CHAPTER EIGHT

Behavior Genetic Research Methods

Testing Quasi-Causal Hypotheses Using Multivariate Twin Data

ERIC TURKHEIMER AND K. PAIGE HARDEN

I find the use of a correlation coefficient a dangerous symptom. It is an enemy of generalization, a focuser on the "here and now" to the exclusion of the "there and then." Any influence that exerts selection on one variable and not the other will shift the correlation coefficient. What usually remains constant under such circumstances is one of the regression coefficients. If we wish to seek for constancies, then, regression coefficients are much more likely to serve us than correlation coefficients.

(Tukey, 1969, p. 89).

INTRODUCTION

The suggested topic of this chapter – methods for behavioral genetic research in personality – is potentially misleading. Behavioral genetics, of course, is simply the science of genetics in its many forms as it is applied to behavior, and for the most part there is no reason for genetic research methods for the study of behavior to be any different than genetic methods applied to nonbehavioral characteristics of organisms. The genetics of behavior can be studied in humans or in nonhuman animals, using correlational or experimental methods; it can be inferred from patterns of familial relations or observed more directly in DNA. Like any characteristic of an organism, behavior can be thought of in terms of individual differences or unvarying species-typical characteristics; it can be studied in the cross-sectional context of the current moment or as an ongoing process in a life span or evolutionary time; and it can be seen as an aspect of normal functioning or as a reflection of disorder and distress. The same is true of personality. There is little about personality that requires it to be studied differently than diabetes, or height, for that matter. The domain of behavioral genetic research methods in personality is in principle no less extensive than the intersection of genetics and personality. As such it is too vast to review in a single chapter.

Research methods in both behavioral genetics and personality are currently at a crossroads. Although the history of the behavioral genetics of personality has its origins in animal breeding, and the foundational work in the field is largely about temperament in dogs (Scott & Fuller, 1965), what has come to be thought of as behavioral genetics are the methods of quantitative genetics, in which genetic and environmental processes are inferred from differences in genetic and environmental relationships in twin and sibling pairs, families, and pedigrees. Over roughly the same period, "personality" has come to refer largely to individual differences in human personality, especially as they are assessed via paper-and-pencil and self-report. The intersection of these more specific paradigms, in which correlations between the self-reported personality scores of family members are analyzed using quantitative genetic statistical models, has defined the behavioral genetics of personality for the last 50 years (Tellegen, Lykken, Bouchard, Wilcox, Segal, & Rich, 1988).

That methodological era is coming to an end. The wider availability of specific genetic markers and the sequencing of the human genome has supplemented, and to some degree supplanted, the classical quantitative methods of the last century (Charney, 2012). Researchers in personality have recognized the limitations of a narrow focus on self-report (Oltmanns & Turkheimer, 2009), and the recent explosion in evolutionary thinking in the behavioral sciences has had a profound effect on the field (Penke, Denissen, & Miller, 2007), in particular by rekindling interest in species-general aspects of personality as opposed

https://doi.org/10.1017/CBO9780511996481.012

to individual differences, and by reminding us of the importance of personality in nonhuman animals (Gosling & John, 1999).

With that in mind, it may seem retrograde to focus a review of behavioral genetic research methods in personality on twin methods for the study of individual differences in self-reported responses in humans. The choice can be defended on several grounds, other than the simple fact that this is where the expertise of the authors happens to lie. First is to counter the too-frequent tendency in the behavioral sciences to move on from one poorly understood method to the next, motivated not by the theoretical completion of the old paradigm but rather by the availability of new technology. This review endeavors to show that both the foundations and implications of classical twin studies of personality have not been fully understood. Related to this motivation, the new molecular genetic methodologies have themselves led to a complex tangle of methodological difficulties, which one of us has recently reviewed elsewhere (Turkheimer, 2012), although not in the context of personality per se (but see Munafo, Clark, Moore, Payne, Walton, & Flint, 2003). Finally, there is the undeniable but somewhat mysterious fact that notwithstanding the thousands of twin studies of personality that have been conducted, and an equally rich history of theoretical writing on the subject, the quantitative genetics of personality remains stubbornly controversial, both widely accepted as foundational yet regularly rejected as misleading or worse. Its merits and implications continue to be debated in the top journals (Charney, 2012). The perhaps unrealistic goal of this chapter is to soften the disagreement about the genetics of behavior by reformulating its methodological foundation of twin and family studies. Later, we also apply our reformulation of older methods to gain realistic understanding of the newer ones that capitalize on the availability of measured DNA.

PERSONALITY AS NONEXPERIMENTAL SCIENCE

Focusing the chapter on individual differences in humans highlights a particularly problematic aspect of scientific inference in the human behavioral sciences: the inference of causality from nonexperimental data (see West, Cham, & Liu, Chapter 4 in this volume). It is, of course, possible to study personality experimentally, using random assignment to experimental conditions to isolate causal effects of manipulations from extraneous variables that might otherwise confound them. The branch of personality psychology that interfaces with social psychology consists largely of this kind of work. It is even possible to conduct randomized experimental research while including genetic information (Burt, 2009), although this is not often attempted in humans.

When studying human individual differences in personality, the observations are usually correlational, beginning with the most fundamental observations in personality, the patterns of association among personality items that have been the basis for factor analytic studies of personality structure since Cattell (1957). Even at a more molar level, the basic observations of personality science usually involve statistical associations, either among the personality traits themselves or with external variables that are indicators of validity. We refer to these relations as phenotypic associations, with "phenotype" denoting a characteristic of an organism at the observational level, as opposed to its underlying causes. Phenotypic associations with personality are easy to observe, but in nonexperimental work the important underlying questions are about cause: What causes individual differences in personality, and what do individual differences in personality cause? For example, does neuroticism cause poor physical health (Shipley, Weiss, Der, Taylor, & Deary, 2007)? Does military service change one's personality, or are men with certain personality characteristics more likely to select military service (Jackson, Thoemmes, Jonkmann, Ludtke, & Trautwein, 2012)? Do changes in impulsivity cause a young adult to "mature out" of alcohol use, or does heavy drinking cause increases in impulsivity (Littlefield, Sher, & Wood, 2009; Quinn, Stappenbeck, & Fromme, 2011)? These causal questions are very difficult to answer, and phenotypic associations alone are causally ambiguous. The only way to demonstrate conclusively that a phenotypic association between heavy drinking and impulsivity is causal would be to assign individuals randomly to different heavy-drinking conditions and see what happens. Random experimentation of this kind is, of course, often impossible for practical or ethical reasons. In the absence of random assignment, how can a social scientist proceed?

As is usually the case in the social sciences, the answer is that social scientists can resort to quasi-experimental methods, in the hope of capturing some of the causal certainty offered by the idealized random experiment (see West et al., Chapter 4 in this volume). Suppose, for example, there existed pairs of children who had been matched for cultural background and

https://doi.org/10.1017/CBO9780511996481.012

genetic predisposition, yet who nonetheless differed in some hypothesized causal factor. To continue with the example of heavy drinking and impulsivity, if there were pairs of children matched for genetic and environmental family background but differing in their drinking habits, and if differences in drinking within pairs of matched children was still related to their personality, this association could not be the result of environmental or genetic family background, because the pairs had been matched for these traits. Associations within matched pairs would not *prove* causation, because matching can never be comprehensive and perfect, but to the extent the matching succeeded in holding constant important confounds, the within-pair associations would strengthen our impression that the association may have a causal basis. We have adopted the qualified term *quasi-causal* to denote associations that have survived analysis using quasi-experimental methods.

Needless to say, the matched pairs we have described do exist: they are called identical (monozygotic; MZ) twins reared together. Other kinds of familial clusters – fraternal (dizygotic; DZ) twins, siblings, half-siblings, twins reared apart, adoptive siblings, and so forth – are also matched, but to a lesser degree than are identical twins reared together. This chapter makes the case that the essential contribution of what is commonly called behavior genetics is the use of such familial clusters to obtain a significant but imperfect degree of quasi-experimental control over nonexperimental phenotypic associations. To illustrate this point, we begin by presenting a regression-based analysis of MZ twin data on religiosity (involvement in organized religious activities) and delinquent behavior during adolescence. This starting-off point differs from what is typically thought of as a "twin study" in two important respects. First, the focus of our analysis is on the relation between two individual differences variables, rather than on dividing the sources of variation in a single behavior into genetic versus environmental components. By the end of this first analysis, we will not know much about how much genes matter for either religiosity or delinquency, but we will know much more about how they are related to each other. Second, we begin by using data from only MZ twins rather than from both MZ and DZ twins. As we describe later, the familiar decomposition of observed variance into genetic and environmental components depends on comparing the relative similarity of MZ versus DZ twins, but we hope to convince the reader that the clearest exposition of the twin method starts elsewhere. Indeed, we request that the reader lay aside what he or she already knows or has heard about twin studies, including the idea that the purpose of behavioral genetics is to estimate the magnitude of genetic and environmental contributions to a trait. In later sections, we expand on our simple MZ-twin regression analysis to show how it intersects with more complex – and perhaps more familiar – methods for analyzing twin data. (The reader interested in a more traditional introduction to the twin method can see Plomin, DeFries, Knopik, & Neiderhiser, 2012 for an exhaustive account, or Neale & Maes, 2007 for a more computationally oriented approach.).

RELIGIOSITY AND DELINQUENCY IN MZ TWINS

To provide a brief substantive background for our example, the incidence of delinquency increases so dramatically in adolescence that some researchers consider it to be developmentally normative (Moffitt, 1993). Adolescents commit more than 30% of major crimes in the United States (Federal Bureau of Investigation, 2004). One potential protective factor against delinquency is religiosity – that is, affiliation with and involvement in religious organizations and activities. Religious involvement may decrease problem behavior by instilling beliefs about divine sanctions, encouraging prosocial ties that foster concern for collective well-being, facilitating the intergenerational communication of conforming values, and buffering against psychological distress that otherwise may be acted out in problem behavior (Alpert, 1939; Smith, 2003). A meta-analysis of 60 studies concluded that there is a moderate negative relationship between religiosity and delinquency (Baier & Wright, 2001). However, the association between religiosity and delinquency is confounded by numerous variables related to both, including genetic factors (Koenig, McGue, Krueger, & Bouchard, 2005; Miles & Carey, 1997).

Twin data on religiosity and delinquency are drawn from the National Longitudinal Study of Adolescent Health (Add Health), a nationally representative study designed to assess adolescent health and risk behavior, collected in four waves between 1994 and 2008. Add Health participants were recruited using a stratified school-based sampling design. A randomly selected subsample of 20,745 participants (randomly selected from school rosters) completed a 90-minute in-home interview between April and December 1995 (Wave I interview; 10,480 female, 10,264 male). Participants ranged in age from 11 to 21 years (M = 16 years, 25th percentile = 14 years, 75th = 17 years). The

design features of the Add Health data set have been extensively described elsewhere (Harris, 2011), and interested readers are referred to the Add Health website (http://www.cpc.unc.edu/addhealth) for additional information. In this chapter we use a subsample of the Add Health participants that comprises all adolescents between 11 and 20 years old who were identified as monozygotic (MZ) or dizygotic (DZ) twins raised together in the same household (N = 289 MZ pairs, 451 DZ pairs). Twin zygosity was determined primarily on the basis of self-report and four questionnaire items concerning how often twins were confused with one another and the similarity of their physical appearance. Eighty-nine pairs of uncertain zygosity were determined to be identical if they shared five or more genetic markers. Details regarding the Add Health twins sample are described by Harris, Halpern, Smolen, and Haberstick (2006).

Religiosity was measured using four items (rated on four-point or five-point ordinal scale) assessing importance of religion, frequency of prayer, attendance at religious services, and attendance at youth groups. Additional religiosity items that focused primarily on type of affiliation (e.g., identification as born-again) or theological beliefs (e.g., divine authorship of sacred texts) were excluded. Individuals who denied any religious affiliation were not assessed for religiosity during the interview; they were assigned scores as appropriate (e.g., "Never" for frequency of prayer; "Not at all important" for importance of religion.) Items scored in reverse direction, such that higher scores reflect less religiosity, were reversed numerically. Religiosity scores were computed by summing responses on the four items (M = 9.44, SD = 4.22, median = 9, range = 4 to 17, alpha = 0.76).

Adolescents were also asked how often in the last 12 months they had engaged in each of 15 antisocial behaviors: Never (0), One or Two Times (1), Three or Four Times (2), or Five or More Times (3). In addition, they were asked how often in the last 12 months each of 4 violent events happened: Never (0), Once (1), More Than Once (2). A previous confirmatory factor analysis of this data indicated that 11 items pertaining to theft, deception, and public rowdiness were indicative of a single factor, labeled here as delinquency (Harden, Mendle, Hill, Turkheimer, & Emery, 2008). (The remaining items were indicative of a factor pertaining to aggressive or violent behavior.) Delinquency factor scores (M = 0.12, SD = 0.81, range = –1.11 to 3.58, 25th percentile = –0.50, 75th percentile = 0.68) were estimated using the program Mplus (Muthén & Muthén, 1998–2010).

