at least) two independent knowledgeable informants, as we did in a study on 1,000 twins pairs. Another problem that is shared by behavior-genetic self-report and peer-report studies is possible contrast effects in descriptions of relatives as the relatives may be compared to each other and not to the population mean. This would result in attenuated correlations between relatives and in underestimates of the importance of the shared environment. The only way to overcome this problem is observational behavior-genetic studies in which the judges know only one of the relatives whose similarities are compared. We therefore ran an observational study on the similarity of 300 monozygotic and dizygotic adult twins pairs, the German Observational Study of Adult Twins (GOSAT). The study and its most important findings for personality are reported.

Key words: Behavior genetics, Contrast effects, Personality assessment.

Author note: Peter Borkenau, Department of Psychology; Alois Angleitner, Department of Psychology; Rainer Riemann, Department of Psychology; Frank M. Spinath, Department of Psychology. The research reported in this article was supported by a grant from the German Research Foundation (Deutsche Forschungsgemeinschaft) to Alois Angleitner, Peter Borkenau, and Rainer Riemann. The authors are indebted to the twins and to the judges for their participation, to the experimenters Susanne Hempel, Veronika Koch, Holger Lorenz, Conny Post, Beatrice Rammstedt, Birgit Schlangen, and Robert Weiss for collecting the data, and to Holger Lorenz and Wolfgang Thiel for their help in the data analysis. This paper is based on an invited lecture given by Peter Borkenau at the Fifth European Conference on Psychological Assessment, August 1999, Patras, Greece.

Address: Peter Borkenau, Institut für Psychologie der Martin-Luther-Universität, D-06099 Halle, Germany. E-mail: p.borkenau@psych.uni-halle.de

Methods and approaches in behavior genetic research

Behavior-genetics is the study of genetic and environmental influences on behavior. Such influences exist in animals as well as in humans, but the available methods to study them differ. Powerful methods like selective breeding or targeted mutation that are frequently employed in animal behavior genetics can not be used with humans. Thus animal and human behavior genetics have developed quite independently. This article is on genetic and environmental influences on behavior in humans.

A further important distinction is that between *molecular behavior genetics* and *quantitative behavior genetics*. Molecular behavior genetics establishes associations between alleles, that is, variants in organisms' desoxyribonucleic acid (DNA), and their behavior, and it tries to explain these links. This research is complicated by the circumstance that the associations between single alleles at a specific gene locus and behavioral traits are generally weak because many alleles each of which contributes a small share to the total genetic variance are involved in behavioral variation in the normal range. For example, an allelic association has been reported between the trait Novelty Seeking and a specific allele of the dopamine type 4 receptor gene (DRD4; Ebstein et al., 1995). But, according to a meta-analysis of 10 studies, the effect size of this genetic polymorphism is $d = .06$ (Bishop & Wahlsten, 1997), implying that the trait level of 52.5 % of the carriers of the one allele exceeds the median of the group with the other allele. Other allelic associations have been reported for anxiety-related traits (Lesch et al., 1996). Such findings are interesting and important, but the effects of single alleles on individual differences in behavior in the normal range seem to be quite small (Plomin & Caspi, 1998).

Nevertheless, the overall effect of the entire genome on human behavioral traits is substantial. This is the subject of quantitative behavior genetics that partitions the variance in human behavior into genetic and environmental contributions. In this approach, the phenotypic variance in a trait is accounted for by several sources, mainly: (a) additive genetic variance, (b) interactive effects of genes, (c) shared environmental influence, and (d) non-shared environmental influence.

Additive genetic variance is that part of the observed variance that reflects the additive effects of single genes, the effects of "gene dose". As first-degree relatives, that is, parents and their offspring, dizygotic (DZ) twins as well as siblings, share half their genes by descent, they also share 50% of those additive genetic effects that contribute to individual differences in the population. Thus if the phenotypic variance in a trait was entirely due to the additive effects of single alleles, the correlations between these first-degree relatives would all be .50 whereas the correlations between monozygotic (MZ) twins would be perfect.

Somewhat different rules apply to the interactive effects of genes that result in lower correlations between first-degree relatives but not between MZ twins. This is because first-degree relatives share 50 % of their genes but less than 50 % of their gene combinations. In this context, the distinction between gene dominance and epistasis becomes important. Gene dominance refers to the interactive effects of the two alleles at the same gene locus, one stemming from one's father and the other from one's mother. These gene combinations cannot be transmitted from parents to their children, and thus parents and their offspring do not share these dominance effects. By comparison, DZ twins and siblings have a 25% probability of having the same gene combination at a specific locus, and therefore they share 25 % of their dominance effects. Epistasis refers to the interactive effects of genes at different loci, and such effects are shared by first-degree relatives to a very low but not clearly specifiable extent.

Most behavior-genetic research relies on comparisons of the similarities between monozygotic and dizygotic co-twins. As MZ twins have identical genomes, they share all additive and interactive genetic influences. By contrast, DZ twins share half their additive and one quarter or less of their interactive genetic influences. Assume that individual differences in a trait entirely reflected additive and interactive effects of genes. Then the trait levels of MZ twins would be identical whereas the trait levels of DZ twins would be correlated at .50 or below, depending on the relative importance of the interactive effects.

