Twin Research and Human Genetics

SUPPLEMENTAL MATERIAL

Genetically adjusted propensity scores: A methodological proposal and simulated comparison to discordant MZ twin models

Ian A. Silver Ph.D.[1 2]*
Hexuan Liu [3]
Joseph L. Nedelec [3]

*** Corresponding Author**:
Rowan University,
Glassboro, NJ, 08028
Ph: 856-256-4828
Email: silveria@rowan.edu

Table of Contents:

Appendix A: Simulations Process in Mathematical Equations
Appendix B: Description of Supplemental Analyses
Appendix C: Supplemental Analysis (Second Execution) Including Statistical Controls for Nonshared Environment.
Appendix D: Simulated Examples of the Effects of Pleiotropy on GAPS
Appendix E: Simulated Example of X and Y Being Unrelated but Colliding on the Shared Environment
Appendix F: Simulated Examples Illustrating How Different Matching Procedures Impact the Performance of GAPS
Appendix G: Simulated Examples Illustrating How Well GAPS Performs on Independent Samples.
Appendix H: Stata Syntax for Real Data Example

[1] Law and Justice Department, Rowan University, Glassboro, NJ
[2] Corrections Institute, University of Cincinnati, Cincinnati, OH
[3] School of Criminal Justice, University of Cincinnati, Cincinnati, OH

Appendix A: Simulations Process in Mathematical Equations

(R-Code Provided)

$n = 10{,}000$
$\boldsymbol{x_1} \sim N(100,1)$
$\boldsymbol{x_2} \sim N(100,1)$
$\boldsymbol{x_3} \sim N(100,1)$
$\boldsymbol{x_4} \sim N(100,1)$
$\boldsymbol{x_5} \sim N(100,1)$
$\boldsymbol{x_6} \sim N(100,1)$
$\boldsymbol{x_7} \sim N(100,1)$
$\boldsymbol{x_8} \sim N(100,1)$
$\boldsymbol{x_9} \sim N(100,1)$
$\boldsymbol{x_{10}} \sim N(100,1)$
$\boldsymbol{x_{11}} \sim N(100,1)$
$\boldsymbol{x_{12}} \sim N(100,1)$
$\boldsymbol{C} = .25\boldsymbol{x_1} + .25\boldsymbol{x_2} + .25\boldsymbol{x_3} + .25\boldsymbol{x_4}$
$\boldsymbol{E} = .25\boldsymbol{x_5} + .25\boldsymbol{x_6} + .25\boldsymbol{x_7} + .25\boldsymbol{x_8}$
$\boldsymbol{A} = .25\boldsymbol{x_9} + .25\boldsymbol{x_{10}} + .25\boldsymbol{x_{11}} + .25\boldsymbol{x_{12}}$
$\boldsymbol{e} = 0\boldsymbol{C} + 0\boldsymbol{E} + 0\boldsymbol{A} \sim N(100,1)$: represents ***rv***
$\boldsymbol{C'} = (\boldsymbol{C} - C_{\min}) / (C_{\max} - C_{\min})$
$\boldsymbol{E'} = (\boldsymbol{E} - E_{\min}) / (E_{\max} - E_{\min})$
$\boldsymbol{A'} = (\boldsymbol{A} - A_{\min}) / (A_{\max} - A_{\min})$
$\boldsymbol{e'} = (\boldsymbol{e} - e_{\min}) / (e_{\max} - e_{\min})$

Data Specification for Treatment ($\boldsymbol{X}$) Without Interactions
$b_{tC'} = b_{tE'} \mid b_{tA'} \to b_{tA'} \in R : 0 \geq b_{tA'} \leq .95$ in .05 increments.
or
$b_{tC'} = 3 * b_{tE'} \mid b_{tA'} \to b_{tA'} \in R : 0 \geq b_{tA'} \leq .95$ in .05 increments.
or
$3 * b_{tC'} = b_{tE'} \mid b_{tA'} \to b_{tA'} \in R : 0 \geq b_{tA'} \leq .95$ in .05 increments.
$\boldsymbol{X_{continuous}} = b_{tC'}\boldsymbol{C'} + b_{tE'}\boldsymbol{E'} + b_{tA'}\boldsymbol{A'} + b_{te'}e\boldsymbol{'}$: Treatment intitially coded continuously
$\boldsymbol{X} = 1 \mid \boldsymbol{X_{continuous}} \geq .50$
$\boldsymbol{X} = 0 \mid \boldsymbol{X_{continuous}} \leq .50$