RANDOM EFFECTS MODELS

The most readily apparent analytic complexity introduced by MZ twin data (as opposed to data on singletons) is that observations may no longer be considered independent: individuals are clustered within twin pairs. Failure to consider this nonindependence may cause serious bias in the estimation of parameters and standard errors. Random effects models, also known as hierarchical linear models or mixed effects models, are a popular approach for the analysis of clustered data. (For a comprehensive introduction to random effects models, see Raudenbush and Bryk (2002) or Schoemann, Rhemtulla, and Little, Chapter 21 in this volume.) A basic mixed effects model for our data, analogous to a simple regression in non-clustered data, is as follows:

$$Y_{ij} = B_{00} + (B_{01} \times X_{ij}) + u_{0J} + e_{ij} \quad (8.1)$$

The subscripts *ij* represent the i^{th} twin within the j^{th} pair. The first part of Model 1 is directly comparable to traditional regression analysis, with B_{00} representing the population intercept and β_{01} representing the expected increase in delinquency given an increase in one unit religiosity. Together this represents the fixed portion of the model. The latter, random portion of the model is composed of u_{0j}, the pair-level error of prediction (i.e., the difference between the population-level intercept B_{00} and the intercept for a given twin pair); and e_{ij}, the individual-level error of prediction. As applied to our example, u_{0j} reflects the extent to which a twin pair is, on average, less or more delinquent than the overall population; e_{ij} reflects the extent to which an individual twin is less or more delinquent than the pair average. Unaccounted-for variation in delinquency is thus divided into two components: one part shared by twins in a pair and another part independent among individual twins. Taken together, the fixed effects and random effects parts comprise a model closely related to traditional regression analysis; the only addition is an estimate of the dependence between observations within a cluster (i.e., the pair-level residual variation).

This model was estimated in MZ twin pairs using PROC MIXED in SAS (see Appendix A at the end of the chapter for code). Results are listed under Model 1 in Table 8.1. Consistent with previous epidemiological research, religiosity was significantly associated with lower delinquency (β = 0.027). Of the total unaccounted-for variation in delinquency (0.324 + 0.343 = 0.667), 51.4% (0.343/0.667) was attributable to genetic and environmental factors that make twins

TABLE 8.1. Results from Mixed Effects Models of Religiosity and Delinquency

	MZ Twins Only			MZ and DZ Twins
	Model 1	Model 2	Model 3	Model 4
Fixed Effects				
Intercept	−.178 (.097)	−.353 (.113)*	−.353 (.113)*	−.404 (.069)*
Religiosity	.027 (.009)*	−.008 (.017)		
Religiosity Deviation			−.008 (.017)	−.008 (.018)
Deviation*Zygosity (DZ Effect – MZ Effect)				.027 (.021)
Religiosity Pair Average		.055 (.020)*	.047 (.011)*	.050 (.007)*
Random Effects				
MZ Twin Pair	.343 (.048)*	.335 (.046)*	.335 (.046)*	.314 (.045)*
Within-MZ Residual	.324 (.029)*	.317 (.028)*	.317 (.028)*	.317 (.028)*
DZ Twin Pair				.225 (.031)*
Within-DZ Residual				.389 (.028)*

* Significant at $P < .05$.

similar. Put another way, the intraclass correlation for twin pairs was 0.51. The remaining 48.6% of the variance existed within twin pairs and can thus be attributed to environmental influences that make twins different and to measurement error.

Besides dividing unaccounted-for variation in delinquency into between-pair and within-pair components (and providing accurate estimates of the standard errors), the results of Model 1 tell us nothing that we could not have known from an ordinary correlational study. There is, however, an additional piece of information that we can include in the model – the average level of religiosity for the twin pair ($\bar{X}_{0j}$):

$$Y_{ij} = B_{00} + (B_{01} \times X_{ij}) + (B_{02} \times \bar{X}_{0j}) + u_{0j} + e_{ij} \quad (8.2)$$

Consider how the addition of this one piece of information changes the meaning of the parameter B_{01}, which now quantifies whether an individual who is more religious is less delinquent, *controlling for the overall level of religiosity in his or her twin pair*. Because religiosity is not randomly assigned, controlling for the average level of religiosity in the pair essentially controls for being from the "type" of family that is religious, including all the between-family genetic and environmental differences that are confounded with average religiosity. Results from this model are listed under Model 2 in Table 8.1. Twin pairs who are more religious, on average, have lower levels of delinquency ($\beta_{02} = 0.055$). However, if Twin A is more religious than Twin B, this within-pair difference does *not* significantly predict delinquency ($\beta_{01} = -0.008$), as would be predicted by a causal hypothesis.

A difficulty with including the pair average as a predictor is that it may be strongly correlated with an individual's score. To ameliorate the problem of multicollinearity, Model 2 may be re-parameterized to yield orthogonal covariates, namely the twin-pair average ($\bar{X}_{0j}$) and the *deviation* of each twin from the twin-pair average ($X_{ij} - \bar{X}_{0j}$):

$$Y_{ij} = B_{00} + (B_B \times \bar{X}_{0j}) + (B_W \times (X_{ij} - \bar{X}_{0J})) + u_{0j} + e_{ij} \quad (8.3)$$

The between-cluster regression coefficient B_B estimates whether pairs with higher average religiosity have lower average delinquency, including the effects of the unmeasured covariates that vary at the pair level. In contrast, the within-cluster regression coefficient B_W estimates whether the MZ twin with higher religiosity than his or her co-twin also has lower delinquency than his or her co-twin. Results from this model (Model 3) are shown in Table 8.1. The regression of delinquency on religiosity within twin pairs is *not* significant ($\beta_W = -0.008$), but the regression on mean pair religiosity *is* significant ($\beta_B = 0.047$). Model 3 is merely a re-parameterization of Model 2, such that B_B in Model 3 equals the sum of the two regression coefficients from Model 2 (.053 – .008 = .047), but because the covariates are orthogonal the standard error for the between-cluster regression is slightly smaller in Model 3.

Again, how to interpret these results? The within-cluster regression is not biased by the exclusion of cluster-level confounds; therefore, the within-cluster regression better approximates the "true" quasi-causal relation between religiosity and delinquency in the population. In other words, B_W is the key parameter of interest for evaluating a quasi-causal hypothesis. In the current analysis, the quasi-causal relation between religiosity and delinquency appears to be nonexistent: Families who are more religious are less delinquent, but a twin who is more religious than his co-twin is not. Another way of making the same point is to observe that to the extent the relationship is ultimately causal, the structural relation between religiosity and delinquency should not differ depending on whether one is comparing means of twin pairs, deviations of individual twins from their pair mean, or unrelated individuals (see the discussion of rat pups in Turkheimer & Waldron, 2000). These comparisons are invariant causally but differ in their confounds: twin comparisons are confounded only by environmental influences unique to each twin, while comparisons between unrelated individuals are confounded by all genetic and environmental differences between families. Inequality of B_W and B_B, therefore, suggests the operation of third-variable confounds operating between families.

Comparison of within- and between-cluster regression coefficients is a strategy found frequently in the medical literature, with the aim of disentangling "maternal" factors from "fetal origins" of disease aetiology (for a review, see Carlin, Gurrin, Sterne, Morley, & Dwyer, 2005). Mixed effects models of twin data have been productively applied to the study of birth weight and cord blood erythropoietin (Morley, Moore, Dwyer, Owens, Umstad, & Carlin, 2005), birth weight and blood pressure (Mann, De Stavola, & Leon, 2004), and tobacco use and bone density (Hopper & Seeman, 1994), to name just a few examples. This statistical method would be just as productively applied to the study of psychological development as to the study of disease.

Differentiating Genetic and Shared Environmental Confounds

Thus far, we have considered average family religiosity and within-twin pair differences in religiosity in pairs of identical twins. The former effect is confounded by all genetic and environmental factors that vary between families and are systematically associated with religiosity. The latter effect is confounded only by those factors that vary within twin pairs.

Inclusion of DZ twin pairs complicates the analysis of between- and within-pair variances to some extent, but ultimately allows the estimation of a third variance, and with the addition of some statistical and biological assumptions leads to the familiar terms of the classical twin model. In MZ twins, who are identical genetically, within-pair confounds are necessarily environmental in origin. In DZ twins, who share only 50% of their genes, there are both genetic and environmental within-pair confounds. Therefore, to the extent that the relation between religiosity and delinquency is attributable to *genetic* confounds, there should be a larger within-pair effect for DZ twin pairs than for MZ twin pairs. To model this, we can specify an additional interaction term to our mixed effects model:

$$\begin{aligned} Y_{ij} = {} & B_{00} + (B_B \times \bar{X}_{0j}) + (B_W \times (X_{ij} - \bar{X}_{0j})) \\ & + (B_{03} \times ZYG) + (B_{04} \times ZYG \\ & \times (X_{ij} - \bar{X}_{0j})) + u_{0j} + e_{ij} \end{aligned} \tag{8.4}$$

When zygosity (abbreviated ZYG) is dummy-coded as 0 in MZ twins and 1 in DZ twins, B_W estimates the the within-pair regression for MZ twins, whereas B_{04} estimates the difference between MZ and DZ twins in the within-pair effect. To the extent that confounding variables are genetic in origin, the within-pair effect will be larger for DZ twins than MZ twins. In contrast, MZ and DZ twin pairs raised together control equally well for shared environmental factors, thus there will be no difference between pair types if the relevant confounds are environmental in origin. There is no reason to expect a main effect of zygosity on the outcome of interest; however, it is necessary to include the main effect of a covariate included in an interaction term.

Results from this model (Model 4), using data from both MZ and DZ twin pairs, are shown in Table 8.1. See Appendix A at the end of the chapter for the corresponding SAS program. The primary result of Model 4 is the same as Model 3: differences in religiosity between MZ twins do not significantly predict delinquency, inconsistent with a causal hypothesis ($B_W = -0.008$). Second, the within-pair effect is not significantly larger in DZ pairs than in MZ pairs ($B_{04} = 0.027$), indicating that the association between religiosity and delinquency is attributable to *environmental* factors that vary between families. Third, the residual variance shared by MZ twins

(0.314) is larger than the variance shared by DZ twins (0.225), indicating that genetic factors account for residual variation in delinquency not accounted for by religiosity.

In the classical twin study, the three pieces of available information – between- and within-pair variances and zygosity – are re-parameterized to yield the three components of the classical twin model: additive genetic influences (*A*), which are assumed to be perfectly correlated in MZ twin pairs and correlated at 0.5 in DZ twin pairs; shared environmental influences (*C*), which are environmental influences that make twins raised in the same home more similar, regardless of zygosity; and non-shared environmental influences (*E*), which are environmental influences that are unique to each twin and thus contribute to within-pair differences, plus measurement error. Differentiating these sources of variation depends on the relative similarity of MZ and DZ twins for a given phenotype. MZ twins are assumed to share all of their genes *and*, by definition, their shared environment; consequently, to the extent that MZ twins are not perfectly correlated for a phenotype, this is reflected in the estimate for *E*. To the extent that MZ twins, who share all of their genes, are more similar than DZ twins, who are assumed to share 50% of their genes, this will be reflected in the estimate of A. Finally, to the extent that the similarity of DZ twins exceeds half that observed in MZ twins, this will be reflected in the estimate of C.

Although most commonly discussed in terms of genetic and shared environmental contributions to an individual phenotype, we can also apply these same concepts to the *association* between two phenotypes, just as we could with the between and within variances in an analysis of MZ twins. To the extent the relation between religiosity and delinquency is truly causal, one would expect the same causal forces to be operating within families as between them. Our MZ twin analysis, however, indicated that the association between delinquency and religiosity was driven by between-family confounds rather than a "true" quasi-causal effect of religious involvement: Twin pairs who were more religious were less delinquent, but twins who differed in their religiosity did not differ in their delinquency. This model could be extended to ask whether the between-family confounds that drive the religiosity-delinquency association are genetic versus shared environmental in origin. It should be noted that both of these extensions are somewhat tangential to the central point of the analysis: testing whether the quasi-causal hypothesis can be excluded. Nevertheless, this extension may be valuable in more fully characterizing the relation between predictor and outcome.

Assuming that the between-pair variance in MZ twins equals all genetic variance and all shared environmental variance, whereas the twin-pair variance in DZ twins equals one-half the genetic variance plus all shared environmental variance, simple arithmetic yields a more precise estimate for the genetic variance (A = 2 × (MZ Twin Pair Var – DZ Twin Pair Var) = 0.178) and shared environmental variance (C = MZ Twin Pair Var − A = 0.136). The within-pair variance in MZ pairs is a direct estimate of the E variance (0.317).

