Estimating Two-Stage Models for Genetic Influences on Alcohol, Tobacco or Drug Use Initiation and Dependence Vulnerability in Twin and Family Data

Andrew C. Heath[1], Nicholas G. Martin[2], Michael T. Lynskey[1,2], Alexandre A. Todorov[1], and Pamela A. F. Madden[1]

[1]*Department of Psychiatry, Washington University School of Medicine, St Louis, U.S.A.*
[2]*Division of Epidemiology and Population Health, Queensland Institute of Medical Research, Brisbane, Australia*

Genetic research on risk of alcohol, tobacco or drug dependence must make allowance for the partial overlap of risk-factors for initiation of use, and risk-factors for dependence or other outcomes in users. Except in the extreme cases where genetic and environmental risk-factors for initiation and dependence overlap completely or are uncorrelated, there is no consensus about how best to estimate the magnitude of genetic or environmental correlations between Initiation and Dependence in twin and family data. We explore by computer simulation the biases to estimates of genetic and environmental parameters caused by model misspecification when Initiation can only be defined as a binary variable. For plausible simulated parameter values, the two-stage genetic models that we consider yield estimates of genetic and environmental variances for Dependence that, although biased, are not very discrepant from the true values. However, estimates of genetic (or environmental) correlations between Initiation and Dependence may be seriously biased, and may differ markedly under different two-stage models. Such estimates may have little credibility unless external data favor selection of one particular model. These problems can be avoided if Initiation can be assessed as a multiple-category variable (e.g. never versus early-onset versus later onset user), with at least two categories measurable in users at risk for dependence. Under these conditions, under certain distributional assumptions, recovery of simulated genetic and environmental correlations becomes possible. Illustrative application of the model to Australian twin data on smoking confirmed substantial heritability of smoking persistence (42%) with minimal overlap with genetic influences on initiation.

Genetic research on risk of dependence on alcohol, tobacco or illicit drugs must take account of the fact that some family members, with no history of use of a particular drug, or perhaps with only a minimal level of exposure, have therefore never been at risk for dependence on that drug. As emphasized in a pioneering paper by Eaves (Eaves & Eysenck, 1980), it is not appropriate simply to make an a priori decision to exclude or include family members with no history of use. If the same genetic and environmental factors that determine variation in risk of dependence among users also determine risk of initiation, then excluding non-users will discard genetic information and, more importantly, in analyses of twin data will lead to biased estimates of genetic and environmental effects on risk of dependence. Conversely, if genetic and environmental influences on risk of becoming a user are uncorrelated with genetic and environmental influences on risk of dependence in those who have become users, including non-users as non-dependent individuals would confound two traits having different modes of inheritance. Using twin data on smoking initiation (ever been a regular smoker) and persistence (whether or not the respondent was still smoking at the time of assessment), Eaves and Eysenck (1980) showed how it is possible to test genetic models for these two extreme cases of a single liability (or 'risk') dimension, or two orthogonal liability dimensions. The models that they presented are potentially applicable not only to genetic analyses of smoking persistence, but also to a broad array of other drug use outcomes, such as dependence, quantity used, or development of a withdrawal syndrome.

A practical limitation of this early work is that there are many risk-factors that might plausibly be expected to influence both initiation of use and drug use outcomes (e.g. a history of antisocial behavior), and that it is also highly likely that there are genetic influences (e.g. relating to drug metabolism) or environmental influences (e.g. quitting by a spouse or partner) that influence the outcomes of drug use but have no influence on whether or not a drug is used in the first place (e.g. Hawkins et al., 1992; Newcomb & Bentler, 1989). Dissecting common versus specific influences on drug use initiation versus outcome measures is important from many perspectives. For the purposes of gene-mapping studies of drug dependence, it would be important to know whether genetic influences observed for drug dependence reflect genetic influences on differences among users in dependence vulnerability, or merely genetic influences on initiation of use that might be explained by personality or other heritable risk-factors that predate the

Address for correspondence: A. C. Heath, Missouri Alcoholism Research Center, Department of Psychiatry, Washington University School of Medicine, 40 N. Kingshighway, Suite One, St Louis, MO 63108, USA. Email:andrew@matlock.wustl.edu

onset of use. For the purposes of prevention research, it would be important to understand whether genotype x environment interaction effects (Heath et al., 2002), whereby the effects of high genetic risk are moderated by an environmental protective factor ('risk-modifier'), are arising through influences on genetic effects associated with initiation of use that may also influence drug use outcomes (e.g. impulsive personality traits), versus genetic effects that specifically influence outcomes in those who have become users.

Attempts to develop a more general genetic model for drug use initiation and outcomes have not been entirely satisfactory. It is not possible to fit a traditional bivariate genetic model to estimate genetic and environmental variances for a binary measure of initiation (never user or ever user), genetic and environmental variances for the outcome (e.g. persistent smoker versus successful quitter, or non-dependent versus alcohol, tobacco or drug dependent), and genetic and environmental correlations between initiation and outcome: the same person cannot be simultaneously a non-user and a persistent or dependent user, hence it is not possible to estimate a correlation between non-shared environmental effects (Heath & Martin, 1993). Heath and Martin (1993) proposed a model that sought to include single liability dimension and orthogonal liability dimension models as special cases. This model has been applied successfully to twin data to show that for measures of smoking initiation and persistence, neither of the extreme cases considered by Eaves and Eysenck were supported (Heath & Martin, 1993; Heath & Madden, 1995). However, this model was unsatisfactory in that if the single liability dimension model was rejected, the assumption of uncorrelated genetic and uncorrelated environmental influences on initiation versus outcome liability dimensions was retained, with a measurement model (allowing some users to be classified as equivalent to never users) used to relax the orthogonality of influences on initiation versus outcome. As an alternative approach, Kendler et al. (1999) have fitted to data on smoking initiation and nicotine dependence a model that may be considered a special case of a direction-of-causation model (cf. Heath, Kessler et al., 1993), modeling genetic and environmental influences on initiation, a unidirectional causal path from initiation to nicotine dependence, and also genetic and environmental influences specific to nicotine dependence. This approach has subsequently been applied by other investigators to data on smoking initiation and persistence (Madden et al., 1999). This model embodies strong assumptions that, if false, may lead to biased inferences about the interrelationship between genetic influences on initiation versus outcome. For example, it implies that if there are shared environmental influences on initiation of use (as has commonly been observed for smoking: Heath & Madden, 1995), then these same shared environmental influences must necessarily also have an effect, albeit attenuated, on dependence risk in those who have become users.

In this paper we revisit the question of how best to model genetic influences on alcohol, tobacco or drug use initiation versus other drug use outcomes. We argue that the unsatisfactory nature of previous models has stemmed from the attempt to operationalize initiation as a purely binary construct. In those cases where it is possible to characterize initiation as a multiple category trait (e.g. never used versus late onset user versus early onset user), the problems of model under-identification that have prevented fitting of a full bivariate genetic model can be overcome, allowing accurate estimation of correlations between genetic effects (or environmental effects) on initiation versus dependence or other outcomes. Futhermore, it will also be possible to fit multivariate genetic factor models that include hypothesized genetic correlates of initiation versus drug use outcome (e.g. hypothesized personality risk-factors: Cloninger (1987)), or to include potential mediators of genetic effects as control variables and test for residual genetic influences on initiation and on outcome, using standard statistical software for genetic model-fitting (e.g. MX: Neale et al., 1999).

