Supplementary material

Genetically informative mediation modeling applied to stressors and personality-disorder traits in etiology of alcohol use disorder

Tom Rosenström[1*], Nikolai Olavi Czajkowski[1,2], Eivind Ystrom[1,2,3], Robert F. Krueger[4], Steven H. Aggen[5], Nathan A. Gillespie[5], Espen Eilertsen[1], Ted Reichborn-Kjennerud[1,6], Fartein Ask Torvik[1,2]

[1]Department of Mental Disorders, Norwegian Institute of Public Health, Oslo, Norway;

[2]Department of Psychology, University of Oslo, Norway;

[3]PharmacoEpidemiology and Drug Safety Research Group, School of Pharmacy, University of Oslo, Norway;

[4]Department of Psychology, University of Minnesota, USA;

[5]Department of Psychiatry, Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, VA, USA;

[6]Institute of Clinical Medicine, University of Oslo, Norway;

[*]Correspondence: tom.rosenstrom@helsinki.fi

Contents

A simulation study of the biometric mediation model

In this section, we investigate accuracy and possible bias in parameter recovery, as well as possible correlated errors of estimation, and compare these statistical properties across the proposed mediation models and a classic mediation approach. To this end, we simulated biometric data adhering to the phenotypic mediation hypothesis using the R function of Appendix A. A hundred distinct datasets were simulated (500 monozygotic and 500 dizygotic simulated twin pairs in each), and the structural equation models (SEMs) specified in Appendix B and Appendix C were fit to each dataset. We made a number of observations.

As shown in the Supplementary Figure S1, estimation of the biometric parameters in a SEM did not inflict mentionable efficacy costs in estimation of phenotypic paths in comparison to the classic, genetically uninformative, regression-based mediation estimates. The phenotypic ‘regression’ paths are much easier to estimate accurately than the specific biometric parameters, however. Estimation accuracy of individual biometric parameters suffers more from a need to use binary- or ordinal-valued variables than accuracy of the phenotypic path coefficients, and ordinal variables generally lead to slightly better accuracy than binary variables despite involving more estimated threshold parameters (Supplementary Figure S2).

While unbiased, some of the parameter estimates had strong negative correlations across the simulated datasets (Supplementary Figure S3). This probably pertains to the parameters being correlated in their asymptotic distributions. While this has been known (though not necessarily well-known) for a long time for biometric estimates of additive genetic and shared environmental variance (Williams 1993) , we also found that the phenotypic *b* and *c* parameter estimates were strongly and negatively correlated (Fig. S3).

We then added normally distributed measurement errors to the latent liabilities underlying the ordinal-valued observed variables so that a reliability of 0.7 ensued. If this level of reliability is accurately modelled using an error-in-variables model (see Appendix B), the mediation model is unbiased, though the added noise has an effect on statistical power (Supplementary Figure S4a). Unaccounted errors in variables attenuate the phenotypic path coefficients and the biometric paths other than the non-shared environmental variance, which includes the error and is therefore inflated (Fig. S4b).

In summary, the biometric mediation model works as intended, but could be further developed to eliminate or decrease the parameter dependencies and to improve estimation accuracy for the biometric parameters (two related goals). Errors in variables do bias the estimates, but the bias can be eliminated using an accurate estimate of reliability. We next turn to theory supporting the main text's classic analysis of statistical power to detect (omnibus) confounding at model level rather than for individual parameters.

Theory for power analysis

In this section, we describe the theory behind our analytic results on statistical power (Figure 2 in the main text). Let $\boldsymbol{V} = [Y, M, X]^T = [b(aX + \varepsilon_M) + cX + \varepsilon_Y, aX + \varepsilon_M, X]^T$ be the vector of variables for a typical phenotypic mediation model (cf. path diagram within the dashed ellipse in Figure 1a in the main text). According to standard properties of expected values and covariances, the model-implied covariance matrix is then

$$Cov(\boldsymbol{V}) = \begin{bmatrix} (ab+c)^2\sigma_X^2 + b^2\sigma_{\varepsilon M}^2 + \sigma_{\varepsilon Y}^2 & \cdots & \cdots \\ b\sigma_{\varepsilon M}^2 + (a^2b + ac)\sigma_X^2 & a^2\sigma_X^2 + \sigma_{\varepsilon M}^2 & \cdots \\ (ab+c)\sigma_X^2 & a\sigma_X^2 & \sigma_X^2 \end{bmatrix},$$