In summary, random effects models of the relation between an environment and putative psychological outcome in twin pairs can be used to assess the following: (1) the extent to which within-twin pair differences in environmental experience predict differences in the outcome of interest, as would be expected under a causal hypothesis; (2) the extent to which the within-pair regression and the between-cluster regression are different, suggesting the operation of confounding variables; (3) the extent to which the within-pair regression in DZ pairs differs from MZ pairs, which suggests that the relevant confounds are genetic in origin; and (4) the extent to which the residual within- and between-pair variance components differ between MZ and DZ pairs, which suggests that genetic factors account for variation in outcome independent of the predictor of interest. The proposed mixed effects model, therefore, provides a rich characterization of how X and Y are related, with very minimal programming (six lines of SAS code).

Nevertheless, translating parameters that are specified in terms of pair means and deviations or between- and within-cluster variation into the components familiar in behavior geneticists – namely, additive genetics, shared environment, and non-shared environment – requires either post hoc arithmetic of the kind we have used here or rather elaborate weighting schemes (McArdle & Prescott, 2005). Moreover, whereas the current example clearly indicates that the relevant confounds were shared environmental in origin, there will obviously be cases in which the confounding variables are both genetic and environmental, and the proposed random effects model does not explicitly quantify genetic versus shared environmental selection effects. We now turn our attention to an analytic approach more commonly used in behavior genetics that addresses some of these shortcomings: structural equation modeling.

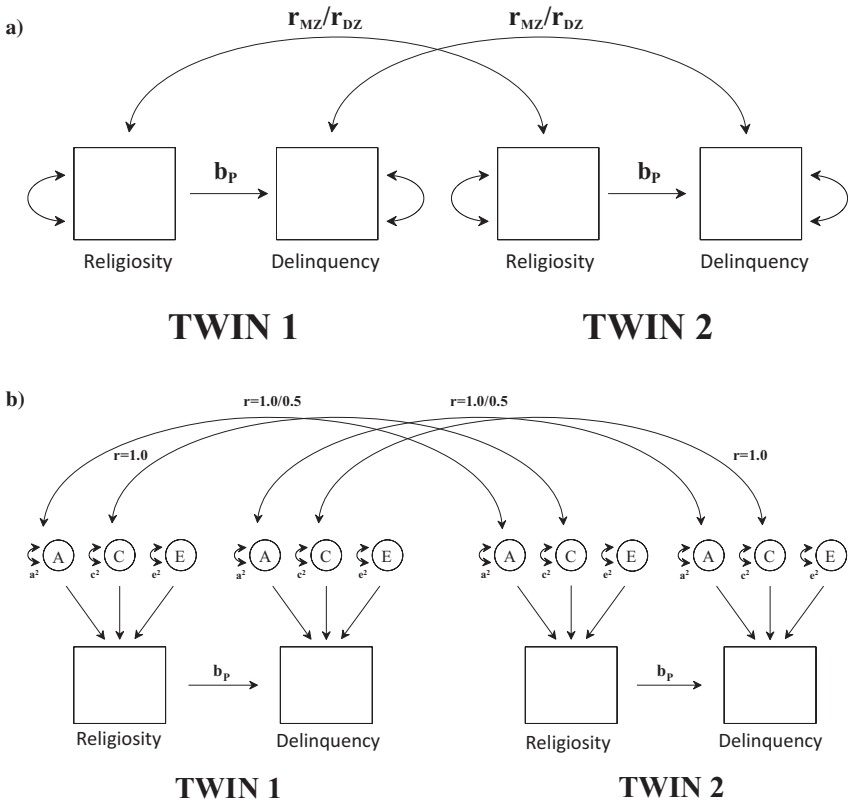

Figure 8.1. Phenotypic regression model in a pair of twins. Delinquency is regressed on religiosity, with equal regression coefficients bP for each member of the pair. Curved paths represent twin correlations for delinquency and religiosity.

STRUCTURAL EQUATION MODELS

Unstandardized ACE Regression

Since the development of powerful and relatively user-friendly structural equation modeling software, such as M*plus* or M*x*, twin data are most commonly analyzed using structural equation models, which, above and beyond offering a graphical means of representing the underlying equations, make it simpler to execute the reparameterizations of variances that we conducted in the preceding section using post hoc arithmetic. Figure 8.1a is a SEM representation of a twin regression, with delinquency predicted from religiosity in twin pairs. As was the case in the random effects regressions employed in the previous section, the dependence of the observations taken from the same twin pair is modeled explicitly, here simply by including estimates of the twin correlations for religiosity and delinquency, separately for MZ and DZ twins. Also similar to the random effect regressions, those MZ and DZ covariances can be re-parameterized as ACE variances, as shown in Figure 8.1b. Notably, the unstandardized regression of delinquency on religiosity, b_P, is the *same* in both figures; partitioning the variance of religiosity and delinquency has no consequences for the regression coefficient between them.

Nevertheless, in bivariate twin data of the kind required to fit the model in Figure 8.1b, one can fit a model, illustrated in Figure 8.2, in which separate regressions are estimated for Y on the three biometric components of X (only one twin is shown). If the true model is represented by Figure 8.1a (i.e., X causes Y), it is easy to see what will happen in a bivariate model. Because the phenotype X is the sum of its three unstandardized components A, C, and E, we will have

$$b_P(A + C + E) = b_P A + b_P C + b_P E \tag{8.5}$$

This shows that when the relationship between X and Y is causal at the level of the phenotype, the

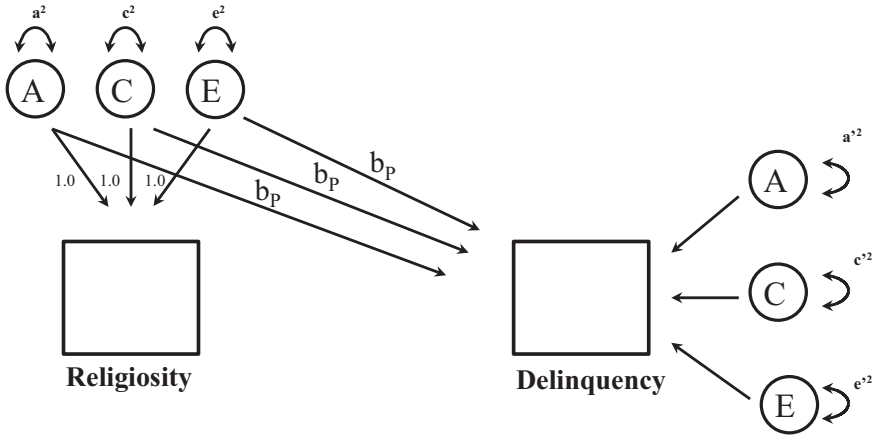

Figure 8.2. Phenotypic regression model in which twin covariances in MZ and DZ twins have been decomposed into biometric components; doing so has no effect on the regression bP.

three unstandardized coefficients in a bivariate ACE regression model will be equal to each other and to the phenotypic causal parameter that underlies them.

The reasons for this equivalence are simple but easily misunderstood, and very important. Intuitions about the parameters of the classical twin model in the context of regression may be better served by temporarily suspending the classical "genes and environment" interpretation of the twin model in which the A term is identified with a latent construct called "the additive effect of genes," the C term is identified as "the shared environment" and the E term as the "nonshared environment" in favor of something more literal, as follows. The classical twin model can use covariances in identical and fraternal twins to partition a phenotype into three components, defined by their covariances between members of twin pairs. One (E) is uncorrelated between members of a pair, one (C) is perfectly correlated between members of a pair, and one (A) is correlated 1.0 in identical twins and 0.5 in fraternal twins. All three are components of the phenotype being analyzed, and their sum, A+C+E, is equal to the phenotype. Understood this way, the invariance of structural regression components in an unstandardized bivariate model of a causal relation is completely unsurprising. If one unit change in the *phenotype* of religiosity causes b_P units of change in delinquency, that will remain the case no matter what component of the phenotype one is examining, and whether one thinks of the origins of the component as genetic or environmental.

The foregoing discussion has not taken into account the problem at the heart of this chapter, namely that individuals are not randomly assigned to religiosity conditions, with the result that religiosity is potentially correlated with a host of unmeasured confounds that will then bias the phenotypic regression of delinquency on religiosity and vitiate the causal interpretation of simple regressions. In a manner entirely analogous to the random effects models developed earlier, with some key assumptions, twin models can provide insight into the nature of the confounding and any causal relations that remain. The key to the analysis is that any unmeasured confounds of religiosity can themselves be partitioned into ACE components, in the same way that confounds in a random effects model can be decomposed into a component correlated with variation in twin pair means and another component correlated with variation within pairs. The ACE components of the unmeasured confounds will then differentially bias the ACE regression coefficients with which they are associated. The model including the unmeasured confounds is illustrated in Figure 8.3

Figure 8.3 is identical to Figure 8.2, except for the addition of unobserved confounds inside the dotted rectangle, and dotted arrows with coefficients b'_A, b'_B and b'_E, which represent the causal effects of the confounds on the outcome. The confounds are perfectly correlated with the predictor of interest, because by definition the predictor and the uncontrolled confounds cannot be distinguished. In actual data, the ACE regression model is fit without reference to the unmeasured confounds, which then bias the unstandardized ACE regressions away from the

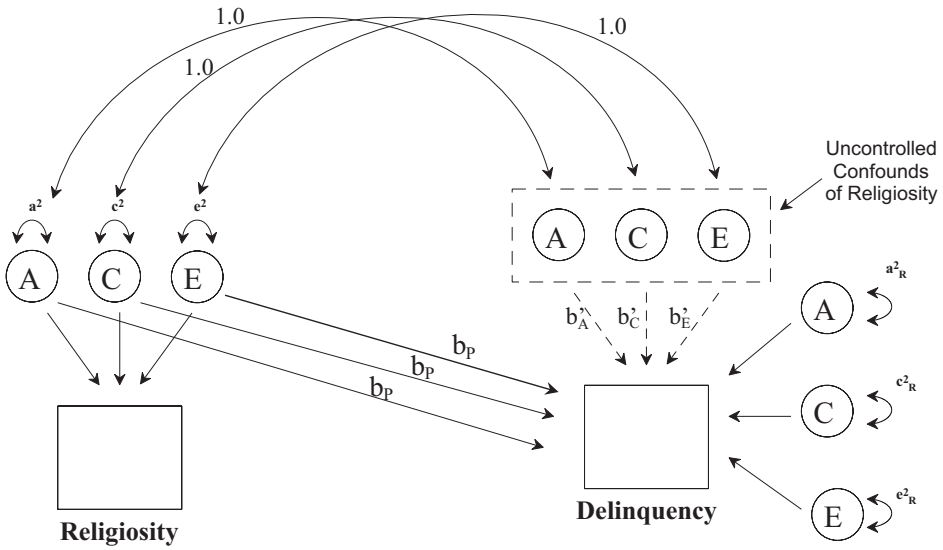

Figure 8.3. Genetically informed phenotypic regression, including unmeasured confounds (only one twin shown). Latent regressions along A, C, and E paths are equal to bP, but observed regressions are biased by unmeasured A and C confounds in dotted box.

true causal value b_P. In the model fit to the observed data,

$$b_A = b_P + b'_A; b_C = b_P + b'_C; b_E = b_P + b'_E \qquad (8.6)$$

This situation would appear to present something of a dilemma, because there are three equations (one for each of the observed ACE regressions) and four unknowns (the true causal parameter and three ACE confounds). This is no small shortcoming, and should serve to remind us of the impossibility of reaching strong causal inferences from nonexperimental data. Nonetheless, the ACE partitioning of the unmeasured confounds offers a partial solution. The A and C terms, representing variation in the confounds shared by identical twins, includes most of the plausible confounds of developmental causal relations. Socioeconomic status, parental education, culture, and, of course, genotype are all shared by identical twins, and therefore are not potential explanations of differences between members of a pair.

In contrast, the E component represents aspects of religiosity and its confounds that are uncorrelated within pairs of identical twins. It is more difficult – not impossible, just more difficult – to posit confounds of the relationship between religiosity and delinquency that are not shared by identical twins. At the very least, we can say that the E regression is free of the many genetic and environmental confounds that identical twins share. The partial solution to the underdetermination of Equation 8.6 (or, alternatively, the twin-based quasi-experimental solution to the unavailability of random assignment in studies of natural variation in humans) is as follows: To the extent we are willing to make the quasi-experimental assumption that relations within identical twin pairs are unconfounded by other variables, which is to say that b'_E is zero, then b_E estimates b_P, the true causal relation, while b_A and b_C estimate the sum of the causal coefficient plus the effects of confounds that vary in each of the domains, respectively.