Actually, however, environmental factors are also important, and quantitative behavior genetics therefore distinguishes between genetic and environmental influences. Moreover, the environment is subdivided into shared and non-shared environment. Shared environment is defined as entirely shared by persons reared in the same family, independent of their genetic relatedness, thus increasing their similarity. Therefore, the correlation between MZ twins, DZ twins, biological siblings, and adoptive siblings reared together would all be r = 1.00 if shared environment was the only source of variance in a trait. Shared environment is inferred from twin studies if the correlation between DZ twins is more than half the correlation between MZ twins as such a finding can not be explained by shared genes alone. In principle, assortative mating, that is, positive correlations between the trait levels of parents, might be a competing explanation as it increases the genetic similarity of DZ but not of MZ twins. Fortunately, however, the extent of assortative mating can be directly observed, and it is known to be substantial for intelligence and social attitudes but negligible for personality.

Non-shared environment are those environmental influences that do not contribute to the similarity of persons reared together (independent of their genetic relatedness), like different treatment by their parents, birth-order effects, different roles in the family, influences from different peer groups, and so on. Such non-shared environmental influences contribute to the variability in the population but not to the correlations between relatives of all kinds, and they are most directly inferred from differences between MZ co-twins reared together. For example, if the intraclass correlation between the trait levels of MZ twins reared together was r = .50 (which is quite a realistic figure), it would be concluded that 50 % of the variance in that trait was due to non-shared environmental influence, simply because MZ twins reared together share all genetic effects as well as their shared environment, implying that all differences between them reflect non-shared environmental influence (and error of measurement).

To summarize: (a) DZ correlations of more than half the MZ correlations suggest shared environmental influence, (b) MZ correlations of more than twice the DZ correlations suggest interactive effects of genes, and (c) MZ correlations below r=1.00 suggest non-shared environmental influence and/or error of measurement. More specifically, Falconer (1960) suggested twice the difference between MZ and DZ correlations as a heritability estimate, and twice the DZ correlation minus the MZ correlation as an estimate of the importance of the shared environment.

More recent behavior-genetic research uses structural equation modeling to partition the genetic and environmental contributions to the phenotypic variance (see Neale & Cardon, 1992, for an introduction). This has two advantages: (a) it allows to test more complex models, and (b) it makes the researcher's assumptions more explicit. Figure 1 depicts the basic twin model in which the correlations between additive genetic effects are fixed to 1.00 for MZ twins and to .50 for DZ twins, the correlations between the genetic dominance effects are fixed to 1.00 for MZ twins and to .25 for DZ twins, shared environmental effects are perfectly correlated independent of the twins' zygosity, and non-shared environmental effects are always uncorrelated. These correlations between latent variables are fixed whereas the size of the paths a, d, c, and e that indicate the influence of the four latent variables on the phenotype are estimated from the data.

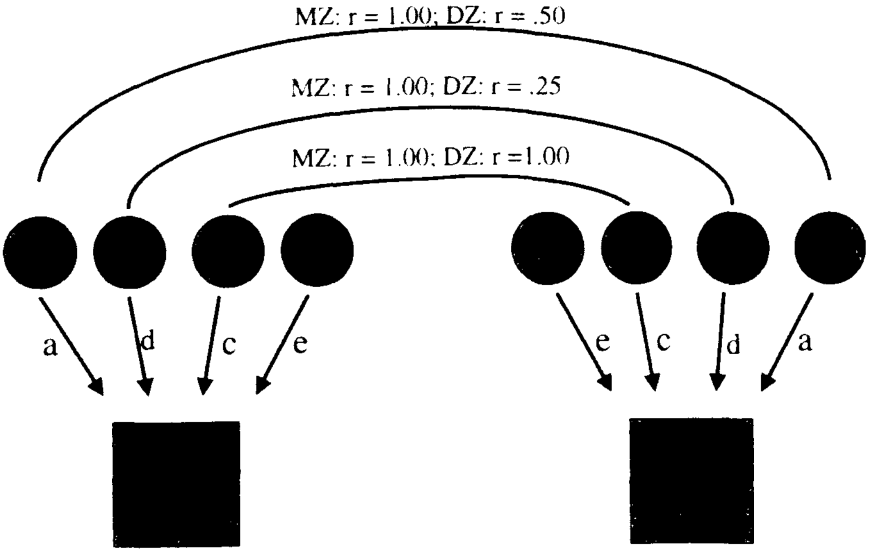
MZ: r = 1.00; DZ: r = .50
MZ: r = 1.00; DZ: r = .25
MZ: r = 1.00; DZ: r =1.00
a
d
c
e
e
c
d
a

Figure 1
Illustration of the basic twin model.

The fit of this full model can not be tested but the fit of reduced models can be tested. What is frequently tested are ADE-models that fix the effects of the shared environment to zero, ACE-models that fix the interactive genetic effects to zero, AE-models that imply no gene interactions and no shared environmental effects, and CE-models that imply shared and non-shared environmental but no genetic effects. Usually, the fit of the different models is compared, the best-fitting model is selected taking parsimony considerations into account, and the strength of the paths *a*, *d*, *c*, and *e* for the best-fitting model is estimated if they are not fixed to zero.