where $\sigma_X{}^2 = \sigma_A{}^2 + \sigma_C{}^2 + \sigma_E{}^2$ is variance of *X*, or sum of the biometric A, C, and E variances contributing to it (same notation for the residual variables). The between-twin covariances are otherwise the same, but the biometric components change according to the know rules of Mendelian inheritance (MZ twins share all genetic variance and DZ twins 50% on average, and both share the rearing environment; e.g., $\sigma_X{}^2 = \sigma_A{}^2 + \sigma_C{}^2$ for MZ variance and $\sigma_X{}^2 = ½\sigma_A{}^2 + \sigma_C{}^2$ for DZ variance). Here, a confounder *u* will add a value $\sigma_u{}^2$ to all the within-twin elements of $Cov(\boldsymbol{V})$ and its effect on the between-twin covariance depends on the specific biometric composition of the source of confounding according to the rules of inheritance. We get the exact expected, confound-dependent log-likelihood ratio, χ_u, by comparing fits of the biometric mediation model and Cholesky model to the expected covariance under *u*. Theoretically, our test statistic is distributed as a non-central chi-squared variate with non-centrality parameter χ_u, and therefore statistical power equals to probability of such a variate exceeding the critical value for the central chi-squared distribution with degrees of freedom corresponding to difference of degrees of freedom between the Cholesky and the biometric mediation models (Neale and Cardon 1992).

Such an omnibus test of 6 degrees of freedom makes most sense here because we want to allow reasonable flexibility for the specific form of confounding, which is unknown in applications. In the power calculations of the main text, we use simple confounders where all the confounding is due to one unknown A, C, or E source of variance. We compare results to those from the same procedure applied to DoC model (cf. Figure 1d of the main text). The difference in degrees of freedom between the DoC model and a bivariate Cholesky model is 2. We plot the statistical power as a function of average ratio of confounder variance to true variance, with the average taken over all the variables involved in the model in question (Figure 2 of the main text).

Supplementary data analyses

Measurement error estimate of the stressful life event count

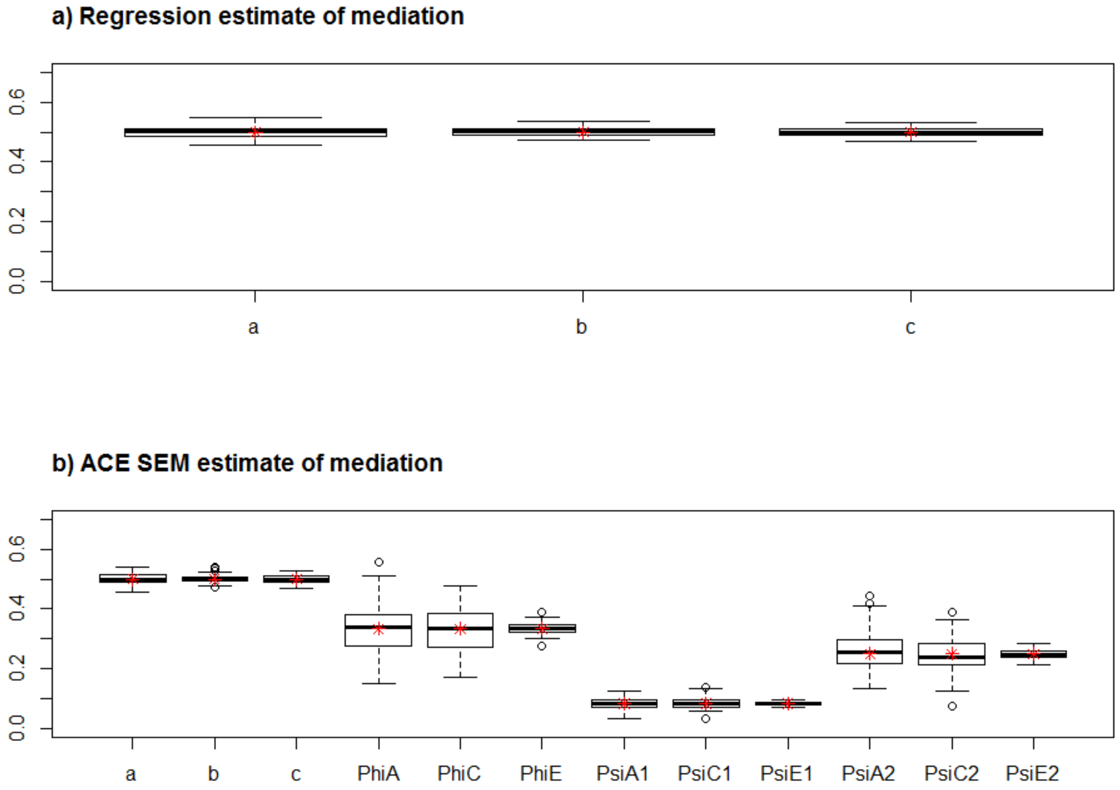
To be able to remove measurement error of stressful life events (SLEs) in the error-in-variables models, we needed an estimate for the relative amount of measurement error variance (i.e., a reliability estimate). To approximate reliability, we made use of the fact that two SLEs, "parental alcohol or mental problem as a child" and "parents divorced of moved apart when child", should have occurred for both the twins even if they report otherwise (possibly excluding very rare cases of adoption etc.). Thus, for these SLEs, the twins can serve as replications for each other's reports, thereby index (polychoric) inter-rater correlation/agreement. Twins' reports of parental divorce or moving apart correlated at r_1 = 0.985, whereas their reports of parental alcohol or mental problem correlated at r_2 = 0.895. We let r_1 index reliability of SLEs that are easy to remember and interpret and relatively free of stigma and r_2 index SLEs that are potentially subject to memory failures, stigma, and/or ambiguity of interpretation (*e.g.*, different people may regard differently the extent an accident is "serious"). These reliabilities were extrapolated to the SLEs from 1 through 18 (see Table 1 in main text) using a reliability vector with respective elements as $v = (r_2, r_2, r_1, r_2, r_2, r_1, r_2, r_2, r_2, r_2, r_2, r_2, r_1, r_1, r_2, r_1, r_2$, and $r_2)$. Then reliability of the SLE count was solved through the following simulation procedure.