The E regression estimates the causal effect of X on Y to the extent we can assume that differences within MZ pairs are not confounded by uncontrolled factors. We have already seen why this is a reasonable assumption to make. MZ twins are, to a first approximation, genetically identical, so differences in delinquency associated with differences in religiosity within pairs cannot be attributed to genetic differences between the pairs. When twins are reared together, they are roughly identical for a host of socioeconomic variables that might otherwise explain differences in delinquency: socioeconomic status, place of residence, and so on. It is, however, quite possible to think of uncontrolled third variables that might confound the within-pair association. For example, parents might choose to send one child to a religious school and the other to a public school. The child in the religious

school would become more religious, and might also make friends that predisposed him or her lower levels of delinquency than the twin in the public school, even in the absence of any direct causal effect of religiosity on delinquency. Accepting the E regression as an estimate of the causal effect is therefore an assumption, or one might better say an approximation, of the true state of affairs. In the absence of random assignment, strict inference of causality is simply impossible, and all twins offer is a quasi-experimental method for approaching it. As before, we prefer the term *quasi-causal* to describe a regression of Y on X (in either random effects or structural equation models) that has survived exposure to a genetically informed design that controls for genetic and shared environmental confounds.

The latent E component of religiosity is analogous to the within-pair deviation in religiosity ($x_{ij} - x_j$), entered as a covariate in mixed effects Models 2 – 4, above, and they offer the same inferential benefits and limitations. To understand this connection, consider again the case of MZ twins. They are necessarily identical for additive genetic and shared environmental factors; any difference between MZ twins is reflected in their scores on the latent E variable. That is, a higher score on the latent E variable indicates that an individual twin has higher levels of religiosity, relative to his or her co-twin. Consequently, the b_E path is directly analogous to the within-pair regression coefficient β_W. The b_E path asks, if Twin A has higher levels of religiosity than his or her co-twin (i.e., higher latent E scores), does Twin A also have lower levels of delinquency? Therefore, the regression on E is the key parameter of interest for testing a causal hypothesis about the relation between religiosity and delinquency.

Some readers may find this interpretation of the regression on E to be counterintuitive. Historically, there have been two important obstacles to a proper understanding of quasi-causal relations in behavior genetics, and the crucial role that is played by "nonshared environmental" differences within pairs of MZ twins. First, the three variance components of the classical twin model – additive genetics, shared and unshared environment – have been reified as "genes," "shared environment," and "nonshared environment" for so long that it is easy to forget that ultimately they are all just components of the variable being analyzed, specifically the predictor in a bivariate regression model, which in this case is phenotypic religiosity. That is to say, although according to a fairly restrictive set of assumptions the A component in a classical twin model of religiosity can be thought of in an abstract way as a stand-in for genetic variance in religiosity, the component itself *is* religiosity, *phenotypic* religiosity. The same is true of the C and E terms, which refer not so much to latent shared and nonshared environments underlying religiosity as to a portion of religiosity itself. By naming the latent variance that is not correlated between twins the non-shared environment, behavior geneticists have promoted as environmental what is really just within-pair variation in the phenotype of interest. Although the *origin* of within-twin pair variation is, by definition, environmental influences (and measurement error) that make twins raised in the same family different, the *effect* of within-twin pair variation consists of the effect of X itself, confounded by the effects of any uncontrolled variables that also vary within pairs.

When describing the relation of a latent E variable with other variables in the model, many authors have mistakenly concluded that the E regression estimates only the role of non-shared environmental confounds, and have neglected that this relation is actually the best estimate of the phenotypic causal effects of X itself. Given that detection of this causal effect is usually the primary goal of the study, misinterpreting the term "non-shared" can sometimes snatch defeat from the jaws of victory in an otherwise successful study. For example, Pike, McGuire, Hetherington, Reiss, and Plomin (1996), although implying that their goal was to evaluate a causal hypothesis ("differential treatment affects adolescent adjustment", p. 599), described their results as follows: "mothers' negativity is significantly associated with depressive symptoms through nonshared environmental processes. (p. 597)" A reader could easily interpret this statement as meaning that some *other* environmental process, unique to each sibling, was responsible for both maternal negativity and depressive symptoms, rather than as evidence for a quasi-causal effect of mothers' negativity on depression. Similarly, Spotts, Pederson, Neiderheiser, Reiss, Lichtenstein, Hansson, & Cederblad (2005) described a previous result (Reiss, Neiderhiser, Hetherington, & Plomin, 2000) as follows: "For example, more than half the correlation between mother's positivity and child's social responsibility is accounted for by genetic influences, with the remainder being accounted for by shared and nonshared environmental influences. (p. 339)" This statement makes it seem as though the association has been carved up into various types of confounds – a little attributable to genes, a little to social class or other between-family environmental differences, a

https://doi.org/10.1017/CBO9780511996481.012

little to peer groups or other within-family environmental differences – with nothing left over to comprise a causal relation. In fact, that "non-shared environmental influences" account for the "correlation between mother's positivity and child's social responsibility" means that within-pair differences in maternal positivity predicted within-pair differences in social responsibility – a quasi-causal relation. A third example is a report by McGue, Iacono, and Krueger (2006), whose stated goal was to evaluate whether early adolescent problem behavior is related to adult disinhibitory psychopathology via a causal mechanism, but who devoted a single sentence to describing, without comment, whether the twin with more problem behavior grew up to have more psychopathology than his or her co-twin – "non-shared environmental factors accounted for the remaining 11% and 5% of the correlation [in females and males, respectively]" (p. 599).

MODELING SEQUENCE

We previously described four ways in which the relation between an environment and putative psychological outcome in twin pairs can be characterized, each of which is explicitly characterized in the unstandardized bivariate ACE regression model. First, the extent to which within-twin pair differences in environmental experience predict differences in the outcome of interest, as would be predicted by a causal hypothesis, will be reflected in the b_e path. If b_e can be fixed to zero without significant loss of model fit, this provides disconfirmatory evidence regarding the quasi-causal hypothesis. Second, to the extent that the association between X and Y is confounded by variables that differ between unrelated individuals, the b_e path will differ from the b_a and b_c paths. If the three paths cannot be fixed to equality, this provides evidence that the quasi-causal association is confounded. Third, the relative magnitudes of b_a and b_c indicate the extent to which the confounding variables are genetic or environmental in origin. Finally, the extent to which genetic, shared environmental, and non-shared environmental factors contribute to residual variation in delinquency is directly estimated by the A, C, and E components of delinquency.

Nested SEMs can be compared using two measures of goodness-of-fit, Bayesian Information Criterion (BIC), and Root Mean Square Error of Approximation (RMSEA), as well as differences in $\chi 2$. BIC is an information-theoretic fit criterion that estimates the Bayes factor, the ratio of posterior to prior odds in comparisons of a model with a saturated one (Raftery & Richardson, 1996; Schwarz, 1978). BIC outperforms other fit criteria in its ability to discriminate between multivariate behavior genetic models, particularly for complex model comparisons in large samples, and is more robust to distributional misspecifications (Markon & Krueger, 2004). Interpretation of BIC values is entirely comparative, with lower values of BIC indicating better model fit. RMSEA measures error in approximating data from the model-per-model parameter (Steiger, 1990). RMSEA values of less than 0.05 indicate a close fit, and values up to 0.08 represent reasonable errors of approximation. Browne and Cudeck (1993) have argued that the RMSEA provides very useful information about the degree to which a given model approximates population values. Differences in model $\chi 2$ are themselves distributed as $\chi 2$, with df equal to the difference between the models' df.

We fit a series of five nested models to data from MZ and DZ twins in the program Mplus (Muthén & Muthén, 1998–2010). Results from the full model (Model 5) are summarized in Table 8.2. Of the total variance in religiosity, additive genetic effects accounted for approximately 24% [4.30/(4.30+9.06+4.64)], shared environmental influences for approximately 50%, and non-shared environmental influences for approximately 26%. Similarly, of the unique variance in delinquency, additive genetic effects accounted for approximately 25%, shared environment for 20%, and non-shared environment for 55%. Notice that the estimate for non-shared environmental variance ($E_{Del} = 0.334$) is equal (to the second decimal place) to the within-pair random effect, θ_{MZ}, estimated in the mixed effects Models 2–4. The regression onto E is not significantly different from zero ($b_e = 0.000$, 95% CI = –0.031, 0.30), a result that falsifies the hypothesis that religiosity causes decreases in adolescent delinquency. The regression onto A is also not significantly different from zero ($b_a = 0.043$, 95% CI = –0.056, 0.142), suggesting that genetic pathways are not a significant confound of the quasi-causal relation between religiosity and delinquency. The regression onto C is significantly different from zero ($b_c = 0.061$, 95% CI = 0.020, 0.102), suggesting that environmental circumstances related to familial religiosity, such as parental education or socioeconomic status, account for the association between religiosity and delinquency.

TABLE 8.2. Results from Structural Equation Models of Religiosity and Delinquency

	Model 5: Full Model		Model 6: Regressions Equal		Model 7: No *A* or *E* Regression**	
Parameters	*Unstandardized*	*Standardized*	*Unstandardized*	*Standardized*	*Unstandardized*	*Standardized*
Variance Components of Religiosity						
Additive Genetic	4.30 (1.21)*	$h^2 = 24\%$	4.19 (1.21)*	$h^2 = 23\%$	4.30 (1.21)*	$h^2 = 24\%$
Shared Environmental	9.06 (.121)*	$c^2 = 50\%$	9.15 (1.21)*	$c^2 = 51\%$	9.04 (1.21)*	$c^2 = 50\%$
Non-Shared Environmental	4.64 (.39)*	$e^2 = 26\%$	4.66 (.40)*	$e^2 = 26\%$	4.64 (.39)*	$e^2 = 26\%$
Regression Paths						
$A_{religion} \rightarrow$ Delinquency	.043 (.050)	$\beta = .111$	.036 (.005)*	$\beta = .091$	[0]	[0]
$C_{religion} \rightarrow$ Delinquency	.061 (.021)*	$\beta = .228$	.036 (.005)*	$\beta = .135$	.078 (.013)*	$\beta = .292$
$E_{religion} \rightarrow$ Delinquency	.000 (.016)	$\beta = -.001$	.036 (.005)*	$\beta = .096$	[0]	[0]
Residual Variance Components of Delinquency						
Additive Genetic	.151 (.073)*	$h^2 = 25\%$	.144 (.073)*	$h^2 = 23\%$	.158 (.072)*	$h^2 = 27\%$
Shared Environmental	.124 (.060)*	$c^2 = 20\%$	.136 (.058)*	$c^2 = 22\%$	.103 (.058)	$c^2 = 17\%$
Non-Shared Environmental	.334 (.026)*	$e^2 = 55\%$	.341 (.027)*	$e^2 = 55\%$	.334 (.026)*	$e^2 = 56\%$
Model Fit Indices						
χ^2 (*df*, *P*)	30.36 (17, .02)		42.09 (19, .002)		31.64 (19, .03)	
$\Delta\chi^2$ (Δdf, *P*)	–		11.73 (2, .003)		1.28 (2, .527)	
CFI / TLI	.979 / .985		.963 /. 977		.980 / .987	
RMSEA	.046		.057		.044	

* Significant at $P < .05$.
** Model accepted as the best representation of the data.

Model 6 tested whether environmental and genetic differences between families confound the association between religiosity and delinquency by fixing the regression paths to equality. This model resulted in a significant increase in $\chi 2$ ($\Delta\chi 2 = 11.73$, $\Delta df = 2$, $P = 0.003$) and an increase in RMSEA, suggesting that the association is indeed confounded by between-family differences. The parameter estimates from Model 5, in which the only significant path between religiosity and delinquency was the regression on *C*, suggested that the relevant between-family confounds were shared environmental in origin. Model 7, then, fixed the regressions of delinquency on the *A* and *E* components to zero, allowing the association between religiosity and delinquency to be accounted for entirely by environmental differences between families. Compared to the full model (Model 5), Model 7 did not fit significantly worse ($\Delta\chi 2 = 1.28$, $\Delta df = 2$, $P = 0.53$), indicating that shared environmental confounds could account for the association between delinquency and religiosity.

