



## ORBIT - Online Repository of Birkbeck Institutional Theses

---

Enabling Open Access to Birkbeck's Research Degree output

Genome-wide analyses to investigate the genetic factors underlying specific psychotic experiences in adolescence and their overlap with psychiatric disorders

<https://eprints.bbk.ac.uk/id/eprint/40326/>

Version: Full Version

**Citation: Pain, Oliver (2018) Genome-wide analyses to investigate the genetic factors underlying specific psychotic experiences in adolescence and their overlap with psychiatric disorders. [Thesis] (Unpublished)**

© 2020 The Author(s)

---

All material available through ORBIT is protected by intellectual property law, including copyright law.

Any use made of the contents should comply with the relevant law.

---

[Deposit Guide](#)  
Contact: [email](#)

**Genome-wide Analyses to Investigate the  
Genetic Factors Underlying Specific Psychotic  
Experiences in Adolescence and Their Overlap  
with Psychiatric Disorders**

Oliver Pain

Birkbeck, University of London

Submitted for the Degree of Doctor of Philosophy, September 2017

## Declaration & Statement of Independent Work

I, Oliver Henry Charles Pain, hereby declare that, except where explicit attribution is made, the work presented in this thesis is entirely my own.

### Exceptions:

My PhD focuses on investigating the common genetic effects underlying adolescent psychotic experiences. Very large sample sizes are required for this type of project, requiring collaboration with previously established studies. Data in this thesis is drawn from three studies: the Twins Early Development Study (TEDS), the Avon Longitudinal Study of Parents and Children (ALSPAC), and the Child and Adolescent Twin Study in Sweden (CATSS). Further details of these studies are provided in Chapters 2 and 4. The details of what was carried out by myself during this thesis are provided throughout the thesis. However, in summary, my thesis involves analysis of existing genotypic or phenotypic data.

Signed:  \_\_\_\_\_

Date: 23/02/2018

# Abstract

Psychotic experiences (PEs) are non-clinical traits, which at the extreme resemble symptoms of psychotic disorders, such as schizophrenia. PEs during adolescence have been associated with a range of psychiatric disorders, including schizophrenia, bipolar disorder, and major depression. Adolescent PEs are moderately heritable, however no genetic variant has been associated with adolescent PEs at genome-wide significance. There are limited and mixed findings regarding a common genetic overlap between adolescent PEs and psychiatric disorders.

Following a systematic review of previous studies using genome-wide genetic data to investigate adolescent PEs, this thesis sets out to improve upon previous research through two main approaches: 1) the use of specific and quantitative measures of adolescent PEs, and 2) the combined analysis of multiple samples. In Chapter 2, a GWAS (genome-wide association study) is performed using specific and quantitative measures of adolescent PEs using the TEDS (Twins Early Development Study) sample. In Chapter 3, the procedure in which phenotypic data is normalised and controlled for covariates is investigated. The remainder of the thesis is based on the combined analysis of three European adolescent samples (TEDS and two others) with available PE data. In Chapter 4, the phenotypic data relating to PEs within each sample are harmonised to create four measures assessing specific PE traits that are comparable across samples. These four traits are Paranoia and Hallucinations, Cognitive Disorganisation, Anhedonia, and Parent-rated Negative Symptoms. In Chapter 5, mega-GWASs of the four specific PE traits (N = 6,297-10,098) are performed across the three samples to highlight associated genetic variation. Chapter 6 then estimates the variance in specific PEs, and the covariance between PEs, that is attributable to common genetic variation. Chapter

7 uses both polygenic risk scoring and LD-score regression to test for common genetic overlap between specific adolescent PEs and schizophrenia, bipolar disorder, and major depression.

This thesis provides evidence that specific PEs during adolescence show common genetic effects, and have a common genetic overlap with psychiatric disorders, specifically schizophrenia and major depression. The findings of this thesis are placed in the context of previous research, with a discussion of the limitations and future directions.

# Table of contents

<b>Declaration &amp; Statement of Independent Work</b> .....	<b>2</b>
<b>Abstract</b> .....	<b>3</b>
<b>Table of contents</b> .....	<b>5</b>
<b>List of figures</b> .....	<b>13</b>
<b>List of tables</b> .....	<b>16</b>
<b>List of supplementary notes</b> .....	<b>20</b>
<b>List of supplementary figures</b> .....	<b>21</b>
<b>List of supplementary tables</b> .....	<b>25</b>
<b>Acknowledgments</b> .....	<b>28</b>
<b>Publications</b> .....	<b>29</b>
<b>Chapter 1 - Introduction</b> .....	<b>30</b>
<b>1.1 - Adolescence</b> .....	<b>30</b>
1.1.1 - Adolescence: A critical period.....	30
1.1.2 - Mental health in adolescence.....	31
<b>1.2 - Psychotic experiences (PEs)</b> .....	<b>32</b>
1.2.1 - Definition and relationship with psychotic disorders.....	32
1.2.2 - PEs as risk factors.....	33
1.2.3 - Assessment of adolescent PEs.....	36
<b>1.3 - Molecular genetics</b> .....	<b>37</b>
1.3.1 - DNA.....	37
1.3.2 - Genetic variation .....	38
1.3.3 - Identification of genetic variants underlying rare and common disease .....	39
<b>1.4 - Genome-wide association studies (GWAS)</b> .....	<b>41</b>
1.4.1 - The premise of GWAS.....	41