Findings for personality

What are the findings for personality that were obtained in this kind of research? To cut a long story short, we refer to a recently published textbook by key researchers in this field (Plomin, DeFries, McClearn, & Rutter, 1997). They concluded that 40% of the variance in personality traits are due to genetic differences whereas the other 60% reflect non-shared environmental influence. What is most surprising in these findings is *not* that genes are moderately important, but that the environment does not contribute to the similarity of persons that are reared in the same family. Rather, it seems that the environment is entirely of the non-shared variety. This implies that parental role models, parenting styles, the home atmosphere, socio-economic status, the neighborhood in which children grow up, and all other environmental circumstances that siblings share, do not contribute to their similarity in personality. Rather, the parent-child and sibling similarities that environmental research has found seem to be entirely accounted for by shared genes.

Scientific psychology has reacted to these findings in several ways. Some authors (Harris, 1995, 1998; Rowe, 1994) have suggested that children are socialized mainly by their peers and not by their parents, whereas other authors showed that there are shared environmental influences on at least some traits like religious orthodoxy (Beer, Arnold, & Loehlin, 1998). Finally, some researchers (Brody, 1993; Miles & Carey, 1997; Rose, 1995) have raised methodological concerns, mainly that the importance of the shared environment may be systematically underestimated in studies that rely on self-reports and ratings of young twins by their parents. I am going to focus on the latter hypothesis here.

The reliability issue

Note that the basic twin model in Figure 1 does not include a measurement model. The phenotype is specified as a directly observed variable rather than a latent trait that is measured by at least two indicators. For twin research on adult personality, this means that what is explained is not individual differences in personality but in personality inventory scores. As Brody (1993) put it, "the behavioral genetics of personality have not been studied, but the behavioral genetics of self-reports about personality have been studied" (p. 162). If a twin model does not account for error of measurement, and less-than-perfect correlations for MZ twins are accounted for by non-shared environment, the importance of the non-shared environment is overestimated at the expense of genetic and shared environmental influences.

This is quite obvious and did not remain unnoticed. Rather, some authors have argued that measurement error was only a minor problem in behavior-genetic research on personality because the internal consistencies of established self-report measures are usually high, frequently exceeding .80. Indeed, less than 20% error variance would not distort the findings of behavior-genetic research very much. Unfortunately, however, coefficients of internal consistency are no appropriate reliability estimates here as they estimate the generalizability of scores across item samples, but error due to item sampling does not attenuate the correlations between relatives who are administered the same personality scale. What actually attenuates the correlations between self-reports by relatives are perceiver effects (Kenny, 1994), as these reflect comparisons between descriptions of different target persons by different perceivers. Thus the similarity between the self-reports of relatives should be corrected for lack of consensus between different perceivers of the same targets. Obviously, this consensus can not be estimated from self-report data alone.

If lack of consensus is the main source of unreliability in behavior-genetic research on personality, however, the lack of a measurement component in the basic twin model becomes a major problem indeed. If target persons are described by intimate acquaintances, the consensus correlations rarely exceed .40, and .60 seems to be an upper limit (Borkenau & Liebler, 1993). Note that correlations between the personality scores of MZ twins also vary around .50 (Loehlin, 1992). Thus the *reliable* variance in self-reports of personality might be entirely accounted for by genetic factors, whereas the large influence that is usually attributed to the non-shared environment might actually reflect perceiver effects. Independent of whether this is actually the case, such considerations show that it is wise to follow the advice by Brody (1993) and Rose (1995) to base behavior-genetic research on adult personality not on self-reports exclusively.

A step forward is to collect twin descriptions by at least two independent peers per twin, employing different judges for co-twins. The independent ratings for the same target person would then allow for appropriate reliability estimates. Such a twin model is illustrated in Figure 2, and a study that used this model has been conducted by Riemann, Angleitner, and Strelau (1997) who administered

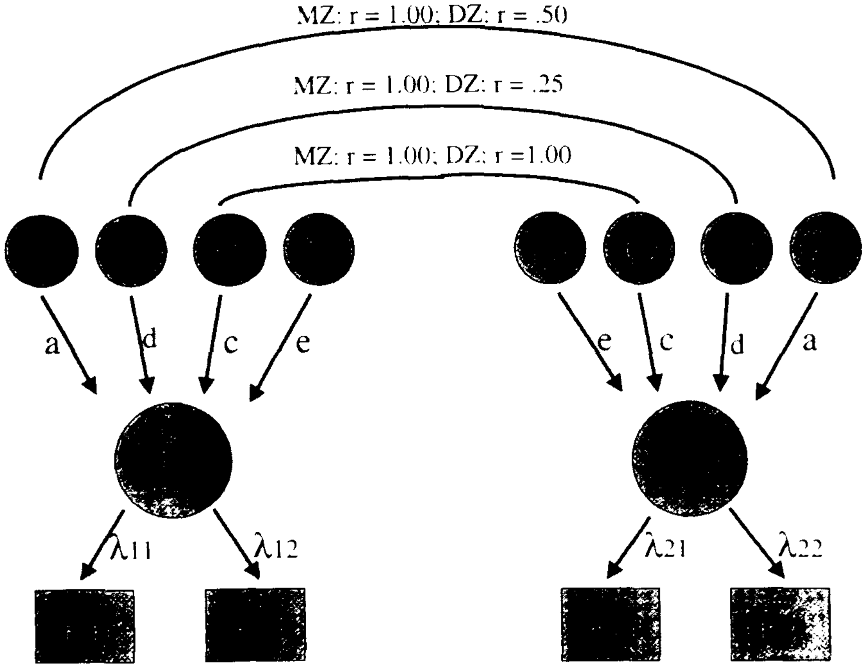
MZ: r = 1.00; DZ: r = .50
MZ: r = 1.00; DZ: r = .25
MZ: r = 1.00; DZ: r =1.00
a
d
c
e
e
c
d
a
$\lambda 11$
$\lambda 12$
$\lambda 21$
$\lambda 22$