Altogether 2000 twin-pair reports for all 18 SLEs were simulated from 18 bivariate normal distributions with correlations corresponding to elements of the reliability vector v. An SLE was recorded whenever a simulated value exceeded the normal-distribution quantile of one minus the frequency of the corresponding SLE in the data; i.e., SLEs were generated with the empirically observed frequencies. SLE counts of 3 or more were collapsed as in the analyses of real data. Then polychoric correlation between simulated twin pairs' counts of SLEs were computed over the 2000 reports, and this quantity was taken as the reliability of the SLE-count variable (*r* = 0.755).

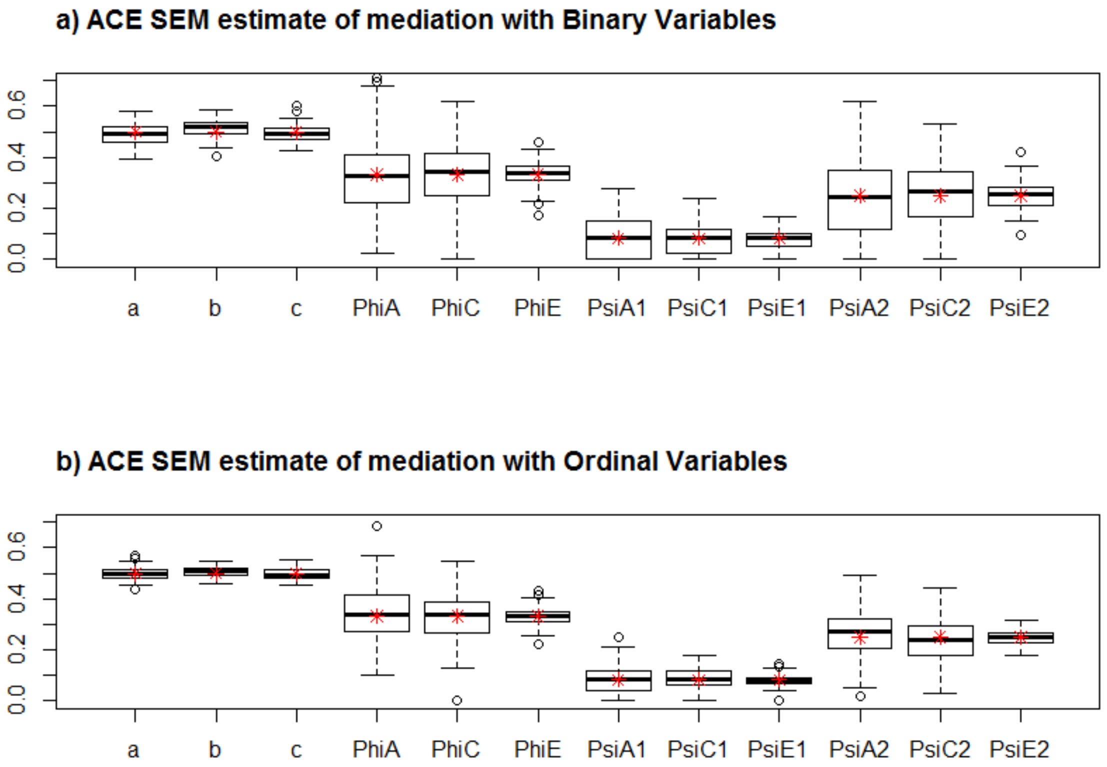
Parameter estimates for biometric mediation models with adulthood SLEs

