Genetically informative designs & Genetic covariance structure analysis: A brief introduction based on the classical twin design

Conor V. Dolan & Michael C. Neale

PPT presentation in 4 parts

PART 1 (11 slides)

- Linear regression
- A covariance structure (based on linear regression)
- The problem: how to infer genetic effects if you have not measured any genes (SNPs)?

PART 2 (19 slides):

- Genetic covariance structure analysis
- Genetically informative design - MZ twins raised together
- The observed covariance matrix vs. the hypothesized covariance matrix (model)
- Representation in path diagram
- Genetically informative design - MZ and DZ twins raised together ... the classical twin design (CTD)
- CTD Illustration height

PART 3 (8 slides):

- CTD multivariate ACE models from 1 to p (p>1) phenotypes - limited to ACE (ADE models also possible)
- Illustration Height and Weight

PART 4 (14 slides):

- The classical twin design (CTD) assumptions
- Other GIDs

Genetically informative design & Genetic covariance structure analysis:

A brief introduction based on the classical twin design

Conor V. Dolan & Michael C. Neale

PPT presentation in 4 parts .... PART 1 (12 slides)

Linear regression

A covariance structure (based on linear regression)

The problem: how to infer genetic effects if you have not measured any genes (SNPs)?

Fitting models to phenotypic data in genetically informative designs (GID) using genetic covariance structure modeling (GCSM)

Aim: infer genetic and environmental contributions to phenotypic variance from the phenotypic covariances (correlations) among family members (no measured genotypes, no measured environmental variables)

Contributions are expressed as “variance components”, so the
the phenotypic variance is decomposed into variance components.

Start with something familiar: the linear regression model
(e.g., as used in GWAS)

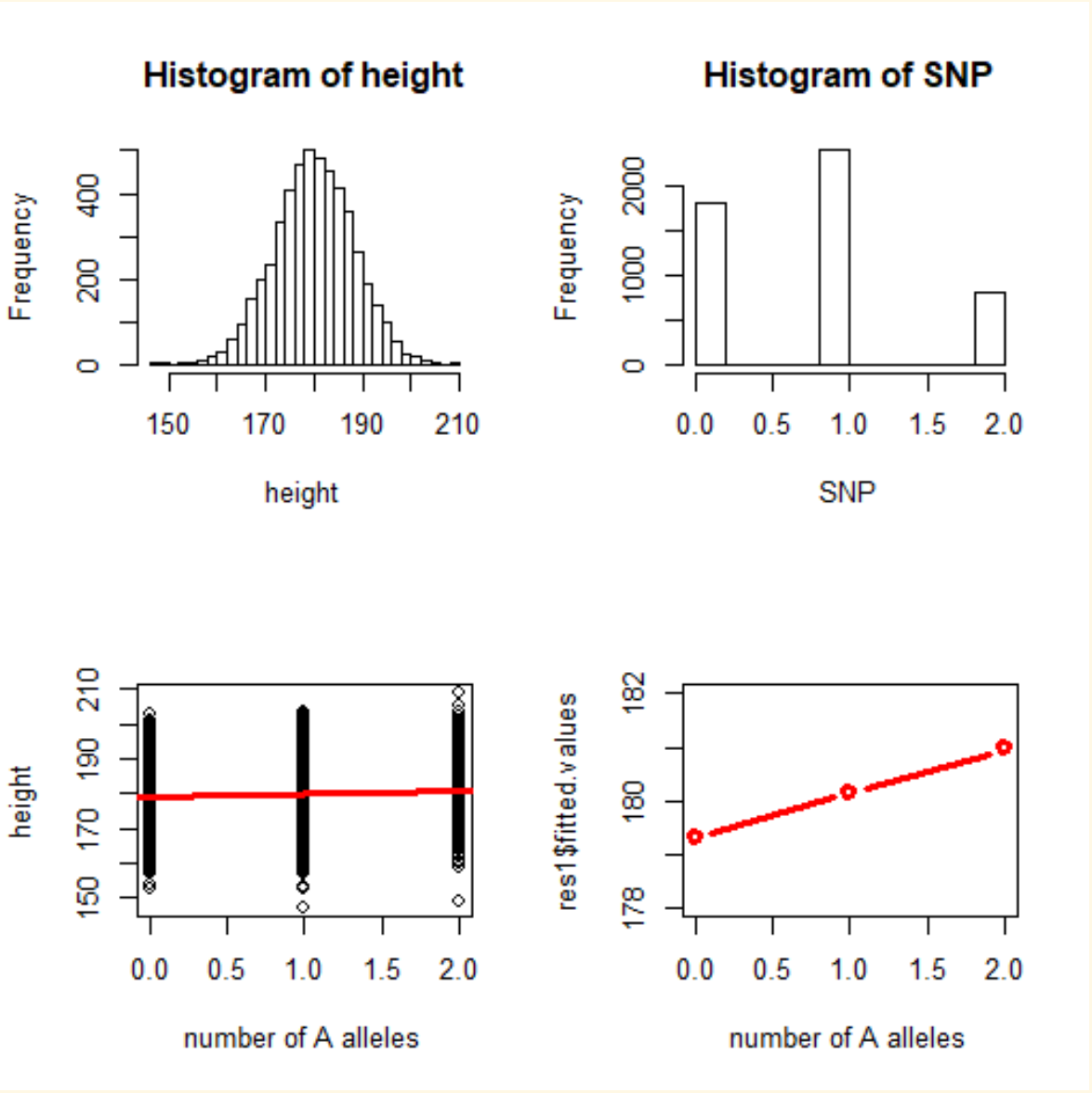
Linear regression model: predict Y from X ....

equation: $Y_i = b_0 + b_1*X_i + e_i$ ... e.g. GWAS: $Height_i = b_0 + b_1*SNP_i + e_i$

variables:	Y_i	dependent (predicted) in participant i (EA, Height, Depression)
	X_i	predictor in participant i (genetic variant: a SNP)
	e_i	residual in participant i
parameters:	b_0	intercept
	b_1	slope or regression coefficient

Y, X and e are <u>variables</u> because their value <u>vary</u> over persons
b_0 and b_1 are (fixed) parameters, with unknown values (in the well defined population)

Are X and Y linearly related? Null hypothesis (H-null): $b_1=0$



N = 20000

Descriptives

height: mean = 180.03 var= 64.01

SNP: mean = 0.80 var = 0.48

Covariance matrix

	height	SNP
height	64.01	0.212
SNP	0.212	0.480

Correlation matrix

	height	SNP
height	1.000	0.038
SNP	0.038	1.000

Alleles A-a, genotypes aa, Aa/aA and AA, coded 0, 1, 2 (Note this is additive coding)

Results of linear regression analysis (in R).

		Estimate	Std. Error	t value	Pr(>|t|)	
b_0	(Intercept)	179.67245	0.08635	2080.67	< 2e-16	***
b_1	SNP	0.44226	0.08160	5.42	6.02e-08	***

Linear association? H-null b1=0, H-alt b1≠0, α=0.01
Conclusion: p< α(p= 6.02e-08) so we reject H-null

Conclusion: **individual differences in height** are linearly related to Individual differences in SNP; or **SNP explains variance of height** or **the SNP is associated with height.**

Linear additive model: the effect of alleles A on height is additive
go from aa (0) to Aa (1) is associated with difference .44226 (b_1)
go from aa (0) to AA (2) is associated with difference .44226 + .44226 (additive: b_1+ b_1)

$Height_i = b_0 + b_1*SNP_i + e_i = 197.67 + .442*SNP_i + e_i$

Increase in the number of A alleles (from aa to Aa and from Aa to AA) is associated with increase in height of b_1 = 0.442 cm. Because the SNP is coded 0 (aa) / 1 (Aa, aA) / 2 (AA) and the model is linear, the explained variance is called <u>additive genetic variance</u>.

R^2: 0.001467 proportion of variance explained or 0.1467%

Covariance structure model:

The linear regression model

1) provides an account of the **covariance** (**correlation**) of Height and SNP
   (remember correlation expression linear association)
2) provide a decomposition of **variance** of Height
   (remember: variance is a measure of the magnitude of individual differences)

covariance matrix

observed numerical cov S		
	Height	SNP
Height	64.01	0.212
SNP	0.212	0.480

in symbols		
	Height	SNP
Height	s^2_H	$s_{H,SNP}$
SNP	$s_{H,SNP}$	s^2_{SNP}

based on linear regression model		
	Height	SNP
Height	$b_1^2 * s^2_{SNP} + s^2_e$	$b_1 * s^2_{SNP}$
SNP	$b_1 * s^2_{SNP}$	s^2_{SNP}

$Height_i = b_0 + b_1 * SNP_i + e_i$ .... $Height_i = 197.67 + .442 * SNP_i + e_i$

Observed numerical		
	Height	SNP
Height	64.01	0.212
SNP	0.212	0.480

linear regression model		
	Height	SNP
Height	$b_1^2 * s^2_{SNP} + s^2_e$	$b_1 * s^2_{SNP}$
SNP	$b_1 * s^2_{SNP}$	s^2_{SNP}

Decomposition: $s^2_H = b_1^2 * s^2_{SNP} + s^2_e =$
$.442^2 * 0.480 = 0.0938$ (i.e. explained additive genetic variance)

Covariance: $b_1 * s^2_{SNP} = .442 * 0.480 = .212$

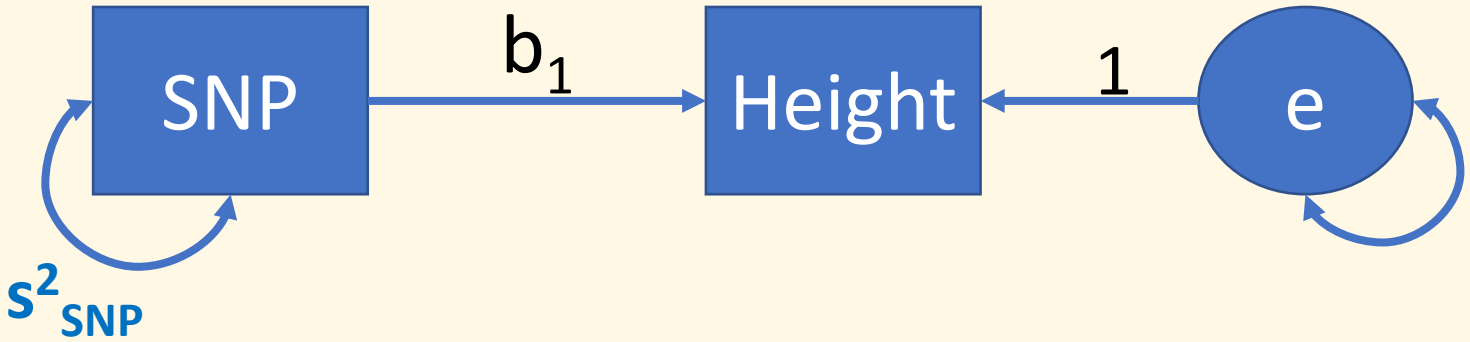
Effect size R^2: $\{b_1^2 * s^2_{SNP}\} / \{b_1^2 * s^2_{SNP} + s^2_e\} =$
$0.0938 / 64.01 = .00146$, or .146%

$$Height_i = b_0 + b_1 * SNP_i + e_i$$

model: linear regression model

linear regression model - implied covariance structure		
	Height	SNP
Height	$b_1^2 * s^2_{SNP} + s^2_e$ 64.01	$b_1 * s^2_{SNP}$ 0.212
SNP	$b_1 * s^2_{SNP}$ 0.212	s^2_{SNP} 0.480

model implied covariance structure



path diagram

Height

an observed variable, a measured variable (phenotype, locus)

A

a latent variable, unobservable variable (additive genetic factor)

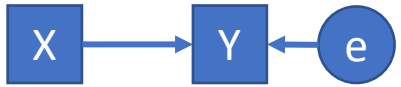
neuro-ticism

a latent variable, unobservable variable (the neuroticism as a latent construct)

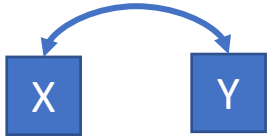
neuro-ticism

neuroticism as measured using a psychometric test, a test score (the test score approximates the latent construct)

regression relationship - linear association (asymmetric)



covariance or correlation - linear association (symmetric)

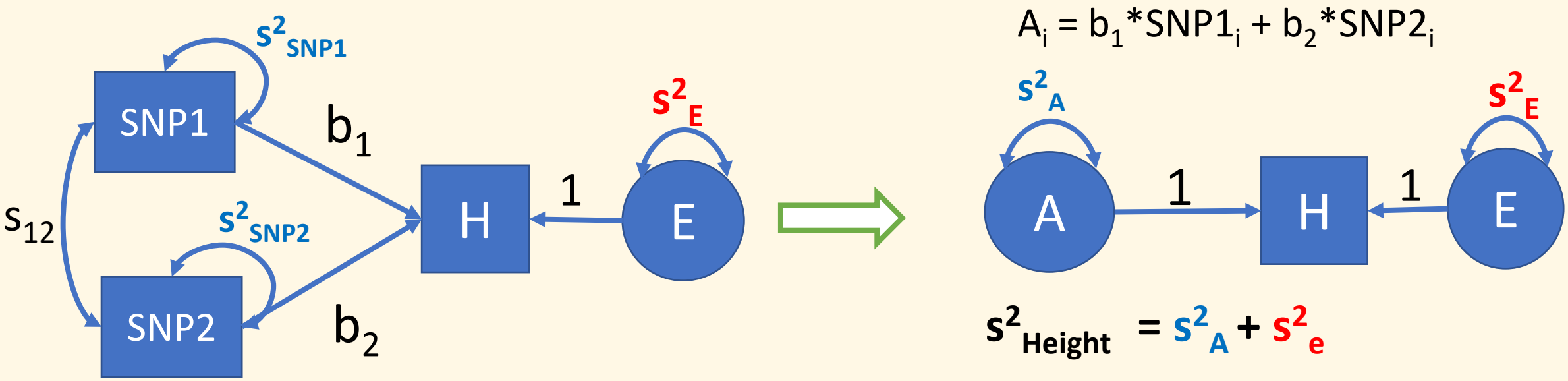


variance

conventions

Suppose we measured all SNPs relevant to height, suppose there are just 2

$Height_i = b_0 + b_1*SNP1_i + b_2*SNP2_i + e_i$

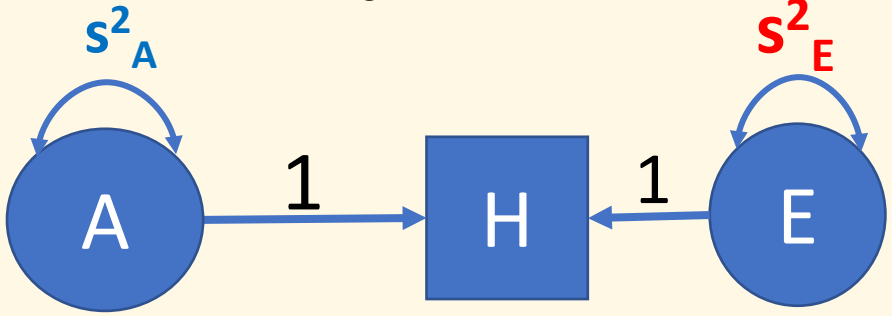


$$s^2_{Height} = \underbrace{b_1^2*s^2_{SNP1} + b_2^2*s^2_{SNP2} + 2*b_1*b_2*s_{12}}_{\text{Additive genetic variance: } s^2_A} + \underbrace{s^2_E}_{\text{(Environmental) variance: } s^2_E}$$

Suppose the following, where s^2_A is attributable to *M* SNPs (*M*>1000, say).