Data Specification for Treatment ($\boldsymbol{X}$) With Interactions
$b_{tC'} = b_{tE'} \mid b_{tA'} \to b_{tA'} \in R : 0 \geq b_{tA'} \leq .95$ in .05 increments.
or
$b_{tC'} = 3 * b_{tE'} \mid b_{tA'} \to b_{tA'} \in R : 0 \geq b_{tA'} \leq .95$ in .05 increments.
or
$3 * b_{tC'} = b_{tE'} \mid b_{tA'} \to b_{tA'} \in R : 0 \geq b_{tA'} \leq .95$ in .05 increments.
$Ib_{tC'A'} = .20b_{tC'} + .20b_{tA'}$: Represents Interaction
$Ib_{tE'A'} = .20b_{tE'} + .20b_{tA'}$: Represents Interaction
$\boldsymbol{X_{continuous}} = (b_{tC'} - .20b_{tC'})\boldsymbol{C'} + b_{tE'}\boldsymbol{E'} + (b_{tA'} - .20b_{tA'})\boldsymbol{A'} + Ib_{tC'A'}(\boldsymbol{A}'*\boldsymbol{C}') + b_{te'}e\boldsymbol{'}$: Treatment intitially coded continuously
or
$\boldsymbol{X_{continuous}} = b_{tC'}\boldsymbol{C'} + (b_{tE'} - .20b_{tE'})\boldsymbol{E'} + (b_{tA'} - .20b_{tA'})\boldsymbol{A'} + Ib_{tE'A'}(\boldsymbol{A}'*\boldsymbol{E}') + b_{te'}e\boldsymbol{'}$
or
$\boldsymbol{X_{continuous}} = (b_{tC'} - .20b_{tC'})b_{tC'}\boldsymbol{C'} + (b_{tE'} - .20b_{tE'})\boldsymbol{E'} + (b_{tA'} - .40b_{tA'})\boldsymbol{A'} + Ib_{tC'A'}(\boldsymbol{A}'*\boldsymbol{C}') + Ib_{tE'A'}(\boldsymbol{A}'*\boldsymbol{E}') + b_{te'}e\boldsymbol{'}$
$\boldsymbol{X} = 1 \mid \boldsymbol{X_{continuous}} \geq .50$
$\boldsymbol{X} = 0 \mid \boldsymbol{X_{continuous}} \leq .50$

Outcome Specification
$\boldsymbol{Y} = \mathit{1.00}\boldsymbol{X} + 1.25\boldsymbol{C'} + 1.25\boldsymbol{E'} + 1.25\boldsymbol{A'} + .005\boldsymbol{e'}$
$\boldsymbol{Y} = \mathit{1.00}\boldsymbol{X} + 1.25\boldsymbol{C'} + 1.25\boldsymbol{E'} + 1.25\boldsymbol{A'} + 1.25(\boldsymbol{A'}*\boldsymbol{E'}) + .005\boldsymbol{e'}$
$\boldsymbol{Y} = \mathit{1.00}\boldsymbol{X} + 1.25\boldsymbol{C'} + 1.25\boldsymbol{E'} + 1.25\boldsymbol{A'} + 1.25(\boldsymbol{A'}*\boldsymbol{C'}) + .005\boldsymbol{e'}$
$\boldsymbol{Y} = \mathit{1.00}\boldsymbol{X} + 1.25\boldsymbol{C'} + 1.25\boldsymbol{E'} + 1.25\boldsymbol{A'} + 1.25(\boldsymbol{A'}*\boldsymbol{E'}) + 1.25(\boldsymbol{A'}*\boldsymbol{C'}) + .005\boldsymbol{e'}$

Appendix B: Description of Supplemental Analyses

Supplemental Analyses

Five supplemental analyses were conducted to further evaluate the validity of the GAPS approach in different scenarios. Besides the first supplemental analysis (Appendix C), the majority of the supplemental analyses were conducted with only the first treatment condition (Example 1; Figure 7).[4] First, Appendix C provides the results of the second execution for the three examples discussed above with the inclusion of discordant MZ-twin estimates, where varying levels of *E* were included in the matching procedure (i.e., 1**A*; 1**C*; .25**E* [App. of discordant MZ twin approach.). The inclusion of *E* in the matching procedures for the approximation of the discordant MZ-twin approach was designed to account for the ability to statistically control for nonshared environmental (*E*) factors when estimating a discordant MZ-twin model. The findings presented in the supplemental analyses (Appendix C) further demonstrate the strength of the discordant MZ-twin approach. Specifically, the inclusion of statistical controls for nonshared environmental (*E*) factors created conditions where the discordant MZ-twin estimates approached the specified slope coefficient ($b = 1.00$) better than the GAPS approach. Nevertheless, the superiority of the discordant MZ-twin approach depended upon the ability to introduce statistical controls for variation in *X* attributed to the nonshared environment. For instance, the discordant MZ-twin models that statistically controlled for 75% of the variation in *X* attributed to the nonshared environment overwhelmingly outperform the GAPS approach. Statistically controlling for 25% or 50% of the variation in *X* attributed to the nonshared environment, however, indicated that the GAPS approach can produce estimates that were similar to the estimates produced by discordant MZ-twin models in particular circumstances.