1.4.2 – Control of population stratification.....	42
1.4.3 – The multiple testing problem and genome-wide significance .....	43
1.4.4 – Factors affecting power in GWAS .....	44
1.4.5 – GWAS consortia .....	45
<b>1.5 - Genetics of adolescent PEs.....</b>	<b>46</b>
1.5.1 – Estimating heritability and co-heritability .....	47
1.5.2 – Heritability of PEs .....	48
1.5.3 – Genetic association between PEs and clinical outcomes.....	49
<b>1.6 - Genome-wide analysis of adolescent PEs – A systematic review.....</b>	<b>49</b>
1.6.1 - Results of systematic review .....	50
<b>1.7 - Future directions for genetic research of adolescent PEs .....</b>	<b>57</b>
<b>1.8 - Aims .....</b>	<b>57</b>

## **Chapter 2 - GWAS of adolescent psychotic experiences in the Twins**

<b>Early Development Study .....</b>	<b>59</b>
<b>2.1 - Introduction.....</b>	<b>59</b>
<b>2.2 - Methods .....</b>	<b>60</b>
2.2.1 - Participants .....	60
2.2.2 - Measure of specific PEs .....	60
2.2.3 - Phenotypic analyses.....	61
2.2.4 - DNA collection and genotyping.....	61
2.2.5 - Association analysis.....	63
<b>2.3 - Results.....</b>	<b>64</b>
2.3.1 - Descriptive statistics.....	64
2.3.2 - Genome-wide association analysis.....	64
<b>2.4 - Discussion .....</b>	<b>81</b>
<b>2.5 – Appendix.....</b>	<b>84</b>

<b>Chapter 3 - Investigating the effect of the procedures used when controlling for the normality assumption and covariates.....</b>	<b>88</b>
<b>3.1 - Introduction .....</b>	<b>88</b>
<b>3.2 - Methods.....</b>	<b>90</b>
3.2.1 - Simulation of phenotypic data.....	90
3.2.2 - Simulation of covariate data.....	91
3.2.3 - Testing the effect of rank-based inverse normal transformation after regressing out covariate effects.....	91
3.2.4 - Testing the effect of applying rank-based inverse normal transformation (randomly splitting ties) before regressing out covariate effects.....	92
3.2.5 - Demonstration using real data.....	93
<b>3.3 - Results .....</b>	<b>94</b>
3.3.1 - Simulated data.....	94
3.3.1.1 - The effect of rank-based inverse normal transformation after regressing out covariate effects using simulated data .....	94
3.3.1.2 - The effect of rank-based inverse normal transformation (randomly splitting ties) before regressing out covariates using simulated data.....	97
3.3.2 - Real data .....	98
3.3.2.1 - Effect of rank-based inverse normal transformation of residuals when using real data.....	98
3.3.2.2 - Effect of rank-based inverse normal transformation before regressing out covariates when using real data .....	98
<b>3.4 - Discussion .....</b>	<b>99</b>
<b>3.5 - Appendix.....</b>	<b>105</b>
 <b>Chapter 4 - Harmonising Subscales of Specific Psychotic Experiences Across TEDS, ALSPAC and CATSS .....</b>	 <b>145</b>

<b>4.1 - Introduction .....</b>	<b>145</b>
<b>4.2 - Methods .....</b>	<b>146</b>
4.2.1 - Samples .....	146
4.2.2 - Exclusion criteria .....	147
4.2.3 - Measures.....	150
4.2.4 - Analyses .....	151
<b>4.3 - Results.....</b>	<b>153</b>
4.3.1 - Stage 1: Identifying PE items in ALSPAC and CATSS that matched SPEQ items.....	153
4.3.2 - Stage 2: Psychometric analysis of identified PE items within ALSPAC and CATSS.....	160
4.3.3 - Stage 3: Adaptation and psychometric analysis of TEDS' SPEQ scale based on the subscales available in ALSPAC and CATSS.....	168
<b>4.4 - Discussion .....</b>	<b>175</b>
<b>4.5 - Appendix.....</b>	<b>179</b>
<b>Chapter 5 - Genome-wide association study of specific psychotic experiences using TEDS, ALSPAC and CATSS samples .....</b>	<b>185</b>
<b>5.1 - Introduction .....</b>	<b>185</b>
<b>5.2 - Methods.....</b>	<b>186</b>
5.2.1 - Samples .....	186
5.2.2 - Measures.....	186
5.2.3 - Handling missing phenotypic data.....	187
5.2.4 - Calculation of phenotypic sum scores .....	187
5.2.5 - Normalisation of phenotypic data and controlling for covariate effects .....	187
5.2.6 - DNA collection and genotyping.....	188
5.2.7 - Genotype imputation and quality control procedure .....	189