Eq 1: Height = $b_0 + b_1*Gene_1 + b_2*Gene_2 + ...b_M*Gene_M + E$

Eq 2: Height = $b_0 + A + E$



$$s^2_{Height} = s^2_A + s^2_E$$

s^2_A attributable to $Gene_1$ to $Gene_M$.

How to estimate the variance components, if we have not measured the SNPs?

Solution: Genetically Informative Design (GID)

+ Genetic covariance structure modelling (GCSM)

Genetically informative design & Genetic covariance structure analysis:

A brief introduction based on the classical twin design

Conor V. Dolan & Michael C. Neale

PPT presentation in 4 parts .... PART 2 (18 slides):

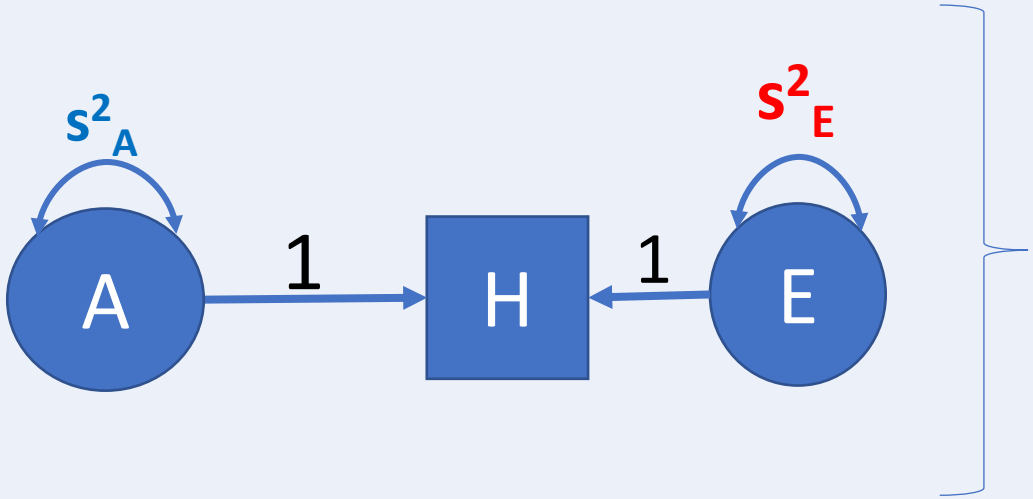
Genetic covariance structure analysis
Genetically informative design - MZ twins raised together
The observed covariance matrix vs. the hypothesized covariance matrix (model)
Representation in path diagram
Genetically informative design - MZ and DZ twins raised together ... the classical twin design (CTD)
CTD Illustration height

Genetic covariance structure model (GCSM)
A model for the linear relationships among phenotype
Phenotypes collected in a **genetically informative design** (GID)
Phenotypes measured in individuals in known genetic / environmental relationships

GID aim: estimate genetic and environmental variance components based only on the phenotype measures, no measured genes (SNPs), no measured environment

Most used GID: MZ and DZ twins raised together: the classical twin design (CTD Polderman et al 2015 - see slide notes for the ref)

Start with a simpler GID: MZ twins raised together (MZT)
Design: collect height in a representative sample of MZ twins
(i.e., representative of the well defined population)



Our hypothesis

A represents genetic effects s^2_A

E represents unshared environmental effects s^2_E

Data: Height measured in 250 twin pairs (500 twins),

Unit of sampling: twin pair

Key to the GID: MZ twins are genetically identical, 100% genetic variance is shared by MZ twins ... implies a **covariance structure model**

MZ tw1 variance $s^2_{Height} = s^2_A + s^2_E$

MZ tw2 variance $s^2_{Height} = s^2_A + s^2_E$

Hypothesis (variance)

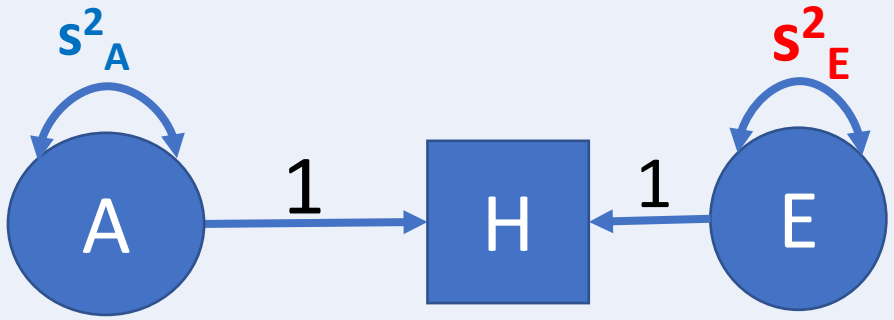
Genes that contribute to height variance, necessarily contribute to MZ covariance, because MZ twins are genetically identical.

MZ tw1-tw2 covariance $s_{H1,H2} = s^2_A$

Hypothesis (covariance)

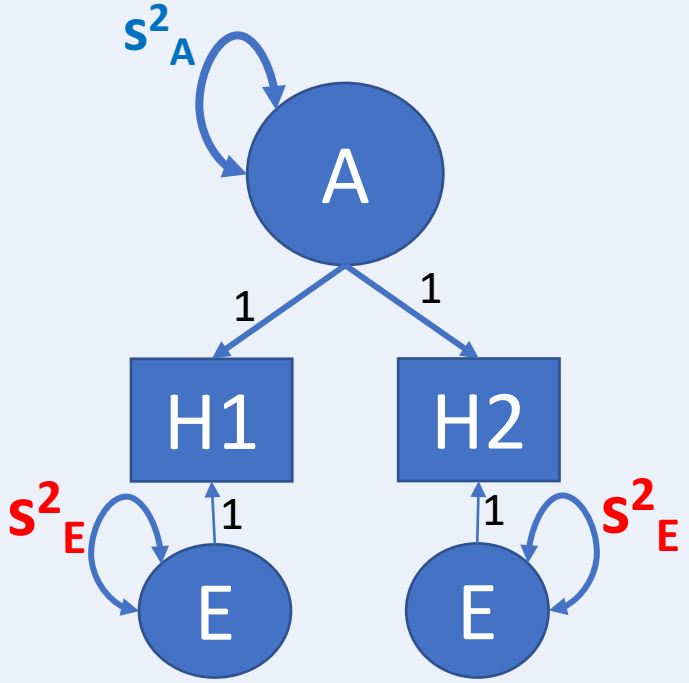
Observed N=250 MZ pairs (**S**)			GCSM (Model) (Σ)		
	MZ1	MZ2		MZ1	MZ2
MZ1	63.891	50.782	MZ1	$s^2_A + s^2_E$	s^2_A
MZ2	50.782	64.150	MZ2	s^2_A	$s^2_A + s^2_E$

Observed N=250 MZ pairs (**S**)			GCSM (Model) (Σ)		
	MZ1	MZ2		MZ1	MZ2
MZ1	63.891	50.782	MZ1	$s^2_A + s^2_E$	s^2_A
MZ2	50.782	64.150	MZ2	s^2_A	$s^2_A + s^2_E$



$s^2_{Height} = s^2_A + s^2_E$

The hypothesis of interest
Estimate s^2_A and s^2_E



The GID: the means to the end of estimating s^2_A and s^2_E

Observed N=250 MZ pairs (**S**)			GCSM (Model) (Σ)		
	MZ1	MZ2		MZ1	MZ2
MZ1	63.891	50.782	MZ1	$s^2_A + s^2_E$	s^2_A
MZ2	50.782	64.150	MZ2	s^2_A	$s^2_A + s^2_E$

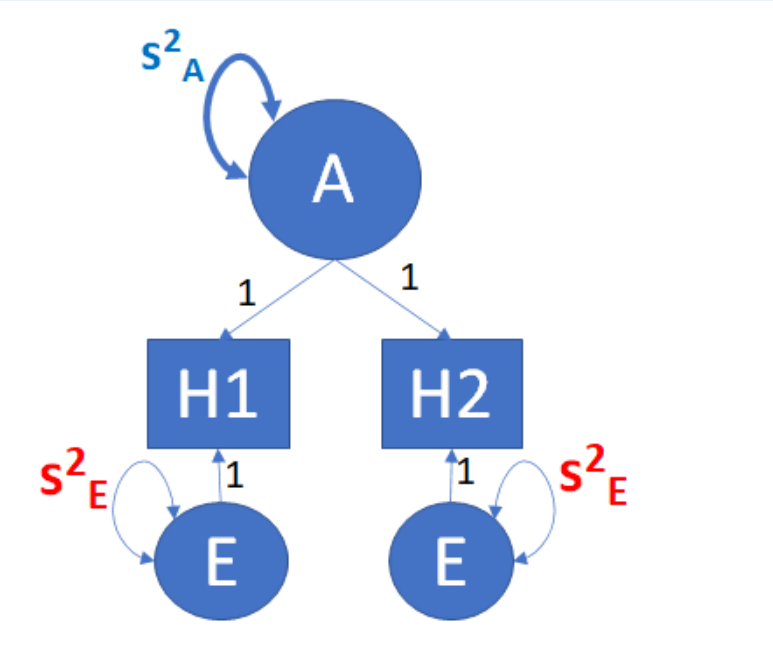
s^2_A = 50.782 (estimate of A variance component)

$s^2_E = s^2_{Ph} - s^2_A$ = 63.891 – 50.782 =13.11 and 64.150 – 50.782 = 13.37

s^2_E = (13.11 + 13.37) /2 = **13.24 (estimate of E variance component)**

GID (MZ)

Graphically: pathmodel



GID (MZ)

Regression Equations

$H_1 = m + A_1 + E_1$
$H_2 = m + A_2 + E_1$

or $(A_1 = A_2)$

$H_1 = m + A + E_1$
$H_2 = m + A + E_2$

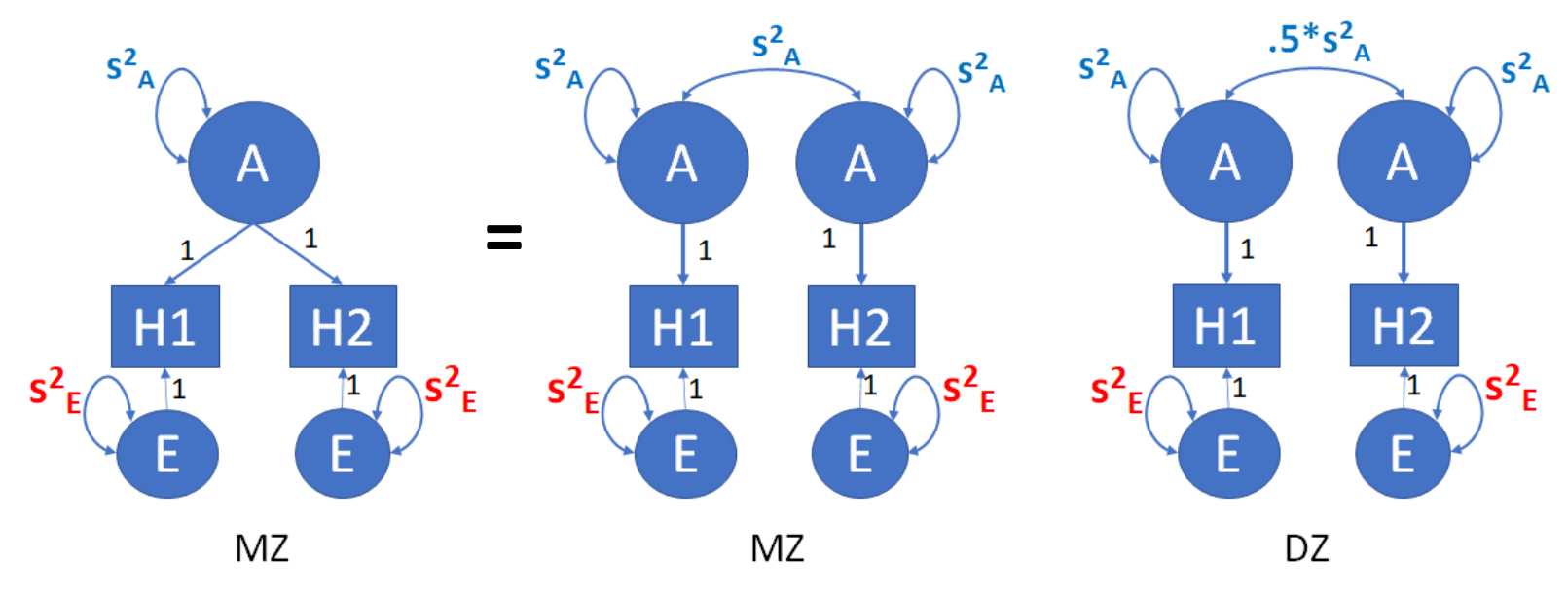
GID (MZ)

Covariance structure (variances and covariance)

GCSM (Model) (Σ)		
	MZ1	MZ2
MZ1	$s^2_A + s^2_E$	s^2_A
MZ2	s^2_A	$s^2_A + s^2_E$

This a weak GID... what have we assumed concerning the environment?
(see the MZ1-MZ2 covariance!)