Second, we examined the potential influence of pleiotropy on the GAPS approach (Appendix D). Briefly, pleiotropy is when a single gene simultaneously influences multiple phenotypes (Stearns, 2010). Pleiotropy is suspected to influence estimates of the associations between genotypes and phenotypes (Socrates et al., 2017; Loika et al., 2020). In particular, research has shown polygenic scores are likely correlated with multiple phenotypes (Belsky et al. 2018; Wertz et al. 2018; Liu 2019). In the context of GAPS, if the polygenic score is correlated with other observed predictors or unobserved predictors, the coefficients in Eq. 2 might be biased. To assess the influence of pleiotropy on the prediction of the GAPS, we conducted three simulations: (1) allow a correlation between the genetic and shared-environmental components when simulating X; (2) allow a correlation between the genetic and non-shared environmental components when simulating X; and (3) allow correlations between the genetic and both shared and non-shared environmental components. Results in Appendix C show that unaddressed pleiotropy can slightly increase the distance between the estimates derived from the GAPS approach and the true estimates. These effects, however, were negligible (~ .01 increase in the estimated slope coefficients).

Third, Appendix E provides a supplemental analysis evaluating the effects of colliders between the independent variable (*X*) and dependent variable (*Y*) on the estimation of GAPS and the post-matching evaluation when no direct path between *X* and *Y* exist. Specifically, for illustrative purposes, components of the shared environment were treated as colliders between *X* and *Y* and the true association between *X* and *Y* was set equal to zero. Overwhelmingly, the

[4] The supplemental R-code will be made available by the authors upon publication and could be adapted to other treatment conditions.

results demonstrated that the inclusion of a collider in the estimation of GAPS and the subsequent matching procedures would generate an association between *X* and *Y* when no association existed in reality. Although not unexpected, these findings support Pearl's (2009) arguments that the inclusion of a collider as a statistical control could result in the observation of an association when no association existed and increase the likelihood of committing a type 1 error. Future users should be cautious when selecting measures of the shared and non-shared environment to include in the calculation of GAPS.

Fourth, Appendix R provides two additional simulation analyses to illustrate the performance of GAPS when different matching procedures are employed. The first simulation analysis employs coarsened exact matching and the second simulation analysis employs optimal matching (Guo and Fraser, 2015). Although the interpretations remained the same, slope coefficients from the simulations using coarsened exact matching and optimal matching, however, were more biased than the slope coefficients presented in the primary text. While the increased bias could exist for a variety of reasons, it appeared that when cases were matched using coarsened exact matching or optimal matching they were more dissimilar from each other than when the cases were matched using nearest neighbor matching.

Finally, two simulated examples are presented in Appendix G to illustrate how effective propensity score matching and GAPS would be when applied to an independent sample. In this simulation, the first dataset was simulated in a manner consistent with the specifications reviewed for the first simulation in Example 1. For the second dataset, the relationship between *X* and *Y* was specified to equal 1.00, while the effects of A, C, and E on *Y* were specified to each equal 1.25 (identical to the first dataset). As illustrated by the simulated examples, propensity score matching and the GAPS approach with matching appeared to adjust for the confounding influence of shared and non-shared environmental effects, as well as the confounding influence of genetic factors on the association between *X* and *Y* in the second dataset. Nevertheless, the distance between the estimated slope coefficients (propensity score matching $b = 1.32$; GAPS $b = 1.21$) and the specified slope coefficient in the second dataset was closer than observed when propensity score matching and the GAPS approach are used on the same sample. These findings suggest that the effectiveness of the GAPS approach might be enhanced when employing one sample to estimate the coefficients and then calculating GAPS in an independent sample.

Appendix C: Supplemental Analysis (Second Execution) Including Statistical Controls for Nonshared Environment.