5.2.8 – Mega-genome-wide association analysis .....	190
5.2.9 – Meta-genome-wide association study .....	191
5.2.10 – Replication analysis of genetic associations.....	193
5.2.11 – Gene-region association analysis.....	193
5.2.12 – Predicted differential gene expression analysis .....	193
<b>5.3 – Results .....</b>	<b>195</b>
5.3.1 – Descriptive of individual psychotic experience scores.....	195
5.3.2 – Mega-genome-wide association study of specific adolescent psychotic experiences .....	200
5.3.3 – Results from mega-analysis and meta-analysis GWAS.....	214
5.3.4 – Evaluating the effect of randomly splitting ties over averaging ties ...	214
5.3.5 – Gene region association from MAGMA.....	215
5.3.6 – Differential predicted gene expression from PrediXcan.....	215
<b>5.4 – Discussion .....</b>	<b>221</b>
<b>5.5 – Appendix.....</b>	<b>228</b>
<b>Chapter 6 - Estimating the SNP-heritability of specific adolescent psychotic experiences using TEDS, ALSPAC and CATSS.....</b>	<b>232</b>
<b>6.1 – Introduction .....</b>	<b>232</b>
<b>6.2 – Methods.....</b>	<b>233</b>
6.2.1 – Samples .....	233
6.2.2 – Measures.....	233
6.2.3 – DNA collection, genotyping, imputation, and quality control.....	234
6.2.4 – Estimation of SNP-heritability .....	234
6.2.4.1 - GREML-methodologies .....	234
6.2.4.1.1 – GREML-SC .....	234
6.2.4.1.2 – GREML-MS .....	234
6.2.4.1.3 – GREML-LDMS .....	235

6.2.4.1.4 – Incorporating related individuals in GREML analysis .....	235
6.2.4.2 – LD-score regression.....	236
<b>6.3 – Results .....</b>	<b>236</b>
6.3.1 – Comparison of within (meta-) and across (mega-) sample estimates of SNP-heritability .....	236
6.3.2 – Comparison of different GREML methodologies.....	237
6.3.2.1 – Comparison of GREML-SC, -MS and –LDMS estimates.....	237
6.3.2.2 – Comparison of constrained and unconstrained GREML estimates ..	237
6.3.2.3 – Effect of phenotypic normalisation on GREML estimates .....	238
6.3.3 – Distribution of genetic effects across MAF bins .....	238
6.3.4 – Comparison of LD-score regression and GREML estimates .....	238
<b>6.4 – Discussion .....</b>	<b>245</b>

**Chapter 7 - Assessing the genetic relationship between specific  
adolescent psychotic experiences in TEDS, ALSPAC and CATSS samples,  
and major psychiatric disorders .....**

<b>7.1 – Introduction .....</b>	<b>250</b>
<b>7.2 – Methods.....</b>	<b>252</b>
7.2.1 – Samples .....	252
7.2.2 – Measures.....	252
7.2.3 – Genotypic data.....	253
7.2.4 – Polygenic risk score analysis.....	253
7.2.5 – Analysis of non-linear polygenic risk score effects .....	254
7.2.6 – Analysis of within sample polygenic risk score effects .....	254
7.2.7 – Estimation of genetic covariance.....	254
7.2.8 – Estimation of genetic correlation .....	255
<b>7.3 – Results .....</b>	<b>258</b>
7.3.1 - Polygenic risk score association .....	258

7.3.2 – Non-linear effects in Paranoia and Hallucinations scale .....	258
7.3.3 - Within sample results .....	259
7.3.4 – Estimates of genetic covariance .....	268
7.3.5 – Estimates of genetic correlation .....	268
<b>7.4 – Discussion .....</b>	<b>272</b>
<b>7.5 – Appendix.....</b>	<b>276</b>
<b>Chapter 8 - Discussion .....</b>	<b>287</b>
8.1 – Motivation for this thesis.....	287
8.2 – Summary of methods and results .....	288
8.2.1 – GWAS of specific adolescent PEs in TEDS.....	288
8.2.2 – Effect of normalising residuals .....	289
8.2.3 – Harmonisation of PE measures of samples .....	289
8.2.4 – GWAS of specific adolescent PEs in TEDS, ALSPAC, and CATSS .....	290
8.2.5 – Estimating SNP-heritability of specific adolescent PEs in TEDS, ALSPAC, and CATSS .....	290
8.2.6 – Estimating genetic association between psychiatric disorders and specific adolescent PEs in TEDS, ALSPAC, and CATSS.....	291
8.3 – Wider implications of research .....	291
8.3.1 – The genetic architecture of adolescent PEs.....	291
8.3.2 – The genetic relationship between adolescent PEs and schizophrenia, bipolar disorder and major depression.....	292
8.3.3 – The pooling of information across multiple samples.....	295
8.4 – Considerations for future research .....	296
8.4.1 – Measurement .....	296
8.4.2 – Models for non-normal data .....	297
8.4.3 – Genetic relationship between adolescent PEs and other traits/disorders .....	297