Method

Model

We begin by assuming a bivariate probit model for risk-factor effects on drug use initiation ('Initiation': defined as having at least 3 categories, at least two of which encompass individuals who can be assessed on the second outcome dimension), and on drug use outcome ('Outcome': e.g. dependence, or persistent smoking). For simplicity, we consider a binary Outcome measure, although to maximize statistical power it would be preferable to define a quantitative or multiple-category Outcome measure where this can be justified empirically. Thus for a sample of unrelated individuals, as in the estimation of a polyserial correlation between two variables (e.g. Joreskog & Sorbom, 1997), we assume normally distributed Initiation and Outcome liability dimensions, which may be correlated. Individuals with liability values less than s_0 on the initiation dimension never become users; between s_0 and s_1 are users with low levels of risk-factors for initiation (e.g. are late-onset users); and above s_1 are users with higher levels of risk-factors for initiation (e.g. are early-onset users). Whether 'late-onset' use and 'early-onset' use adequately characterizes individuals with low versus high levels of risk-factors for Initiation, and can be placed on the same normal liability dimension as non-users, will need to be tested empirically, as discussed below. Individuals with liability values less than t_0 on the Outcome dimension will not become dependent if they become users, whereas individuals with liability values greater than or equal to t_0 will become dependent. For simulations, we assumed that 33.4% of individuals were never users, 33.3% of individuals were late onset users, and 33.3% of users were early-onset users, corresponding to threshold values for a standardized normal distribution of $s_0 = -0.4289$ and $s_1 = 0.4316$. We also assumed a threshold value for the outcome dimension of $t_0 = 0$, implying that if the entire population were users, or if Initiation and Outcome liability dimensions were uncorrelated, 50% of users would become dependent.

Figure 1 summarizes the two-way contingency table, cross-classifying Initiation status and Outcome status, that is implied by our model, including explicitly two cells of

the table that are not observable, i.e. may be considered to summarize structural missing data, because never-users cannot be characterized on the Outcome dimension. The cell frequencies predicted for these two cells, and therefore the predicted prevalence of observed cases of dependence under this model, will depend upon the magnitude of the correlation between the Initiation and Outcome liability dimensions (the so-called 'polychoric correlation': Joreskog & Sorbom, 1989). For polychoric correlations of 0.2, 0.4 and 0.6, respectively, corresponding expected proportions of observed dependent cases in the population may be derived, by integrating the bivariate normal distribution, as 54.3%, 58.9% and 63.8% respectively. For simulated data-sets, we assumed a polychoric correlation between initiation and dependence dimensions of either 0.6 or, in one case, 0.3.

The missing data structure implied by this model has the property that data are Missing at Random (MAR) in the technical sense defined in statistical treatments of missing data (e.g. Little & Rubin, 1987). In this case, the probability of structural missing data on the second, Outcome variable is solely determined by values on the first observed variable, Initiation, being 100% for never users, 0% otherwise. Provided that (i) there are at least two categories on the Initiation dimension for which data are available on the second Outcome dimension, (ii) the missing data are indeed MAR, and (iii) a bivariate normal liability distribution can be assumed for the Initiation and Outcome dimensions, enough information is available to estimate the polychoric correlation between these dimensions. If the two liability dimensions are uncorrelated, then whether a respondent falls in the 'late onset' or 'early onset' categories on the Initiation dimension conveys no information about whether that person will be non-dependent or dependent on the second Outcome dimension. As the correlation between the two dimensions increases, the relative risk of dependence in 'early onset' compared to 'late onset' users will increase. An implication of this is that statistical software that is designed to handle data that are MAR, such as MX (Neale et al., 1999) or MPlus (Muthen & Muthen, 1998), when presented with two-way frequency tables corresponding to Figure 1, with a missing data code used to indicate missing data for the Outcome variable in those coded 0 on the Initiation variable, will appropriately recover the true polychoric correlation between initiation and outcome (simulated as 0.6 in our data), and threshold value for the outcome dimension (simulated as 0.0). Appendix 1 gives an example MX script and input data file for this case. In contrast, ignoring the structural missing data by using standard software such as PRELIS to estimate tetrachoric correlations will lead to biased estimates of the polychoric correlation (e.g., using the threshold values of Figure 1, for simulated values of 0.6, 0.4 or 0.2, estimated values of the polychoric correlation of 0.446, 0.283 and 0.1378 are obtained) and of the threshold value for the outcome dimension (instead of a simulated value of 0.0, estimated values of –0.3535, –0.2255 and –0.1101 respectively). By analyzing Initiation and Outcome dimensions jointly with other predictor variables, using standard LISREL-type models, software such as MX or MPlus can be further used to test structural equation models for relationships between drug use initiation and outcomes and other hypothesized predictors. Except for the case of genetic models, we shall not consider such applications further here.

(i) Incorporation of Covariates

The model of the previous section is an intercept-only bivariate probit model, in the sense that it estimates only threshold values and the correlation between variables, without control for potential covariates. Statistical packages such as MX make it possible to include control variables by modeling mean liability as a function of covariates that will vary from individual to individual. In the context of research on variables such as smoking persistence which might be expected to show strong age effects, such control variables might include dummy variables for age category, to adjust for age differences in risk of being a continued smoker. We simulated data with 3 uncorrelated binary covariates, each with 50% prevalence, and a residual correlation between Initiation and Outcome dimensions of 0.6. Probit regression coefficients for the Initiation dimension were 0.2, 0.4 and 0; and those for the Outcome dimension 0, 0.2, 0.4. For the Initiation dimension, with no missing data, these regression coefficients

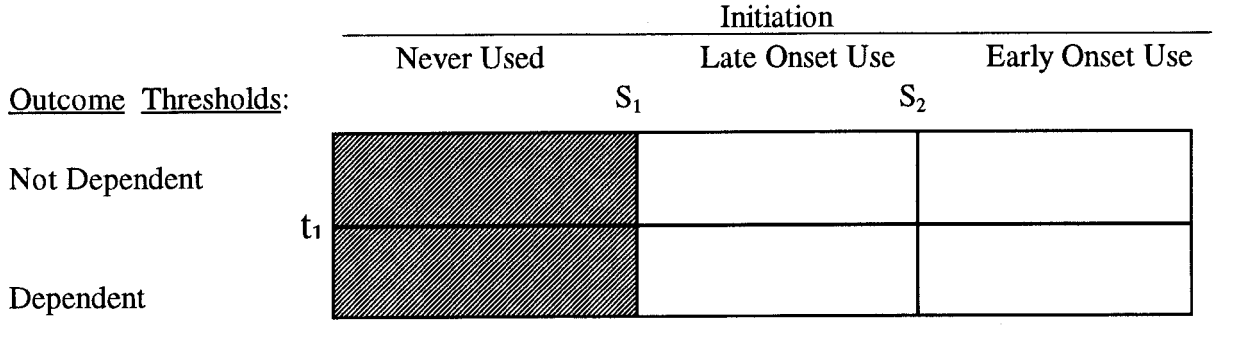

Figure 1

Two-way contingency table for drug use initiation, cross-classified by drug use outcome defined in those who have become users. The shaded cells of the table are not observable (i.e. represent structural missing data) because individuals who have never used a given drug cannot be classified with respect to drug use outcome.

will be estimated appropriately using standard statistical software for Probit regression analysis. Using MX (see Appendix 2 for example script) to handle the structural missing data for the Outcome dimension, regression coefficients were estimated correctly and the correct residual correlation obtained. In contrast, when structural missing data were ignored and standard statistical software used, biased estimates were obtained (–0.06, 0.09 and 0.42 respectively).