Classical twin design: MZ twins and DZ twins (raised / growing up together in the same household) .... AE model



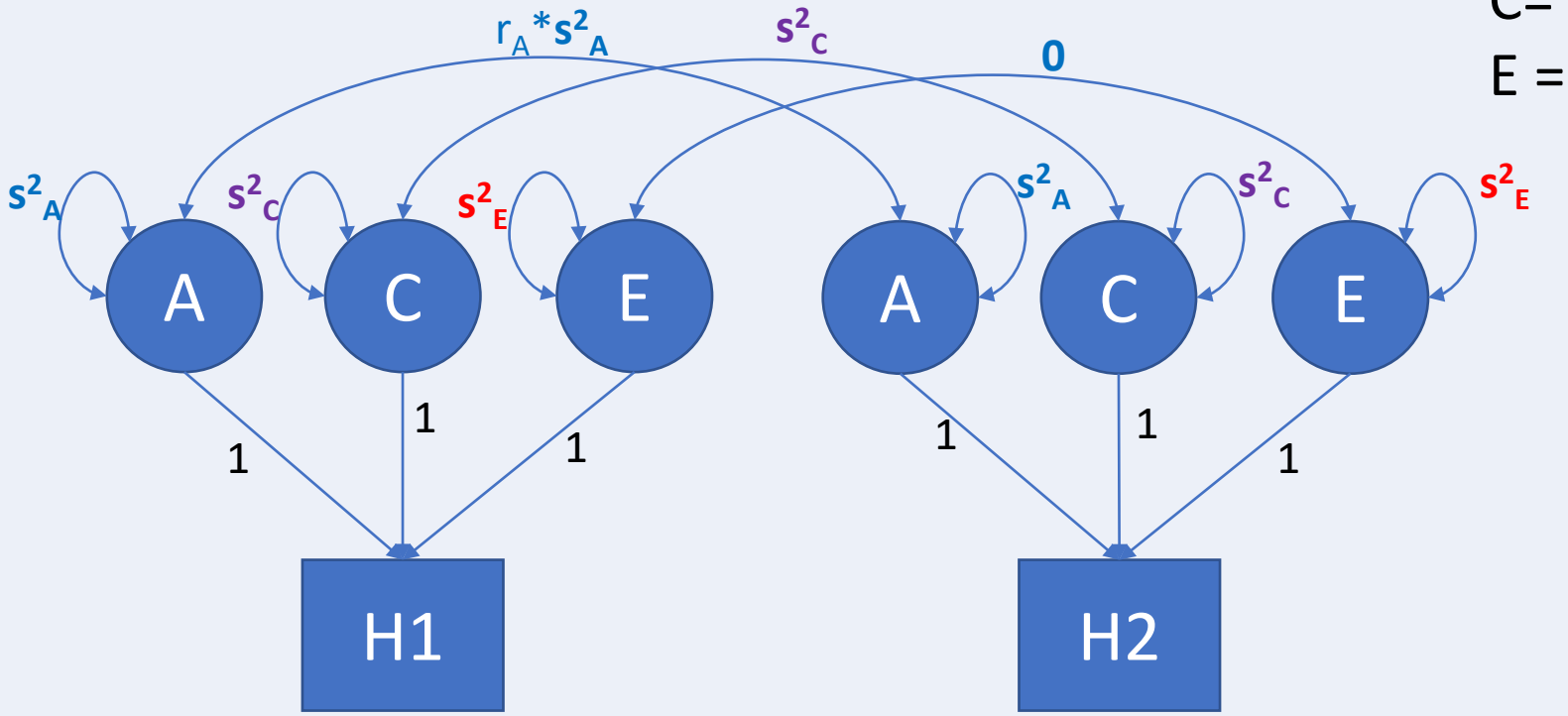
GCSM (Model) (Σ_{MZ})		
	MZ1	MZ2
MZ1	$s^2_A + s^2_E$	s^2_A
MZ2	s^2_A	$s^2_A + s^2_E$

GCSM (Model) (Σ_{DZ})		
	DZ1	DZ2
DZ1	$s^2_A + s^2_E$	$.5*s^2_A$
DZ2	$.5*s^2_A$	$s^2_A + s^2_E$

Why add DZ twins? To extend the model (add variance component): ACE model or ADE model

ACE model

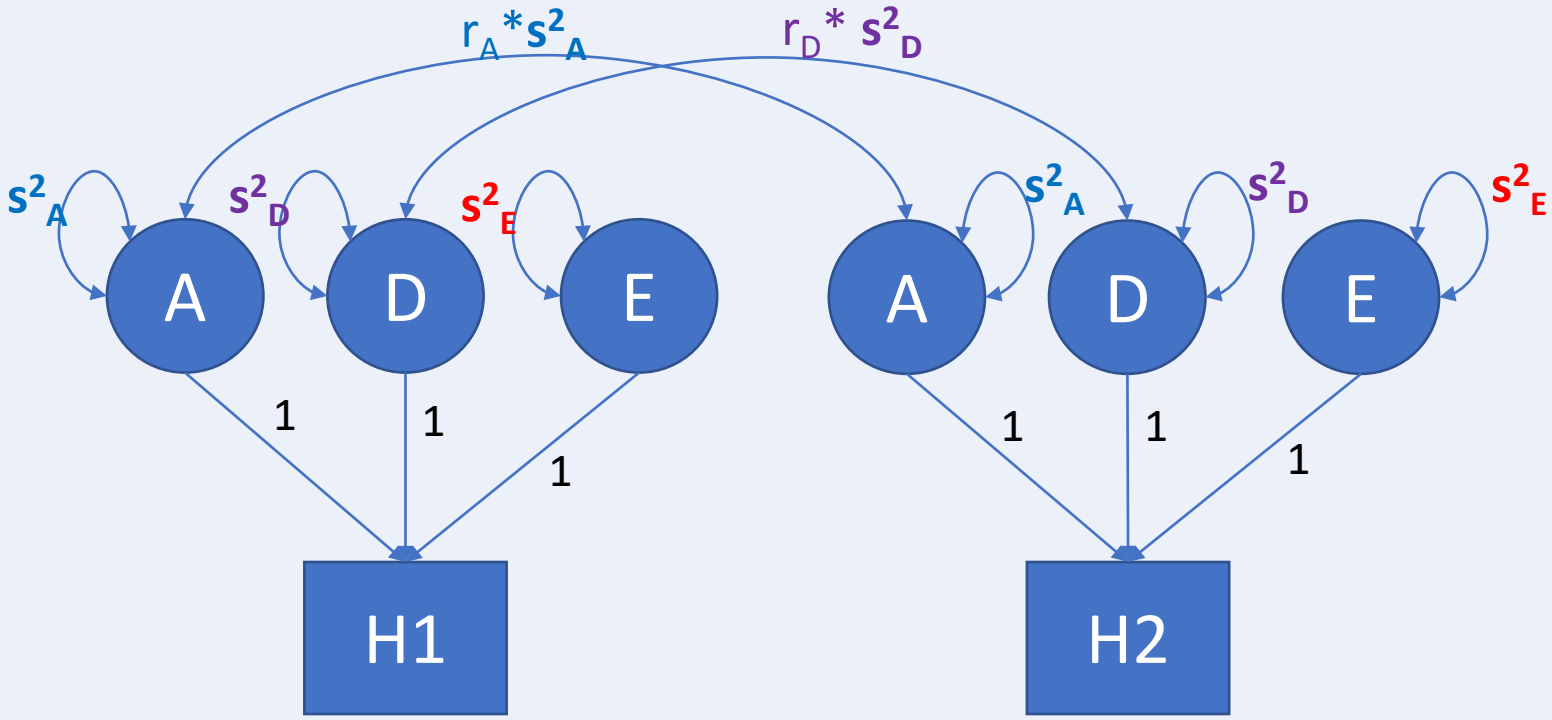
A = additive genetic
C= common (shared) environmental
E = unshared environmental
(+ measurement error)



GCSM (Model) (Σ_{MZ}) $r_A = 1$		
	MZ1	MZ2
MZ1	$s^2_A + s^2_C + s^2_E$	$1*s^2_A + s^2_C$
MZ2	$1*s^2_A + s^2_C$	$s^2_A + s^2_C + s^2_E$

GCSM (Model) (Σ_{DZ}) $r_A = ½$		
	DZ1	DZ2
DZ1	$s^2_A + s^2_C + s^2_E$	$½*s^2_A + s^2_C$
DZ2	$½*s^2_A + s^2_C$	$s^2_A + s^2_C + s^2_E$

ADE model



A = additive genetic
D= dominance genetic
E = unshared environmental
(+ measurement error)

GCSM (Model) (Σ_{MZ}) $r_A = 1$ $r_D = 1$		
	MZ1	MZ2
MZ1	$s^2_A + s^2_D + s^2_E$	$s^2_A + s^2_D$
MZ2	$s^2_A + s^2_D$	$s^2_A + s^2_D + s^2_E$

GCSM (Model) (Σ_{DZ}) $r_A = ½$ $r_D = ¼$		
	DZ1	DZ2
DZ1	$s^2_A + s^2_D + s^2_E$	$½ * s^2_A + ¼ * s^2_D$
DZ2	$½ * s^2_A + ¼ * s^2_D$	$s^2_A + s^2_D + s^2_E$

Illustration - height in females twins (mean age 23; std age 3.6)

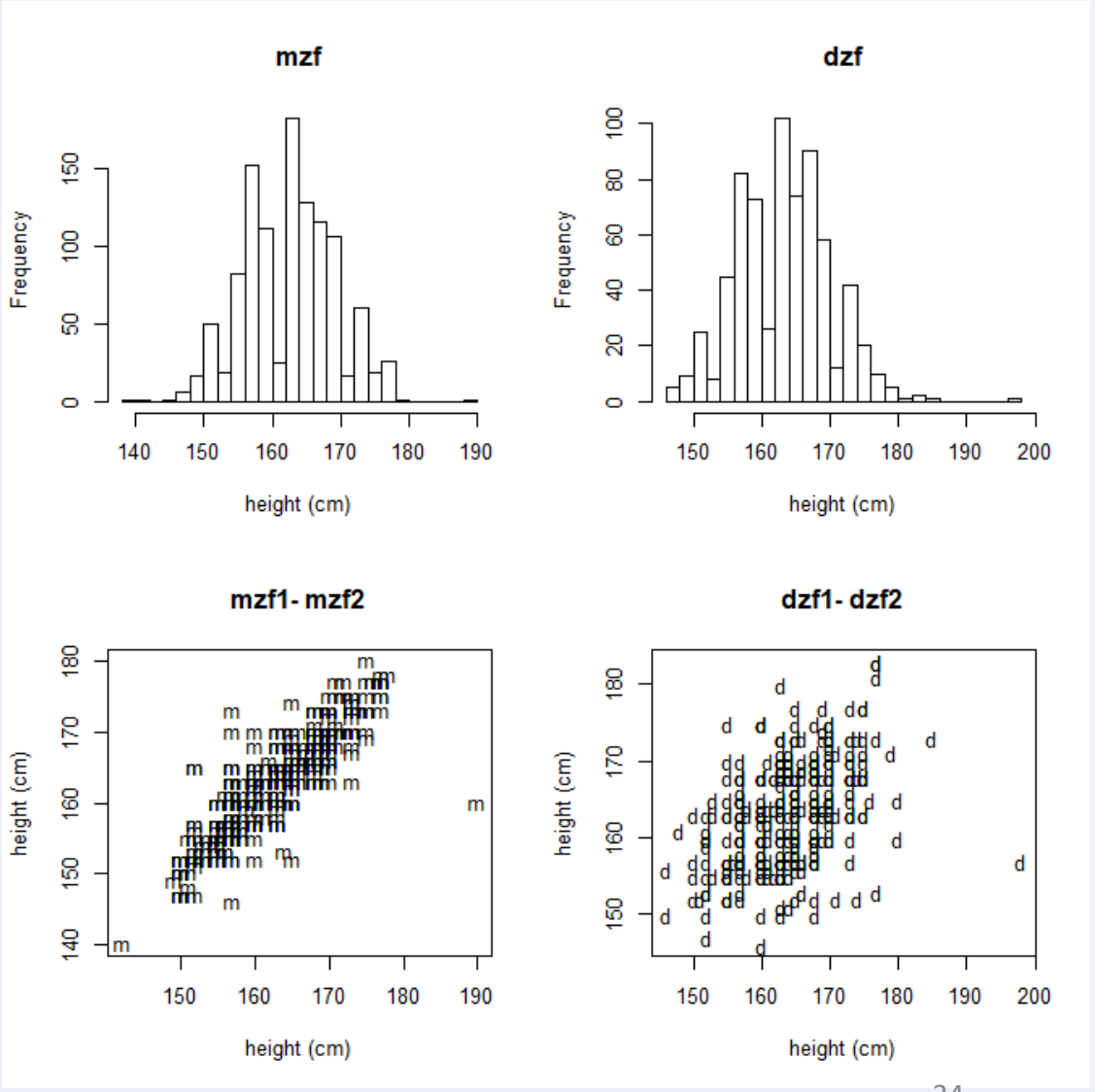
(twinData in OpenMx R library ... R code in slide notes)

```
MZF Data descriptives
    vars   N*  mean   sd    min    max
ht1    1 556 162.97 6.64 141.99 189.99
ht2    2 560 162.93 6.65 139.99 179.98
```

mzf correlation r_{MZ} = .878

```
DZF Data descriptives
    vars   N*  mean   sd min    max
ht1    1 348 164.09 6.94 146 198.00
ht2    2 343 163.28 6.73 146 182.98
```

dzf correlation r_{DZ} = .439



*note: variation in N is due to missing data

GID: the classical twin design.

Decomposing phenotypic variance based on ACE Model: $s^2_{Height} = s^2_A + s^2_C + s^2_E$

GCSM (Model) (Σ_{MZ})		
	MZ1	MZ2
MZ1	$s^2_A + s^2_C + s^2_E$	$s^2_A + s^2_C$
MZ2	$s^2_A + s^2_C$	$s^2_A + s^2_C + s^2_E$

GCSM (Model) (Σ_{DZ})		
	DZ1	DZ2
DZ1	$s^2_A + s^2_C + s^2_E$	$½*s^2_A + s^2_C$
DZ2	$½*s^2_A + s^2_C$	$s^2_A + s^2_C + s^2_E$

Observed data (variances, covariances, correlations)

Observed $\mathbf{S}_{MZ}$ ($\mathbf{R}_{MZ}$) (N=569)		
	MZ1	MZ2
MZ1	**44.068 (1)**	**38.721 (.878)**
MZ2	**38.721 (.878)**	**44.177 (1)**

GCSM (Model) $\mathbf{S}_{DZ}$ ($\mathbf{R}_{DZ}$) (N=351)		
	DZ1	DZ2
DZ1	**48.175 (1)**	**20.519 (.439)**
DZ2	**20.519 (.439)**	**45.319 (1)**

Quick method based on standardized phenotypes .... Falconer's equations

$s^2_A + s^2_C + s^2_E$ = variance = 1
$s^2_A + s^2_C$ = r_{MZ} = .878
$½ * s^2_A + s^2_C$ = r_{DZ} = .439

three equations,
three unknowns, three knowns

solve for the unknowns....

ACE model, if $(2*r_{DZ}) \geq r_{MZ}$

s^2_A = $2*(r_{MZ}-r_{DZ})$ = 2*(.878 - .439) = .878
s^2_C = $2*r_{DZ}-r_{MZ}$ = 2*.439-.878 = 0.0
s^2_E = $1- s^2_A - s^2_C$ = 1-.0-.878 = .122

Solution

Conclusion given $s^2_A + s^2_C + s^2_E$ =.878 + 0 + .122 = 1. In young females adults, **#1)** 87.8% of variance is genetic (87.8% of phenotypic differences due to genetic differences); **#2)** No contribution of shared environment; #3) **12.2%** of variance is environmental (+ measurement error).

Note: ADE model equations in slide notes (used if $(2*r_{DZ}) < r_{MZ}$)

In practice, we use genetic covariance structure modeling to fit models to data collected in genetically informative design.

1) Optimal estimates of parameters + information about precision of estimates (95% CIs)
2) Overall goodness of fit testing: does the specified model fit the observed covariance matrices?
3) Statistical testing of individual parameters (ACE vs AE; ACE vs CE; ADE vs AE).

and

4) Generalizes from 1 phenotype to P phenotypes (multivariate phenotype / repeated measures)
5) Accommodates missing data
6) Can handle binary / dichotomous phenotypic data
7) Can handle any (multivariate) Genetically Informative Design
(e.g. twins + parents; twins + siblings; children of twin design; extended pedigree design)

GCSM

We have the data (MZ and DZ twin phenotypic data)

The data summary (linear relationship): two covariance matrices and the 4 means

We have a linear model for the data (pathmodel), which implies covariance structure(s)

The covariance structure(s) are covariance matrices expressed in terms of unknown (to be estimated) and known parameters (GID!).