8.4.4 – Developmental stages of PEs.....	298
8.4.5 – Rare genetic variation .....	298
8.5 – Conclusion and future directions.....	299
<b>Bibliography .....</b>	<b>301</b>

## List of figures

Figure 2.1. Manhattan plot and QQ-plot of specific psychotic experiences in adolescence in TEDS.....	71
Figure 2.2. Mean Cognitive Disorganisation scores by genotype at rs7830364. ....	73
Figure 2.3. Mean Cognitive Disorganisation scores by genotype at rs7845752. ....	74
Figure 2.4. Mean Negative Symptoms scores by genotype at rs7587811. ....	75
Figure 2.5. Mean Negative Symptoms scores by genotype at rs16876921.....	76
Figure 2.6. Regional association plot of variation surrounding rs7830364 for Cognitive Disorganisation during adolescence. ....	77
Figure 2.7. Regional association plot of variation surrounding rs7845752 for Cognitive Disorganisation during adolescence. ....	78
Figure 2.8. Regional association plot of variation surrounding rs7587811 for Parent-reported Negative Symptoms during adolescence.....	79
Figure 2.9. Regional association plot of variation surrounding rs16876921 for Parent-reported Negative Symptoms during adolescence.....	80
Figure 3.1. The relationship between the number of available responses (x-axis) and correlation between normalised residuals and covariate (y-axis) for different values of the skew in the raw phenotypic data. ....	96
Figure 3.2. The effect of applying a rank-based INT to residuals of questionnaire-type data, i.e. after regressing out covariates. ....	100
Figure 3.3. The effect of applying a rank-based INT to questionnaire-type data before regressing out covariates.....	102

Figure 4.1. Schematic representation of the phenotypic harmonisation process..	153
Figure 4.2. Scree plot of ALSPAC psychotic experience items.....	162
Figure 4.3. Scree plot of CATSS psychotic experience items.....	165
Figure 4.4. Scree plot of SPEQ items in TEDS that match psychotic experience items available in the ALSPAC and CATSS samples.....	172
Figure 5.1. LocusZoom plot of genome-wide significant SNP for Anhedonia.....	202
Figure 5.2. Manhattan plot of Paranoia and Hallucinations mega-GWAS.....	205
Figure 5.3. Manhattan plot of Anhedonia mega-GWAS.....	206
Figure 5.4. Manhattan plot of Cognitive Disorganisation mega-GWAS.....	207
Figure 5.5. Manhattan plot of Parent-rated Negative Symptoms mega-GWAS.....	208
Figure 5.6. Quantile-quantile plot of psychotic experience domain mega-GWASs. .....	209
Figure 7.1. Polygenic risk scores for schizophrenia, bipolar disorder, and major depression predict adolescent psychotic experience domains.....	260
Figure 7.2. Schizophrenia polygenic risk score predicting psychotic experience domains in adolescence.....	262
Figure 7.3. Bipolar disorder polygenic risk score predicting psychotic experience domains in adolescence.....	263
Figure 7.4. Major depression polygenic risk score predicting psychotic experience domains in adolescence.....	264

Figure 7.5. Schizophrenia, bipolar disorder, and major depression polygenic risk score mean differences between low- and high-scoring psychotic experience domain groups.....265

Figure 7.6. Local polynomial regression of schizophrenia polygenic risk score ( $p$ -value threshold of  $p<0.3$ ) and Paranoia and Hallucinations.....266

Figure 7.7. Mean schizophrenia polygenic risk score (SCZ PRS) in six quantiles of Paranoia and Hallucinations scores.....267