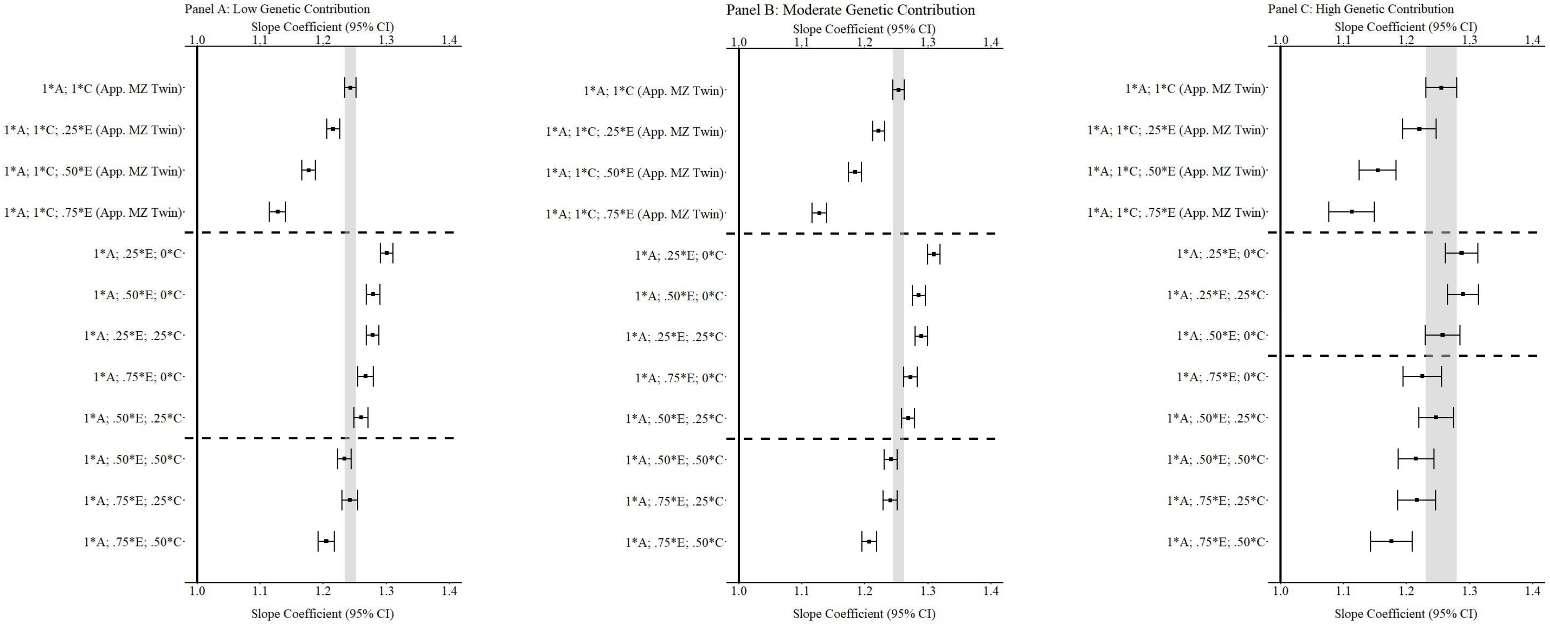


Fig. C1 (Corr. Ex. 1): Slope coefficients of *Y* regressed on *X* including statistical controls for nonshared environment in App. MZ twin.
Notes: A = genetics, E = nonshared environment, C = shared environment. The true association between *Y* and *X* is 1.00 (Starting N = 10,000). For the current example, Low Genetic Contribution: 10 % genetic, 64.5% nonshared environment, 21.5% shared environment, and 4% error. Moderate Genetic Contribution: 45 % genetic, 38.2% nonshared environment, 12.8% shared environment, and 4% error. High Genetic Contribution: 80 % genetic, 12% nonshared environment, 4% shared environment, and 4% error. Genetic, shared environmental, and nonshared environmental factors equally contributed to the variation in *Y*. All estimates were derived from a post matching OLS model. Matching was completed using nearest neighbor matching with a caliper of .05.

The dashed lines separate the approximation of an MZ model, the specifications that performed worse, and the specifications that performed better than the MZ model (without statistical controls). The proportions on Y-axis represent the proportion of the variation in *X* contributed by the specified component (A, C, or E) that is adjusted for by the model.