The classical twin design: unknown parameters (s^2_A , s^2_C, s^2_e) and known parameters (1, ½, ¼)

GID: the classical twin design.

Decomposing phenotypic variance based on ACE Model: $s^2_{Height} = s^2_A + s^2_C + s^2_E$

GCSM (Model) (Σ_{MZ})		
	MZ1	MZ2
MZ1	$s^2_A + s^2_C + s^2_E$	$s^2_A + s^2_C$
MZ2	$s^2_A + s^2_C$	$s^2_A + s^2_C + s^2_E$

GCSM (Model) (Σ_{DZ})		
	DZ1	DZ2
DZ1	$s^2_A + s^2_C + s^2_E$	$½*s^2_A + s^2_C$
DZ2	$½*s^2_A + s^2_C$	$s^2_A + s^2_C + s^2_E$

Observed data (variances, covariances, correlations)

Observed $\mathbf{S}_{MZ}$ ($\mathbf{R}_{MZ}$) (N=569)		
	MZ1	MZ2
MZ1	**44.068 (1)**	**38.721 (.878)**
MZ2	**38.721 (.878)**	**44.177 (1)**

GCSM (Model) $\mathbf{S}_{DZ}$ ($\mathbf{R}_{DZ}$) (N=351)		
	DZ1	DZ2
DZ1	**48.175 (1)**	**20.519 (.439)**
DZ2	**20.519 (.439)**	**45.319 (1)**

25

How to obtain estimates? Maximum likelihood estimation

Illustration of ML estimation in GCSM with MZ and DZ twin design (phenotype height)

Observed $\mathbf{S}_{MZ}$ ($\mathbf{R}_{MZ}$) (N=569)		
	MZ1	MZ2
MZ1	**44.068 (1)**	**38.721 (.878)**
MZ2	**38.721 (.878)**	**44.177 (1)**

GCSM (Model) $\mathbf{S}_{DZ}$ ($\mathbf{R}_{DZ}$) (N=351)		
	DZ1	DZ2
DZ1	**48.175 (1)**	**20.519 (.439)**
DZ2	**20.519 (.439)**	**45.319 (1)**

μ **the phenotypic mean**
s^2_A **genetic variance**
s^2_C **shared env variance**
s^2_E **unshared env variance**

Parameters associated with the hypothesis ACE model

OpenMx ML estimates ACE model ... Hypothesis ACE model

free parameters:

	name	matrix	row	col	Estimate	Std.Error	4 Parameters	
1	mean	meanH	1	1	163.296892	0.2045165	μ	the phenotypic mean
2	VA11	VA	1	1	41.162844	4.1639267	s^2_A	genetic variance
3	VC11	VC	1	1	-1.093649	4.1465672	s^2_C	shared env variance
4	VE11	VE	1	1	5.419555	0.3276949	s^2_E	unshared env variance

Model Statistics:

	Parameters	Degrees of Freedom	Fit (-2lnL units)
Model:	4	1803	11135.91 .... $-2*f_{ML}(\theta)$

OpenMx ML estimates AE model i.e., $s^2_C = 0$ (fixed to zero) ... Hypothesis AE model

free parameters:

	name	matrix	row	col	Estimate	Std.Error	3 Parameters	
1	mean	meanH	1	1	163.29555	0.2051183	μ	the phenotypic mean
2	VA11	VA	1	1	40.18631	1.8228455	s^2_A	genetic variance
3	VE11	VE	1	1	5.42909	0.3269049	s^2_E	unshared env variance

Model Statistics:

	Parameters	Degrees of Freedom	Fit (-2lnL units)
Model:	3	1804	11135.99 .... $-2*f_{ML}(\theta)$

Likelihood ratio test. ACE vs AE .... can we "drop" C (i.e., set $s^2_C = 0$)?

A statistical test of the hypothesis $s^2_C = 0$ **based on the values of the likelihood functions**

Model Statistics:

	Parameters	Degrees of Freedom	Fit (-2lnL units)
Model:	4	1803	11135.91

Model Statistics:

	Parameters	Degrees of Freedom	Fit (-2lnL units)
Model:	3	1804	11135.99

Test statistic called the (log-)Likelihood Ratio test (LRT):

11135.99 - 11135.91 = .08

If H-null: $s^2_C = 0$ is true, the LRT is distributed chi2(1), where 1 (df) = 4-3, difference in the number of parameters

.08, df=1, p-value = .777 (in R pchisq(.08,1,lower=F)
If p-value < alpha (e.g. .01), we would reject $s^2_C = 0$.
Here we conclude $s^2_C = 0$ ... So there is no shared environmental variance
no shared environmental contributions to the phenotype variance in height

$$s^2_{Height} = s^2_A + \cancel{s^2_C} + s^2_E$$

$s^2_{Height} = s^2_A + s^2_E = 40.186 + 5.429 = 45.615$

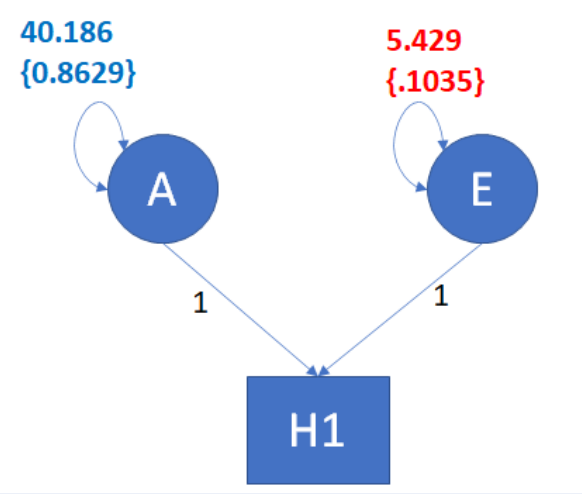
standardized variance components

$s^2_A / \{s^2_A + s^2_E\} = 40.186 / 45.615 = .881$ **(a.k.a "narrow-sense" heritability, a proportion like R^2)**
$s^2_E / \{s^2_A + s^2_E\} = 5.429 / 45.615 = .119$

95% confidence intervals of the standardized variance components:

	lbound	estimate	ubound
$s^2_A / \{s^2_A + s^2_E\}$	0.8629	0.881	0.8964
$s^2_E / \{s^2_A + s^2_E\}$	0.1035	0.119	0.1370

95% CI tell us how precise the estimate are ...



Genetically informative design & Genetic covariance structure analysis:

A brief introduction based on the classical twin design

Conor V. Dolan & Micheal C. Neale

PPT presentation in 4 parts .... PART 3 (8 slides):
CTD multivariate ACE models from 1 to p (p>1) phenotypes - limited to ACE (ADE models also possible)
Illustration Height and Weight

Univariate ACE model (one phenotype: s^2_A and s^2_C and s^2_E are variances)

GCSM (Model) (Σ_{MZ})		
	MZ1	MZ2
MZ1	$s^2_A + s^2_C + s^2_E$	$s^2_A + s^2_C$
MZ2	$s^2_A + s^2_C$	$s^2_A + s^2_C + s^2_E$

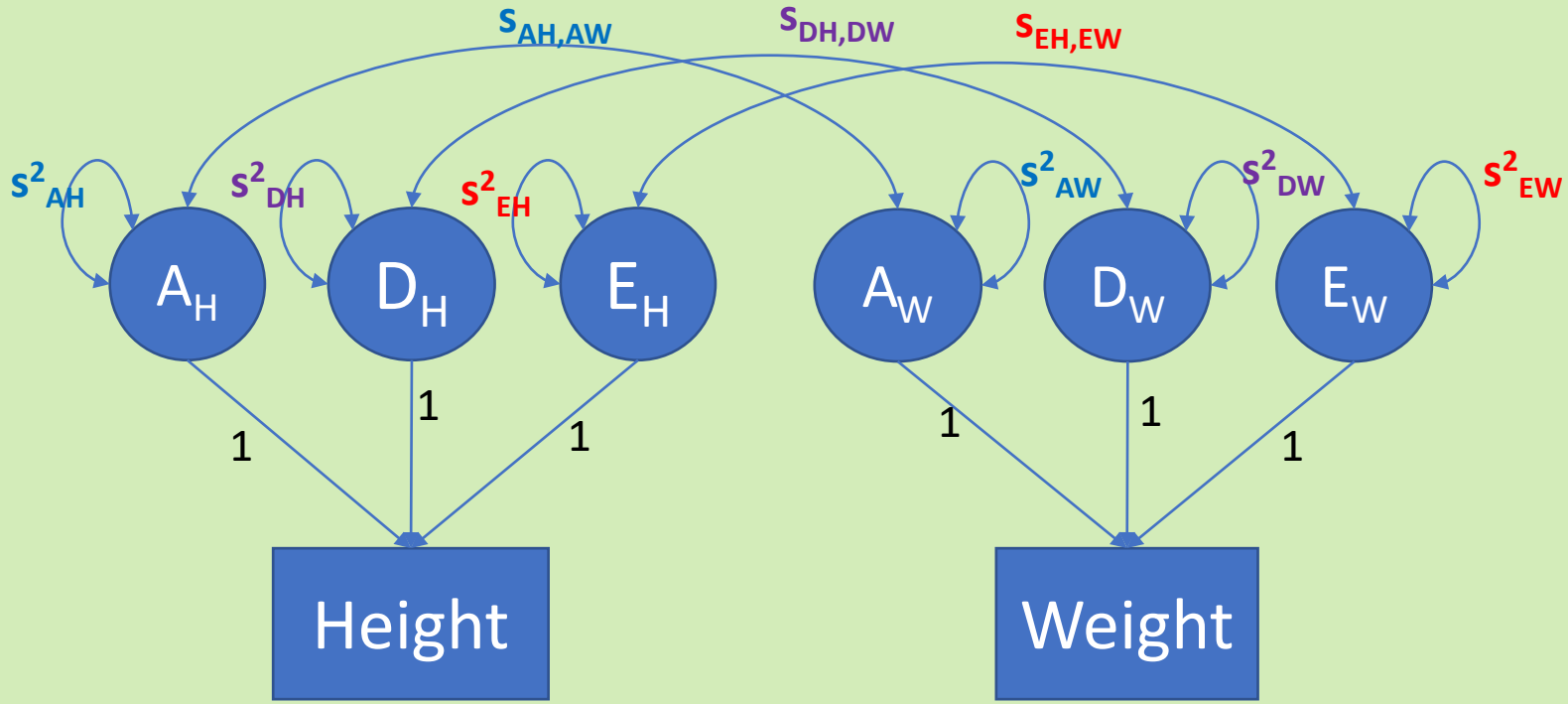
GCSM (Model) (Σ_{DZ})		
	DZ1	DZ2
DZ1	$s^2_A + s^2_C + s^2_E$	$\frac{1}{2}*s^2_A + s^2_C$
DZ2	$\frac{1}{2}*s^2_A + s^2_C$	$s^2_A + s^2_C + s^2_E$

Classical twin model generalizes readily to the multivariate case.
p-phenotypes: S_A and S_C and S_E are pxp covariance matrices in the ACE model

GCSM (Model) (Σ_{MZ}) $r_A = 1$		
	MZ1	MZ2
MZ1	$S_A + S_C + S_E$	$S_A + S_C$
MZ2	$S_A + S_C$	$S_A + S_C + S^2_E$

GCSM (Model) (Σ_{DZ}) $r_A = \frac{1}{2}$		
	DZ1	DZ2
DZ1	$S_A + S_C + S_E$	$\frac{1}{2}*S_A + S_C$
DZ2	$\frac{1}{2}*S_A + S_C$	$S_A + S_C + S_E$

Path diagram of 2 phenotypes: height and weight. Hypothesis / aim: $S_{PH} = S_A + S_D + S_E$



S_{PH} = S_A + S_D + S_E

S_{PH}	H	W
H	s^2_H	$s_{H,W}$
W	$s_{H,W}$	s^2_W

=

S_A	H	W
H	s^2_{AH}	$s_{AH,AW}$
W	$s_{AH,AW}$	s^2_{AW}

+

S_D	H	W
H	s^2_{DH}	$s_{DH,DW}$
W	$s_{DH,DW}$	s^2_{DW}

+

S_E	H	W
H	s^2_{EH}	$s_{EH,EW}$
W	$s_{EH,EW}$	s^2_{EW}

Covariance matrices (correlations)

MZ	H1	W1	H2	W2
H1	44.068 (1)	28.066	38.721	24.283
W1	28.066 (.493)	73.441 (1)	27.702	63.359
H2	38.721 (.878)	27.702 (.486)	44.177 (1)	26.909
W2	24.283 (.415)	63.359 (.839)	26.909 (.459)	77.662 (1)

DZ	H1	W1	H2	W2
H1	48.175 (1)	26.426	20.519	14.952
W1	26.426 (.441)	74.632 (1)	10.158	26.773
H2	20.519 (.439)	10.158 (.175)	45.319 (1)	28.205
W2	14.952 (.234)	26.773 (.337)	28.205 (.456)	84.564 (1)

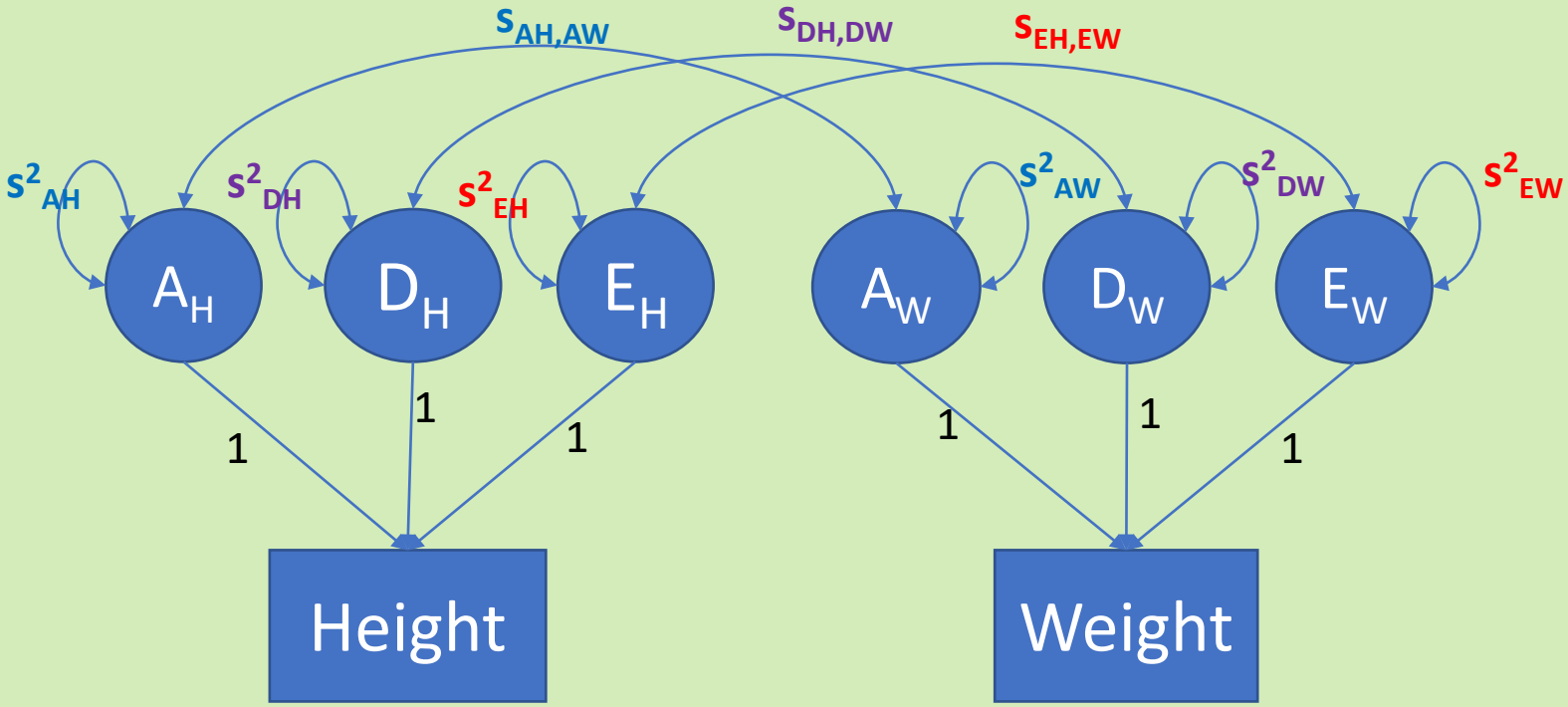
Covariance matrices (m)