(ii) Extension to Twin or Family Data: Simulated Data

To investigate estimation of genetic parameters for Initiation and Outcome dimensions, we simulated data under a full bivariate genetic model (Heath et al., 1989; Neale & Cardon, 1992), estimating additive genetic variances VA (correlated 1.0 in MZ pairs, 0.5 in DZ or full sibling pairs), shared environmental variances VC (assumed equally correlated in MZ and DZ or full sibling pairs), and non-shared environmental variances VE (uncorrelated between family members). We refer to this as a 'two-stage' model since there will be structural missing data on the Outcome dimension for individuals with liability values below some threshold on the Initiation dimension (e.g. never users, as simulated here, but potentially also minimal users). We assumed a quadrivariate normal distribution of Initiation and Outcome dimensions, with the same threshold values used in Figure 1. Under the assumption of normally distributed liability dimensions, with at least 3 categories assessed for the Initiation dimension, including at least 2 that are characterized in individuals for whom data are available about the second dimension, a general bivariate genetic model for twin data is fully identified. Twin pairs who are discordant for drug use still provide information about genetic influences on Initiation, and about the genetic and environmental correlations between Initiation and Outcome dimensions. Testing the assumption that Initiation dimensions in twin (or sib) pairs have a bivariate normal distribution will be critical in the application to real data of the two-stage model that we propose here. If substantially different sets of traits influence early onset of drug use versus later onset of drug use, then we might expect the assumption of a single liability dimension underlying the 'No use' versus 'Early Onset Use' versus 'Later Onset Use' scale to be rejected. This would be indicated by finding that a model estimating a separate polychoric correlation plus separate row and column thresholdss to the 3x3 twin pair contingency tables for each zygosity group would give a poor fit to the observed data. Testing the fit of such a model will therefore be important in all applications to real data.

Four data sets were simulated (with subscripts I and O used to distinguish variance components for Initiation and Outcome dimensions, and correlations between additive genetic effects, shared environmental effects, and non-shared environmental effects on Initiation and Outcome dimensions denoted rG, rC and rE respectively). These are summarized in Table 1. (i) VA_I = 30%, VA_O = 60%, rG = 0.7071; VC_I = 30%, VC_O = 0%, rC = 0; VE_I = 40%, VE_O = 40%, rE = 0.75; (ii) VA_I = 30%, VA_O = 60%, rG = 0.7071; VC_I = 30%, VC_O = 0%, rc = 0; VE_I = 40%, VE_O = 40%, rE = 0; (iii) VA_I = 30%, VA_O = 49.2%, rG = 0.4685; VC_I = 30%, VC_O = 10.8%%, rC = 1.0; VE_I = 40%, VE_O = 40%, rE = 0.6; (iv) VA_I = 20%, VA_O = 40%, rG = 0.7071; VC_I = 40%, VC_O = 20%, rC = 0.3536; VE_I = 40%, VE_O = 40%, rE = 0.75. In the first two simulated data-sets, genetic effects on Initiation also account for one-half the total genetic variance in the Outcome dimension, whereas shared environmental effects influence Initiation only. In data-set (i), there is also a substantial non-shared environmental correlation between Initiation and Outcome dimensions, with non-shared environmental influences on Initiation accounting for 75% of the total non-shared environmental variance in the Outcome dimension; whereas in data-set (ii), non-shared environmental influences are assumed uncorrelated. The third simulated data-set uses the same genetic and environmental variance estimates for the Initiation dimension as data-sets (i) and (ii), and also the same total familial variance (i.e. the sum of additive genetic and shared environmental variances) for the Outcome dimension, but is based on the unidirectional causal model of Kendler et al. (1999), with a regression of 0.6 of the Outcome dimension on the Initiation dimension. Thus while we assumed no direct shared environmental influence on the Outcome dimension for data-set (iii), shared environmental influences on the Initiation dimension under this model will also have an attenuated effect on the Outcome dimension, because of the causal path from Initiation to Outcome. For data-set (iv), we allowed for both additive genetic and shared environmental correlations between Initiation and Outcome dimensions, as well as a non-shared environmental correlation. For all data-sets, we simulated 2000 MZ and 2000

Table 1

Genetic and Environmental Variances and Correlations Used in Simulations Under Two-stage Model. VA Denotes Additive Genetic Variance, VC Shared Environmental Variance, and VE Non-shared Environmental Variance, with Subscripts I for Initiation, O for Outcome Dimension. Genetic, Shared Environmental and Non-shared Environmental Correlations Between Initiation and Outcome are Denoted by r_G, r_C and r_E Respectively

	VA_I (%)	VA_O (%)	r_G	VC_I (%)	VC_O (%)	r_C	VE_I (%)	VE_O (%)	r_E
Data-set 1	30	60	0.7071	30	0	0	40	40	0.75
Data-set 2	30	60	0.7071	30	0	0	40	40	0
Data-set 3	30	49.2	0.4685	30	10.8	1.0	40	40	0.6
Data-set 4	20	40	0.7071	40	20	0.3536	40	40	0.75

DZ pairs, and assumed no genotype x sex interaction and no genotype x environment interaction. To ensure satisfactory recovery of parameter estimates when a general bivariate genetic model was fitted to these data, we fitted models by maximum-likelihood to the simulated data using MX. An example MX script for the analysis of such data is included in Appendix 3.

To investigate the potential bias that would arise when fitting models to drug use and dependence data when only binary data were available to characterize the Initiation dimension, we collapsed the simulated 3-level data on Initiation to 2 levels, by combining the 'early-onset' and 'late-onset' categories of Figure 1. Five 'false' models were then fitted to the simulated data: (1) assuming a single normal liability dimension underlying Initiation and Outcome categories, with ordered categories 'never used', 'non-dependent user' and 'dependent user' (the first, single liability dimension model considered by Eaves & Eysenck, 1980); (2) assuming two orthogonal liability dimensions underlying Initiation and Outcome dimensions, with genetic and environmental correlations fixed to zero (the second model considered by Eaves & Eysenck, 1980); (3) analyzing the Outcome data using only data from pairs concordant for drug use; (4) jointly analyzing the Initiation and Outcome data with the non-shared environmental correlation between these two dimensions (rE) fixed at zero; or (5) jointly analyzing the Initiation and Outcome data using a unidirectional causation model as applied by Kendler et al. (1999).

In our consideration of two-stage genetic models for drug use Initiation and Outcomes, we have not incorporated control variables in our simulations. However, combining the elements of the program for estimating polyserial correlations with control for covariates (Appendix 2) and the program for fitting a bivariate genetic model (Appendix 3) is straightforward.