Supplementary Figure S5 provides parameter estimates for the consistent biometric mediation models with a PD as an exposure, adulthood SLEs as the mediator, and AUD as the outcome. One observes that the PDs had a genetic and (non-shared) environmental effect on AUD and adulthood SLEs, but no strong mediated effect through the SLEs. Phenotypic association between the SLEs and AUD was low and statistically non-significant. In contrast to childhood SLEs, shared environment played a negligible role in adulthood SLEs.

Supplementary Figures



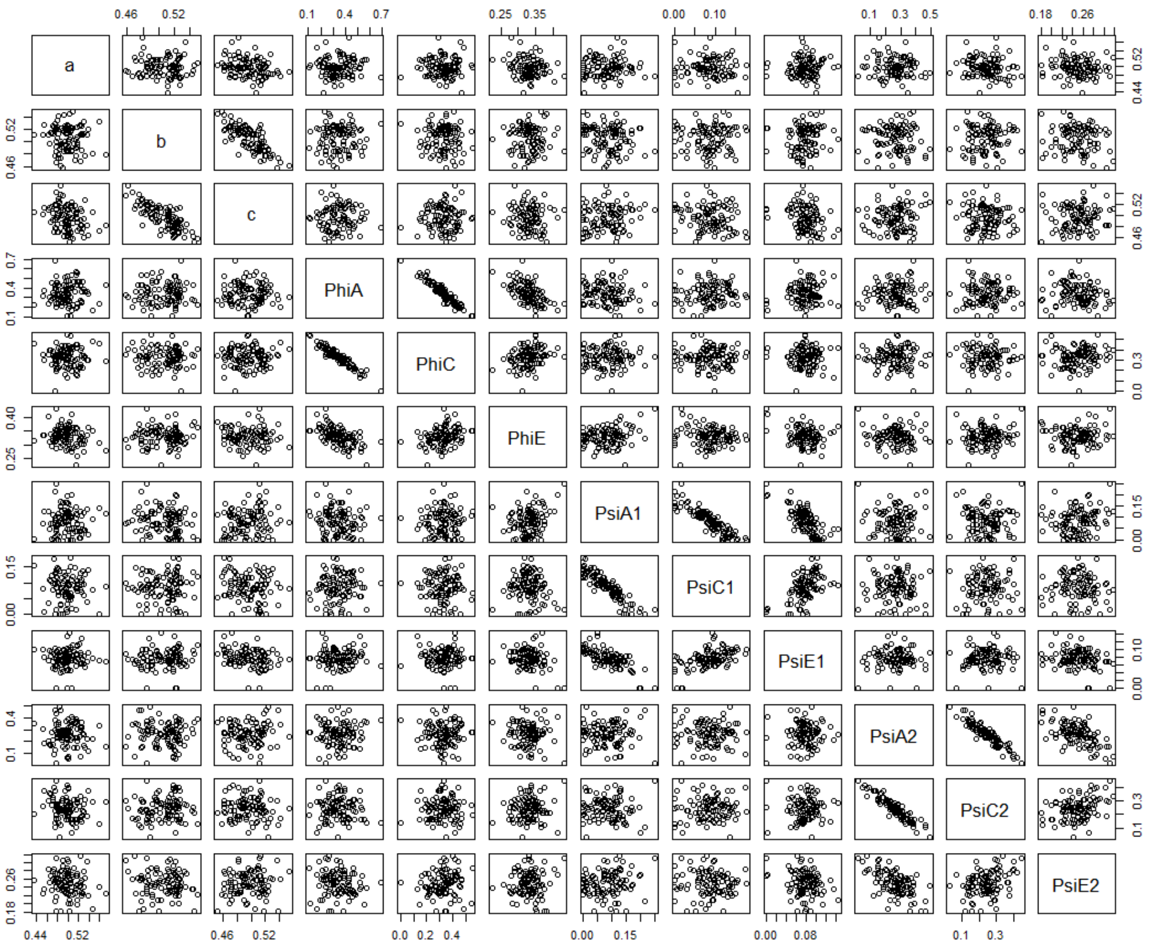
Supplementary Figure S1. *Boxplots of parameter estimates for 100 simulated datasets. Each dataset contained 2000 observations, consisting of 500 monozygotic and 500 dizygotic simulated twin pairs. The data adhered to the mediation model according to Appendix A, and the parameters are described in Appendix B. The red crosses represent the true parameters underlying the simulated data.* a) *Mediation parameters estimated using the classic regression approach.* b) *Mediation and biometric ACE parameters estimated using structural equation modeling (SEM) approach.*