$S_A + S_D + S_E$	$S_A + D$
$S_A + S_D$	$S_A + S_D + S_E$

$S_A + S_D + S_E$	$.5*S_A + .25*S_D$
$.5*S_A + .25*S_D$	$S_A + S_D + S_E$

S_A, S_D, and S_E are 2x2 matrices (2 phenotypes)

Path diagram of 2 phenotypes: height and weight. Hypothesis / aim: $S_{PH} = S_A + S_D + S_E$



LRT results
ADE model vs AE model:
LRT stat = 6.62, df=3, p=0.085
Drop D, reduce model to AE

S_{PH} = S_A + ~~S_D~~ + S_E

S_{PH}	H	W
H	s^2_H	$s_{H,W}$
W	$s_{H,W}$	s^2_W

=

S_A	H	W
H	s^2_{AH}	$s_{AH,AW}$
W	$s_{AH,AW}$	s^2_{AW}

+

S_C	H	W
H	s^2_{DH}	$s_{DH,DW}$
W	$s_{DH,DW}$	s^2_{DW}

+

S_E	H	W
H	s^2_{EH}	$s_{EH,EW}$
W	$s_{EH,EW}$	s^2_{EW}

S_{PH} = S_A + S_E

S_{PH}	H	W
H	s^2_H	$s_{H,W}$
W	$s_{H,W}$	s^2_W

=

S_A	H	W
H	s^2_{AH}	$s_{AH,AW}$
W	$s_{AH,AW}$	s^2_{AW}

+

S_E	H	W
H	s^2_{EH}	$s_{EH,EW}$
W	$s_{EH,EW}$	s^2_{EW}

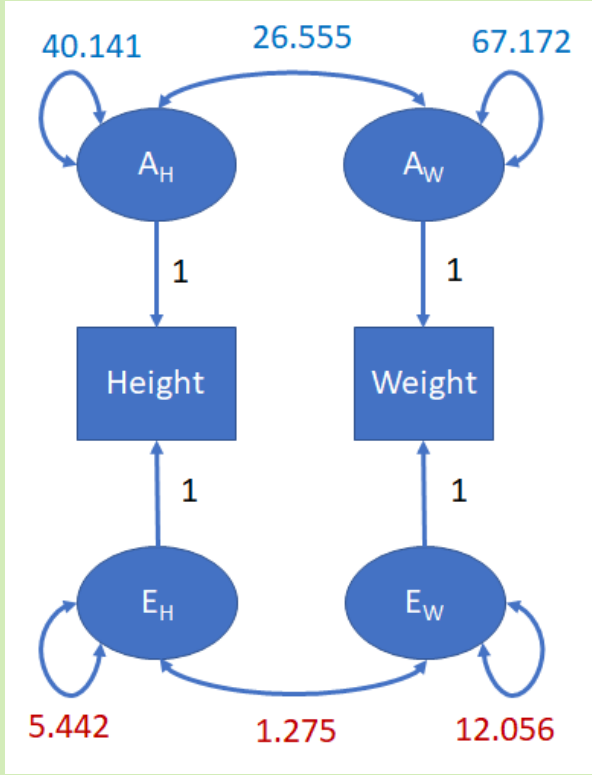
S_{PH}	H	W
H	45.584	27.829
W	27.829	79.228

=

S_A	H	W
H	40.141	26.555
W	26.555	67.172

+

S_E	H	W
H	5.442	1.275
W	1.275	12.056



Bivariate AE model reveals:

1) Contribution of A and E to phenotypic height variance (s^2_{AH} s^2_{EH})

(40.141 / 45.484 = .881; 5.442 / 45.484 = .119)

2) Contribution of A and E to phenotypic weight variance (s^2_{AW} s^2_{EW})

(67.172 / 79.228 = .848; 12.056 / 79.228 = .152)

3) Contribution of A and E to phenotypic height - weight covariance ($s_{AH,AW}$ $s_{EH,EW}$)

(26.555 / 27.829 = .954; 1.275/27.829 = .046)

.... **Pleiotropy** is used to denote genetic effects common to 2 or more phenotypes.

R script in slide notes

The p-variate twin model based on the CTD represents the following hypothesis

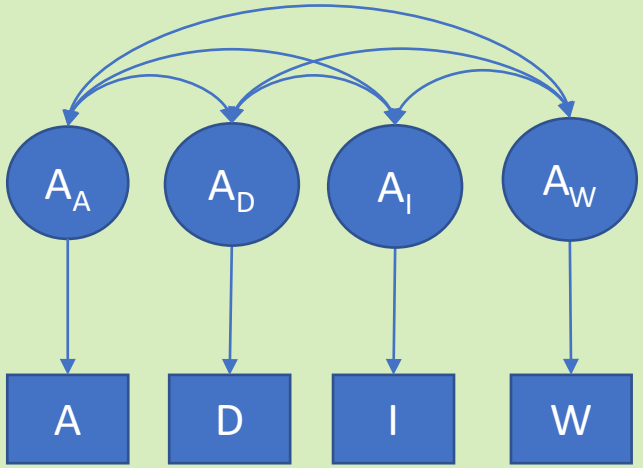
$\mathbf{S_{PH} = S_A + S_C + S_E}$ (or $\mathbf{S_{PH} = S_A + S_D + S_E}$)

where $\mathbf{S_A}$, $\mathbf{S_C}$ and $\mathbf{S_E}$ are pxp covariance matrices

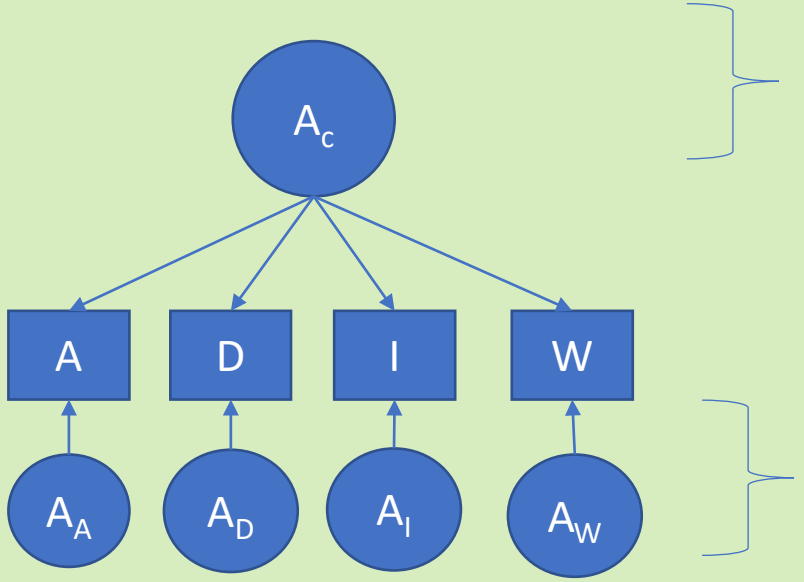
In genetic covariance structure pxp covariance matrices $\mathbf{S_A}$, $\mathbf{S_C}$ and $\mathbf{S_E}$ may themselves be subject to a covariance structure model. Well known models in standard phenotypic covariance structure modeling, can be applied to $\mathbf{S_A}$, $\mathbf{S_C}$ and $\mathbf{S_E}$.

Example: **common factor model**

Suppose $\mathbf{S_{PH}} = \mathbf{S_A} + \mathbf{S_E}$ where S_{PH} is the phenotypic covariance of p=4 phenotypes: anxiety, depression, introversion, withdrawnness



S_A as a 4x4 covariance matrix with 10 parameters (estimated not modeled)



Common set (source of pleiotropy) source of variance and covariance

Unique set source of variance, not covariance

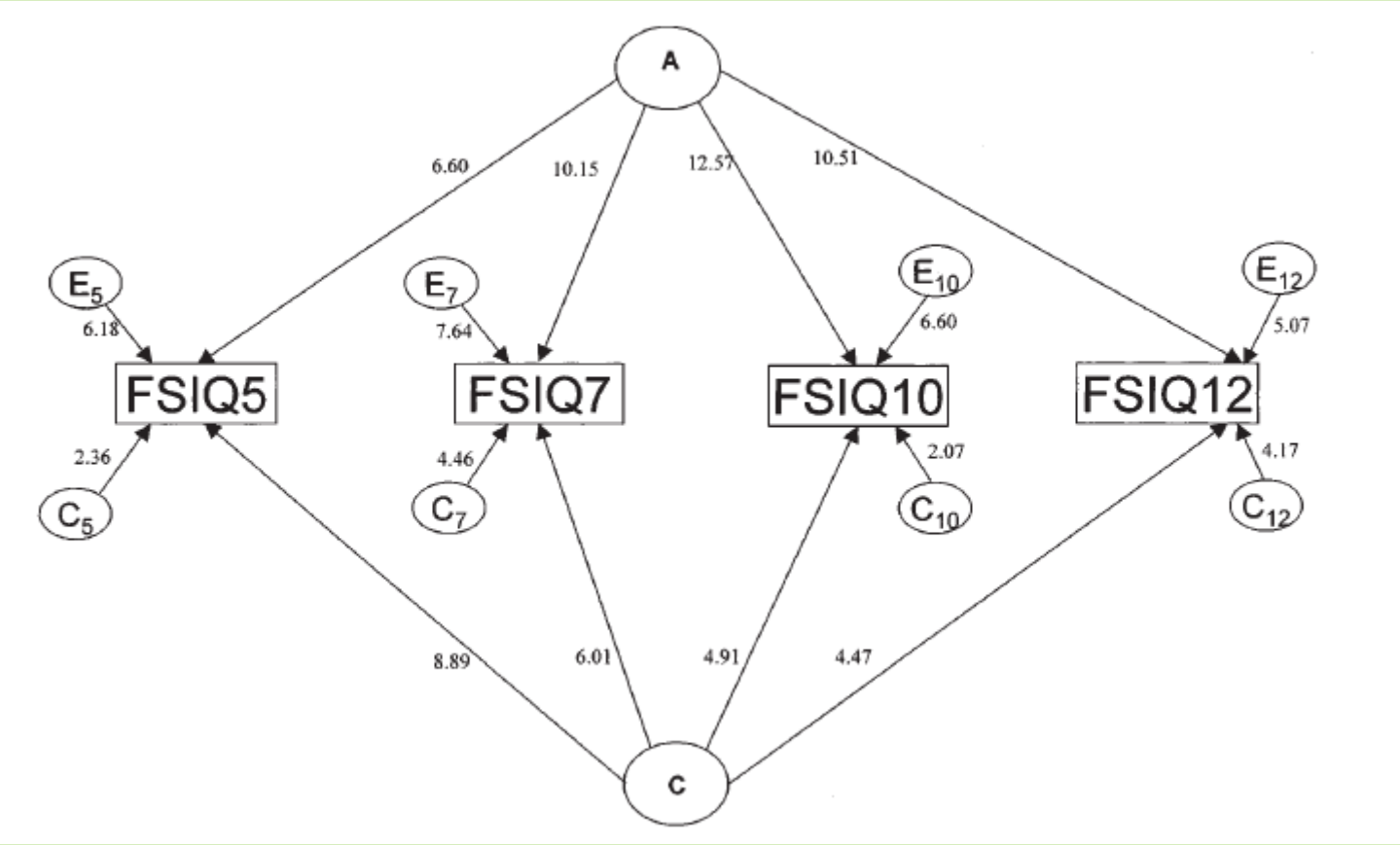
S_A as a 4x4 covariance matrix with 8 parameters (estimated subject to specified structure: 1 common factor model)

We know that these phenotypes are phenotypically correlated (correlations between .4 and .6).
Hypothesis: The phenotypic correlations are due to a set of genes common to the four phenotypes (A_c).
In addition each phenotype has its own unique set of genes.... ($A_A A_D A_I A_W$)

S_C 4x4 covariance matrix ... a 1 common factor model

S_A 4x4 covariance matrix ... a 1 common factor model without phenotype specific genetic residuals

S_E 4x4 covariance matrix ... a diagonal matrix (E does not contribute to the phenotypic covariance)



Bartels M, Rietveld MJH, Baal van GCM, Boomsma DI. (2002) *Behavior Genetics, 32, 237-249.*

Genetically informative design & Genetic covariance structure analysis:

brief introduction based on the classical twin design

Conor V. Dolan & Michael C. Neale

PPT presentation in 4 parts .... PART 4 (13 slides):

The classical twin design (CTD) assumptions

Other GIDs

Assumptions of the CTD: generalizability.

The CTD is a means to an end .... $\mathbf{S_{PH}} = \mathbf{S_A} + \mathbf{S_C} + \mathbf{S_E}$, cognitive abilities in **12 year olds**

The MZ and DZ twins are representative of the "target" population (i.e., 12 year olds).