Application: Sample, Assessment, Analyses

To illustrate the application of two-stage models to real data, we have reanalyzed smoking data on smoking initiation and persistence (Heath, Cates et al., 1992; Heath & Martin, 1993; Madden et al., 1999) from the 1981 survey of the older Australian twin cohort, using data from only twin pairs born 1951 or earlier. Pairs younger than age 30 were excluded from the analysis since it was considered that the younger age-group would be relatively uninformative about genetic influences on smoking persistence. Data were available from both members of 692 monozygotic female (MZF) pairs, 312 MZM, 420 dizygotic female (DZF) pairs, 157 DZM and 427 DZ unlike-sex (DZFM) pairs. In addition, in the final model-fitting analyses we included data from 49 MZ female twins, 38 MZ male twins, 68 DZ female like-sex twins, 50 DZ male like-sex twins, 99 DZ female twins from unlike-sex pairs, and 21 DZ male twins from unlike-sex pairs, whose cotwin did not return a questionnaire. Inclusion of these twins will increase the precision of prevalence estimates, and reduce any bias that would arise if smoking status is predictive of non-response.

For Initiation, twins were coded 0 if they reported never having smoked cigarettes; 1 if they reported an onset of smoking after age 18; and 2 if they reported onset of smoking at age 18 or younger. For Persistence, twins were coded 1 if they reported still being a smoker, 0 if they reported that they had quit smoking, or were given a missing value if they had never smoked.

Data-analysis proceeded in four stages. First, we fitted a single liability dimension model to the Initiation data, to determine whether the assumption of a single underlying normally distributed liability dimension was supported in these data. This also yielded preliminary estimates of genetic and environmental variance components, and threshold values, for the Initiation dimension. Second, we fitted a univariate genetic model to twin pair contingency tables for smoking persistence, using only data from twin pairs who were concordant smokers (i.e. implicitly assuming an orthogonal liability dimensions model), to provide a preliminary estimate of the total genetic and environmental variances for the Persistence dimension. Third, we used the MX script of appendix 1 to estimate a polychoric correlation between Initiation and Persistence dimensions, separately for women versus men, ignoring the twin structure of our data. Unless Initiation and Persistence dimensions were uncorrelated, this third step would be expected to yield an improved initial estimate of thresholds for the Persistence dimension, compared to the previous step, since it takes into account structural missing data. Finally, using starting values generated in the previous stages, we then fitted a full bivariate genetic model to the Initiation and Persistence data. This phased approach to generating starting values was used because of the imperfect numerical accuracy of software for integrating the 4-variate normal distribution, which might otherwise lead to convergence problems. All models were fitted using MX software, using the option for ordinal data-analysis, by the method of maximum-likelihood (Neale et al., 1999).

Results

Simulated Genetic Data-sets

Table 2 summarizes parameter estimates, and 95% confidence intervals, obtained when models were fitted to the 4 simulated genetic data-sets. Within rounding error, analyses using the full bivariate genetic model, with Initiation operationalized as a 3-level variable, successfully recovered estimates of simulated threshold values and genetic and environmental variances and correlations. Even with 2000 MZ and 2000 DZ twin pairs, and 66.7% of the population assumed to be users, however, 95% confidence intervals for the genetic correlation between Initiation and Outcome dimensions were quite broad (e.g. 0.20–0.73 for data-set 3).

The remaining models that we fitted all used data-sets where Initiation was defined as a binary variable, so that there would be insufficient information to recover parameter estimates under a full bivariate genetic model. Not surprisingly, all of these models recovered the correct threshold estimate for Initiation: there are no structural missing data for the Initiation dimension. Fitting a single liability dimension model, with inclusion of non-users as the lowest category, yielded estimates of genetic and shared environmental variance components that were intermediate between those simulated for the Initiation and Outcome

Table 2

Maximum-likelihood Estimates of Model Parameters, and Their 95% Confidence Intervals, Obtained Using Simulated Data-sets (i) – (iv). Parameter Estimates Used for Simulation are Described Under Methods, but are Closely Approximated by the Estimates Obtained When a Full Bivariate Genetic Model Was Fitted.

Full bivariate genetic model	Additive Genetic Variance (%)	95% CI	Shared Environmental Variance (%)	95% CI	Non-shared Environmental Variance (%)	95% CI	Threshold Values
Data-set 1:							
Initiation	30.0	17.8–39.8	30.1	18.7–40.2	39.9	38.3–43.3	$s_o = -0.44$, $s_i = 0.41$
Outcome	59.9	59.5–65.3	0.0	0.0–20.2	40.0	33.7–40.1	$t_o = -0.01$
Genetic/Environmental Correlations	0.70	0.44–0.76	0.0	—	0.75	0.74–0.83	
Data-set 2:							
Initiation	30.0	18.4–41.6	30.0	20.1–39.5	40.0	36.5–43.8	$s_o = -0.43$, $s_i = 0.43$
Outcome	59.9	37.5–66.1	0.0	0.0–18.8	40.0	33.9–47.2	$t_o = 0.00$
Genetic/Environmental Correlations	0.71	0.42–1.00	0.0	—	0.0	–0.11–0.11	
Data-set 3:							
Initiation	30.0	24.2–41.7	29.9	21.4–39.5	40.0	36.4–43.9	$s_o = -0.43$, $s_i = 0.43$
Outcome	49.2	28.3–60.5	10.8	2.4–28.1	40.0	34.1–45.3	$t_o = 0.00$
Genetic/Environmental Correlations	0.47	0.20–0.73	0.99	0.51–1.00	0.60	0.51–0.68	
Data-set 4:							
Initiation	20.0	0.0–21.5	40.0	23.1–46.3	40.0	38.6–40.4	$s_o = -0.40$, $s_i = 0.40$
Outcome	40.0	31.9–48.5	20.0	19.0–24.0	40.0	37.2–55.8	$t_o = 0.01$
Genetic/Environmental Correlations	0.70	0.25–0.84	0.35	–0.01–0.43	0.75	0.73–0.84	
False models — binary definition of initiation							
(1) Single liability dimension							
Data-set 1:	40.3	26.6–53.9	13.4	1.8–24.6	46.4	42.1–50.9	$s = -0.43$, $s_i = 0.19$
Data-set 2:	41.5	28.7–54.4	15.3	4.4–26.0	43.2	39.2–47.3	$s = -0.43$, $s_i = 0.31$
Data-set 3:	32.9	19.8–46.0	22.6	11.6–33.4	44.5	40.3–48.9	$s_o = -0.43$, $s_i = 0.19$
Data-set 4:	27.0	13.8–40.3	26.6	15.5–37.4	46.4	42.1–50.9	$s_o = -0.43$, $s_i = 0.19$
(2) Orthogonal liability dimension							
Data-set 1:							
Initiation	30.0	13.7–46.1	30.0	16.3–43.2	40.0	34.9–45.6	$s_o = -0.43$
Outcome	56.7	33.1–66.5	3.1	0.0–22.7	40.2	33.5–47.8	$t_o = -0.35$
Data-set 2:							
Initiation	30.0	13.8–46.3	29.9	16.2–47.6	40.0	35.3–45.5	$s_o = -0.43$
Outcome	57.6	40.7–64.0	0.0	0.0–18.4	42.5	36.0–49.9	$t_o = -0.15$
Data-set 3:							
Initiation	30.0	13.7–46.5	29.9	16.7–43.1	40.0	34.9–45.6	$s_o = -0.43$
Outcome	53.4	28.8–63.5	3.0	0.0–13.4	43.6	36.6–51.2	$t_o = -0.34$
Data-set 4:							
Initiation	20.1	4.1–36.1	39.9	26.6–52.6	40.0	34.9–45.6	$s_o = -0.43$
Outcome	38.1	15.4–58.5	21.7	2.2–40.2	40.2	33.4–47.8	$t_o = -0.35$
(3) Bivariate model, $r_E = 0$							
Data-set 1:							
Initiation	29.9	13.7–46.2	30.0	16.4–43.2	40.0	34.9–45.6	$s_o = -0.43$
Outcome	56.8	34.1–65.5	3.1	0.0–12.5	40.1	33.6–47.5	$t_o = -0.28$
Genetic/Environmental Correlations	0.54	0.07–0.98	–1.0	–1.0–1.0	0.0		
Data-set 2:							
Initiation	30.0	16.4–39.5	29.9	19.1–40.5	40.0	35.5–45.6	$s_o = -0.43$
Outcome	60.0	55.0–66.1	0.0	0.0–18.5	40.0	34.0–46.5	$t_o = 0.00$
Genetic/Environmental Correlations	0.71	0.49–1.00	–0.21	–1.0–1.0	0.0		
Data-set 3:							
Initiation	30.1	13.9–46.4	29.9	16.2–42.3	40.0	34.9–45.5	$s_o = -0.43$
Outcome	52.3	28.6–62.8	5.3	0.0–20.5	42.4	35.7–50.2	$t_o = -0.22$
Genetic/Environmental Correlations	0.27	–0.16–0.77	0.98	–1.0–1.0	0.0		
Data-set 4:							
Initiation	19.8	4.3–36.0	40.1	26.6–52.6	40.1	34.9–45.5	$s_o = -0.43$
Outcome	38.6	15.9–61.2	21.5	2.2–39.9	40.0	33.2–47.5	$t_o = -0.28$
Genetic/Environmental Correlations	0.53	–0.16–1.00	–0.07	–0.71–0.42	0.0		