## List of tables

Table 1.1. Summary of all publications performing genome-wide analysis of psychotic experiences during adolescence.....	53
Table 2.1. Descriptive statistics for TEDS sample of related and unrelated individuals and six dimensions of adolescent psychotic experiences assessed using SPEQ.....	67
Table 2.2. List of independent variants with $p < 1 \times 10^{-5}$ for each of the specific psychotic experiences.....	68
Table 3.1. Skew, range, and correlation with covariates for dependent variables derived from TEDS sample. ....	94
Table 4.1. Exclusion variables applied in each sample. These exclusion criteria are standard practice for genome-wide association studies of behavioural and cognitive traits (Docherty et al., 2010).....	149
Table 4.2. Number of individuals from TEDS, ALSPAC and CATSS. Figures shown before and after exclusion criteria have been applied.....	149
Table 4.3. List of items assessing adolescent psychotic experiences in the ALSPAC sample. ....	155
Table 4.4. List of items assessing adolescent psychotic experiences in the CATSS sample. ....	158
Table 4.5. Principal component loadings of psychotic experience items in ALSPAC. ....	163
Table 4.6. Cronbach's alpha for the paranoia and hallucinations, anhedonia and parent-rated negative symptoms subscales identified within ALSPAC. ....	164

Table 4.7. Principal component loadings of psychotic experience items in CATSS. .....	166
Table 4.8. Cronbach's alpha for the paranoia and hallucinations, cognitive disorganisation and parent-rated negative symptoms subscale identified within CATSS. ....	167
Table 4.9. List of SPEQ items that match items available in ALSPAC and CATSS. ..	170
Table 4.10. Principal component loadings of psychotic experience items in TEDS. .....	173
Table 4.11. Cronbach's alpha of paranoia and hallucinations (after converting hallucinations into two items), cognitive disorganisation, anhedonia and parent-rated negative symptom scales in the TEDS sample using matched items.....	174
Table 5.1. Power to detect association at genome-wide significance for mega- genome-wide association analyses of psychotic experience traits.....	192
Table 5.2. Descriptive statistics for raw psychotic experience domain sum scores in each sample.....	196
Table 5.3. Pearson's correlation between raw PE sum scores and age. ....	197
Table 5.4. Mean sex differences for untransformed psychotic experience sum scores.....	198
Table 5.5. Pearson correlation between PEs and anxiety symptoms, depressive symptoms and cognitive ability in TEDS. ....	199
Table 5.6. Pearson correlation between PEs and self-reported depressive symptoms in TEDS, ALSPAC and CATSS.....	199

Table 5.7. Pearson’s correlations between raw sum scores and scores after inverse-rank based normalisation splitting ties randomly.....	200
Table 5.8. Independent loci achieving suggestive significance ( $p < 1 \times 10^{-5}$ ) in mega-genome-wide association study of psychotic experience domains. ....	203
Table 5.9. Combined sample (TEDS, ALSPAC and CATSS) association results for genome-wide significant SNPs in TEDS only GWAS of Chapter 2. ....	210
Table 5.10. Suggestive loci from psychotic experience GWASs in the latest schizophrenia GWAS. ....	211
Table 5.11. Suggestive loci from psychotic experience GWASs in the latest bipolar disorder GWAS.....	212
Table 5.12. Suggestive loci from psychotic experience GWASs in the latest major depression GWAS.....	213
Table 5.13. Skew of psychotic experiences after normalisation when averaging tied observations. ....	215
Table 5.14. Top ten genes associated with psychotic experience domains using MAGMA. ....	217
Table 5.15. Top ten differentially-expressed genes for psychotic experience domains based on predicted gene expression levels.....	219
Table 5.16. Annotation of suggestive loci with prior evidence of association in neuropsychiatric phenotypes.....	225
Table 6.1. Mega- and meta- SNP-heritability estimates for specific PEs from GREML and LD-score regression.....	240

Table 6.2. Within sample GREML-SC estimates of SNP-heritability for specific PEs. .....	240
Table 6.3. Within sample LD-score regression estimates of SNP-heritability for specific PEs.....	241
Table 6.4. SNP-heritability estimates for specific psychotic experiences from mega- GREML-SC, -MS and -LDMS.....	241
Table 6.5. Constrained and unconstrained estimates of SNP-heritability from GREML analyses.....	242
Table 6.6. Constrained and unconstrained meta-GREML-SC estimates of SNP- heritability for specific psychotic experiences.....	242
Table 6.7. Within sample GREML-SC SNP-heritability estimates for normalised and untransformed specific psychotic experiences.....	243
Table 6.8. MAF-breakdown of mega-GREML-MS SNP-heritability estimates.....	244
Table 7.1. Parameters used in AVENGEME analysis.....	257
Table 7.2. Schizophrenia, bipolar disorder, and major depression polygenic risk scores predicting psychotic experience domains in adolescents.....	261
Table 7.3. Genetic covariance between each psychotic experience domain and schizophrenia, bipolar disorder, and major depression.....	270
Table 7.4. Estimates of genetic correlation between specific adolescent psychotic experiences, and schizophrenia, bipolar disorder, and major depression.....	271