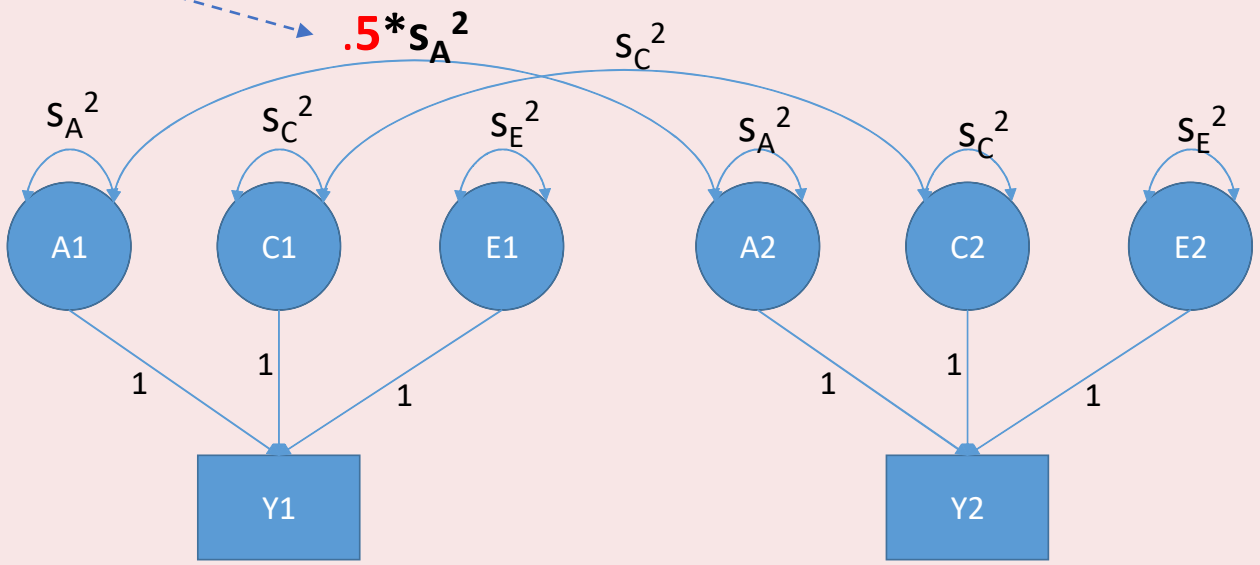
12 year old Dutch urban MZ twins are representative of 12 year old Dutch urban children.
12 year old Dutch urban DZ twins are representative of 12 year old Dutch urban children.

In a study of IQ (say), this means *statistically* ...

phenotypically: same mean, same variance,
genetically: same genetic influences / genetic variants
environmentally: same environmental influences

Assumption: random mating (testable if you have parental data .... $r_{spouse} = 0$).

The **.5 in** $.5*s_A^2$ is based on the assumption of random mating



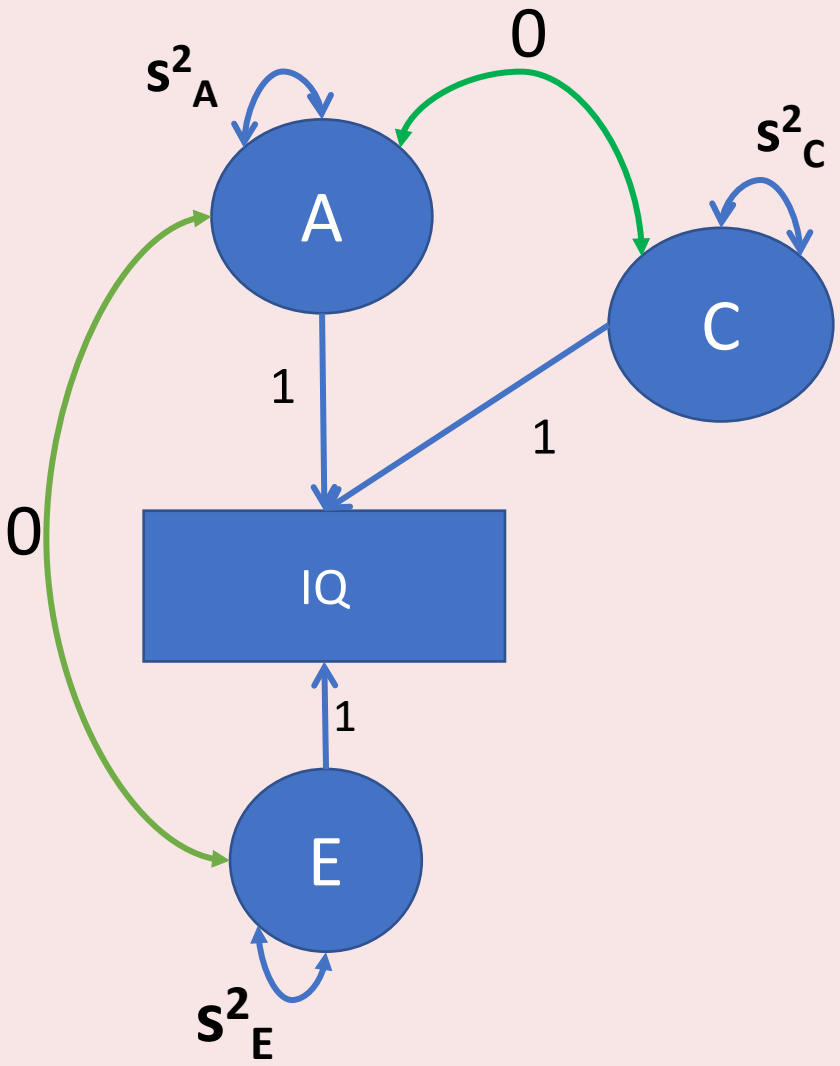
Positive non-random mating (a.k.a. *assortative mating*) may result in $r_A * s_A^2$, where $r_A > .5$.

Simple test: what is the phenotypic spousal correlation r_{spouse}?

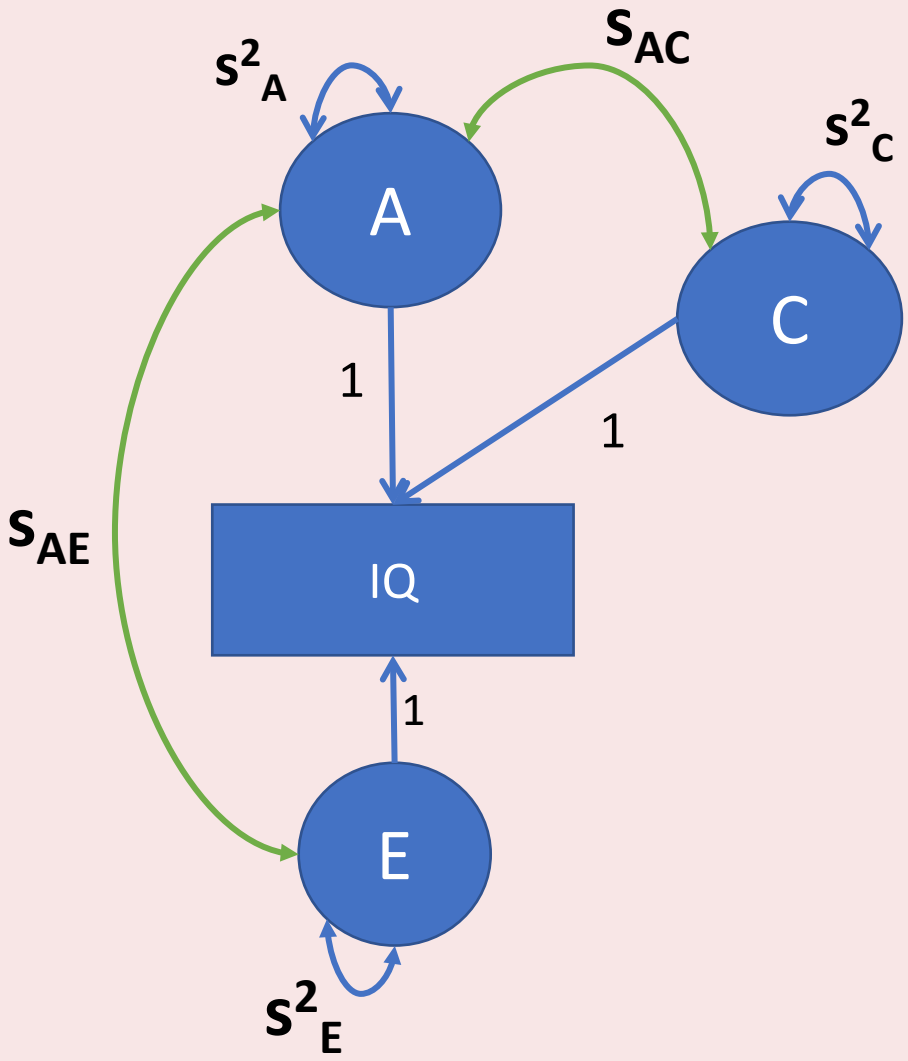
If $r_{spouse} > 0$, then we acknowledge assortative mating ..... This raises the question: what process underlies assortative mating?

Is mating random? Height r_{spouse} = ~ .2 ... IQ r_{spouse} = ~ .3 to ~.4

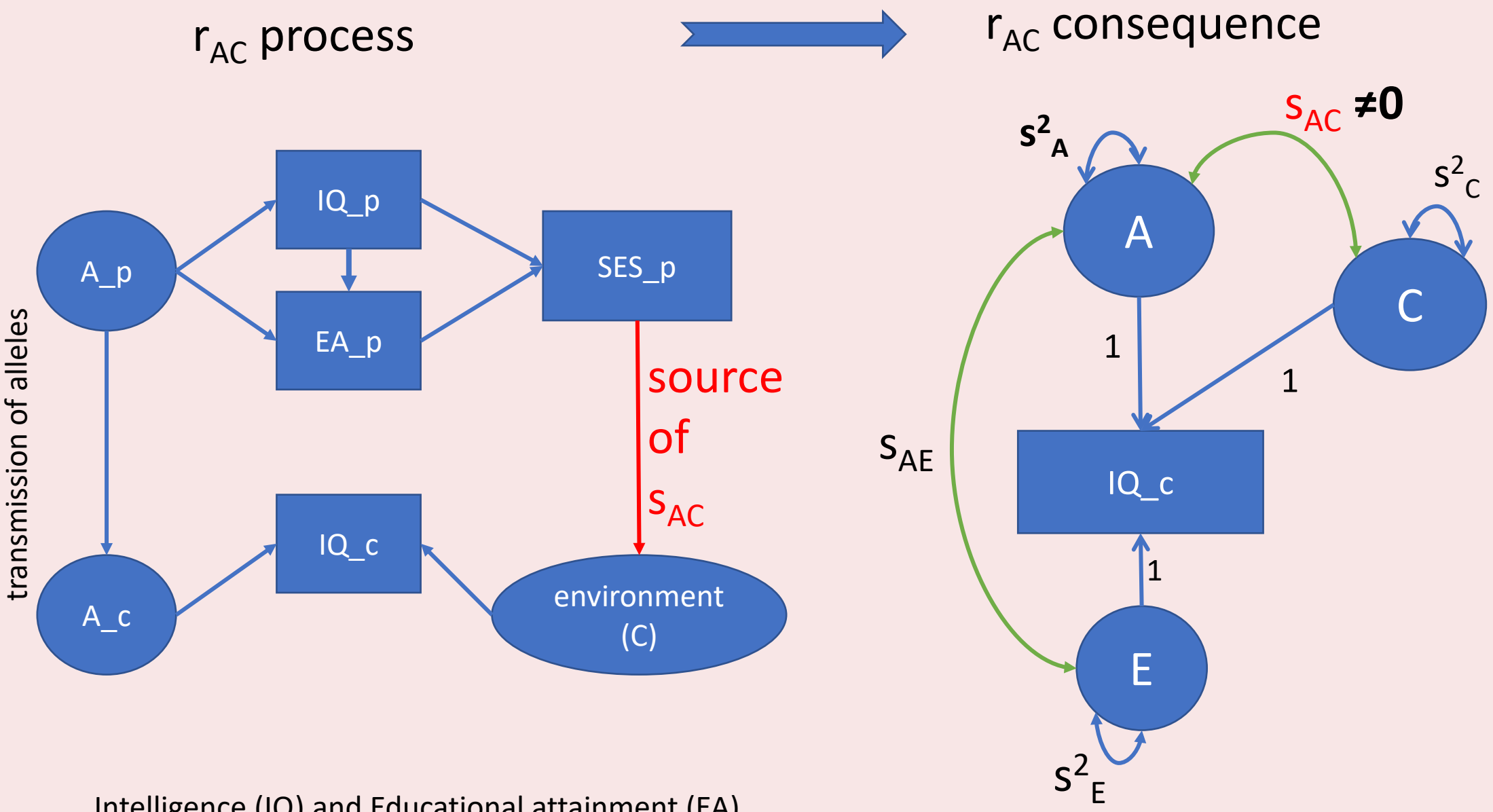
no A-E (r_{AE}) and A-C (r_{AC}) correlation



A-E (s_{AE}≠0) and A-C (s_{AC} ≠0) covariance



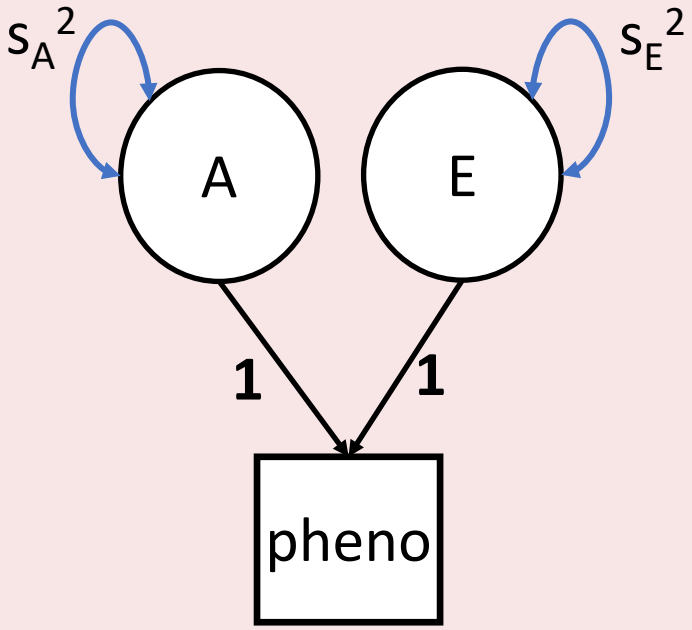
Crucial: What process gives rise to rAC, rAE?