TABLE 2 CONTINUED	Additive Genetic Variance (%)	95% CI	Shared Environmental Variance (%)	95% CI	Non-shared Environmental Variance (%)	95% CI	Threshold Values
(4) Bivariate model, unidirectional causation							
Data-set 1:							
Initiation	31.0	14.9–47.4	29.1	15.5–42.3	39.9	34.8–45.3	s_o = –0.43
Outcome	56.2	33.0–65.8	3.8	0.0–23.1	40.0	33.4–47.5	t^o_o = –0.25
Genetic/Environmental Correlations	0.13	–0.03–0.32	0.49	–1.0–1.0	0.18	–0.03–0.39	
Data-set 2:							
Initiation	35.1	21.3–50.2	25.7	12.5–31.4	39.1	34.2–44.4	s_o = –0.43
Outcome	51.2	34.9–58.9	10.0	4.4–23.6	38.9	34.1–45.0	t^o_o = 0.18
Genetic/Environmental Correlations	0.52	0.33–0.60	1.00	0.99–1.00	0.60	0.42–0.80	
Data-set 3:							
Initiation	29.9	18.9–45.0	29.9	29.9–41.8	40.2	36.0–40.4	s_o = –0.43
Outcome	48.5	30.3–56.0	11.7	6.2–26.1	39.8	39.0–43.8	t^o_o = 0.00
Genetic/Environmental Correlations	0.48	0.36–0.63	0.97	0.61–1.00	0.61	0.53–0.79	
Data-set 4:							
Initiation	21.0	5.1–36.5	39.1	25.9–51.9	39.9	34.8–45.4	s_o = –0.43
Outcome	37.6	15.5–52.8	22.6	3.8–40.6	39.8	33.2–45.9	t^o_o = –0.22
Genetic/Environmental Correlations	0.18	0.02–0.25	0.31	0.04–0.42	0.24	0.03–0.29	

dimensions, and therefore did not adequately describe the inheritance of either dimension.

When an orthogonal liability dimensions model was fitted to the simulated data, estimates of genetic and environmental variances for the Initiation dimension were the same, within rounding error, as those used for the simulation. The confidence intervals were of course wider, since for these cases we were collapsing the two highest Initiation categories into a single category, with corresponding loss of statistical precision. Despite the false assumption under this model of uncorrelated genetic effects on Initiation versus Outcome dimensions, and uncorrelated environmental effects, the biased estimates of genetic and environmental variances for the Outcome dimension were not too discrepant from the actual values used for simulation. The only serious bias was for the threshold for the Outcome dimension, which would lead to underestimation of the proportion of the population at risk of dependence. Results obtained when a single liability dimension model was fitted to the data, with all pairs where at least one twin was a non-user excluded from the analysis, are not shown in Table 2. Within rounding error, these estimates and their confidence intervals were (predictably) identical to those obtained for the Outcome dimension under the Orthogonal Liability Dimensions model: under the Orthogonal Liability Dimensions model, information about genetic and environmental effects on Outcome is solely derived from pairs who are concordant users.

Fitting a bivariate genetic model with the correlation between non-shared environmental effects on Initation and on Outcome dimensions (rE) fixed at zero also recovered appropriate estimates for the Initiation dimension, and yielded estimates for genetic and environmental variances for the Outcome dimension that were not too discrepant from those used in the simulation, and that were less biased than in the case of the Orthogonal Liability Dimensions model. In the case of the 2nd data-set, which was simulated with a zero non-shared environmental correlation, parameter estimates used for simulation were recovered within rounding error. In the remaining cases, the estimated threshold for the Outcome dimension was less biased than when an Orthogonal Liability Dimensions model was fitted. In the remaining cases, however, the genetic correlation between Initiation and Outcome dimensions was consistently underestimated compared to the values used for simulation.

The third data-set was simulated under the unidirectional causation model used by Kendler et al. (1999). Not surprisingly, therefore, parameter estimates used in the simulation were recovered, within rounding error, when this model was fitted. For the remaining data-sets, however, this model underestimated the genetic correlation between Initiation and Outcome to a more serious degree than did the bivariate genetic model with rE = 0; and yielded parameter estimates that were more biased than under either of the other bivariate models, discrepancies being particularly notable for data-set 2, where a zero non-shared environmental correlation between Initiation and Outcome, but a substantial genetic correlation, had been simulated.

Application to Smoking Status

Table 3 summarizes the distribution of smoking status in the Australian 1981 twin cohort. Although asked about

Table 3

Smoking Status in the 1981 Questionnaire Survey of Australian Twins Born 1951 or Earlier

	Women		Men	
	N	%	N	%
Never smoked regularly	1742	60.9	598	40.4
Ex-smoker, started after 18	204	7.1	134	9.0
Ex-smoker, started by 18	286	10.0	347	23.4
Current smoker, started after 18	274	9.6	136	9.2
Current smoker, started by 18	354	12.4	267	18.0
TOTAL	2860		1482	

whether they had ever smoked, respondents answered as though they were indicating whether they had ever been regular smokers, with 61% of women, and 40% of men, reporting that they had never smoked. Median age at onset of smoking was 17 for men, 18 for women. Early onset of smoking was classified as smoking by age 18, reported by 57% of women smokers, and 69% of men smokers.