Figure 2
The twin model combined with a measurement model.

the peer-rating version of the NEO-Five Factor Inventory (Costa, & McCrae, 1992; in the German adaptation by Borkenau, & Ostendorf, 1993) to 660 pairs of MZ and 200 pairs of same-sex DZ twins. Averaged across the five factors, the agreement between two individual peers who described the same target was *r* = .44, and therefore the mean reliability of their composite score was *α* = .61, according to the Spearman-Brown formula. The latter coefficient estimates how strongly the averaged ratings by two-judge panels should correlate if they describe the same targets. Thus 61% of the variance were reliable, whereas the other 39% reflected perceiver effects and perceiver-target interactions. When these averaged judgments by two peers were correlated between MZ twins, the mean correlation was *r* = .40, whereas the mean correlation was *r* = .21 when they were correlated between DZ twins. That the MZ correlation was lower than the inter-judge agreement indicated non-shared environmental influence. Moreover, model-fitting tests that are not reported in detail here showed that a model including only additive genetic and non-shared environmental effects fitted the data for all traits, although a model that included nonadditive rather than additive genetic effects fitted better for Neuroticism. Of the total variance, 39% were due to error of measurement, 40% to additive genetic effects, and 21% to the non-shared environment, whereas the estimates for the reliable variance were 66% genetic and 34% non-shared environmental. Thus if non-shared environment had been left confounded with error of measurement, 40% genetic and 60% non-shared environmental influence or exactly the figures suggested by

Plomin et al. (1997) would have been estimated. In contrast, inclusion of a measurement model and partitioning of the reliable variance only changed this ratio from 2:3 to 2:1. These findings show the importance of being aware of psychometric issues in behavior-genetic research.

Contrast effects

Unfortunately, another possible problem in behavior-genetic research on personality, namely contrast effects, is shared by self-report and peer report studies. Contrast effects may occur in behavior-genetic research in two ways: First, twins may mutually influence each other in such a way that the differences between their actual personalities increase. An example would be different roles of co-twins to emphasize their unique identities. Second, apart from twins' actual similarity, contrast effects may reduce the similarity of their personality *descriptions* because the twins are compared (and compare themselves) to each other and not to the population mean. Indeed, self-reports of personality are subject to comparisons with salient other persons (Schwarz, 1999), and a particularly salient other person for twins may be their co-twin. For example, if twins respond to the item "Do you like to go to parties?", they may endorse it if they like parties more than their co-twin, and deny it if the co-twin likes parties more. Such a response set would result in higher variances within twin pairs and lower variances between pairs, implying reduced twin correlations.

The first kind of contrast effect is not an assessment problem. If environmental influences did *not* only *not* contribute to twin similarity, but actually made the behavior of MZ and DZ twins different from one another, this would be appropriately reflected in reduced estimates of shared environmental and higher estimates of non-shared environmental influence. The response-set variant of contrast effects, however, would distort the parameter estimates: The importance of the shared environment would be underestimated if such a contrast effect operated in MZ twins and DZ twins alike or in DZ twins more strongly than in MZ twins, whereas genetic influences would be underestimated if it operated more strongly in MZ than in DZ twins.

Both types of contrast effects might result in negative correlations between relatives, a phenomenon that is inconsistent with the standard behavior-genetic models that predict positive correlations between relatives. But negative correlations have been repeatedly found for parental descriptions of young DZ twins who share half their genes plus their family environment. One example is a study by Spinath and Angleitner (1998) who administered Buss and Plomin's EAS to 184 MZ and 170 same-sex DZ twin pairs aged two to twelve years. If age and sex of the twins were controlled, the average correlation between the mothers' ratings was r = .56 for MZ twins and r = .00 for DZ twins, and the average correlation between the fathers' ratings was r = .55 for MZ twins and r = -.01 for DZ twins. These correlations are misleading, however, because co-twins were nested within judges, implying that the twin correlations were inflated by perceiver effects. Therefore, Spinath and Angleitner (1998) also calculated cross-correlations, that is, mothers' ratings of Twin A were correlated with fathers' ratings of Twin B, and vice versa. The averages of these cross-correlations were r = .40 for MZ twins and r = -.08 for DZ twins, and a correction of these correlations for attenuation would further increase the negative correlations between DZ twins. Thus contrast effects did definitely occur, although it is not clear whether this was a contrast effect in the twins' actual behavior or a contrast effect in the descriptions of their behavior.