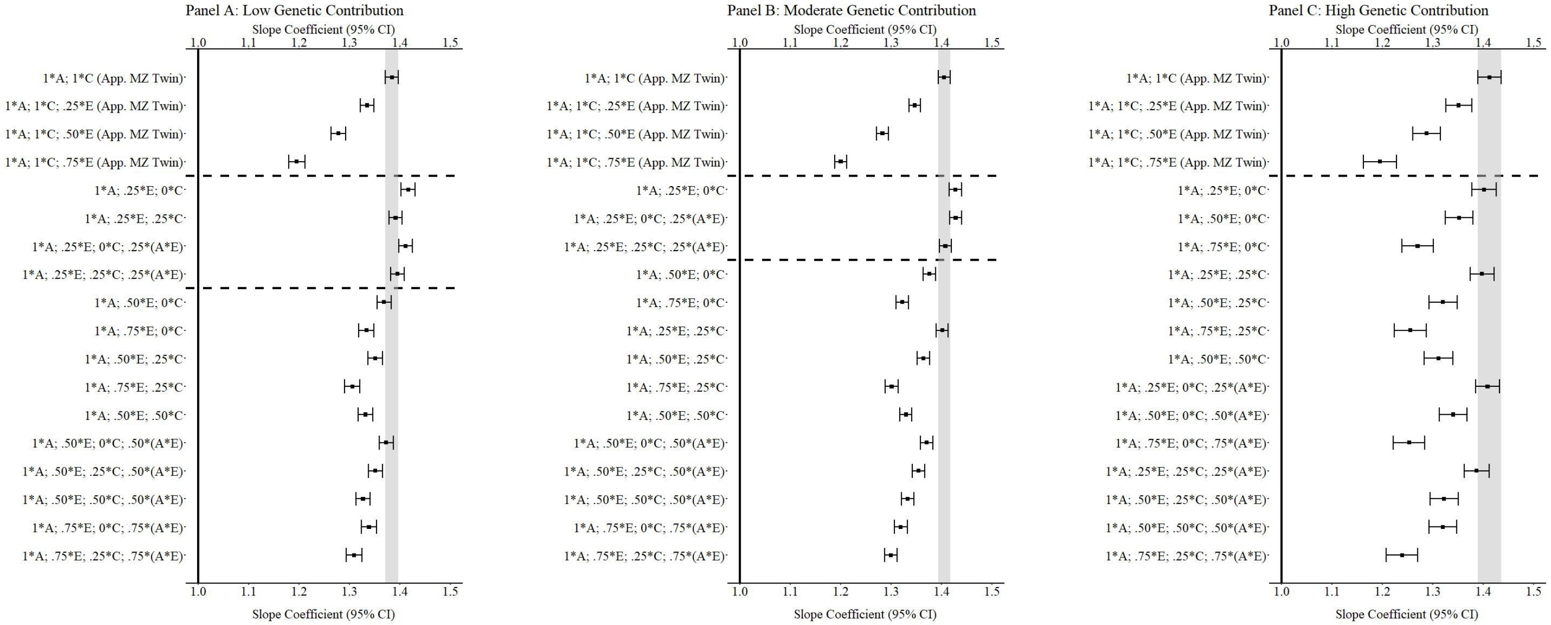
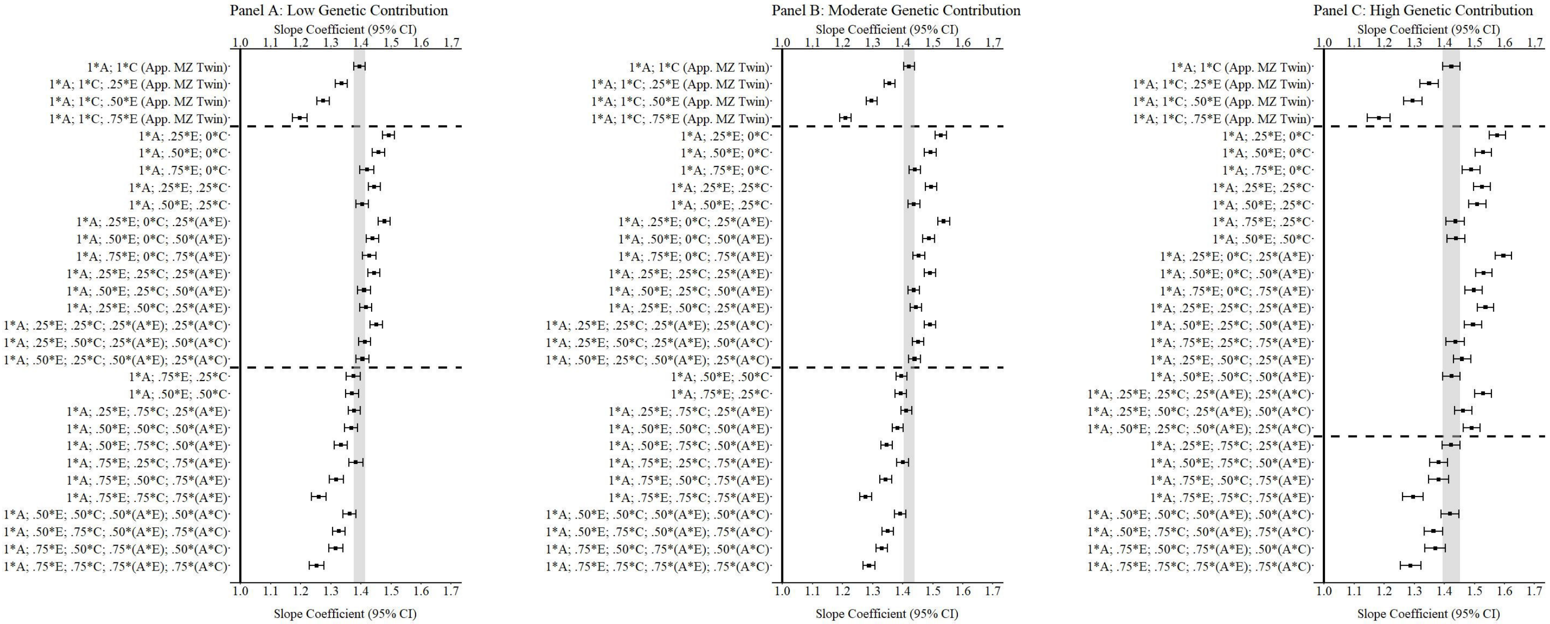


Fig. C2 (Corr. Ex. 2): Slope coefficients of *Y* regressed on *X* including statistical controls for nonshared environment in App. MZ twin. *Notes*: A = genetics, E = nonshared environment, C = shared environment. The true association between *Y* and *X* is 1.00 (Starting N = 10,000). , Low Genetic Contribution: 8 % genetic, 51.6% nonshared environment, 21.5% shared environment, 14.9% genetic* nonshared environment, and 4% error. Moderate Genetic Contribution: 36 % genetic, 30.6% nonshared environment, 12.8% shared environment, 16.7% genetic* nonshared environment, and 4% error. High Genetic Contribution: 64 % genetic, 9.6% nonshared environment, 4% shared environment, 18.4% genetic* nonshared environment, and 4% error. Genetic, shared environmental, and nonshared environmental factors equally contributed to the variation in *Y*. All estimates were derived from a post matching OLS model. Matching was completed using nearest neighbor matching with a caliper of .05.

The dashed lines separate the approximation of an MZ model, the specifications that performed worse, and the specifications that performed better than the MZ model (without statistical controls). The proportions on Y-axis represent the proportion of the variation in *X* contributed by the specified component (A, C, or E) that is adjusted for by the model.