Twin pair concordance or discordance for smoking status as a function of zygosity is summarized in Table 4. Rather than reporting full two-way contingency tables, we have pooled data from like-sex pairs, so that pairs where the first-born twin was an early-onset persistent smoker and the second born twin was an early-onset successful quitter, and pairs where the statuses of first and second-born twins were reversed, are combined in the table. Data from discordant unlike-sex pairs are however reported separately. Table 4 shows acceptable observed cell frequencies for most concordant or discordant twin pair statuses, exceptions being in unlike-sex pairs where the female twin is an early onset persistent smoker and her male cotwin a late onset successful quitter (N = 1), or the female twin is an early onset successful quitter and her male cotwin either a late onset persistent smoker (N = 1) or a late onset successful quitter (N = 0). These rare cases of low observed cell frequencies would not be expected to impact adversely genetic model-fitting analyses.

In analyses of the Initiation data, a single liability dimension model gave a poor fit to the data from MZ female pairs (χ^2 = 18.28, *d.f.*= 3, $p < 0.001$), but gave an acceptable fit to the other four zygosity groups (p = 0.04 – p = 0.15). Estimated twin pair polychoric correlations (+/- standard errors) were: MZF: 0.73 +/– 0.03; DZF: 0.41+/– 0.06; MZM: 0.63+/–0.05; DZM: 0.54+/– 0.08; DZ unlike-sex: 0.29 +/– 0.06. Thus there was evidence for strong genetic effects on smoking initiation in women, with only modest shared environmental influences, but for strong shared environmental influences on smoking initiation in men, with only modest genetic influences. Estimated additive genetic variances for Initiation for women and men (and 95% confidence intervals) under the single liability dimension model were 69.9% (48.8–78.4%) and 15.8% (0.0–55.4%); shared environmental variances were 3.6% (0.0–22.3%) and 46.5% (10.3–67.0%) and non-shared environmental variances were 26.5% (21.1–32.8%) and 37.7% (28.0–48.6%).

When we ignored information from twin pairs where at least one twin had not smoked (i.e. implicitly assuming an orthogonal liability dimensions model), we obtained consistent evidence for a strong genetic contribution to risk of smoking persistence. Estimated twin pair polychoric correlations (+/– standard errors) were : MZF: 0.53 +/– 0.09; DZF: 0.32+/– 0.15; MZM: 0.57 +/– 0.11; DZM: 0.21 +/– 0.18 and DZFM: 0.28 +/– 0.14. There was no evidence for genotype x sex interaction effects for Persistence (χ^2 = 0.35, *d.f.*= 3, p = 0.95). Genetic model-fitting analyses yielded point estimates (and 95% confidence intervals) of 53.2% (8.9–66.3%) for the additive genetic variance; 1.3% (0.0–37.3%) for the shared environmental variance; and 45.6% (33.6–60.4%) for the non-shared environmental variance. Estimated threshold values for the persistence dimension were –0.22 for women, 0.04 for men.

Table 4

Twin Pair Smoking Status for Onset of Smoking, and Smoking Persistence, in the 1981 Questionnaire Survey, for Complete Pairs Born 1951 or Earlier.

Twin Status		Cotwin Status		Number of Pairs				
INIT.[a]	PERSIST.[b]	INIT.[a]	PERSIST.[b]	MZF	DZF	MZM	DZM	DZFM[c]
≤ 18	Y	≤18	Y	38	12	20	12	22
≤ 18	Y	≤ 18	N	26	18	19	12	13/9
≤ 18	Y	> 18	Y	24	13	8	5	7/11
≤ 18	Y	> 18	N	7	8	5	2	1/8
≤ 18	Y	N	_	37	43	12	15	7/48
≤ 18	N	≤ 18	N	23	12	35	12	13
≤ 18	N	> 18	Y	12	4	8	9	1/9
≤ 18	N	> 18	N	11	6	13	5	0/9
≤ 18	N	N	_	30	34	38	12	12/62
> 18	Y	> 18	Y	20	9	5	2	10
> 18	Y	> 18	N	14	7	6	7	4/4
> 18	Y	N	–	40	39	17	8	7/22
> 18	N	> 18	N	12	5	7	3	3
> 18	N	N	–	39	32	17	11	12/24
N	–	N	–	359	178	102	42	109
			TOTAL	692	420	312	157	427

Notes:[a]INIT.: N indicates never smoked; > 18: onset of smoking at 19 or older; ≤ 18: onset of smoking at 18 or younger.[b]PERSIST.: Y indicates continuing smoker; N indicates successful quitter; — indicates structural missing data (i.e., never smoked).[c]For unlike-sex pairs discordant for smoking status, we first list the number of pairs where the female twin has smoking status given under 'Twin' and the male twin has smoking status given under 'Cotwin,' and then list the converse case. This is necessary because contingency tables for unlike-sex twin pairs are not expected to be symmetric.

Table 5
Estimated Genetic and Environmental Variances and Correlations, and 95% Confidence Intervals, Under a Two-stage Bivariate Genetic Model for Smoking Initiation and Persistence

	Additive Genetic Variance				Shared Environmental Variance				Non-shared Environmental Variance			
	Females		Males		Females		Males		Females		Males	
	%	95% CI	%	95% CI	%	95% CI	%	95% CI	%	95% CI	%	95% CI
Initiation	62.6	37.5–78.3	21.7	0.0–62.0	10.5	0.0–33.1	41.5	5.1–65.6	26.8	21.4–33.3	36.8	27.4–48.1
Persistence	42.2	0.6–64.1	(42.2	0.6–64.1)	10.4	0.0–50.4	8.7	0.0–50.1	47.4	32.2–65.9	49.2	31.3–70.6
Genetic/environmental correlation	r_g: 0.28	–0.21–1.00	0.11	—	r_c: –1.0	—	–0.07	—	r_e: –0.22	–0.48–0.07	–0.39	–0.68–0.03

When we estimated the polychoric correlation between Initiation and Persistence dimensions, for women we obtained a small negative estimate (–0.06), with the estimated threshold for the Persistence dimension (–0.22) identical to that obtained when an orthogonal liability dimensions model was implicitly assumed. Likewise in males the estimated polychoric between Initiation and Persistence was small and negative (–0.18) and the estimated threshold value for Persistence essentially unchanged (0.00).