## List of supplementary notes

Supplementary Note 3.1. 'SimCont' – Function to simulate continuous variables.	105
Supplementary Note 3.2. 'SimQuest' – Function to simulate questionnaire-type variables. ....	106
Supplementary Note 3.3. 'SimContNorm' – Function to simulate continuous variables with skew and kurtosis equal to zero.....	107
Supplementary Note 3.4. 'SimQuestNorm' – Function to simulate questionnaire-type variables with skew and kurtosis equal to zero. ....	108
Supplementary Note 3.5. 'CovarCreator' – Function to create correlated covariates for continuous and questionnaire-type variables.....	109
Supplementary Note 3.6. 'rntransform_random' – Function to perform rank-based INT whilst randomly splitting tied observations.....	110
Supplementary Note 4.1. Example of PLIKS-Q item in ALSPAC' Life of a 16+ Teenager Questionnaire - Section D: Your Current Feelings.....	184

## List of supplementary figures

Supplementary Figure 2.1. Regional association plot for observed variation only surrounding rs7830364 for Cognitive Disorganisation during adolescence. ..	84
Supplementary Figure 2.2. Regional association plot for observed variation only surrounding rs7845752 for Cognitive Disorganisation during adolescence. ..	85
Supplementary Figure 2.3. Regional association plot for observed variation only surrounding rs7587811 for Parent-reported Negative Symptoms during adolescence.....	86
Supplementary Figure 2.4. Regional association plot for observed variation only surrounding rs16876921 for Parent-reported Negative Symptoms in adolescence.....	87
Supplementary Figure 3.1. Effect of regressing out covariate effects from questionnaire-type data with a range of 5 before rank-based non-parametric analyses.....	119
Supplementary Figure 3.2. Effect of regressing out covariate effects from questionnaire-type data with a range of 10 before rank-based non-parametric analyses.....	120
Supplementary Figure 3.3. Effect of regressing out covariate effects from questionnaire-type data with a range of 20 before rank-based non-parametric analyses.....	121
Supplementary Figure 3.4. Effect of regressing out covariate effects from questionnaire-type data with a range of 40 before rank-based non-parametric analyses.....	122

Supplementary Figure 3.5. Effect of regressing out covariate effects from questionnaire-type data with a range of 80 before rank-based non-parametric analyses.....	123
Supplementary Figure 3.6. Effect of regressing out covariate effects from questionnaire-type data with a range of 160 before rank-based non-parametric analyses.....	124
Supplementary Figure 3.7. Effect of regressing out covariate effects from continuous data before rank-based non-parametric analyses.....	125
Supplementary Figure 3.8. Effect of rank-based INT of questionnaire-type data residuals (after regressing out covariates) with range of 5.....	126
Supplementary Figure 3.9. Effect of rank-based INT of questionnaire-type data residuals (after regressing out covariates) with range of 10.....	127
Supplementary Figure 3.10. Effect of rank-based INT of questionnaire-type data residuals (after regressing out covariates) with range of 20.....	128
Supplementary Figure 3.11. Effect of rank-based INT of questionnaire-type data residuals (after regressing out covariates) with range of 40.....	129
Supplementary Figure 3.12. Effect of rank-based INT of questionnaire-type data residuals (after regressing out covariates) with range of 80.....	130
Supplementary Figure 3.13. Effect of rank-based INT of questionnaire-type data residuals (after regressing out covariates) with range of 160.....	131
Supplementary Figure 3.14. Effect of rank-based INT of continuous data residuals (after regressing out covariates).....	132

Supplementary Figure 3.15. Effect of rank-based INT when kurtosis and skew are equal to zero. ....	133
Supplementary Figure 3.16. Effect of proportion of ties and magnitude of original covariate correlation on the confounding effect of normalising the dependent variable residuals. ....	134
Supplementary Figure 3.17. The relationship between the number of available responses in the dependent variable (x-axis) and the absolute Spearman rank-based correlation between normalised residuals and covariate (y-axis) for different values of the skew. ....	135
Supplementary Figure 3.18. Effect of proportion of ties and magnitude of original covariate correlation on the Spearman correlation between dependent variable residuals and the covariate.....	136
Supplementary Figure 3.19. Difference in covariate correlation with dependent variable after rank-based INT when original dependent-covariate correlation is 0.5. ....	137
Supplementary Figure 3.20. Difference in covariate correlation with dependent variable after rank-based INT when original dependent-covariate correlation is 0.25.....	138
Supplementary Figure 3.21. Difference in covariate correlation with dependent variable after rank-based INT when original dependent-covariate correlation is 0.12.....	139
Supplementary Figure 3.22. Difference in covariate correlation with dependent variable after rank-based INT when original dependent-covariate correlation is 0.06.....	140