The same conclusion is apparent, regardless of whether we use mixed effects models or structural equation models – religiosity is not causally related to delinquency, but families whose environmental circumstances make them the "type" of family to be

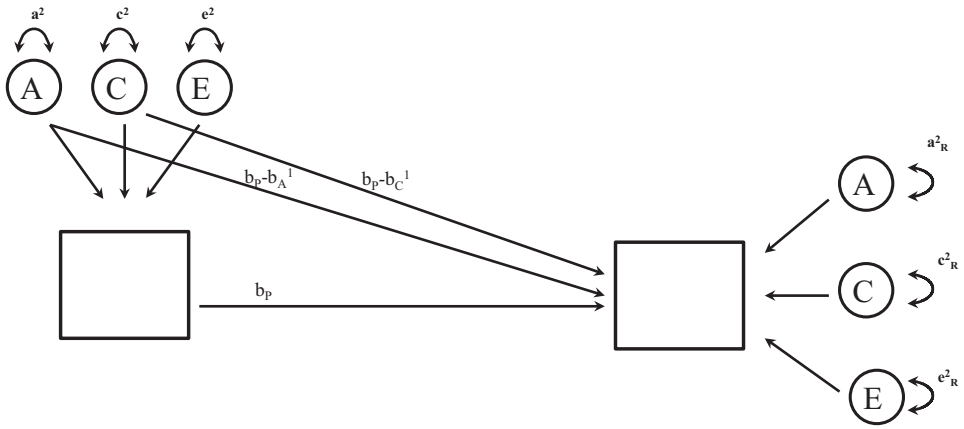

Figure 8.4. Genetically informed phenotypic regression, including unmeasured confounds (only one twin shown). Delinquency is regressed on religiosity along A, C and E pathways. E pathway estimates bP.

religious have less delinquent children. This example, in particular, demonstrates the usefulness of ostensibly "genetic" designs in evaluating hypotheses about *phenotypic* causation in the presence of genetic and environmental confounds. The structural equation models have the added benefit of explicitly parameterizing genetic and shared environmental variance components, as well as separate genetic and shared environmental regressions. However, fitting these structural equation models requires substantially more programming than the mixed effects models (see Appendix B at the end of the chapter for program).

ALTERNATIVE PARAMETERIZATIONS

The genetically informed regression model we have described provides a straightforward test of hypotheses regarding environmental causation, yet a scan of the behavior genetics literature reveals a plethora of alternative parameterizations for bivariate twin or sibling data. While these parameterizations are, in most cases, mathematically indistinguishable (Loehlin, 1996), there are important conceptual differences among them that have implications for interpretative clarity. We include discussion of two reparameterizations that were not included in Loehlin (1996). One of these we recommend, the other not.

Genetically Informed Phenotypic Regression

A re-parameterization of the bivariate regression model that we find to be interpretively useful is illustrated in Figure 8.4. As before, the outcome variable is regressed on the unstandardized A and C components of the predictor, but instead of also being regressed on the E term, it is regressed on the full phenotype of the predictor. This parameterization most closely approximates the substantive goal of research of this kind, which is to examine the phenotypic regression of Y on X while controlling for the between-pair genetic and environmental confounds contained in A and C, respectively.

The interpretation of the coefficients in the genetically informed phenotypic regression model is to multiply through the coefficients for the ACE regression Equation 8.6. Redistributing, we obtain,

$$Y = b_{A'}A + b_{C''}C + (b_P b_{E'})(A + C + E) \qquad (8.7)$$

The sum of *A*, *C*, and *E*, of course, is simply *X*. So in this parameterization, the regression on the phenotype tests for the quasi-causal effect, on the assumption that there are no confounds within pairs of identical twins. The A and C regressions test the difference between the quasi-causal effect and the effects of the A and C confounds, respectively. They are equal to zero when the quasi-causal relation is unconfounded. In this parameterization, the sequence of inferences is to test the phenotypic regression for the quasi-causal effect, then test whether b_A and b_C differ from 0 to test for the presence of confounding, then test the difference between b_A and b_C to determine if confounds are genetic or shared environmental in origin, and finally to test the residual variation in Y, as before.

The results of fitting the genetically informed phenotypic regression model are given in Table 8.3. The phenotypic regression, analogous to the non-shared environmental regression in the previous models, is equal to zero, suggesting that once shared familial

TABLE 8.3. Results from Alternative Parameterization of Bivariate Cholesky (with Phenotypic Path from Religiosity to Delinquency)

Parameters	Model 8: Alternative Full Model (Phenotypic Path) *Unstandardized*	*Standardized*
Variance Components of Religiosity		
Additive Genetic	4.30 (1.21)*	$h^2 = 24\%$
Shared Environmental	9.06 (.121)*	$c^2 = 50\%$
Non-Shared Environmental	4.64 (.39)*	$e^2 = 26\%$
Regression Paths		
$A_{religion} \rightarrow$ Delinquency	.043 (.050)	$\beta = .112$
$C_{religion} \rightarrow$ Delinquency	.061 (.021)*	$\beta = .229$
Religion $\rightarrow$ Delinquency	.000 (.016)	$\beta = -.002$
Residual Variance Components of Delinquency		
Additive Genetic	.151 (.073)*	$h^2 = 25\%$
Shared Environmental	.124 (.060)*	$c^2 = 20\%$
Non-Shared Environmental	.334 (.026)*	$e^2 = 55\%$
Model Fit Indices		
χ^2 (*df*, *P*)	30.36 (17, .02)	
$\Delta\chi^2$ (Δdf, *P*)	–	
CFI / TLI	.979 / .985	
RMSEA	.046	

* Significant at $P < .05$.

confounds have been controlled, there is no reason to hypothesize a quasi-causal relation between religiosity and delinquency. The values of the A and C regressions are the same as they were in the previous models, although this is not the usual result. In general, the A and C regressions in a phenotypic model will be equal to the *differences* between the A and C regressions and the E regression in an ACE regression model, which is to say they measure the magnitude of the bias introduced by the A and C confounds. In this example, however, the quasi-causal parameter is equal to exactly zero, so we conclude that the phenotypic regression is *all* confound, and the A and C regressions are the same in the ACE and phenotypic regression models.

INTERPRETATION AND STANDARDIZATION

The reader experienced in fitting SEM models to twin data may have noticed that we have chosen to parameterize our models by estimating A, C, and E variances while fixing to 1.0 the paths from A, C, and E to the phenotype. We refer to this as an unstandardized parameterization. It is far more common to standardize the A, C, and E components of the predictor variable to a variance of 1, and estimate the paths to the observed variable, which we call a standardized parameterization. In univariate behavioral genetics, the choice between the standardized and unstandardized parameterizations is trivial. In the unstandardized parameterization, the variance accounted for by A, for example, is equal to the estimated A variance; in the standardized parameterization, it is equal to the square of the estimated path. The fit of the models is identical, and in a univariate design they are identical in interpretation.

In the context of multivariate regression models, however, these parameterizations have important conceptual differences. The complication arises from the fact that the variances of the latent components of X have consequences for the regressions of Y on X. Return to Equation 8.7. Under conditions of no confounding, the unstandardized regression b_E estimates the quasi-causal regression b_P, and the three

ACE regressions b_A, b_C, and b_E will be equal to each other. If we consider instead the standardized regressions β_A, β_C, and β_E, these relations no longer hold. We have, instead,

$$\beta_A = b_p \text{var}(A); \beta_c = b_p \text{var}(C); \beta_E = b_p \text{var}(E) \quad (8.8)$$

When the shared and non-shared variances are standardized, the regression of Y on the latent variables no longer depends solely on the structural coefficient relating the phenotypes, but is a function of this coefficient and the magnitude of the *A*, *C*, and *E* components of the predictor variable. Even when the phenotypic relation between *X* and *Y* is invariant – a structural, causally determined property – the magnitude of the latent variances can be expected to vary from population to population and study to study. In particular – as would be considered obvious in a typical regression context – the amount of variance in *Y* accounted for by the ACE components of *X* depends on the relative magnitudes of the ACE components.

Although the issue of standardization may at first seem to be a technical issue of biometric structural equation modeling, in fact it cuts to the heart of the behavior genetic enterprise. We will therefore make the point in a few different ways. One way to understand the difference between standardized and unstandardized regressions is as the difference between structural regression parameters (estimated by unstandardized regression coefficients) and variance explained (estimated by standardized regressions). Consider a concrete physical system that has been designed in a way that a change of one unit of *X* causes a change of two units of *Y*, with no other causes and no error. In any set of observations in which *X* varies, the unstandardized regression of *Y* on *X* will estimate the causal parameter – that is, two. (The standard error of the estimate will increase as the variance of *X* decreases, and of course no estimate can be computed when the variance of X becomes zero.) But what is the amount of variance in *Y* that is accounted for by *X*, or equivalently, the standardized regression coefficient of *Y* on *X*? That quantity is equal to the square of the causal parameter multiplied by the variance of *X*. In conditions of high variability in *X*, it will account for a large amount of variability in *Y*, and in conditions of low variability in *X* it will account for a small amount of variability; but the unstandardized parameter, two units of *Y* per unit of *X*, remains constant regardless of how variance in *X* may change.

The same kind of thinking applies to regressions with error of the kind we are considering here. Suppose that data on religiosity and delinquency were collected in the Netherlands rather than the United States, and that the phenotypic causal relationship between religiosity and delinquency was the same there as here, but for whatever reason in the Netherlands twins varied less in their religiosity. If the unstandardized model were fit, the investigator would gain insight into the similarities and differences between the American and Dutch populations: the quasi-causal relations represented by the unstandardized regressions (presumably the point of the study) would be identical, but the *A*, *C*, and *E* variance components would be generally be different from each other, and different in the Netherlands than they were in the United States. If the standardized model were fit, the structural invariance of the unstandardized parameter would be lost. The standardized regression parameters would be equal to the product of the invariant unstandardized parameters and the respective ACE variances. The investigator would observe that the regression coefficient linking adolescent rule-breaking to the *E* component of religiosity is larger in the Dutch study than in the American one, and, in the opaque language commonly encountered in multivariate behavior genetic research, conclude that the relation between religiosity and rule-breaking appears to be mediated along non-shared environmental pathways.

Another way of understanding the issue is in terms of metrics. When a phenotype is partitioned into unstandardized ACE components with estimated variances, all three components are expressed in the same units of *X*, so regression coefficients of *Y* on *X* are all expressed in the same units, that is, units of *Y* per units of *X*. If the latent components are standardized to have unit variance, however, their metric is no longer in units of *X*, but instead is equal to the standard deviations of each particular component. The A regression is in units of *Y* per standard deviation of *A*, the *C* regression in terms of the standard deviation of *C*, and so forth. In a biometric study, the magnitudes of these standard deviations are going to differ from each other, and across studies they will vary even more. Therefore the standardized regression coefficients relating Y to the biometric components of X will no longer be the same metric, and cannot be meaningfully compared.

Suppose a social psychologist interested in the fundamental attribution error set herself the task of determining the percentage of variance that the fundamental attribution error (FAE; the FAE refers to a tendency

to emphasize internal as opposed to situational explanations of the behavior of other people.) accounts for in some outcome. The research program would be a lost cause, because even assuming a fixed causal effect of the FAE that can be observed across situations, there is no invariant percentage of variance accounted for. In situations in which reliance on the FAE varies a great deal, it will account for a lot of variance in *Y*; in situations where it hardly varies at all, it will account for very little. On the other hand, although the question of how much variance is accounted for by the fundamental attribution error is ill-conceived and not useful scientifically, it would be incorrect to conclude on this basis that the FAE itself was causally unimportant or that its effect could not be quantified. Based on either randomized experimentation or whatever quasi-experimentation can be cobbled together, causation is quantifiable by unstandardized regression coefficients, which are invariant against changes in the variance of *X*. The percentage of variance explained, quantified by standardized regression parameters, is not invariant, even when the underlying causal processes are.

To summarize: The structural constant underlying covariation between a cause and an effect is the unstandardized regression coefficient, expressing the units of *Y* caused by each unit of change in *X*. When regressions are standardized, they no longer estimate this quantity. Instead they estimate the variance in *Y* that is accounted for by *X* (if *Y* is also standardized, the variance explained will be a proportion), a quantity that depends on the structural parameter and the variance of *X*. Questions of how many units of reduction in delinquency are caused by an increase in one unit in religiosity, and whether that value is the same for religiosity itself as for the uncontrolled genetic and environmental traits that confound it, despite their myriad methodological and interpretive complexities, at least have an invariant correct answer that a diligent investigator can hope to estimate. It is true that relying on unstandardized coefficients forces us to take seriously the sometimes arbitrary units in which our constructs are measured, but that is ultimately a good thing, and in any event throwing the units away by standardizing makes regression analyses even harder to interpret. How much variance in delinquency is accounted for by religiosity and its ACE components has no invariant answer: It depends on the variability of delinquency and its biometric decomposition in a particular situation, and is not a worthwhile scientific question.

Heritability

We have saved until last the most important reason for the misapprehension of research methods in behavioral genetics: the concept of heritability, which we have intentionally delayed mentioning until this sentence. We have conducted a reasonably extensive review of behavioral genetic research methods without once referring to the construct that most defenders or critics of the field view as its central idea. (The most recent biologically oriented broadside against behavior genetics [Charney (2012)] refers to the object of its derision as "heritability studies.") A full evaluation of heritability would take us far afield, but note that heritability is the proportion of phenotypic variability accounted for by the total effect of genotype (broad heritability) or by the additive effect of genotype (narrow heritability). Heritability is thus a standardized variance component, and as such it is not invariant as the genetic or phenotypic variance changes, so it is not a meaningful indicator of the causal effect of genotype on phenotype.