Panel A: Low Genetic Contribution

Panel B: Moderate Genetic Contribution

Panel C: High Genetic Contribution

Fig. C3 (Corr. Ex. 3): Slope coefficients of *Y* regressed on *X* including statistical controls for nonshared environment in App. MZ twin.
Notes: A = genetics, E = nonshared environment, C = shared environment. The true association between *Y* and *X* is 1.00 (Starting N = 10,000). For the current example, Low Genetic Contribution: 6% genetic, 51.6% nonshared environment, 17.2% shared environment, 14.9% genetic* nonshared environment, 6.3% genetic*shared environment, and 4% error. Moderate Genetic Contribution: 27 % genetic, 30.6% nonshared environment, 10.2% shared environment, 16.7% genetic* nonshared environment, 11.6% genetic*shared environment, and 4% error. High Genetic Contribution: 48% genetic, .96% nonshared environment, 3.2% shared environment, 18.4% genetic* nonshared environment, 16.8% genetic*shared environment, and 4% error. Genetic, shared environmental, and nonshared environmental factors equally contributed to the variation in *Y*. All estimates were derived from a post matching OLS model. Matching was completed using nearest neighbor matching with a caliper of .05.

The dashed lines separate the approximation of an MZ model, the specifications that performed worse, and the specifications that performed better than the MZ model (without statistical controls). The proportions on Y-axis represent the proportion of the variation in *X* contributed by the specified component (A, C, or E) that is adjusted for by the model.

Appendix D: Simulated Examples of the Effects of Pleiotropy on GAPS

Table D1: Demonstration of the Effects of Pleiotropy on Example 1 (Simulation Starting N = 10,000)

Variation in X = *A*: 45%; *E*: 38.2%; *C*: 12.8%	Slope Coefficient (A correlated with E)	Slope Coefficient (A correlated with C)	Slope Coefficient (A correlated with E & C)
Specified slope Coefficient of *Y* on *X*	1.00	1.00	1.00
Proportion of Variation in X			
1*A; 25*E; .0*C	1.30	1.30	1.29
1*A; 50*E; .0*C	1.28	1.27	1.27
1*A; 75*E; .0*C	1.27	1.25	1.25
1*A; 0*E; .25*C	1.30	1.31	1.29
1*A; 0*E; .50*C	1.28	1.29	1.27
1*A; 0*E; .75*C	1.26	1.27	1.25
1*A; .25*E; .25*C	1.28	1.28	1.27
1*A; .50*E; .25*C	1.26	1.25	1.24
1*A; .50*E; .50*C	1.23	1.23	1.22
1*A; .75*E; .25*C	1.24	1.22	1.22
1*A; .75*E; .50*C	1.21	1.20	1.19

Notes: A = genetics, E = nonshared environment, C = shared environment. For the current example, variation in *X* was specified as 45 % genetic, 38.2% nonshared environment, 12.8% shared environment, and 4% error. Genetic, shared environmental, and nonshared environmental factors equally contributed to the variation in *Y.* All estimates were derived from a post matching OLS model. Matching was completed using nearest neighbor matching with a caliper of .05.

The proportions represent the proportion of the variation in *X* contributed by the specified component (A, C, or E) that is adjusted for by the model. For example, .25*E indicates that 9.56% of the variation of *X* (contributed by the nonshared environment) is adjusted for in the model.

Appendix E: Simulated Example of X and Y Being Unrelated but Colliding on the Shared Environment

Table E1: Demonstration of the Effects X and Y being unrelated but colliding on the shared environment C on Example 1 (Simulation Starting N = 10,000)

Variation in X = *A*: 45%; *E*: 38.2%; *C*: 12.8%	Slope Coefficient
Specified slope Coefficient of *Y* on *X*	.00
Discordant MZ-twin Slope Coefficient	-.63
Proportion of Variation in X	
1*A; 0*E; .25*C	-.28
1*A; 0*E; .50*C	-.45
1*A; 0*E; .75*C	-.54
1*A; .25*E; .25*C	-.27
1*A; .50*E; .25*C	-.27
1*A; .50*E; .50*C	-.43
1*A; .75*E; .25*C	-.27
1*A; .75*E; .50*C	-.56

Notes: A = genetics, E = nonshared environment, C = shared environment. For the current example, variation in *X* was specified as 45 % genetic, 38.2% nonshared environment, 12.8% shared environment, and 4% error. Genetic and nonshared environmental factors equally contributed to the variation in *Y*. All estimates were derived from a post matching OLS model. Matching was completed using nearest neighbor matching with a caliper of .05.

The proportions represent the proportion of the variation in *X* contributed by the specified component (A, C, or E) that is adjusted for by the model. For example, .25*E indicates that 9.56% of the variation of *X* (contributed by the nonshared environment) is adjusted for in the model.

Appendix F: Simulated Examples Illustrating How Different Matching Procedures Impact the Performance of GAPS

Table F1: Demonstration of the Effects of Matching Procedure on the Effectiveness of GAPS (Simulation Starting N = 10,000).