Since the single liability dimension model had given an acceptable fit to the data for four out of the five zygosity groups, we proceeded finally to fit two-stage bivariate genetic models to the data of Table 4. Even though the net phenotypic correlation between Initiation and Persistence was estimated as small and negative at the previous step, it would theoretically be possible for a significant positive genetic correlation to be masked by a negative environmental correlation. Parameter estimates and their 95% confidence intervals, estimated under a model allowing for genotype x sex interaction, are summarized in Table 5. The total genetic variance for Persistence was constrained equal in males and females in these analyses. The previously noted higher heritability of Initiation in women than in men, and a stronger shared environmental influences on Initiation in men than in women was confirmed, albeit with broad confidence intervals for these sex-dependent parameters. Significant heritability was confirmed for Persistence, albeit only marginally so (95% Confidence Interval 0.6%–64.1%). The estimated genetic correlation between Initiation and Persistence dimensions in women was small (0.28, implying that genetic influences on Initiation are accounting for approximately 8% of the genetic variance in Persistence) and non-significant, but with a wide confidence interval that included unity. The estimated genetic correlation in males was also small (0.11), with a 95% confidence interval that is undefined because genetic effects on Initiation were not significant in males. The estimated non-shared environmental correlations between Initiation and Persistence dimensions were negative in both females (–0.22) and males (–0.39), although only in males did this correlation differ significantly from zero.

Conclusions

We have illustrated, by computer simulation, the problems that can arise when trying to draw inferences about the overlap of genetic (or environmental) influences on initiation of alcohol, tobacco or other drug use, and genetic or environmental influences on outcomes observed in those who have become users. We have seen that when only a binary measure of Initiation can be defined, while estimates of total genetic and environmental variances for the Outcome measures may be not too seriously biased, a more serious bias may arise for estimates of correlations between genetic (or environmental) effects on Initiation versus Outcome. If only a binary operationalization of Initiation is available, a sensitivity analysis will be needed to explore variation in point estimates of genetic and environmental correlations under different simplified models (e.g. fixing to zero the non-shared environmental correlation; or, following

Kendler et al. (1999), using a unidirectional causation model), and their associated 95% confidence intervals. For substances for which high MZ concordances are observed, such as cigarette smoking (Heath & Madden, 1995; Kendler et al., 1999; Madden et al., 1999), these confidence intervals will in many cases be extremely broad, so that very large sample sizes will be needed to achieve acceptable precision of parameter estimates.

The problems arising when using a binary Initiation measure can be reduced if several ordered categories can be defined for Initiation, as we have illustrated using a classification of smoking initiation into never smokers, late-onset smokers, and earlier onset smokers. If the assumption of a single normal liability dimension underlying these ordered categories can be empirically justified, it becomes possible to estimate genetic and environmental variances for Initiation and Outcome measures, as well as their genetic and environmental correlations, under a full bivariate genetic model. This approach appears preferable to relying on untested assumptions such as (i) the absence of a non-shared environmental correlation between the two measures (in fact, in our smoking data, a significant negative non-shared environmental correlation was observed in males) or (ii) attenuated effects on the Outcome measure of genetic and environmental influences on Initiation (which assumption also cannot accommodate genetic and environmental correlations that are opposite in sign, such as we observed in the smoking data).

Substantively, our analyses confirm a much stronger genetic influence, and correspondingly weaker shared environmental influence, on smoking initiation in women than men in this cohort, but equal importance of genetic effects on smoking persistence in both genders. This may reflect the fact the much greater role of social factors in determining whether onset of regular smoking occurred in this cohort of men. Given the relatively weak genetic influence on smoking initiation in men, our sample sizes were too small to permit an accurate determination of the degree of correlation between genetic influences on persistent smoking, and genetic influences on initiation. A meta-analytic approach (cf. Madden et al., 1999) is likely to be necessary for this purpose.

In the present paper we have focused on the application of two-stage models specifically to initiation of substance use, and outcomes of substance use such as dependence or smoking persistence. A much broader range of problems in substance abuse and psychiatric genetic research could potentially be addressed using these methods, however, such as whether there are genetic risk-factors specific to alcohol withdrawal, in addition to those that contribute more generally to risk of alcohol dependence; or whether there is a strong overlap of genetic influences on risk of suicidal ideation, and genetic influences on risk of suicide attempt (cf. Statham et al., 1997; Fu et al., 2002); or the extent of overlap of genetic influences on childhood conduct disorder that persists as adult antisocial personality disorder, versus transient childhood conduct problems (cf. Moffitt, 1993). Likewise, while we have emphasized genetic applications, these methods apply much more broadly to structural equation modeling methods where two-stage models can reasonably be hypothesized. With the availability of increasingly powerful software for genetic and other structural equation modeling analyses, such as MX (Neale et al., 1999) or MPlus (Muthen & Muthen, 1998), critical questions about mediators and moderators of genetic and environmental influences on different stages in the onset and progression of alcohol, tobacco, or other drug use and dependence (Jacob et al., 2001) are becoming amenable to analysis in ways that were not previously possible.

Acknowledgments

Supported by NIH grants P50-AA11998, R37-AA07728, R01-AA10248, R01-AA09022 from the U.S. National Institute of Alcoholism and Alcohol Abuse; P01-CA75581 from the U.S. National Cancer Institute; and R01-DA12540 and R01-DA12854 from the U.S. National Institute of Drug Abuse.

References

Cloninger, C. R.(1987). Neurogenetic adaptive mechanisms in alcoholism. *Science, 236,* 410–416.

Eaves, L. J., & Eysenck, H. J. (1980). New approaches to the analysis of twin data and their application to smoking behavior. In H. J. Eysenck (Ed.), *The causes and effects of smoking* (pp. 158–235). London: Maurice Temple Smith.

Fu, Q., Heath, A. C., Bucholz, K. K., Nelson, E. C., Glowinski, A., Goldberg, J., Lyons, M. J., Tsuang, M. T., Eisen, S. A., Jacob, T., True, W. R., & Eisen, S. A. (2002). A twin study of genetic and environmental influences on suicidal behavior in male veterans. *Psychological Medicine, 32,* 11–24.

Hawkins, J. D., Catalano, R. F., & Miller, J. Y. (1992). Risk and protective factors for alcohol and other drug problems in adolescence and early adulthood: Implications for other substance abuse prevention. *Psychological Bulletin, 112,* 64–105.

Heath, A. C., Cates, R. C., Martin, N. G., Meyer, J., Hewitt, J. K., Neale, M. C., & Eaves, L .J. (1993). Genetic contribution to risk of smoking initiation: Comparisons across birth cohorts and across cultures. *Journal of Substance Abuse, 5,* 221–246.

Heath, A. C., Kessler, R. C., Neale, M. C., Hewitt, J. K., Eaves, L. J., & Kendler, K. S. (1993). Testing hypotheses about direction of causation using cross-sectional family data. *Behavior Genetics, 23,* 29–50.

Heath, A. C., & Madden, P. A. F. (1995). Genetic influences on smoking behavior. In J. R. Turner, L. R. Cardon, & J. K. Hewitt (Eds.), *Behavior genetic approaches in behavioral medicine*. New York: Plenum Press.

Heath, A. C., & Martin, N. G. (1993). Genetic models for the natural history of smoking: Evidence for a genetic influence on smoking persistence. *Addictive Behaviors, 18,* 19–34.

Heath, A. C., Neale, M. C., Hewitt, J. K., Eaves, L. J., & Fulker, D. W. (1989). Testing structural equation models for twin data using LISREL. *Behavior Genetics, 19,* 9–35.

Heath, A. C., Todorov, A. A., Nelson, E. C., Madden, P. A. F., Bucholz, K. K. & Martin, N. G. (2002). Gene-environment interaction effects on behavioral variation and risk of complex disorders: The example of alcoholism and other psychiatric disorders. *Twin Research, 5*(1), 30–37.