A good way to characterize the unstandardized multivariate models we have described in this chapter is that rather than being focused on the estimation of heritability, they are designed to estimate relationships between variables that are *invariant* as regards heritability. Religiosity and delinquency are heritable – everything is heritable or potentially so (Turkheimer, 2000) – and it would not be difficult to estimate heritability coefficients using the models we have described. The goal of our analyses, however, is not to estimate these coefficients, but rather to estimate the part of regression of delinquency on religiosity that is independent of their heritabilities. We want to estimate the extent to which delinquency is causally associated with religiosity above and beyond any variation in genetic background they may share. In the same way, the models control for any shared environmental variability that is common to religiosity and delinquency, again without attending to the magnitude of the shared environmental variance component. The only way to accomplish such invariance is to estimate unstandardized regressions, because the standardized alternative will depend in part on magnitudes of the biometric variances.

MOLECULAR GENETIC APPROACHES

Twenty years ago, most scientists studying the genetics of personality would not have predicted that the most crucial questions regarding causation would involve

complex issues in the design and analysis of twin studies. As the Human Genome Project neared its completion, it was widely anticipated that the availability of data from actual DNA, as opposed to the statistical inferences of quantitative genetics, would provide the causal, biologically based foundation that twin and family studies lacked. From the beginning, personality has played a signature role in the development of what are called "molecular" genetic methods for the study of behavior, providing some of the earliest successes but also some of the greatest frustrations. On the assumption that the molecular genetic methods are not yet as widely incorporated into the general body of research methods in personality, we review some of the basic techniques that are available. The review suggests that, in fact, the problems of scientific inference facing DNA-based studies of behavior turn out to have much in common with traditional family studies: The core scientific problem is still the inference of causation in a nonexperimental setting, and the contrasting of comparisons within and between family members continues to play a crucial role.

Linkage Analysis

The first DNA-based method that was applied to personality, linkage analysis operates within families. The word "linkage" refers the nonindependence of genetic loci that occur close to each other on a chromosome, a phenomenon called "linkage disequilibrium" (LD). In general, genes on different chromosomes are passed on independently, and crossover processes lead to independence for genes well separated on a single chromosome. Genes close together on the same chromosome will tend to be transmitted together, however. If within a family (either a complete pedigree or a pair of siblings and their parents), individuals who share a behavioral trait are also identical by descent (IBD) for a particular gene, it can be inferred that the trait in question is related to the gene or to another gene close to it on the chromosome.

Linkage analysis has been the earliest molecular method to be adopted in the study of behavior because it requires minimal knowledge of actual genetic sequence. The first linkage study of personality to be reported (Benjamin, Press, Maoz, & Belmaker, 1993) looked for linkage between the 16 PF and phenotypic color-blindness in 17 pairs of brothers of whom at least one was color-blind. Because color-blindness was known to be caused by a single X-linked locus, to the extent brothers concordant for color-blindness were more similar for personality, it would locate some of the variance in personality somewhere on the X chromosome. The results of the study foreshadowed much of what has happened since in the molecular genetics of personality. Of the 16 scales of the 16PF, one (Q2, self-sufficiency vs. group adherence) was more similar in the brothers concordant for color-blindness compared to the discordant brothers: The authors conclude with a recommendation that the finding be replicated. The authors do not even mention statistical power, but with 17 sibling pairs it is obviously quite limited. More advanced linkage methods (e.g., Fullerton et al., 2003) use multiple markers to evaluate the probability of linkage continuously across the genome.

Linkage analysis has an important disadvantage as a means of studying personality. Although linkage can be detected with reasonable power using plausible sample sizes for single genes of large effect, it is almost certain that no such genes exist for normal variation in personality. For multiple genes of very small effect, which is just as certainly the situation that does obtain, the power to detect linkage is very small (Risch, 1990). At the same time, when genetic multiple markers are used in combination with multiple potential outcomes, the number of significance tests employed increases rapidly, meaning that both Type 1 and Type 2 error rates are often severely inflated. Another problem is that linkage analysis does not identify a single genetic locus that is associated with a phenotype, but rather a region of a chromosome within which variation in a gene is likely to be in LD with an outcome. Further analysis must be conducted (using association methods described later in the chapter) to establish the particular gene that is responsible for the linkage.

Candidate Gene Studies

In contrast to linkage analysis, which is conducted within families, most *candidate gene* or *association* studies are conducted between families. In its most basic form, the candidate gene study is about as simple as a study can be: A candidate genetic locus is mapped and a personality variable measured in a sample of unrelated individuals, and the outcome of the study is the correlation between the two. Most of the well-known molecular genetic studies of personality have been association studies. In what is still the most widely cited study of this kind, Ebstein et al. (1996) studied the relation between novelty seeking as measured by Tridimensional Personality Questionnaire (Cloninger, Svrakic, & Przybeck, 1993) and the D4 dopamine

receptor gene (DRD4) in a sample 124 normal volunteers. Thirty-four participants who had at least one copy of the seven repeat allele scored about half a standard deviation higher than 90 who did not.

Whether or not the relation between DRD4 and novelty seeking has held up is a matter of meta-analytic controversy that we cannot review here (Kluger, Siegfried, & Epstein, 2002; Munafo, Yalcin, Saffron, & Flint, 2008). What has become crystal clear, however, is that the effect size of half a standard deviation was unrealistically large. In the years since the heady early days of the genome project, it has become clear that associations between individual genetic loci and psychological outcomes are small and context-dependent, even when they appear to be statistically reliable. It has proven extremely difficult to screen reports of association studies for what is known as the "winner's curse" (Xiao & Boehnke, 2009): Faced with thousands of potential alleles, thousands of potential outcomes, intense pressure to publish, and a stringent peer-review system that prefers positive results, even a well-intentioned and honest community of scientists will produce effect sizes that are severely biased upward. The winner's curse is not exclusive to genomics; it is rampant in the behavioral sciences generally. It is just that the modern technology of genomics has brought with it an expectation of scientific rigor and replicability that the social sciences have long since gotten used to doing without (Turkheimer, 2011).

Genome-Wide Association Studies

The Gordian knot of methodology in association studies – myriad hypotheses, small effects, inadequate power, and results that seem to depend on context – has combined with the next wave of technological possibility in genomics to launch a new paradigm of DNA-based research. Genome-wide association studies, or GWAS, capitalize on the existence of single nucleotide polymorphisms (SNPs), individual segments of DNA that take only two values of the four (ATCG) that are possible. Upward of a million SNPs can be inexpensively assessed in a dense array across the genome. SNPs are not genes – they are indicators of genes, and associations between SNPs and outcomes are indicative of corresponding associations with genes at some location on the chromosome close enough to be in LD with the SNP.

With GWAS, it is possible to search for associations between personality and literally millions of locations in the genome. It quickly became apparent that for any trait of psychological interest, any one association would be tiny at best. The extremity of the inferential problems brought on by this situation has led to a radical restructuring of the way GWAS-based science is conducted. When candidate gene studies were first introduced, the theory-dependent process by which genes became candidates seemed like a tonic for the atheoretical gene searches of the early linkage era. But it turned out that the theory by which genes were selected could not be separated from the chaotic problems of the winner's curse and publication bias. GWAS finally made these problems intractable, and the field has reversed course: rather than being guided by theory-driven hypotheses, GWAS is now conducted completely atheoretically, using highly stringent ($p < 10–^{8)}$ significance levels to guard against Type I error. GWAS has produced important discoveries in the medical domain (Visscher & Montgomery, 2009), but it has been disappointing at best in the behavioral sciences, and particularly so for personality (Munafo, Clark, Moore, Payne, Walton, & Flint, 2003).

There is another threat to the validity of association studies, usually identified with GWAS but in fact relevant to all studies of genetic association: population stratification, sometimes called the "chopsticks gene" problem (Hamer, 2000). Here is how one could find a gene associated with the use of chopsticks. Assemble a sample consisting of half North American and half Japanese participants, and identify a gene – any gene will do – that occurs more frequently in the Japanese than in the North Americans. Assuming the Japanese are more likely to use chopsticks, eating style will necessarily be correlated with variation in the gene. Population stratification is usually controlled either by ensuring ethnic homogeneity in the sample or by using statistical methods like principal component analysis to identify ethnic dimensions in the SNPs and then controlling for them statistically. The problem, however, extends beyond the confines of ethnic identification (Turkheimer, 2011). In the context of GWAS, culture of origin is an instance of the core problem that we identified in the analysis of twin studies – a shared environmental confound. In the chopsticks example, the problem is not that the candidate gene and chopsticks may not actually be associated, because their correlation is a simple statistical fact. The problem is that the statistical association is not an indicator of an actual causal pathway from the gene to the chopstick, in exactly the same sense that holds for an association between religiosity and delinquency. In fact, in a mixed sample of Japanese and North American youth in which the Japanese are more religious

and less delinquent, there would be a statistical association that would probably not be indicative of a causal relation.

What can be done to rescue causal inference under such extreme nonexperimental conditions? Resisting the temptation to say "nothing," one interesting possibility is to return to within- and between-families approach that we endorsed as a structure for causal inference in twin studies, and which formed the basis for linkage studies in the early days of molecular genetics. An association between a gene and chopstick use within families (the sibling with the gene used chopsticks; the one without the gene did not) effectively rules out population stratification as an alternative explanation. Designs combining within- and between-families genetic associations and the statistical methods to analyze them are available and predate GWAS (Fulker, Cherny, Sham, & Hewitt, 1999), but have not been widely employed for personality. One exception is a study by Middeldorp, de Geus, Beem, Lakenberg, Hottenga, Slagboom, and Boomsma (2007), which studied the relation between the serotonin transporter gene (5-HTTLPR) and neuroticism, anxiety, and depression in a sample of 1,804 twins, both sibling and parents. Only two of the eighteen within-family tests reached a significance level of $p<.05$, leading the authors to reject the hypothesis of a causal relation between 5-HTTLPR and the outcomes.

The molecular genetics of personality has reached a conundrum. One can design "theory-driven" studies within and between families, which control for a subset of potential confounds of genomic causation, but which are unavoidably contaminated by data exploration and the winner's curse, cherry picking of results, and publication bias. These studies wind up looking like non-genomic social science: locally interesting but frustratingly noncumulative. Or, one can opt for GWAS of massive populations with tiny *p* levels, atheoretical by design and blind to the possibility of noncausal confounds, hoping for a few reliably significant effects that collectively account for a few percent of the variance at best, and which have not, in the behavioral sciences at any rate, produced substantive causal science. What would seem to be the logical compromise – GWAS of enormous samples of siblings – simply isn't practical.

Genome-Wide Complex Trait Analysis

We close this section with a consideration of the newest molecular genetic method. Genome-wide complex trait analysis (GCTA) uses GWAS data in a novel way that closes the methodological gap between quantitative and molecular genetics (Visscher, Yang, & Goddard, 2010; Yang et al., 2010; see also Turkheimer, 2012). In a sample of individuals from whom SNP chips have been obtained, pairwise coefficients are computed that quantify the degree of genetic similarity between pairs of individuals across the SNPs. These coefficients, which are analogous to the coefficients of genetic relatedness in twin and family studies (e.g., siblings are on average 50% genetically related), are generally close to zero, and in fact pairs with coefficients higher than 2.5% are usually eliminated as genetically related. Once the genetic similarity matrix is obtained, one can compute the relationship between the degree of SNP-based genetic similarity and the degree of similarity in the trait of interest, obtaining a proportion of variance that is essentially a heritability coefficient computed in a sample of unrelated individuals. These heritabilities are generally smaller than those computed from family members, but considerably larger than the percentage of variance that can be obtained by adding up the effects of individual SNPs or genes.

Vinkhuyzen et al. (2012) conducted GCTA of neuroticism and extraversion scores in a sample of approximately 12,000 individuals collected from several research centers. Across the centers, the traits had quantitative heritabilities (computed in the usual way) of .4 to .45. In contrast, 6% of the variation in neuroticism and 12% of the variation in extraversion could be explained by SNP-based similarity. These proportions are somewhat smaller than they have been for other traits, like height and intelligence, for which almost half of the phenotypic heritability has been recovered from the SNPs, although 6% and 12% are still significantly more than the 1% that can be recovered by quantifying the effects of individual genes. It should be emphasized that the methodology of GCTA is in fact much more similar to a twin study than it is to a GWAS. No distinction is made between the effects of individual SNPs, and no inference of the causal effects of individual SNPs is even attempted (Turkheimer, 2012).