Supplementary Figure 3.23. Difference in covariate correlation with dependent variable after rank-based INT when original dependent-covariate correlation is 0.03.....	141
Supplementary Figure 3.24. Difference in covariate correlation with dependent variable after rank-based INT when original dependent-covariate correlation is 0.01.....	142
Supplementary Figure 3.25. Correlation between the dependent variable before and after rank-based INT (randomly splitting tied observations).....	143
Supplementary Figure 3.26. Magnitude of skew reintroduced when regressing out covariate effects from normalised dependent variables.....	144
Supplementary Figure 4.1. Scree plot of CATSS PE items after removal of excess parent-reported asociality items.....	183
Supplementary Figure 7.1. Schizophrenia (SCZ) PRS predicting specific adolescent PEs within TEDS, ALSPAC and CATSS samples. ....	284
Supplementary Figure 7.2. Bipolar disorder (BD) PRS predicting specific adolescent PEs within TEDS, ALSPAC and CATSS samples. ....	285
Supplementary Figure 7.3. Major depression (MDD) PRS predicting specific adolescent PEs within TEDS, ALSPAC and CATSS samples. ....	286

## List of supplementary tables

Supplementary Table 3.1. Difference in covariate correlation with the dependent variable before and after rank-based INT when splitting tied observations randomly. This is based on simulated data.....	111
Supplementary Table 3.2. Outcome of rank-based INT after regressing effect of a continuous covariate (age) from real questionnaire data.....	112
Supplementary Table 3.3. Outcome of regressing effect of a continuous covariate (age) from real questionnaire data on Spearman correlation.....	113
Supplementary Table 3.4. Outcome of rank-based INT after regressing effect of a dichotomous covariate (sex) from real questionnaire data.....	114
Supplementary Table 3.5. Outcome of regressing effect of a dichotomous covariate (sex) from real questionnaire data on Spearman correlation.....	115
Supplementary Table 3.6. Effect of rank-based INT (randomly ranking tied observations) on the relationship between real questionnaire variable and continuous covariate (age). This table also shows to what extent regressing covariate effects reintroduces skew.....	116
Supplementary Table 3.7. Effect of rank-based INT (randomly ranking tied observations) on the relationship between real questionnaire variable and binary covariate (sex). This table also shows to what extent regressing covariate effects reintroduces skew.....	117
Supplementary Table 3.8. Effect of rank-based INT of real questionnaire variable when randomly ranking tied observations. Shows Pearson correlation between dependent variable before after normalisation.....	118

Supplementary Table 4.1. List of SPEQ items capturing adolescent PEs. ....	179
Supplementary Table 4.2. Loadings of CATSS PE items from principal components analysis after removal of excess parent-reported asociality items. ....	182
Supplementary Table 5.1. Measures of Paranoia and Hallucinations in TEDS, ALSPAC, and CATSS.....	228
Supplementary Table 5.2. Measures of Anhedonia in TEDS and ALSPAC. ....	229
Supplementary Table 5.3. Measures of Cognitive Disorganisation in TEDS and CATSS.....	230
Supplementary Table 5.4. Measures of Parent-rated Negative Symptoms in TEDS, ALSPAC, and CATSS.....	231
Supplementary Table 7.1. Power calculations for genetic covariance analysis between PEs and schizophrenia.....	276
Supplementary Table 7.2. Power calculations for genetic covariance analysis between PEs and bipolar disorder. ....	277
Supplementary Table 7.3. Power calculations for genetic covariance analysis between PEs and major depression.....	278
Supplementary Table 7.4. Schizophrenia polygenic risk score predicting psychotic experience domains at 8 <i>p</i> -value thresholds.....	279
Supplementary Table 7.5. Bipolar disorder polygenic risk score predicting psychotic experience domains at 8 <i>p</i> -value thresholds.....	281
Supplementary Table 7.6. Major depression polygenic risk score predicting psychotic experience domains at 8 <i>p</i> -value thresholds.....	282

Supplementary Table 7.7. Comparison of schizophrenia, bipolar disorder, and  
major depression polygenic risk scores in low and high psychotic experience  
domain groups.....283

## **Acknowledgments**

I would like to give thanks to everyone that has supported me throughout the completion of this thesis.

First, I would like to thank my supervisors, Professor Angelica Ronald and Professor Frank Dudbridge. They have given me excellent advice relating to the analysis of behavioural and genetic data, and the logistics of research. I am also very grateful for their continual encouragement during this thesis. I would also like to thank Dr Emma Meaburn for her temporary supervision during Angelica's maternity leave and Frank's research sabbatical.

Second, this project would not have been possible without the amazing data from the Twins Early Development Study (TEDS), the Avon Longitudinal Study of Parents and Children (ALSPAC), and the Child and Adolescent Twin Study in Sweden (CATSS). I would like to give thanks to the researchers and participants that contributed to these studies, and particularly to Robert Plomin, George Davey Smith and Paul Lichtenstein for their collaboration on this project.