Whereas reliable negative correlations between co-twins indicate contrast effects of one sort or the other, positive correlations do not indicate lack of contrast effects because contrast effects may attenuate otherwise positive correlations between relatives without depressing them below zero. It is therefore desirable to use personality measures in twin research that may not be subject to contrast effects in personality descriptions. The German Observational Study of Adult Twins was initiated by the present authors to investigate whether the control of contrast effects in twin descriptions affects the estimates of genetic and environmental influences in general, and of the shared environment in particular.

The German Observational Study of Adult Twins (GOSAT)

The German Observational Study of Adult Twins is a multimethod twin study on the sources of individual differences in personality and intelligence. Concerning personality, the study makes use of the Five-factor model of personality. Although we are aware that the Five-factor model is not without problems, we regard it as the currently most suitable model for our purposes. For most of the twin pairs, self- and peer reports were available from the already-reported peer rating study by Riemann et al. (1997). From the approximately 1,000 twin pairs of that study, 300 pairs could be invited to the University of Bielefeld for an entire day. More female (233) than male (67) pairs actually participated whereas the proportion of the two zygosity groups (168 MZ and 132 DZ pairs) was quite balanced. The zygosity of 283 pairs was diagnosed by blood typing of genetic markers, whereas the remaining 17 pairs had to be diagnosed by other methods. The overall rate of misclassifications was estimated as approximately 1%.

During the investigation day, the twin pairs were separated most of the time and worked on various tasks for about six hours. Co-twins were always taken care of by different experimenters and interacted with different experimental confederates. For an extensive description of the data, the reader is referred to Spinath et al. (1999). In the present article, we focus on two kinds of data: On-line behavior counts and ratings of videotaped behavior sequences.

On-line behavior counts

Unbeknown to the twins, the experimenters recorded the frequencies of several of their behaviors, particularly the number of questions they asked, the number of comments they gave, and other kinds of utterances they made. These variables were recorded in seven different settings, for example, while the twins were administered Raven's Advanced Progressive Matrices. Moreover, questions and comments were separately recorded before and during, and questions also separately after each task. The behavior counts were first regressed on the twins' age and sex as these variables inflate MZ as well as DZ twin correlations if the behavioral traits under study correlate with these demographic variables (McGue & Bouchard, 1984). Moreover, they were regressed on the experimenters to control for systematic experimenter effects. The residualized behavior counts were then analyzed at four levels of data aggregation: The unaggregated behavior counts constituted Level I of our analyses, whereas the aggregate of the behavior counts before, during, and after the same task constituted Level II. At Level III, the behavior counts were also aggregated across the seven settings, yielding composite scores for the overall number of questions, comments, and utterances. Finally, as the frequencies of these three behaviors were correlated at .35 and beyond, they were combined into a Talkativeness trait score that constituted Level IV.

At Level I, the twin correlations were very low for MZ as well as DZ twins, indicating lack of reliability and non-shared environmental effects almost exclusively. At Level II the mean MZ and DZ correlations were .15 and .12, at Level III the mean MZ and DZ correlations were .26 and .24, and at Level IV the MZ correlation was .31 whereas the DZ correlation was .23. Thus the twin correlations were modest at all levels, and they were not much higher for MZ than for DZ twins. That the highest twin correlation was .31 probably reflected a lot of error variance in these data. This may seem trivial, but it is at variance with reports from the Minnesota studies on MZ twins reared apart where MZ twins were reported to show spectacular concordances in very specific behaviors. By contrast, our findings do not suggest strong genetic influences on specific behaviors. Another interesting finding in GOSAT is that the difference between the MZ and the DZ correlations was higher at the trait level than at the level of its constituent behaviors, suggesting that genetic influences operate mainly at the level of traits (top-down model of genetic influence).

To pursue the latter question more systematically, we tested the fit of a *Common Pathway Genetic Model* that is illustrated in Figure 3. This model distinguishes between a common latent trait and its specific indicators, and it distinguishes between genetic and environmental influences at both levels. In such a model, the cross-correlations between one indicator in Twin A and another indicator in Twin B become important, as such cross-correlations suggest genetic or shared environmental influences at the common trait level. Moreover, if the cross-correlations are higher for MZ twins than for DZ twins, they indicate genetic influence at the common trait level. By contrast, genetic and shared environmental influences at the specific level contribute to the correlations but not to the cross-correlations between co-twins, whereas the specific non-shared environmental influences do not contribute to any of the correlations or cross-correlations at all.

We fitted a Common Pathway Genetic Model to our on-line behavior counts and found that an ACE model which included genetic and shared environmental influences at the common trait level fitted the data well ($\chi^2 = 33.8$, $df = 28$, $p = .21$). These analyses are reported in more detail by Borkenau, Riemann, Spinath, and Angleitner (2000). Most interesting was that genetic influences were identified at the trait level but not at the level of its specific indicators, suggesting that the genetic influences on the specific behaviors were all mediated by the common Talkativeness trait. This might reflect that genetic influences tend to operate via nerve cells and hormone levels that are more likely to affect global dispositions than specific behaviors. By contrast, specific behaviors are subject to situation-specific influences (Epstein, 1979), and they are more likely to be shaped by learning processes. Our data base is too narrow to confirm such wide-ranging conclusions, but our on-line behavior counts are at least consistent with such a view. But a far more extensive data base is available from GOSAT concerning ratings of videotaped behavior sequences.