Joreskog, K. G., & Sorbom, D. (1989). *PRELIS: A preprocessor for LISREL* (2nd ed.). Mooreseville, IN: Scientific Software.

Kendler, K. S., Neale, M. C., Sullivan, P., Corey, L. A., Gardner, C. O., & Prescott, C. A. (1999). A population-based twin study in women of smoking initiation and nicotine dependence. *Psychological Medicine, 29,* 299–308.

Little, R. J. A., & Rubin, D. (1987). *Statistical analysis with missing data.* In Series in Probability and Mathematical Statistics: Applied Probability and Statistics. New York: Wiley

Madden, P. A. F., Heath, A. C., Pedersen, N., Kaprio, J., Koskenvuo, M. J., & Martin, N. G. (1999). The genetics of smoking persistence in men and women: A multi-cultural analysis. *Behavior Genetics, 29,* 421–429.

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

Muthen, L. K., & Muthen, B. O. (1998). *Mplus user's guide.* Los Angeles, CA: Muthen & Muthen.

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

Neale, M. C., Boker, S. M., Xie, G., & Maes, H. H. (1999). *MX: Statistical modeling* (5th ed.). Richmond, VA: VCU, Department of Psychiatry.

Newcomb, M. D., & Bentler, P. M. (1989). Substance use and abuse among children and teenagers. *American Psychologist, 44,* 242–248.

Statham, D. J., Heath, A. C., Madden, P. A. F., Bucholz, K. K., Bierut, L .J., Dinwiddie, S. H., Slutske, W. S., Dunne, M. P., & Martin, N. G. (1998). Suicidal behaviour: An epidemiologic and genetic study. *Psychological Medicine, 38,* 839–855.

Appendix 1.

MX program for estimation of tetrachoric correlation when data for 2nd variable are Missing at Random if respondent scores at lowest level on 1st variable. The input data file used for simulation where tetrachoric correlation between initiation and drug use outcome is 0.6 is also given. This script takes advantage of the option in MX to analyze weighted data, defining the variable WT as a definition variable (see Neale et al., 1999, for details).

Program file:

```
Estimation of tetrachoric correlation when data are MAR for 2nd variable
DA NI=3 NG=1
LA VARA VARB WT
ORDINAL FI=INPUT.DAT
DEFINITION_VARIABLES WT / ! Used because we are going to read in weighted (simulated) data  BEGIN MATRICES;
M FU 2 1 FR !  Thresholds (2 rows because 2 needed for VARA)
L LO 2 2 !  Allows 2nd threshold for VARA to be estimated as increment over  1st threshold
R LO 1 1 FR !  Tetrachoric correlation to be estimated
Z LO 1 1      !  Will be weight matrix
END MATRICES;
SP Z -1
MAT L 1 1 1
fi M (2,2) !  There is no 2nd threshold to be estimated for VARB
MAT M
-0.4 0.0
0.8 5.0
MAT V 1.0 !  Variance fixed to unity - we are estimating a 2x2 correlation matrix
MAR R 0.45 !  Starting value for tetrachoric correlation
FREQ Z;
TH L*M;
CO V I R _ R I V /
INTERVAL R(1,1) !  Get 95% confidence interval for R.
BO -0.999 0.999 R(1,1)
BO 0.001 0.999 M(2,1)
BO -5.0 5.0 M(1,1) M(1,2)
OPT FUNC=1.E-12
OPT RS
END
```

Input data-file (VarA, VarB, followed by weight variable). '.' is the default missing value indicator for MX.

```
0 .  196.1941
0 .  137.8058
1 0 166.4651
1 1 166.5351
2 0 137.3408
2 1 195.6592:
```

Appendix 2.

MX script for estimation of tetrachoric correlation with statistical control for effects of covariates on mean liability.

```
! input variables are drug use initiation, drug use outcome (conditional on init)
! initiation is 3-level, e.g. no use, early-onset use, later onset use
#define nvar 1
#define nvar2 2
#define maxthres 2
Estimation of tetrachoric correlation when data are MAR for 2nd variable
data NI=6 NG=1
la vara varb wt cova covb covc
Ordinal fi=tetrasim.dat
definition_variables wt cova covb covc /
Begin matrices;
M FU maxthres nvar2 fr
L LO maxthres maxthres
V LO nvar nvar fi
R LO nvar nvar fr
Z LO 1 1 ! weight matrix
K FU 1 3 ! matrix of covariates (control variables)
B FU 3 2 FR ! matrix of probit regression coefficients (to be estimated)
end matrices;
```

```
SP Z
-1
SP K
-2 -3 -4
MAT L
1.0
1.0 1.0
SP M
1 2
3 0
MATRIX M
-0.4 0.0
0.8 5.0
MAT V 1.0
MAT R 0.5
FREQ z;
TH (L*M)-(K_K)*B; ! Thresholds adjusted for covariates
CO V | R _
  R' | V ;
interval r(1,1)
bo -0.999 0.999 r(1,1)
bo 0.001 0.999 m(2,1)
bo -5.0 5.0 m(1,1) m(1,2) ! m(1,3) m(1,4)
OPT func=1.E-12
OPT RS
END
```

Appendix 3.

MX script for fitting full bivariate genetic model to simulated Initiation and Dependence data.

Script:

```
#define nvar 2
#define nvar2 4
#define maxthres 2
Analysis of simulated initiation and dependence data: MZM
data NI=5 NG=3
LA twina1 twina2 twinb1 twinb2 wt
Ordinal fi=twostage32.mzm
definition_variables wt/
Begin matrices;
M FU maxthres nvar fr
L LO maxthres maxthres
W LO nvar nvar fr
X LO nvar nvar fr
Y LO nvar nvar fr
z FU 1 1
end matrices;
SP Z -1
MAT L
1.0
1.0 1.0
MATRIX M
-0.4 0.6433
 0.8 0.0
fi M(2,2)
MAT W
0.7
0.105 0.70
mat x
0.1
0.1 0.1
mat y
0.7
0.105   0.7
Begin algebra;
A=W*W';
C=X*X';
E=Y*Y';
P=A+C+E;
end algebra;
FREQ Z;
TH L*(M|M);
CO P | A + C _
  A' + C'   | P ;
bo 0.001 1.0 y(1,1) y(2,2) x(1,1) x(2,2)
bo 0.0001 0.999 w(1,1) w(2,2)
bo -0.999 0.999 w(2,1) x(2,1) y(2,1)
bo 0.001 3.0 m(1,1)-m(2,2)
bo -5.0 5.0 m(1,1) m(1,2)
OPT func=1.E-12
OPT RS
END
Analysis of ordinal simulated data
data NI=5
LA twina1 twina2 twinb1 twinb2 wt
Ordinal fi=twostage32.dzm
definition_variables wt/
Begin matrices = group 1;
g fu 1 1
z FU 1 1
end matrices;
SP Z -1
mat g
0.5
FREQ Z;
TH L*(M|M);
CO P | g@A + C _
  g@A' + C'   | P ;
OPT RS
END
Constraint function - constrain phenotypic variances to unity
CO NI=1
Begin matrices = group 1;
U unit 1 nvar
end matrices;
CO\d2v(P) = u;
end
```