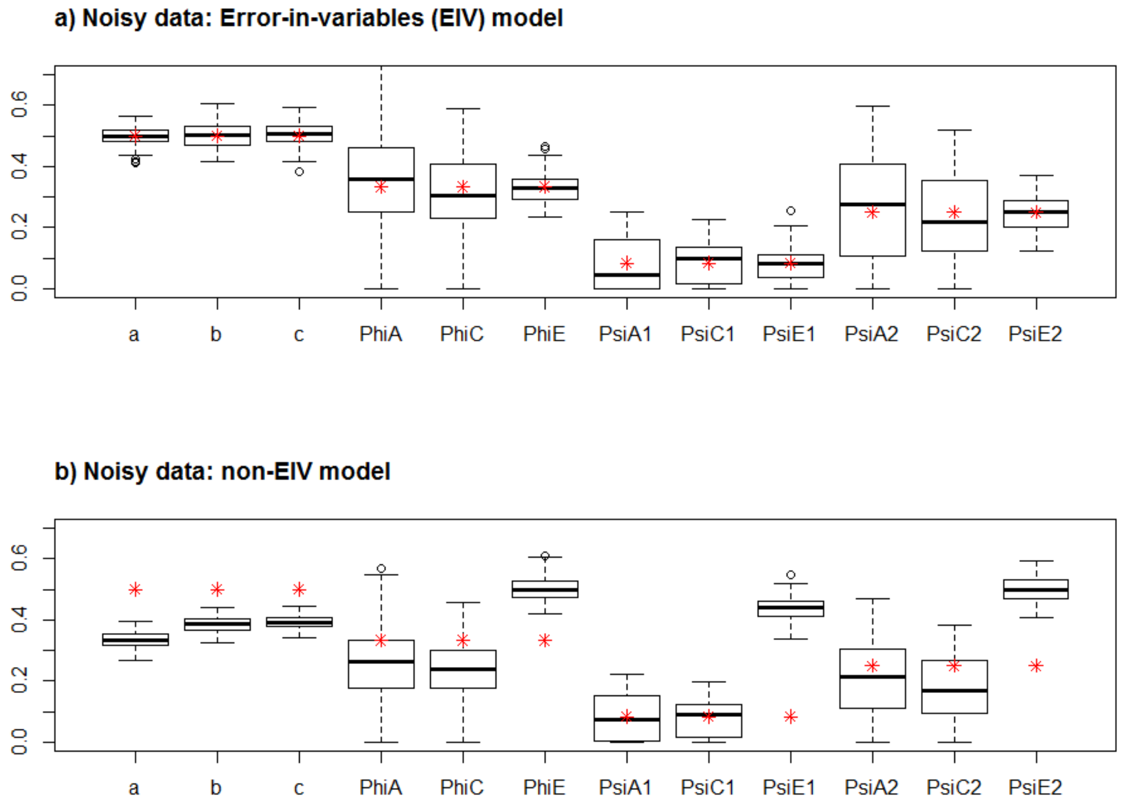
a) ACE SEM estimate of mediation with Binary Variables