CONCLUSIONS AND RECOMMENDATIONS

Historically, it is undeniable that behavior genetics has progressed from Galtonian ideas about "nature and nurture," by way of supportable notions of heritability in animal breeding, to a long era of concern, if not

outright obsession, with the values of heritability coefficients for human individual differences. Other than the important task of disconfirming any remnants of blank-slate environmentalism mistakenly held over from previous eras of behaviorism or psychoanalysis, this effort was in our view not especially productive. Heritability is greater than zero for all individual differences, and takes a determinate value for none of them. Figuring out how "genetic" traits are, either in absolute terms or relative to each other, is a lost cause: Everything is genetic to some extent and nothing is completely so. There is little more to be said.

But despite the endless assertions of heritability and the similarly endless denunciations of behavior genetic studies and their conclusions, both of which continue unabated to this day, we contend that heritability was never the most important motivation for human of behavioral genetics. Instead, behavioral genetics is justified by the simple observation that there is more than one reason why differences among people are correlated with each other, either within individual lives or across genetically related individuals. There are genetic as well as environmental reasons why extraverted mothers have extraverted children, or why religiously committed youth are less likely to become delinquent. Any question worth asking about the behavioral genetics of personality comes down to a question about *why* two or more traits are related to each other, and like any other kind of association-based psychology, such questions are ultimately about whether and how one trait causes another. Once that point is conceded, a huge segment of nonexperimental human psychology threatens to collapse unless genetically informative designs can be called on to support it, and such designs turn out not to depend on point estimates of heritability at all; indeed, their correct analysis relies on methods that are invariant with regard to changes in the genetic and environmental variability of individual differences.

Our analysis leads to a several specific recommendations for the conduct of genetically informed research in personality, and we will close by enumerating them.

1. Behavioral genetic investigations of relations among personality variables or between personality and exogenous variables should begin with an observation of a phenotypic association, which will usually be uncontrolled by random assignment. The goal of the genetically informed part of the analysis is to expose the causal basis of the phenotypic association to risk of disconfirmation.
2. Regression-based genetically informative analyses can be conducted more or less equivalently using multilevel models, structural equation models, or a combination of the two. Multilevel models usually have the advantage of being easier to code, whereas structural equation models have the advantage of greater flexibility, especially in their ability to re-parameterize random variances into the familiar biometric components.
3. Although behavior genetic designs are commonly thought of as a means of identifying and controlling genetic effects, shared environmental confounds are often equally important threats to causal hypotheses. If neuroticism is associated with poorer school performance, but living in a violent neighborhood contributes to both, the shared environmental effect of neighborhood is an alternative, noncausal environmental explanation of the phenotypic association.
4. Causal hypotheses are almost always about phenotypic relations among variables, not relations among abstract variance components presumably representing genetic and environmental processes underlying observed behavior. Non-shared environmental regressions are usually the best available estimate of causal relations among uncontrolled variables, because neither genes nor shared environments can account for them. The non-shared environment plays a special role in genetically informed social scientific methodology because it encompasses associations among variables that cannot be accounted for by shared genes or environments, and are thus more plausible instances of phenotypic causation.
5. Notwithstanding the aforementioned, uncontrolled associations within identical twin pairs are not immune from confounds, and behavior genetic methodology is ultimately just another quasi-experimental tool in the social scientific workshop. Once phenotypic associations have survived exposure to analyses of genetic and shared environmental confounds, confidence in the causal relation may increase, but it is not proven. We prefer the term "quasi-causal" to describe the hypothesis that remains.
6. From the point of view of understanding the relationship between behavior genetics and the rest of naturalistic developmental psychology, the inferential imperfection of genetically

informed designs is a good thing. Too often, proponents and detractors of behavior genetics describe the discipline as though it were somehow alien to the rest of social scientific methodology, generating either robustly scientific or falsely reductionist genetic counterhypotheses to psychological theories of human development. Neither is true. Behavior genetics is only a threat to psychological theories in the same sense that the cross-lagged panel design is. Yes, behavior genetic designs can sometimes make it harder to believe in causal hypotheses (Turkheimer & Waldron, 2000), but that is as it should be, and ultimately behavior genetics is no more probative of causal relations than any other quasi-experimental method.

7. To a surprising degree, issues of standardization are crucial to placing behavior genetic methodology on a strong foundation. Since Tukey, it has been well-understood that only unstandardized regression coefficients provide invariant estimates of causal relations in the face of changes in the variances of predictor and outcome, but unfortunately those old insights have been lost in contemporary practice that still relies on correlations and standardized "beta weights." Accounting for variance and explaining causation are two different and ultimately independent enterprises, and science is almost always properly concerned with the latter.
8. As Tukey famously said about correlation coefficients, we believe that the world would be a less confusing and contentious place without heritability coefficients, at least if one is concerned with a more complex and uncontrollable aspect of behavior than, say, milk production in cows. As with the heritability of milk production, the heritability of neuroticism informs us that we could selectively breed for human neuroticism if we wanted to, but fortunately we do not. Genes, in the very abstract sense in which the term is used in human quantitative genetics, influence neuroticism, and this will generally ensure that the heritability of neuroticism is not zero. Beyond that the numerical value of heritability is indeterminate, and the question of "how important" genes are to differences in neuroticism has no meaningful answer.
9. Skepticism about the utility of heritability coefficients should not be a basis for believing that genetic variance in neuroticism does not matter. It does matter, because familial variance in neuroticism and its familial covariance with other traits are alternative explanations of causal hypotheses about its phenotypic origins and consequences.
10. Molecular genetic methods have added to the tools available to behavior geneticists, but they have not replaced twin studies and quantitative genetic statistical methods. Just as we have contended that the goal of twin studies was never to quantify the magnitude of genetic effects on phenotypes, the goal of molecular genetics is not to discover the individual genes that underlie differences in personality. Instead, the goal of both quantitative and molecular genetics is to aid in the identification of causal processes in development, and in that regard molecular genetics faces many of the same problems as quantitative genetics, often in even more intractable forms.
11. Viewed in this way – as a quasi-experimental method that sits alongside many others – behavior genetic research methods can be seen for what they truly are rather than as the threatening or naïve stereotypes that are often represented. Behavior genetics is not a radical, reductive alternative to psychological explanation of behavior, as earlier critics once feared (Lewontin, Rose, & Kamin, 1984), and it is not a poorly specified, dumbed-down version of the astounding understanding of genomics we have achieved at a biological level of analysis, as more recent critics have contended (Charney, 2012). Behavior genetics is not oversimplified genomics any more than nonexperimental developmental psychology is oversimplified developmental biology. Behavior genetics is ordinary social science, with all the problems that come with a lack of experimental control; it is, however, social science conducted a little more carefully, analyzed with a little more realism about why individual differences are associated with each other, and interpreted with a little more skepticism about the vagaries of correlation and causation.

REFERENCES

Alpert, H. (1939). Emile Durkheim and sociologismic psychology. *American Journal of Sociology*, *45*(1), 64–70.

Baier, C. J., & Wright, B. R. E. (2001). "If you love me, keep my commandments": A meta-analysis of the effect of religion on crime. *Journal of Research in Crime and Delinquency*, *38*(1), 3–21.

Benjamin, J., Press, J., Maoz, B., Belmaker, R. H. (1993). Linkage of a normal personality trait to the color-blindness gene: Preliminary evidence. *Biological Psychiatry*, *34*(8), 581–583.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen, & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage Publications.

Burt, S. A. (2009). A mechanistic explanation of popularity: Genes, rule breaking, and evocative gene-environment correlations. *Journal of Personality and Social Psychology*, *96*(4), 783–794.

Carlin, J. B., Gurrin, L. C., Sterne, J. A., Morley, R., & Dwyer, T. (2005). Regression models for twin studies: A critical review. *International Journal of Epidemiology*, *34*(5), 1089–1099.

Cattell, R. B. (1957). *Personality and motivation structure and measurement*. Oxford: World Book Company.

Charney, E. (2012). Behavior genetics and postgenomics. *Behavioral and Brain Sciences*, *35*(5), 331–358.

Cloninger, C. R., Svrakic, D. M., & Przybeck, T. R. (1993). A psychobiological model of temperament and character. *Archives of General Psychiatry*, *50*(12), 975–990.

Ebstein, R. P., Novick, O., Umansky, R., Priel, B., Osher, Y., Blaine, D., Bennett, E. R., Nemanov, L., Katz, M., & Belmaker, R. H. (1996). Dopamine D4 receptor (D4DR) exon III polymorphism associated with the human personality trait of Novelty Seeking. *Nature Genetics*, *12*, 78–80.

Ellison, C. G., & Sherkat, D. E. (1995). The "semi-involuntary institution" revisited: Regional variations in church participation among Black Americans. *Social Forces*, *73*(4), 1415–1437.

Farrington, D. P. (2005). Childhood origins of antisocial behavior. *Clinical Psychology & Psychotherapy*, *12*(3), 177–190.

Federal Bureau of Investigation. (2004). *Crime in the United States: 2004*. Washington, DC: U.S. Government Printing Office.

Fulker, D. W., Cherny, S. S., Sham, P. C. & Hewitt, J. K. (1999). Combined linkage and association sib-pair analysis for quantitative traits. *American Journal of Human Genetics*, *64*(1), 259–267.

Fullerton, J., Cubin, M., Tiwan, H., Wang, C., Bomhra, A., Davidson, S., Miller, S., Fairburn, C., Goodwin, G., Neale, M. C., Fiddy, S., Mott, R., Allison, D. B., & Flint, J. (2003). Linkage analysis of extremely discordant and concordant sibling pairs identifies quantitative-trait loci that influence variation in the human personality trait neuroticism. *American Journal of Human Genetics*, *72*(4), 879–890.

Gosling, S. D., & John, O. P. (1999). Personality dimensions in nonhuman animals: A cross-species review. *Current Directions in Psychological Science*, *8*(3), 69–75.

Hamer, D. (2000). Beware the chopsticks gene. *Molecular Psychiatry*, *5*(1), 11–13.

Harden, K. P., Mendle, J., Hill, J. E., Turkheimer, E., & Emery, R. E. (2008). Rethinking timing of first sex and delinquency. *Journal of Youth and Adolescence*, *37*(4), 373–385.

Harris, K. M. (2011). *Design features of Add Health*. Chapel Hill, NC: Carolina Population Center at the University of North Carolina at Chapel Hill. Available at http://www.cpc.unc.edu/projects/addhealth/data/guide.

Harris, K. M., Halpern, C. T., Smolen, A., & Haberstick, B. C. (2006). The National Longitudinal Study of Adolescent Health (Add Health) twin data. *Twin Research and Human Genetics*, *9*, 988–997.

Hopper, J. L., & Seeman, E. (1994). The bone density of female twins discordant for tobacco use. *New England Journal of Medicine*, *330*(6), 387–392.

Jackson, J. J., Thoemmes, F., Jonkmann, K., Lüdtke, O., & Trautwein, U. (2012). Military training and personality trait development does the military make the man, or does the man make the military? *Psychological Science*, *23*(3), 270–277.

Kazdin, A. E. (1997). Parent management training: Evidence, outcomes, and issues. *Journal of the American Academy of Child and Adolescent Psychiatry*, *36*(10), 1349–1356.

Koenig, L. B., McGue, M., Krueger, R. F., & Bouchard, T. J. (2005). Genetic and environmental influences on religiousness: Findings for retrospective and current religiousness ratings. *Journal of Personality*, *73*(2), 471–488.

Laub, J. H., & Sampson, R. J. (1993). Turning points in the life course: Why change matters to the study of crime. *Criminology*, *31*(3), 301–325.

Lewontin, R. C., Rose, S. R., & Kamin, L. J. (1984). *Not in our genes: Biology, ideology, and human nature*. New York: Pantheon Books.

Littlefield, A. K., Sher, K. J., & Wood, P. K. (2009). Is "maturing out" of problematic alcohol involvement related to personality change? *Journal of Abnormal Psychology*, *118*(2), 360.

Loehlin, J. C. (1996). The Cholesky approach: A cautionary note. *Behavior Genetics*, *26*(1), 65–69.

Mann, V., De Stavola, B. L., & Leon, D. A. (2004). Separating within and between effects in family studies: An application to the study of blood pressure in children. *Statistics in Medicine*, *23*(17), 2745–2756.

Markon, K. E., & Krueger, R. F. (2004). An empirical comparison of information-theoretic selection criteria for multivariate behavior genetic models. *Behavior Genetics*, *34*(6), 593–610.

McArdle, J. J., & Prescott, C. A. (2005). Mixed-effects variance components models for biometric family analyses. *Behavior Genetics*, *16*(1), 163–200.

McGue, M., Iacono, W. G., & Krueger, R. (2006). The association of early adolescent problem behavior and adult psychopathology: A multivariate behavioral genetic perspective. *Behavior Genetics*, *36*(4), 591–602.

Middeldorp, C. M., deGeus, E. J. C., Beem, A. L., Lakenberg, N., Hottenga, J., Slagboom, P. E., & Boomsma, D. I. (2007). Family based association analyses between the serotonin transporter gene polymorphism (5-HTTLPR) and neuroticism, anxiety and depression. *Behavior Genetics*, *37*(2), 294–301.