Behavior ratings

During the investigation day, the twins were individually videotaped in 15 different settings that were diagnostic of individual differences in personality. For example, the twins had to: (a) introduce themselves, (b) tell an experimental confederate a joke, (c) persuade an ostensible obstinate neighbor (actually a confederate) on the phone at 11 p.m. to reduce the loudness of her stereo, (d) build a high and stable paper tower, or (e) sing a song of their choice.

In this way, approximately 60 min. of videotapes per participant or about 600 hours of videotapes altogether were collected. These were then rated by judges who never met the targets and provided trait ratings of the twins, relying on these videotapes only. To increase the reliability of these trait ratings, each twin was observed in each setting by four independent judges. Moreover, the behavior in different settings was rated by different panels of four judges to secure independence of the ratings for different settings. Finally, different panels of judges were employed for co-twins to prevent any assimilation or contrast effects in co-twin perceptions. Thus a total of 4 (parallel judgments) x 15 (number of settings) x 2 (co-twins) = 120 judges were employed. Each of these judges provided ratings of one twin of each pair, that is, of 300 persons.

The judges provided their ratings via a

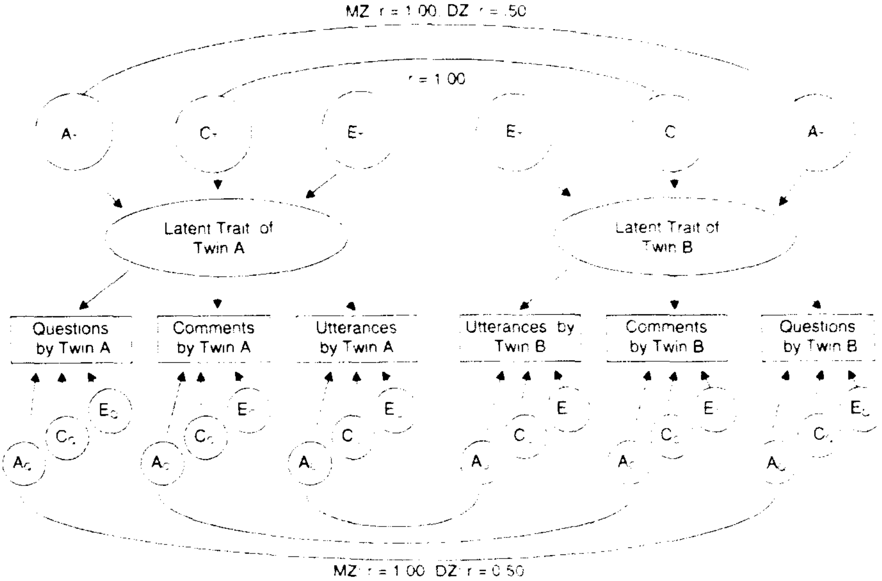

Figure 3
Illustration of a common pathway ACE model.

Note: A_T = genetic influences at the common trait level, C_T = shared environmental influences at the common trait level, E_T = non-shared environmental influences at the common trait level, A_Q = specific genetic influences on the number of questions, C_Q = specific shared-environmental influences on the number of questions. E_Q = specific non-shared-environmental influences on the number of questions, A_C = specific genetic influences on the number of comments, C_C = specific shared-environmental influences on the number of comments, E_C = specific non-shared-environmental influences on the number of comments, A_U = specific genetic influences on the number of utterances, C_U = specific shared-environmental influences on the number of utterances, E_U = specific non-shared-environmental influences on the number of utterances. The *arrows indicating perfect correlations between the specific shared-environmental influences on co-twins* have been omitted.

computer on 35 bipolar trait rating scales. Each of Goldberg's (1990) Big Five factors Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Intellect was measured with four bipolar scales, and four additional scales were included to measure McCrae and Costa's (1987) Openness to Experience conceptualization of Factor V. The selection of the specific adjective scales relied on a large trait-taxonomic study by Ostendorf (1990) of the German personality-descriptive language. Two of the four scales per factor were reverse scored to control for acquiescence response bias. Moreover, ratings of the targets' attractiveness and likeability were included, mainly to control for the higher expected similarity of MZ twins in physical attractiveness. In addition to these 26 adjectives that were used in all 15 settings, nine setting-specific rating scales were included. Altogether, 1.26 million behavior ratings were collected this way, taking more than 4,100 hours for observation and judgments. Although the ratings were simultaneously collected in Bielefeld and in Halle, the rating procedure took more than one year to be completed.

The ratings thus collected were regressed on the twins' age and sex and on systematic effects of the experimenter who took care of them. As co-twins were always taken care of by different experimenters, any experimenter effects would attenuate the correlations between them. Moreover, the scores were regressed on the twins' perceived physical attractiveness as attractiveness might substantially affect ratings by strangers if they rely on a stereotype like "who is beautiful is good" (Dion, Berscheid, & Walster, 1972). As the MZ correlation for attractiveness was .68, whereas the DZ correlation for attractiveness was .27, such a stereotype might have inflated the perceived MZ-DZ differences and thus the heritability estimates.