b) ACE SEM estimate of mediation with Ordinal Variables

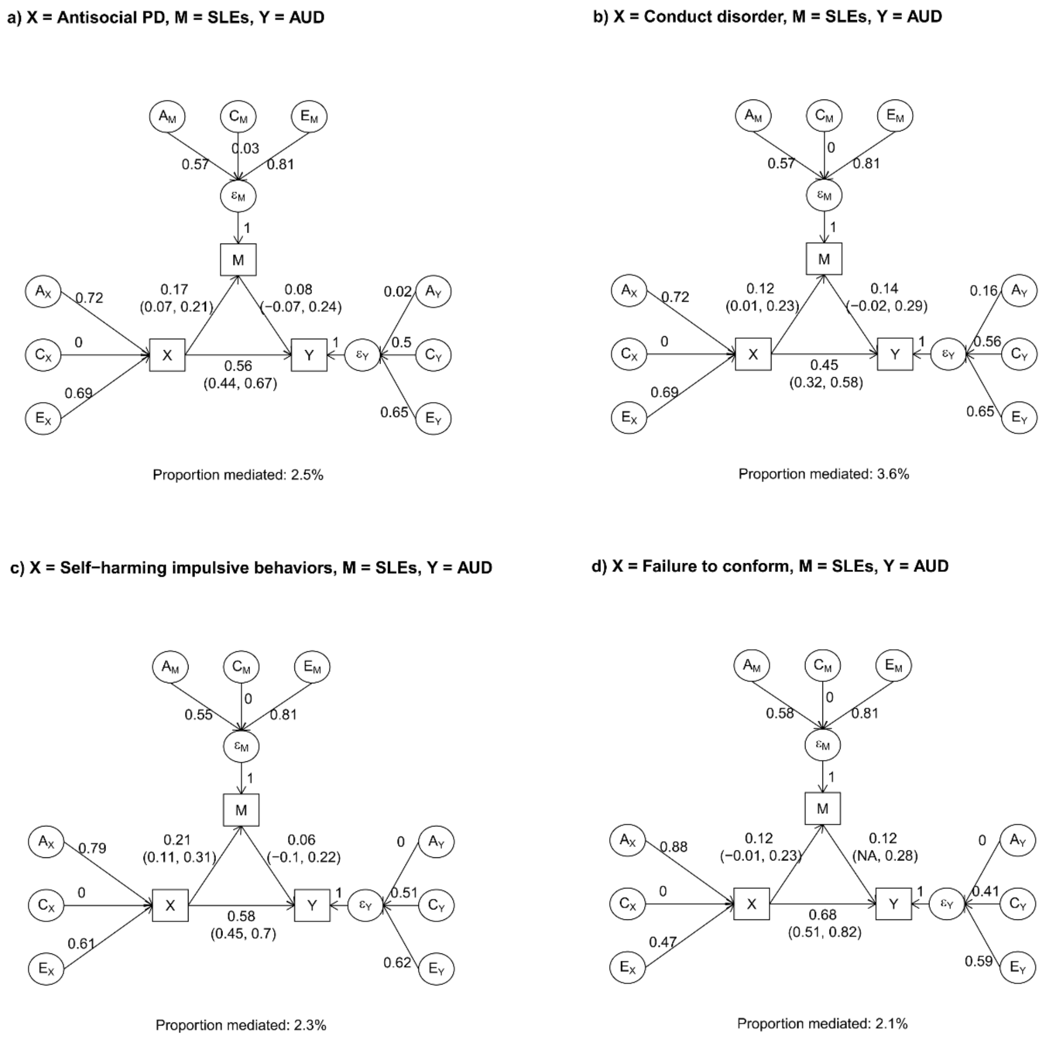
Supplementary Figure S2. *Boxplots of parameter estimates for 100 simulated datasets with binary (panel* a*) and ordinal (panel* b*) variables.* a) *The unit-variance simulated liabilities had a threshold at 0.8, which resulted in endorsement of the variable in question.* b) *The liabilities had two thresholds, at 0.3 and 0.8, leading to an ordinal variable.*



Supplementary Figure S3. Scatterplot matrix of the 100 simulated estimates for all the model parameters.



Supplementary Figure S4. *Effect of measurement errors in the ordinal-valued biometric mediation model. Reliability 0.7 was used in the 100 simulations shown in the boxplots.* a) *Estimates are unbiased when the amount of error is known and modeled.* b) *Non-modelled error in variables attenuates the phenotypic path coefficients and the biometric paths other than the non-shared environmental variance, which includes error.*



a) X = Antisocial PD, M = SLEs, Y = AUD

Proportion mediated: 2.5%

b) X = Conduct disorder, M = SLEs, Y = AUD

Proportion mediated: 3.6%

c) X = Self−harming impulsive behaviors, M = SLEs, Y = AUD

Proportion mediated: 2.3%

d) X = Failure to conform, M = SLEs, Y = AUD

Proportion mediated: 2.1%

Supplementary Figure S5. *Convergent mediation models for personality disorder (PD) traits, adulthood stressful life events (SLEs), and alcohol use disorder (AUD). Panels a-d show estimates for different PD traits. In one case, a lower confidence interval estimate was unreliable according to Open Mx software and therefore "NA" is shown instead of the estimate.*