Miles, D. R., & Carey, G. (1997). Genetic and environmental architecture of human aggression. *Journal of Personality and Social Psychology*, *72*(1), 207–217.

Moffitt, T. E. (1993). Adolescence-limited and life-course-persistent antisocial behavior: A developmental taxonomy. *Psychological Review*, *100*(4), 674–701.

Morley, R., Moore, V. M., Dwyer, T., Owens, J. A., Umstad, M. P., & Carlin, J. B. (2005). Association between erythropoietin in cord blood of twins and size at birth: Does it relate to gestational factors or to factors during labor or delivery? *Pediatric Research*, *57*, 680–684.

Munafo, M. R., Clark, T. G., Moore, L. R., Payne, E., Walton, R., & Flint, J. (2003). Genetic polymorphisms and personality in healthy adults: A systematic review and meta-analysis. *Molecular Psychiatry*, *8*(5), 471–484.

Munafo, M. R., Yalcin, B., Saffron, A. W., & Flint, J. (2008). Association of the dopamine D4 receptor (DRD4) gene and approach-related personality traits: Meta-analysis and new data. *Biological Psychiatry*, *63*(2), 197–206.

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

Neale, M. C., & Maes, H. H. (2007). *Methodology for genetic studies of twins and families*. Dordrecht, The Netherlands: Kluwer Academic.

Oltmanns, T. F., & Turkheimer, E. (2009). Person perception and personality pathology. *Current Directions in Psychological Science*, *18*(1), 32–36.

Pearce, L. D., & Axinn, W. G. (1998). The impact of family religious life on the quality of mother-child relations. *American Sociological Review*, *63*(6), 810–828.

Penke, L., Denissen, J. J. A., & Miller, G. F. (2007). The evolutionary genetics of personality. *European Journal of Personality*, *21*(5), 549–587.

Pike, A., McGuire, S., Hetherington, E. M., Reiss, D., & Plomin, R. (1996). Family environment and adolescent depressive symptoms and antisocial behavior: A multivariate genetic analysis.*Developmental Psychology*, *32*(4), 590–603.

Plomin, R., DeFries, J. C., Knopik, V. S., & Neiderhiser, J. M. (2013). *Behavioral genetics* (6th ed.). New York: Worth Publishers.

Quinn, P. D., Stappenbeck, C. A., & Fromme, K. (2011). Collegiate heavy drinking prospectively predicts change in sensation seeking and impulsivity. *Journal of Abnormal Psychology*, *120*(3), 543.

Quinsey, V. L., Skilling, T. A., Lalumière, M. L., & Craig, W. M. (2004). *Juvenile delinquency: Understanding the origins of individual differences*. Washington, DC: American Psychological Association.

Raftery, A. E., & Richardson, S. (1996). Model selection for generalized linear models via GLIB, with application to epidemiology. In D. A. Berry & D. K. Stangl (Eds.), *Bayesian biostatistics* (pp. 321–354). New York: Marcel Dekker.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage Publications.

Reiss, D., Neiderhiser, J. M., Hetherington, E. M., & Plomin, R. (2000). *The relationship code: Deciphering genetic and social influences on adolescent development*. Cambridge, MA: Harvard University Press.

Risch, N. (1990). Linkage strategies for genetically complex traits. II. The power of affected relative pairs. *American Journal of Human Genetics*, *46*(2), 229–241.

Sampson, R. J., Morenoff, J. D., & Gannon-Rowley, T. (2002). Assessing "neighborhood effects": Social processes and new directions in research. *Annual Review of Sociology*, *28*, 443–478.

Schwarz, G. (1978). Estimating the dimension of a model.*The Annals of Statistics*, *6*(2), 461–464.

Scott, J. P., & Fuller, J. L. (1965).*Genetics and the social behavior of the dog*. Chicago: University of Chicago Press.

Shipley, B. A., Weiss, A., Der, G., Taylor, M. D., & Deary, I. J. (2007). Neuroticism, extraversion, and mortality in the UK Health and Lifestyle Survey: A 21-year prospective cohort study. *Psychosomatic Medicine*, *69*(9), 923–931.

Smith, D. N. (2003).*Hinduism and modernity*. Malden, MA: Blackwell Publishing.

Spitz, E., Moutier, R., Reed, T., Busnel, M. C., & Marchaland, C. (1996). Comparative diagnoses of twin zygosity by SSLP variant analysis, questionnaire, and dermatoglyphic analysis. *Behavior Genetics*, *26*(1), 55–63.

Spotts, E. L., Pederson, N. L., Neiderhiser, J. M., Reiss, D., Lichtenstein, P., Hansson, K., & Cederblad, M. (2005). Genetic effects on women's positive mental health: Do marital relationships and social support matter? *Journal of Family Psychology*, *19*(3), 339–349.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25*(2), 173–180.

Tellegen, A., Lykken, D. T., Bouchard, T. J., Wilcox, K. J., Segal, N. L., & Rich, S. (1988). Personality similarity in twins reared apart and together. *Journal of Personality and Social Psychology*, *54*(6), 1031–1039.

Tukey, J. W. (1954). Causation, regression, and path analysis. In O. Kempthorne, T. A. Bancroft, J. W. Gowen, & J. L. Lush (Eds.), *Statistics and mathematics in biology* (pp. 35–66). Ames: Iowa State University Press.

Tukey, J. W. (1969). Analyzing data: Sanctification or detective work*? American Psychologist*, *24*(2), 83–91.

Turkheimer, E. (2000). Three laws of behavior genetics and what they mean. *Current Directions in Psychological Science*, *9*(5), 160–164.

Turkheimer, E. (2011). Genome wide association studies of behavior are social science. In K. S. Plaisance & T. A.

C. Reydon (Eds.), *Philosophy of behavioral biology* (pp. 43–64). New York: Springer.
Turkheimer, E. (2012). Still missing. *Research in Human Development*, *8*(3–4), 227–241.
Turkheimer, E., & Waldron, M. (2000). Nonshared environment: A theoretical, methodological, and quantitative review. *Psychological Bulletin*, *126*(1), 78–108.
Vinkhuyzen, A. A. E., Pedersen, N. L., Yang, J., Lee, S. H., Magnusson, P. K. E, et al. (2012). Common SNPs explain some of the variation in the personality dimensions of neuroticism and extraversion. *Translational Psychiatry*, *2*, e102. doi:10.1038/tp.2012.27
Visscher, P. M., & Montgomery, G. W. (2009). Genome-wide association studies and human disease: From trickle to flood. *Journal of the American Medical Association*, *302*(18), 2028–2029.
Visscher, P. M., Yang, J., & Goddard, M. E. (2010). A commentary on 'common SNPs explain a large proportion of the heritability for human height.' *Twin Research and Human Genetics*, *13*, 517–524.
Wallace, J. M., Brown, T. N., Bachman, J. G., & Laveist, T. A. (2003). The influence of race and religion on abstinence from alcohol, cigarettes and marijuana among adolescents. *Journal of Studies on Alcohol*, *64*(6), 843–848.
Wilcox, W. B. (1998). Conservative protestant childrearing: Authoritarian or authoritative? *American Sociological Review*, *63*(6), 796–809.
Wilson, J., & Sherkat, D. E. (1994). Returning to the fold. *Journal for the Scientific Study of Religion*, *33*(2), 148–161.
Xiao, R., & Boehnke, M. (2009). Quantifying and correcting for the winner's curse in genetic association studies. *Genetic Epidemiology*, *33*(5), 453–462.
Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K. et al. (2010).Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, *42*(7), 565–569.

APPENDIX A: SAS CODE

`f11` = Delinquency factor score at Wave I
`religsum` = Individual religiosity sum score
`religavg` = Twin pair average religiosity
`religdev` = Individual deviation from twin pair average religiosity
`MZ` = 0 or 1; Monozygotic twin pair

MODEL 1

```
proc mixed data=mztwin12 method= ml covtest noclprint;
class pair;
model f11 = religsum / solution;
random intercept / subject=pair type=vc;
run;
```

MODEL 2

```
proc mixed data=mztwin12 method= ml covtest noclprint;
class pair;
model f11 = religsum religavg / solution;
random intercept / subject=pair type=vc;
run;
```

MODEL 3

```
proc mixed data=mztwin12 method = ml covtest noclprint;
class pair;
model f11 = religdev religavg / solution;
random intercept / subject=pair type=vc;
run;
```

MODEL 4

```
proc mixed data=twin12 method = ml covtest noclprint;
class pair MZ twinid;
```

```
model f11 = religdev religavg religdev*MZ / solution;
random intercept / subject=pair group=MZ type=vc;
repeated / group=MZ type=vc;
run;
```

SAS OUTPUT FOR MODEL 4

```
The SAS System      12:43 Thursday, December 1, 2011  75
                                   The Mixed Procedure
                                    Model Information
                      Data Set                      WORK.TWIN12
                      Dependent Variable            f11
                      Covariance Structure          Variance
Components
                      Subject Effect                PAIR
                      Group Effects                 MZ, MZ
                      Estimation Method             ML
                      Residual Variance Method      None
                      Fixed Effects SE Method       Model-Based
                      Degrees of Freedom Method     Containment
                                        Dimensions
                                Covariance Parameters
4
                                Columns in X
5
                                Columns in Z Per Subject
2
                                Subjects
644
                                Max Obs Per Subject
2
                                     Number of Observations
                           Number of Observations Read
1370
                           Number of Observations Used
1286
                           Number of Observations Not Used
84
                                      Iteration History
                    Iteration    Evaluations        -2 Log Like
Criterion
                            0              1      3035.96177444
                            1              2      2906.05165136
0.00000000
                                  Convergence criteria met.
                                Covariance Parameter Estimates
                                                          Standard
Z
            Cov Parm      Subject    Group    Estimate       Error
Value       Pr > Z
            Intercept     PAIR       MZ 0       0.2112     0.03214
```

```
(CONTINUED ON NEXT PAGE)
6.57      <.0001
          Intercept     PAIR       MZ 1       0.3354     0.04633
7.24      <.0001
          Residual                 MZ 0       0.3889     0.02778
14.00      <.0001
          Residual                 MZ 1       0.3172     0.02831
11.20      <.000

The SAS System      12:43 Thursday, December 1, 2011  76
                                    The Mixed Procedure
                                      Fit Statistics
                               -2 Log Likelihood
2906.1
                               AIC (smaller is better)
2922.1
                               AICC (smaller is better)
2922.2
                               BIC (smaller is better)
2957.8
                                     Solution for Fixed Effects
                                                Standard
             Effect          MZ    Estimate       Error      DF
t Value    Pr > |t|
             Intercept             -0.4036     0.06920     641
-5.83      <.0001
             religdev             -0.00848     0.01681     641
-0.50      0.6138
             religavg              0.05042    0.006817     641
7.40      <.0001
             religdev*MZ    0      0.02717     0.02068     641
1.31      0.1894
             religdev*MZ    1            0           .       .
.         .
                                  Type 3 Tests of Fixed Effects
                                         Num     Den
                           Effect         DF      DF    F Value
Pr > F
                           religdev        1     641       0.24
0.6221
                           religavg        1     641      54.69
<.0001
                           religdev*MZ     1     641       1.73
 0.189
```

APPENDIX B: MPLUS CODE

```
data: file = model6.txt;
variable:
names = pair zygo f11a f11b religa religb;
missing =.;
grouping = zygo (1=mz 2=dz);
```

```
usevariable = religa religb f11a f11b;
analysis:
type = missing h1;
model = nocovariances;
model:
a11 by religa@1;
c11 by religa@1;
e11 by religa@1;
a21 by religb@1;
c21 by religb@1;
e21 by religb@1;
[religa* religb*] (relmean);
[a11-e21@0];
a11*4 (amz);
c11*8 (c);
e11*4 (e);
a21*4 (amz);
c21*8 (c);
e21*4 (e);
a11 with a21*4 (amz);
c11 with c21*8 (c);
religa@0;
religb@0;
religa with religb@0;
f11a on a11*.01 (areg)
c11*.04 (creg)
e11*.01 (ereg);
f11b on a21*.01 (areg)
c21*.04 (creg)
e21*.01 (ereg);
a12 by f11a@1;
c12 by f11a@1;
e12 by f11a@1;
a22 by f11b@1;
c22 by f11b@1;
e22 by f11b@1;
[f11a* f11b*] (delmean);
[a12-e22@0];
a12* (xmz);
c12* (y);
e12* (z);
a22* (xmz);
c22* (y);
e22* (z);
a12 with a22* (xmz);
c12 with c22* (y);
f11a@0;
f11b@0;
f11a with f11b@0;
model dz:
a11 with a21*4 (adz);
```

```
a12 with a22* (xdz);
model constraint:
amz= 2*adz;
xmz =2*xdz;
output:
sampstat tech1 cinterval standardized;
```