Using the residualized ratings, we then estimated the inter-rater reliabilities for judgments of the twins' personality. These analyses were run at the level of the 24 adjective scales as well as at the factor level, and they were run at the level of each setting as well as averaged across the 15 settings. Here, we report findings at the factor level only. For more details, the reader is referred to Borkenau, Riemann, Angleitner and Spinath (2001). At the level of the individual settings, the average reliability of the mean rating of the four relevant judges (ICC 2, 4; according to Shrout & Fleiss, 1979) was .67. Thus if four other judges had watched the same videotaped behavior, a correlation of .67 between the mean ratings by the two four-judge panels would have been expected. This is an appropriate standard of comparison for the twin correlations that rely on judgments of co-twins by non-overlapping panels of four judges. The average correlation between the ratings of MZ twins was r = .30, whereas the average correlation between the ratings of DZ twins was r = .18. This suggested substantial contributions of the non-shared environment (37%) and of error of measurement (33%). A large part of the remaining variance (24%, according to Falconer's formula) was due to additive genetic effects, whereas the effects of the shared environment on trait expressions in a particular situation were weak (6%). Correcting these estimates for unreliability of measurement and analyzing the reliable variance only yielded estimates of 36% genetic, 9% shared environmental, and 55% non-shared environmental influence.

Note, however, that the behavior of the twins in a specific setting may have depended on subtle situational circumstances, among them interactions between the target and the experimental confederate, that contributed to differences in the behavior of co-twins and thus raised the estimates of non-shared environmental influence. More dependable measures were the composite trait ratings across all 15 observational settings. As four different judges observed each twin in each of the 15 settings, these composite scores were the averaged judgments by 60 different perceivers, four perceivers being nested within the same setting. The coefficients of rater agreement thus estimated the correlation of the composite score of a panel of 60 judges with the theoretical composite score of ratings by another hypothetical panel of 60 judges who observed the same target persons. These composite ratings could therefore be expected to be highly reliable. Indeed, the inter-rater reliability at this level of data aggregation was .94, averaged across the six trait domains under study.

Moreover, the average MZ correlation was r = .59 whereas the average DZ correlation was r = .38. Thus 6% of the variance reflected perceiver effects and 35% reflected non-shared environmental influence. Moreover, according to Falconer's formula, 42% of the variance were genetic and 18% were due to the shared environment. These estimates differ from those by Plomin et al. (1997), but not with respect to the importance of genes. Rather, the discrepancy concerns the decomposition of the environmental variance into shared and non-shared sources. We obtained higher estimates of shared environmental influence than have usually been obtained in self-report and peer report studies. This discrepancy may reflect the control of contrast effects in the behavioral ratings.

As already reported, the measurement of personality traits by ratings of videotaped behavior sequences yielded highly reliable scores. But were these scores also valid? Unfortunately, this could not be checked for most of the traits under study because an appropriate validity criterion was lacking. But one of the 35 adjective scales in GOSAT referred to the twins' intelligence, and we also administered two tests of psychometric intelligence, Raven's Advanced Progressive Matrices and a short version of the Leistungsprüfsystem (LPS) by Horn, a popular German intelligence test. So perceived intelligence could be compared to measured intelligence. The correlation between the Raven and the LPS was r = .60, the correlation between the Raven and perceived intelligence was r = .33, and the correlation between the LPS and perceived intelligence was r = .49. Thus the correlation between the LPS and judgments of intelligence by multiple perceivers was not much lower than the correlation between the LPS and the Raven. Moreover, these correlations suggest that the intelligence ratings reflected cristallized aspects of intelligence that are measured by the LPS but not by the Raven. Obviously, we can only speculate that the substantial validity of our intelligence ratings generalizes to the ratings of the other traits in our study.

Conclusions

Research on individual differences frequently blurs the distinction between the level of specific measures and the level of the constructs that are the target of measurement. This has been the rule rather than the exception in behavior-genetic research on personality. In this article, we argue that the neglect of psychometric considerations in behavior-genetic research on personality resulted in overestimates of the importance of the non-shared environment and in underestimates of the importance of the shared environment. Indeed, several years ago, Rose (1995) has warned that "perhaps the obituary for the shared environment effect has been written too soon" (p. 646). We agree and add that this obituary has been written too early because psychometric aspects have not been appropriately considered.

At a more general level, we believe that psychometricians work too frequently in isolation to pursue their traditional research areas like structure and measurement of personality and abilities. Although these are important fields of research, we feel that psychometricians should also consider to work more frequently in tandem with specialists in other fields like, for example, behavior geneticists. In a similar vein, Wahlsten (1999) has recently argued that behavioral testing has been severely neglected in animal behavior genetics. To be clear, we do not limit this suggestion to behavior genetics; there are other fields of research that suffer from psychometrically poor designs as well. But this article is on assessment issues in behavior-genetic research on personality, and so we hope to have made a convincing case that this is a field in which extremely important psychometric considerations have been largely neglected for a long time.

References

Beer, J. M., Arnold, R. D., & Loehlin, J. C. (1998). Genetic and environmental influences on MMPI factor scales: Joint model fitting to twin and adoption data. *Journal of Personality and Social Psychology*, *74*, 818-827.