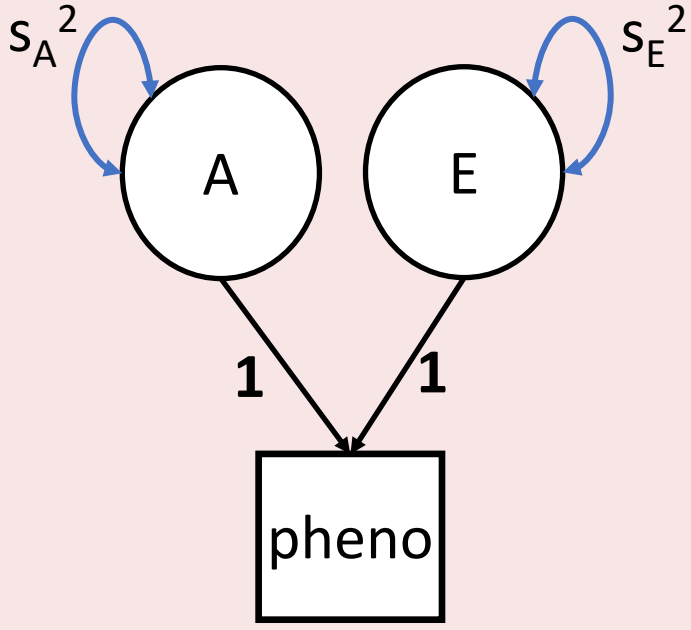
Intelligence (IQ) and Educational attainment (EA)

Moderation / interaction

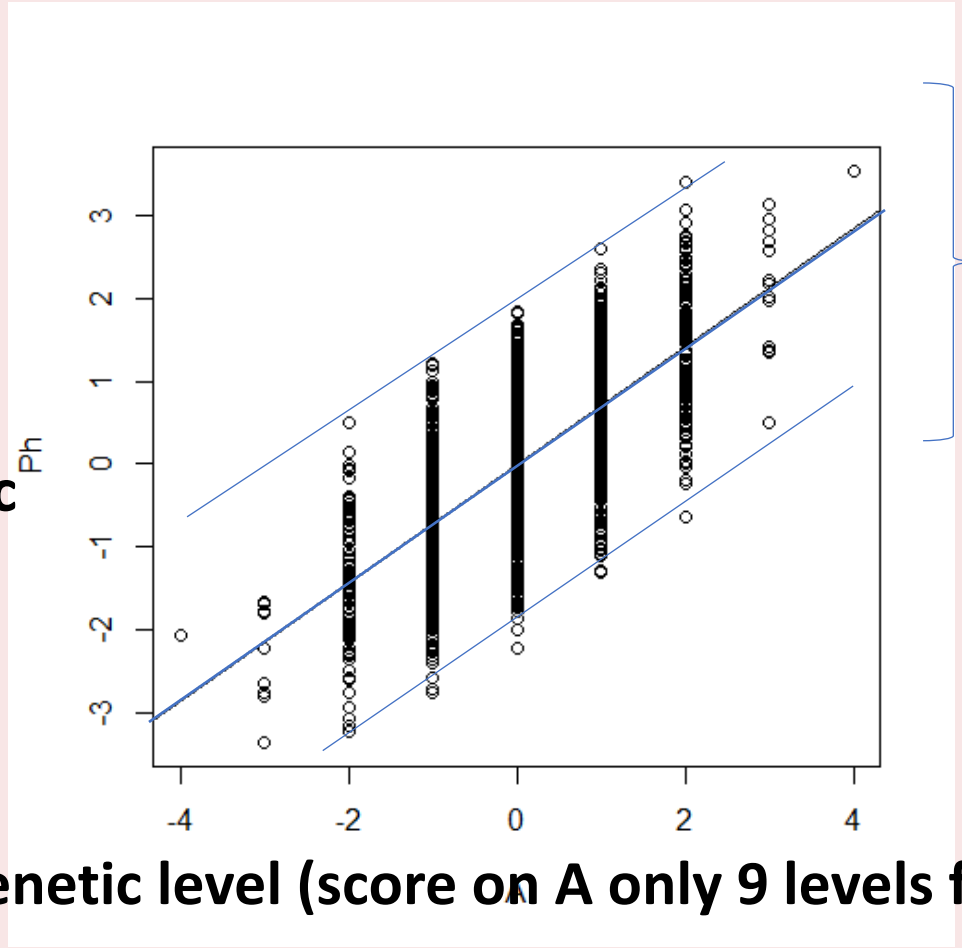
The effect of A is expressed as s_A^2 and quantified as $s_A^2/(s_A^2 + s_E^2)$. The effect of A does not depend on E: E does not moderate the effect of A. There is no AxE interaction.



The effect of E is expressed as s_E^2 and quantified as $s_E^2/(s_A^2 + s_E^2)$. The effect of E does not depend on A: A does not moderate the effect of E. There is no AxE interaction.



NO AxE interaction



Phenotypic scores

Genetic level (score on A only 9 levels for clarity)

Environmental Dispersion / variance

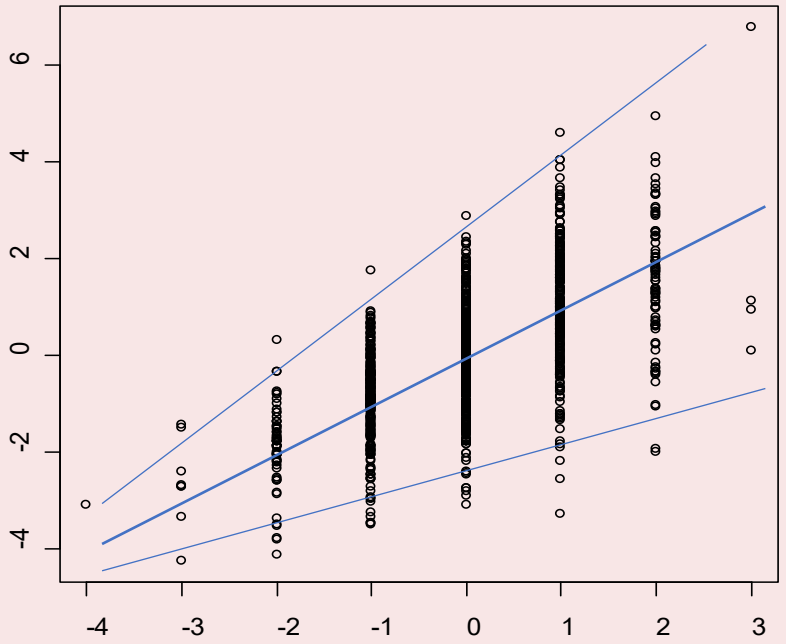
Variance of E given A score ... does not depend on A

a.k.a. homoskedasticity

s_E^2 is **constant** over levels of A:
environmental effects (s_E^2) are the same given any value of A

AxE interaction

Phenotypic scores

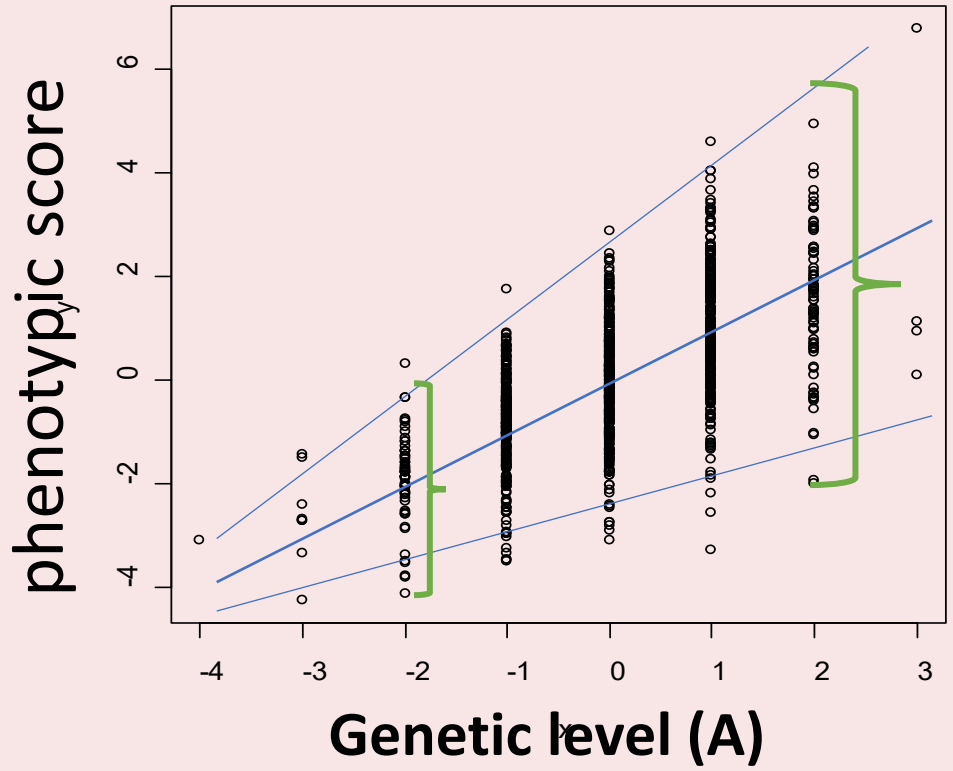


Conditional variance of E given A

a.k.a. heteroskedasticity

G x E as "genetic control" of E effects: The effect of E, expressed as s_E^2 is a function of A: $s_E^2 = f(A)$

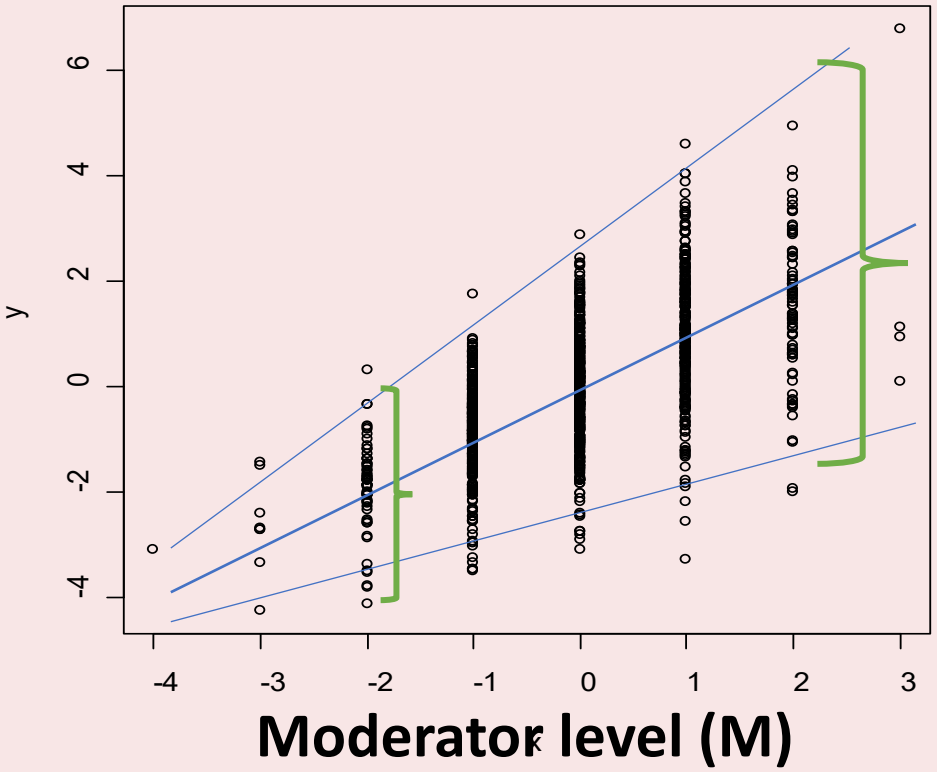
Interaction AxE



s^2_{Ei}

Environmental effects (variance) depend on A-level

Moderation



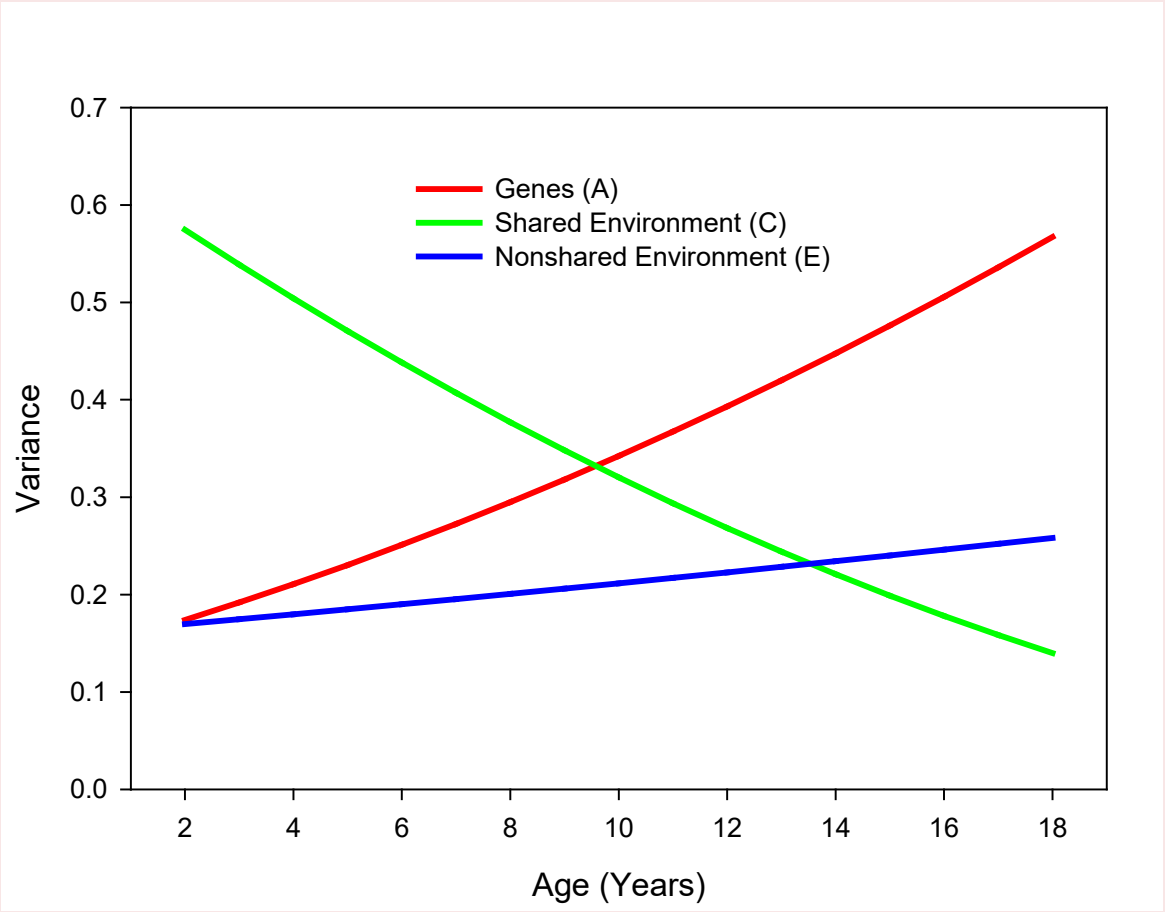
any one or more

s^2_{Ai}
s^2_{Ci}
s^2_{Ei}

Environmental effects and /or genetic effects (variances) depend on M-level (linear increase)

$s^2_E = f(M)$ $s^2_A = f(M)$ $s^2_C = f(M)$

S. Purcell (2002). Variance Components Models for Gene–Environment Interaction in Twin Analysis Twin Research Volume 5 Number 6 pp. 554-571



Psychological Science OnlineFirst, published on December 15, 2015 as doi:10.1177/0956797615612727

Research Article

Large Cross-National Differences in Gene × Socioeconomic Status Interaction on Intelligence

Elliot M. Tucker-Drob[1,2] and Timothy C. Bates[3]
[1]Department of Psychology, University of Texas at Austin; [2]Population Research Center, University of Texas at Austin; and [3]Department of Psychology, University of Edinburgh