Variation in X = *A*: 45%*; E*: 38.2%; *C*: 12.8%	Slope Coefficient Coarsened Exact Matching	Slope Coefficient Optimal Matching
Specified slope Coefficient of *Y* on *X*	1.00	1.00
Proportion of Variation in X		
1*A; 25*E; .0*C	1.37	1.45
1*A; 50*E; .0*C	1.33	1.45
1*A; 75*E; .0*C	1.28	1.43
1*A; 0*E; .25*C	1.38	1.46
1*A; 0*E; .50*C	1.35	1.45
1*A; 0*E; .75*C	1.32	1.45
1*A; .25*E; .25*C	1.34	1.44
1*A; .50*E; .25*C	1.29	1.44
1*A; .50*E; .50*C	1.26	1.43
1*A; .75*E; .25*C	1.25	1.43
1*A; .75*E; .50*C	1.21	1.42

Notes: A = genetics, E = nonshared environment, C = shared environment. For the current example, variation in *X* was specified as 45 % genetic, 38.2% nonshared environment, 12.8% shared environment, and 4% error. Genetic, shared environmental, and nonshared environmental factors equally contributed to the variation in *Y*. All estimates were derived from a post matching OLS model.

The proportions represent the proportion of the variation in *X* contributed by the specified component (A, C, or E) that is adjusted for by the model. For example, .25*E indicates that 9.56% of the variation of *X* (contributed by the nonshared environment) is adjusted for in the model.

Appendix G: Simulated Examples Illustrating How Well GAPS Performs on Independent Samples.

Table G1: Demonstration of the Effects of PSM and GAPS in Independent Samples (Dataset 1 N = 10,000; Dataset 2 = 5,000).

Variation in X =	Propensity Score Matching		GAPS	
A: 45%*; E*: 38.2%; *C*: 12.8%	Data 1	Data 2	Data 1	Data 2
Specified slope Coefficient of *Y* on *X*	1.00	1.00	1.00	1.00
Proportion of Variation in X				
.50*E; .75*C	1.35	1.33	--	--
1*A; .50*E; .75*C	--	--	1.23	1.21

Notes: A = genetics, E = nonshared environment, C = shared environment. For the current example, variation in *X* was specified as 45 % genetic, 38.2% nonshared environment, 12.8% shared environment, and 4% error. Genetic, shared environmental, and nonshared environmental factors equally contributed to the variation in *Y*. All estimates were derived from a post matching OLS model.

The proportions represent the proportion of the variation in *X* contributed by the specified component (A, C, or E) that is adjusted for by the model. For example, .25*E indicates that 9.56% of the variation of *X* (contributed by the nonshared environment) is adjusted for in the model.