Bishop, K., & Wahlsten, D. (1997). Sex differences in the human corpus callosum: Myth or reality? *Neuroscience and Biobehavioral Reviews*, *21*, 581-601.

Borkenau, P., & Liebler, A. (1993). Convergence of stranger ratings of personality and intelligence with self-ratings, partner ratings, and measured intelligence. *Journal of Personality and Social Psychology*, *65*, 546-553.

Borkenau, P., & Ostendorf, F. (1993). *Neo-Fünf-Faktoren-Inventar* (NEO-FFI) [NEO Five-Factor Inventory]. Göttingen, Germany: Hogrefe.

Borkenau, P., Riemann, R., Spinath, F. M., & Angleitner, A. (2000). Behavior-genetics of personality: The case of observational studies. In S. Hampson (Ed.), *Advances in personality psychology* (pp. 107-137). London: Routledge.

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

Brody, N. (1993). Intelligence and the behavioral genetics of personality. In R. Plomin, & G. E. McClearn (Eds.), *Nature, nurture, and psychology* (pp. 161-178). Washington, DC: American Psychological Association.

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory: Professional manual.* Odessa, FL: Psychological Assessment Resources.

Dion, K., Berscheid, E., & Walster, E. (1972). What is beautiful is good. *Journal of Personality and Social Psychology, 24,* 285-290.

Ebstein, R. P., Novick, O., Umansky, R., Priel, B., Osher, Y., Blaine, D., Bennett, E. R., Nemanov, L., Katz, M., & Belmaker, R. H. (1995). Dopamine D4 receptor (D4DR) exon III polymorphism associated with the human personality trait of novelty seeking. *Nature Genetics, 12,* 78-80.

Epstein, S. (1979). The stability of behavior: I. On predicting most of the people much of the time. *Journal of Personality and Social Psychology, 37,* 1097-1126.

Falconer, D. S. (1960). *Introduction to quantitative genetics.* Edinburgh, Scotland: Oliver & Boyd.

Goldberg, L. R. (1990). An alternative "description of personality": The big-five factor structure. *Journal of Personality and Social Psychology, 59,* 1216-1229.

Harris, J. R. (1995). Where is the child's environment? A group socialization theory of development. *Psychological Review, 102,* 458-489.

Harris, J. R. (1998). *The nurture assumption: Why children turn out the way they do.* New York: The Free Press.

Kenny, D. A. (1994). *Interpersonal perception: A social relations analysis.* New York: Guilford.

Lesch, K. P., Bengel, D., Heils, A., Sabol, S. Z. et al. (1996). Association of anxiety-related traits with a polymorphism in the serotonin transporter gene. *Science, 274,* 1527-1531.

Loehlin, J. C. (1992). *Genes and environment in personality development.* Newbury Park, CA: Sage.

McCrae, R. R., & Costa, P. T. (1987). Validation of the five factor model of personality across instruments and observers. *Journal of Personality and Social Psychology, 52,* 81-90.

McGue, M., & Bouchard, T. J. (1984). Adjustment of twin data for the effects of age and sex. *Behavior Genetics, 14,* 325-343.

Miles, D. R., & Carey, G. (1997). Genetic and environmental architecture of human aggression. *Journal of Personality and Social Psychology, 72,* 207-217.

Neale, M. C., & Cardon, L. R. (1992). *Methodology for genetic studies of twins and families.* Dordrecht, The Netherlands: Kluwer.

Ostendorf, F. (1990). *Sprache und Persönlichkeitsstruktur: Zur Validität des Fünf-Faktoren-Modells der Persönlichkeit* [Language and personality structure: On the validity of the five-factor model of personality]. Regensburg, Germany: Roderer.

Plomin, R., & Caspi, A. (1998). DNA and personality. *European Journal of Personality, 12,* 387-407.

Plomin, R., DeFries, J. C., McClearn, G. E., & Rutter, M. (1997). *Behavioral genetics.* New York: Freeman.

Riemann, R., Angleitner, A., & Strelau, J. (1997).

Genetic and environmental influences on personality: A study of twins reared together using the self- and peer report NEO-FFI scales. *Journal of Personality*, *65*, 449-475.

Rose, R. (1995). Genes and human behavior. *Annual Review of Psychology*, *46*, 625-654.

Rowe, D. C. (1994). *The limits of family influence: Genes, experience, and behavior.* New York: Guilford.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, *54*, 93-105.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420-428.

Spinath, F. M., & Angleitner, A. (1998). Contrast effects in Buss and Plomin's EAS questionnaire: A behavioral genetic study on early developing personality traits assessed through parental ratings. *Personality and Individual Differences*, *25*, 947-963.

Spinath, F. M., Riemann, R., Hempel, S., Schlangen, B., Weiss, R., Borkenau, P., & Angleitner, A. (1999). A day in the life: Description of the German Observational Study of Adult Twins (GOSAT) assessing twin similarity in controlled laboratory settings. In I. Mervielde, I. Deary, F. de Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe* (Vol. 7, pp. 311-333). Tilburg, The Netherlands: Tilburg University Press.

Wahlsten, D. (1999). Single-gene influences on brain and behavior. *Annual Review of Psychology*, *50*, 599-624.