Supplementary Tables

Supplementary Table S1. **Genetic and environmental correlations for cases where Cholesky model fit better than the biometric mediation models (cases involving borderline PD)**

Exposure (X) and ACE components[a]		**Correlations**			**Genetic variance%**[c]
X = childhood SLEs		SLEs	BPD	AUD	*h*^{2}%
A	SLEs	1	-0,208	-0,37	3,5
	BPD	-0,208	1	0,986	26,7
	AUD	-0,37	0,986	1	29,7
C	SLEs	1	0,893	0,735	73,3
	BPD	0,893	1	0,351	8,7
	AUD	0,735	0,351	1	14,2
E	SLEs	1	0,136	-0,111	23,3
	BPD	0,136	1	0,065	64,6
	AUD	-0,111	0,065	1	56
X = borderline PD		BPD	SLEs	AUD	*h*^{2}%
A	BPD	1	0,21	0,97	26,5
	SLEs	0,21	1	0,443	10,5
	AUD	0,97	0,443	1	33
C	BPD	1	0,918	0,299	8,8
	SLEs	0,918	1	0,653	12,3
	AUD	0,299	0,653	1	11,4
E	BPD	1	0,03	0,067	64,7
	SLEs	0,03	1	-0,049	77,2
	AUD	0,067	-0,049	1	55,6

a) Variables were temporally order starting from exposure “X” (two different models are shown for two different exposure). Each biometric component is similarly ordered, with genetic (A) influences shown first and followed by shared environmental (C) and non-shared environmental (E) influences.
b) Measured constructs were stressful life events (SLEs), borderline personality disorder (BPD), and alcohol use disorder (AUD).
c) Proportion of total variance explained by genetic influences (per variable, not including measurement errors that were removed in “EIV” modeling).

Appendix A: an R function for data simulation

```
medDat <- function() {
  # Data parameters, simulate data with rbind(DZ1,MZ1,DZ2,MZ2)
  K = 500   # Number of DZ twins
  J = 2*K   # Number of twin pairs
  N = 2*J   # Number of observations

  # mediation parameters
  a = sqrt(1/4)
  b = sqrt(1/4)
  cdot = sqrt(1/4)

  # simulate data
  pairno <- rep(1:J,2); dzs <- rep(c(rep(1,K),rep(0,K)),2)
  gmz = rnorm(K); gdz = rnorm(K)
  xg <- c(gdz*sqrt(1/2) + rnorm(K)*sqrt(1/2), gmz, gdz*sqrt(1/2) + rnorm(K)*sqrt(1/2), gmz)
  xc = rep(rnorm(J),2); xe = rnorm(N)
  X <- sqrt(1/3)*(xg + xc + xe)

  gmz = rnorm(K); gdz = rnorm(K)
  mg <- c(gdz*sqrt(1/2) + rnorm(K)*sqrt(1/2), gmz, gdz*sqrt(1/2) + rnorm(K)*sqrt(1/2), gmz)
  mc <- rep(rnorm(J),2); me <- rnorm(N)
  M <- a*X + sqrt(3/4)*sqrt(1/3)*(mg + mc + me)

  gmz = rnorm(K); gdz = rnorm(K)
  yg <- c(gdz*sqrt(1/2) + rnorm(K)*sqrt(1/2), gmz, gdz*sqrt(1/2) + rnorm(K)*sqrt(1/2), gmz)
  yc <- rep(rnorm(J),2); ye <- rnorm(N)
  Y <- b*M + cdot*X + sqrt(1 - 4*a*b*cdot)*sqrt(1/2)*sqrt(1/3)*(yg + yc + ye)

  ( data.frame(X = X, M = M, Y = Y, pairno = pairno, dzyg = dzs, id = 1:length(Y)) )
}
```

Appendix B: Mathematical details of the biometric mediation model

The classic mediation model is a three-variable system involving an exposure variable x, a mediator variable m, and an outcome variable y. The model-predicted relationships among the variables are captured by the path diagram in Figure 1a of the main manuscript, or alternatively, by the equations

$$m = \mu_m + ax + \xi_m,$$

$$y = \mu_y + bm + cx + \xi_y,$$

where μ_m and μ_y are fixed constants known as “intercepts” and ξ_m and ξ_y are normally distributed independent residual variables with mean of zero. The parameters a, b, and c are constant regression slopes, with c representing the direct effect of x on y and the product ab representing the indirect (i.e., mediated) effect of x on y through m. To derive a structural equation model (SEM), we then further assume that all variables (or their latent liabilities; see Methods in main text) are normally distributed and have a zero mean (i.e., $\mu_m = \mu_y = 0$).