Appendix H: Stata Syntax for Real Data Example

```
*** Written By Ian Silver & Joseph Nedelec
*** Cleaning Add Health for GAPS Paper


*data: Restricted version of Add Health (all waves)

*UPDATED: Jan.3/2022

*~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~*

*** Primary IV: Educational Attainment (Wave 3)
*____________________________________________________________*
tab h3ed1

recode h3ed1 (6/15 = 0) (16/22 = 1) (96 98 99 = .) , gen(college_4yr)

tab h3ed1 college_4yr, m


*** Primary DV: Income (Wave 4) Personal Earnings
*____________________________________________________________*

tab h4ec2

recode h4ec2 (9999996/max = .) , gen(Per_Earn)

fre Per_Earn


* logged version of personal earnings
gen LPer_Earn = log(Per_Earn)


*~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~*
* Predictor Variables



* Parental Separation (Wave 1)
*____________________________________________________________*

tab pa44
tab pa43

recode pa44 (1 2 3 = 1)(4 5 = 0)(6 8 9= .), gen(PD1)
label variable PD1 "parental separation"

tab PD1

recode PD1 (7 = 0) if (pa43 == 1)

tab PD1
```

```
recode PD1 (7 = .)

tab PD1



** Parents Worked; full time employment past year (Wave 1)  **
** Rename: PW(1)
          **
** Dichotomous construct                                                              **
*________________________________________________________________*

tab pa13
tab pa14
tab pa15
tab pa16
tab pa17


recode pa17 (6 = .) , gen (PW1)

tab PW1

recode PW1 (7 = 1) if (pa14 == 1) & (pa15 == 1)

tab PW1

recode PW1 (7 = 0) if (pa14 == 0)

tab PW1

recode PW1 (7 = 0) if (pa14 == 1) & (pa15 == 0)

tab PW1

recode PW1 (7 = 0) if (pa16 == 0)

tab PW1

recode PW1 (7 = 0) if (pa16 == 1)

tab PW1

recode PW1 (7 = .) if (pa13 == 6)

tab PW1

recode PW1 (7 = .) if (pa14 == 6)

tab PW1


** Cognitive Abilities (Wave 1)                                   **
** Rename: CA(1)                                                              **
** Continuous construct                                           **
*________________________________________________________________*
```

```
tab ah_raw

gen CA1 = ah_raw

tab CA1


** Houshold Income (Wave 1)                                              **
** Rename: HI(1)                                                                     **
** Continuous construct                                            **
*________________________________________________________________*


tab pa55

recode pa55 (9996 = .) , gen (HI1)

tab HI1




** Parents Education (Wave 1)                                       **
** Rename: PE(1)                                                                     **
** Ordinal construct                                                          **
*________________________________________________________________*


* Coding scheme
** 1 = didnt graduate high school (1,2,3, 10)
** 2 = graduated high school (4,5)
** 3 = didnt graduate college (6, 7)
** 4 = graduated college (8, 9)
** . = (11, 12, 96, 97, 98)


** Mother
*________________________________________________________________*

tab h1rm1

recode h1rm1 (1 = 1) (2 = 1) (3 = 1) (10 = 1) ///
(4 = 2) (5 = 2) (6 = 3) (7 = 3) (8 = 4) (9 = 4) ///
(11 = .) (12 = .) (96 = .) (97 = .) (98 = .) (99 = .) , gen (ME1)

fre  ME1


** Father
*________________________________________________________________*

tab h1rf1

recode h1rf1 (1 = 1) (2 = 1) (3 = 1) (10 = 1) ///
(4 = 2) (5 = 2) (6 = 3) (7 = 3) (8 = 4) (9 = 4) ///
```

```
(11 = .) (12 = .) (96 = .) (97 = .) (98 = .) (99 = .) , gen (FE1)

fre  FE1

** Combined
*                                                                        *

egen PE1 = rowmean(ME1 FE1)

tab PE1



** Biological Sex & ancestry
*                                                                        *
tab sex_w1

tab ancestry


** Smoking Cigarettes (Wave 1)                                            **
** Rename: SC(1)                                                                          **
** Dichotomous construct                                                          **
*                                                                        *

tab h1to1
recode h1to1 (6 = .) (7 = .)(8 = .)(9 = .), gen (SC1)
tab SC1 h1to1,m



* Major Injury (Wave 1)
*                                                                        *

tab h1gh54

recode h1gh54 (1 = 0) (2 = 0) (3 = 0) (4 = 1) (5 = 1) (6 = .) (7 = .)(8 = .)(9 = .), gen (MI1)

tab h1gh54 MI1,m


* Polygenic Risk Scores *
*                                                             *

*BMI (pgsbmi)

tab pgsbmi

*Height
tab pgshgt

*Number of cigs per day
tab pgscpd10

*Extraversion
tab pgsexv15
```

```
*ADHD
tab pgsadd17

*Bipolar disorder
tab pgsbpr11

*Major depressive disorder
tab pgsmdd18

*Schizophrenia
tab pgsscz11

*Mental health cross disorder
tab pgsmhx13

*Alzheimers
tab pgsad13

*Educational attainment
tab pgsedu18


*** Propentsity Score Matching ***

* Change Log Location and Name
        *log using "BLR_PP matching.txt", text replace

        log using "[filepath]\GAPS_real data eg_LOG.txt", text replace


* PSM

set seed 1000
generate x=uniform()
sort x

psmatch2 college_4yr PD1 PW1 CA1 HI1 ME1 FE1 sex_w1 ancestry SC1 MI1 ///
                pgsbmi pgshgt pgscpd10 pgsexv15 pgsadd17 pgsbpr11 pgsmdd18 pgsscz11 pgsmhx13 pgsad13
pgsedu18 ///
                ,       caliper(0.05) ///
                        noreplacement ///
                        descending ///
                        neighbor(1) ///
                        logit

/*
sort _id
g BLR_MID=Id_Number[_n1]
g BLR_TID=Id_Number if _nn==1
g BLR_Cases=_weight
list BLR_MID BLR_TID in 1/100
summarize BLR_MID BLR_TID
list BLR_MID BLR_TID in 1/100
*/
```

```
*Matched sample*
reg LPer_Earn college_4yr if !missing(_nn)

    /*MATCHED SAMPLE WITHOUT ANY TWINS (MZ or DZ)*/
        preserve
            drop if pair_type4==1 | pair_type4==2
                reg LPer_Earn college_4yr if !missing(_nn)
        restore


*Full sample*
reg LPer_Earn college_4yr

    /*FULL SAMPLE WITHOUT ANY TWINS (MZ or DZ)*/
        preserve
            drop if pair_type4==1 | pair_type4==2
                reg LPer_Earn college_4yr
        restore


*MZ twins only*
reg LPer_Earn college_4yr if g==1

*---------------------------------------*

    *Sibling comparison*
    destring famid, replace
    sort famid

    *Check for level 1 variability (within-twin)*
    mixed college_4yr || famid: if g==1, vce(robust)
    estat icc

    *Creates family mean for IV*
    egen fam_college = mean(college_4yr), by(famid)

    *Creates deviation score for each twin's score from the family mean*
    gen twin_college = fam_college - college_4yr


    *Sibling comparison -- MZ TWINS ONLY*
        reg LPer_Earn fam_college  twin_college if g==1

    *Sibling comparison -- MZ & DZ TWINS*
        reg LPer_Earn fam_college  twin_college if  pair_type4==1 | pair_type4==2

*---------------------------------------*

log close
```