Psychological Science
1–12

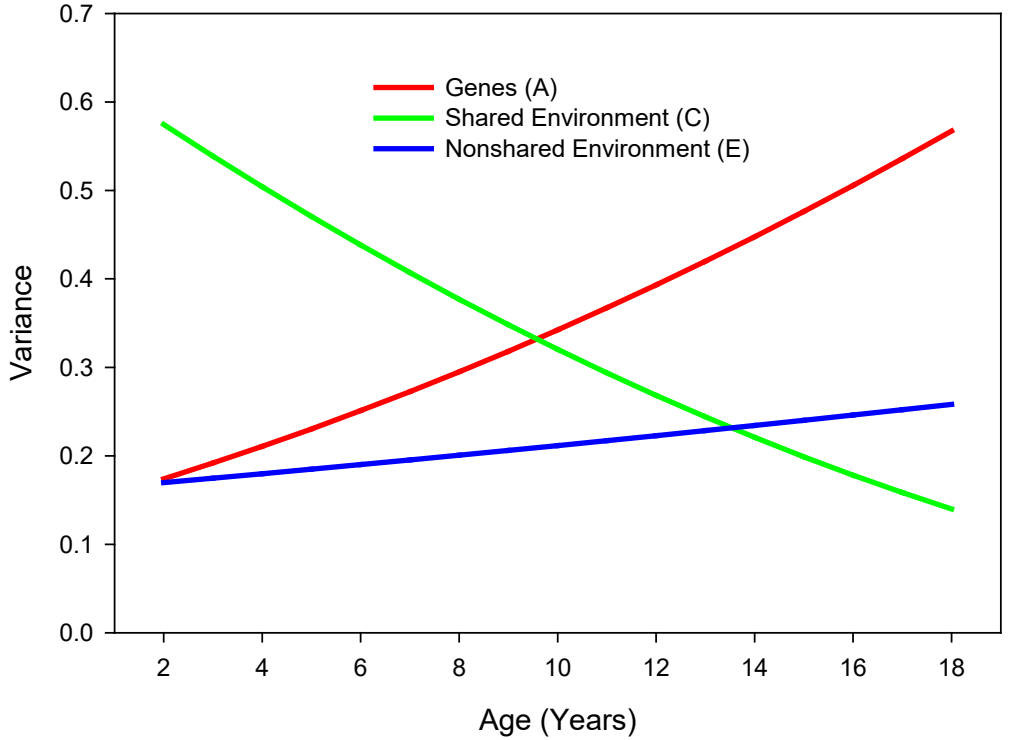
DOI: 10.1177/0956797615612727
pss.sagepub.com

Age a continuous moderator of genetic and environmental effects on IQ

Tucker-Drob & Bates (2015): Variance components. Moderation model with measured moderator (age) …. *A increase and C decreases with age*

Suppose that there is GxE interaction or G-E covariance, but you fit the standard ACE twin model... how are the variance components biased?

		consequence in the CTD
Assumption	violation	Bias
AxE	AxE>0 mimics E	E overestimated
AxC	AxC>0 mimics A	A overestimated
cov(A,C)	cov(A,C) >0 mimics C	C overestimated
cov(A,E)	cov(A,E) >0 mimics A	A overestimated



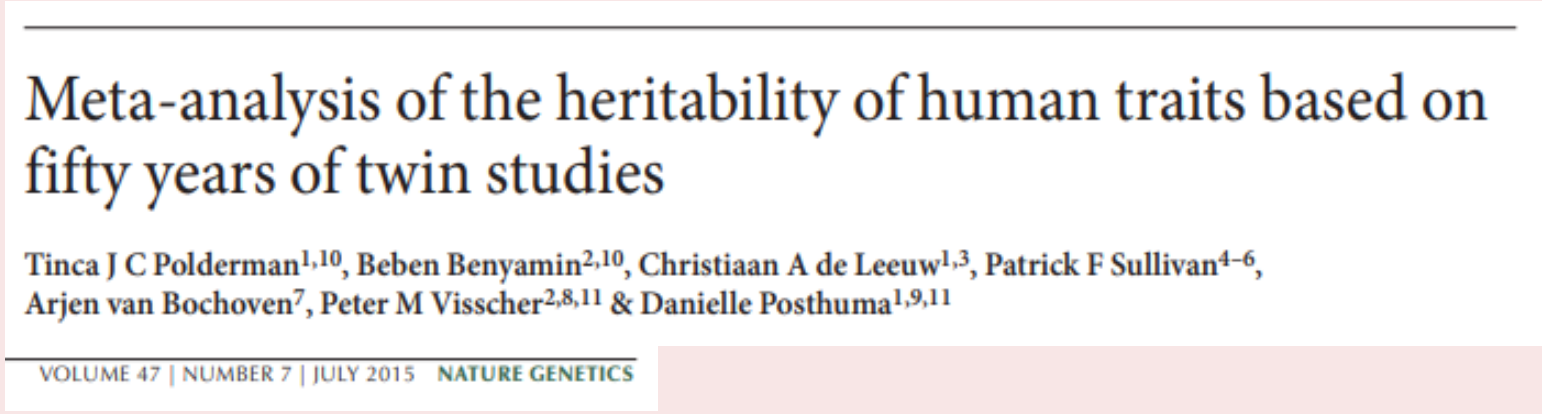
		consequence in the CTD
Assumption	violation	Bias
cov(A,C)	cov(A,C) >0 mimics C	C overestimated
cov(A,E)	cov(A,E) >0 mimics A	A overestimated

C variance declines with age .... Effect of C declines with age ....
However, cov(AC) results in C overestimation in the twin model....
So the large C variance in early years could be due in part to cov(AC).

So the apparent C variance may - in part - be cov(AC) and the decline in C effects, may be due - in part - to a change in process that gives rise to the cov(AC)

.... just sayin' interpret your variance components carefully .... bearing in mind possible violations of model assumptions

CTD is ONE genetically informative design (GID) that has proven to be highly productive and very influential... (changed the view of genetics in psychology!)



Meta-analysis of the heritability of human traits based on fifty years of twin studies

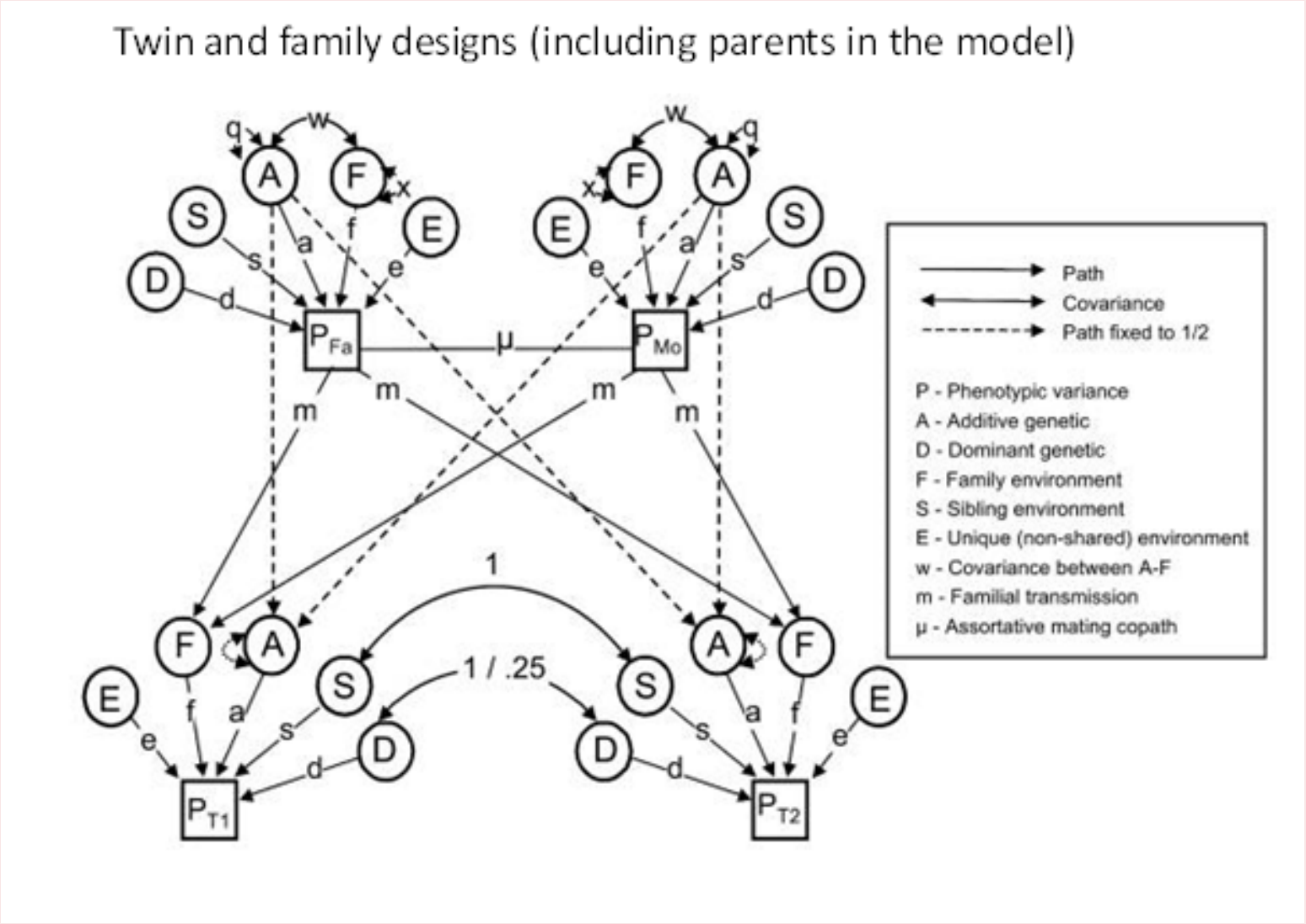
Tinca J C Polderman[1,10], Beben Benyamin[2,10], Christiaan A de Leeuw[1,3], Patrick F Sullivan[4–6], Arjen van Bochoven[7], Peter M Visscher[2,8,11] & Danielle Posthuma[1,9,11]

VOLUME 47 | NUMBER 7 | JULY 2015 NATURE GENETICS

It is generally recognized to be a useful design that includes strong assumptions.

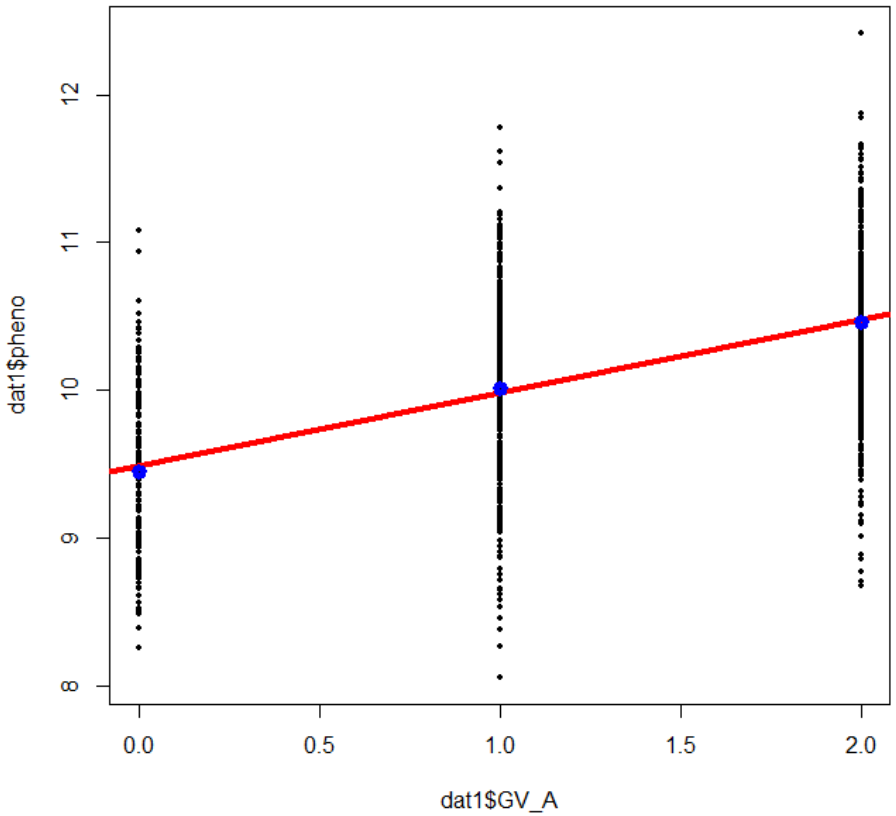
However the CTD is not the only GID ... more extended designs are less dependent on assumptions or allow one to test assumptions. Examples of extended GIDs:

Nuclear Twin Family Design (e.g., Keller et al 2009 Twin Research and Human Genetics)
Children of Twin Designs (e.g., McAdams et al 2018 Behavior Genetics)
Twins and spouses Designs (e.g. Reynolds et al 2006 Behavior Genetics)



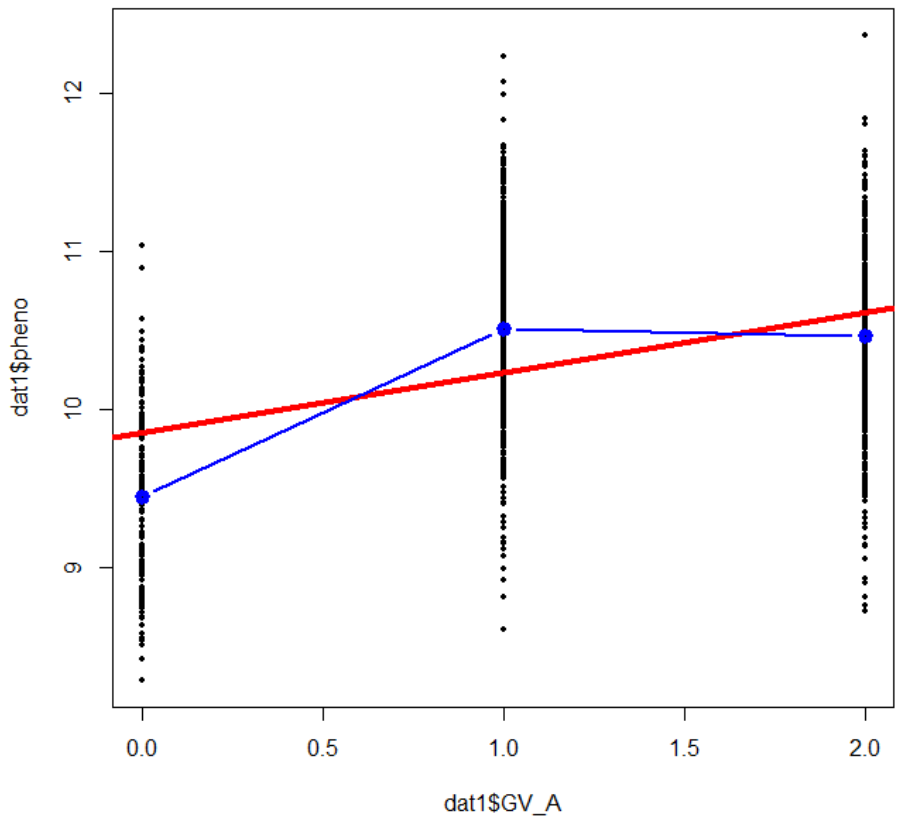
Includes assortative mating (μ), s_{AC} (A-C covariance) stemming from cultural transmission (m)

Keller et al (2009) Twin Res Hum Genet. 2009 Feb;12(1):8-18. doi: 10.1375/twin.12.1.8.



blue dots are the conditional mean
read line is the linear regression line

data are consistent with linear regression



blue dots are the conditional mean
read line is the linear regression line

data are not consistent with linear regression
see difference between conditional means and the regression
.... that is dominance