We place the mediation model to the SEM framework using equations provided in a classic book by Bollen (1989). We let vector $u = [y, m]^T$, where T denotes transpose of a matrix or vector, collect the outcome y and the mediating variable m of the mediation model. These are the “endogenous” variables of the system to which the “exogenous”, or input, variable x affects. Vector $\xi = [\xi_y, \xi_m]^T$ contains their residuals. The general SEM is captured by the equation

$$u = \mathrm{B}u + \Gamma x + \xi \ ,$$

and this becomes the mediation model by setting

$$\mathrm{B} = \begin{bmatrix} 0 & b \\ 0 & 0 \end{bmatrix}$$

and $\Gamma = [c, a]^T$. Covariance matrix of the residual term ξ is denoted by Ψ. The matrix Ψ is assumed to be a diagonal matrix (representing independent residuals). Variance of x is denoted by Φ, and basically corresponds to the population variance in the exposure. The SEM version of the mediation model then yields the following expected population covariance matrix for $[u, x]$:

$$\Sigma(\theta) = \begin{bmatrix} \Sigma_{uu}(\theta) & \Sigma_{ux}(\theta) \\ \Sigma_{xu}(\theta) & \Sigma_{xx}(\theta) \end{bmatrix} = \begin{bmatrix} (I - \mathrm{B})^{-1}(\Gamma\Phi\Gamma^T + \Psi)(I - \mathrm{B})^{-T} & (I - \mathrm{B})^{-1}\Gamma\Phi \\ \Phi\Gamma^T(I - \mathrm{B})^{-T} & \Phi \end{bmatrix},$$

where θ collects the free/estimable parameters of the model; i.e., $\theta = (a, b, c, \Psi_{1,1}, \Psi_{2,2}, \Phi)$.

However, this is only phenotypic part of the covariance and cannot explain why twins are more similar to each other than randomly sampled representatives of the population. To that end, one needs to extend the model to a twin-study design (Neale and Cardon 1992). We did this by partitioning Φ as $\Phi = \Phi_A + \Phi_C + \Phi_E$, and Ψ as $\Psi = \Psi_A + \Psi_C + \Psi_E$, where A stands for additive genetic factors, C for shared environmental factors, and E for non-shared environmental factors. This standard partition is further discussed in the main manuscript and in pertinent literature (Neale and Cardon 1992).

Even when the data-generating process assumed by the mediation model holds true in the nature, the expected covariance $\sum(\theta)$ does not correspond to covariance of data measured or observed with error. If one has an estimate for the reliable variance, say $\alpha = [\alpha_y, \alpha_m, \alpha_x]^T$, the situation can be remedied using various forms of ‘error-in-variables’ modeling (Carroll et al. 2006). A simple solution in the SEM context is to let δ contain square roots of the elements in α and then let the expected covariance be

$$\Sigma_{EIV}(\theta) = \delta\delta^T \cdot \Sigma(\theta) + diag(1 - \delta \cdot \delta),$$

where "·" refers to element-wise multiplication instead of matrix product and the *diag*(1 – δ·δ) operator makes a diagonal matrix with the vector $[1 - \alpha_y, 1 - \alpha_m, 1 - \alpha_x]^T$ in the diagonal. Now, fitting $\Sigma_{EIV}(\theta)$ to the observed 'error-in-variables' covariance matrix yields unbiased, or error-free, estimates for the parameters in θ. Hybrid versions of the above model can be created by manipulating the covariance components according to established rules (Bollen 1989; Neale and Cardon 1992).

Appendix C: Scripts for fitting biometric mediation models using Open Mx R package

Scripts for fitting the models of this study can be found either from the URL:

www.iki.fi/tom.rosenstrom/codes.html

Or, from the Bitbucket-hosted URL:

https://bitbucket.org/rosenstroem/biometric_mediation_scripts

We hope that other researchers find these useful. If so, please cite this paper when using the code.

Supplementary references

Bollen KA (1989) Structural Equations with Latent Variables. John Wiley & Sons, Inc., New York, USA

Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM (2006) Measurement error in nonlinear models: a modern perspective. Chapman & Hall/CRC, Boca Raton, USA

Neale MC, Cardon LR (1992) Methodology for Genetic Studies of Twins and Families. Kluwer Academic Publishers, Dordrecht, The Netherlands

Williams CJ (1993) On the covariance between parameter estimates in models of twin data. Biometrics 49:557–568