Third, I would like to thank Dr Alastair Cardno and Professor Daniel Freeman for their involvement in this thesis. Their experience in both clinical and research settings provided me with insight into adolescent psychotic experiences and helped to ensure the content validity of the measures used in this thesis.

Last but not least, I would like to thank my family and friends for their encouragement during this thesis and prior education.

# Publications

## Chapter 3:

Pain, O., Dudbridge, F., Ronald, A. (under review). Investigating the effect of rank-based normalization after regressing out covariates. *European Journal of Human Genetics*. Pre-print: <http://www.biorxiv.org/content/early/2017/05/15/137232>

## Chapters 4-7:

Pain, O., Dudbridge, F., Cardno, A., Freeman, D., Lu, Y., Lundstrom, S., Lichtenstein, P., Ronald, A. (under review). Genome-wide analysis of adolescent psychotic experiences shows genetic overlap with psychiatric disorders. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*.

# Chapter 1 - Introduction

In this chapter, an in-depth overview of molecular genetic research on adolescent psychotic experiences (PEs) will be provided. The importance of adolescence as a developmental stage will be demonstrated, focusing on mental health. Adolescent PEs will be described in terms of behavioural features, presence in the general population, and their relationship with psychiatric traits and disorders occurring in adolescence and adulthood. The field of molecular genetics will then be introduced, focusing on the development of methods for identifying genetic factors underlying common traits/disorders and the nuances of genome-wide association studies. Evidence that adolescent PEs are heritable will be described before providing a systematic review of all studies investigating adolescent PEs using genome-wide molecular genetic data. The findings and limitations of previous genome-wide studies will present considerations from which the aims of this thesis will be derived.

## 1.1 - Adolescence

### *1.1.1 - Adolescence: A critical period*

Major physical and psychological changes occur during adolescence, the developmental stage between childhood and adulthood which is typically defined as 11-20 years of age (Dahl, 2004). These changes are in conjunction with substantial social changes driven in part by separation from ones parents, forming peer relationships, and taking on responsibility. The simultaneous transformation of these aspects of life makes adolescence a highly sensitive and complex developmental period (Steinberg, 2005). Due to the relatively low mortality during adolescence (World Health Organization, 2009b), adolescents are often thought to be in good health. However, adolescents are exposed to many risk factors, such as

substance use, and are commonly affected by non-lethal health problems, particularly relating to mental health (Patton et al., 2010), both of which can be influential for health outcomes later in life (World Health Organization, 2009a). For example, the World Health Organisation reported almost 35% of the global burden of disease occurs during or is attributable to adolescence (World Health Organisation, 2012). The importance of understanding adolescent health is further justified by its great potential to influence many aspects of society in the future (Sawyer et al., 2012).

### 1.1.2 - Mental health in adolescence

Mental health issues during adolescence have been reported as a major health concern (Patel, Flisher, Hetrick, & McGorry, 2007). More than 1 in 5 individuals 11-16 years old have a psychiatric disorder, the most prevalent being anxiety and conduct disorder (Green, McGinnity, Meltzer, Ford, & Goodman, 2005). The debilitating nature of these psychiatric disorders during adolescence has harmful downstream consequences for mental and physical health. Psychiatric issues starting in adolescence, account for half of the psychiatric disorders in adulthood (P. B. Jones, 2013). Furthermore, a number of behavioural traits in healthy adolescents have been associated with the development of psychiatric disorders in adulthood (Hemphälä & Hodgins, 2014; Poulton et al., 2000). Many adult mental disorders are highly prevalent and have a high comorbidity with physical disorders such as diabetes, cardiovascular disease, asthma, and musculoskeletal disease (Australian Institute of Health and Welfare, 2011), suggesting that an improved understanding of adolescent mental health could lead to major health benefits.

## **1.2 - Psychotic experiences (PEs)**

This section will first describe the term PEs and how they relate to the symptomatology of psychotic disorders. Then, previous research investigating PEs as a risk factor will be discussed, highlighting findings relating to adolescence.

### *1.2.1 – Definition and relationship with psychotic disorders*

PEs (sometimes referred to as psychotic-like experiences or psychosis proneness) are defined as traits in the general population that at the extreme resemble symptoms of psychotic disorders, such as schizophrenia (American Psychiatric Association, 2013b; Ronald, 2015). Similar to the symptomatology of psychotic disorders (Siever, Kalus, & Keefe, 1993), PEs contain a number of separable behavioural domains (Chen, Hsiao, & Lin, 1997; Ericson, Tuvblad, Raine, Young-Wolff, & Baker, 2011; Fossati, Raine, Carretta, Leonardi, & Maffei, 2003; Ronald et al., 2014; Vollema & Hoijtink, 2000). Previous investigation of PEs in several general population samples using principal components analysis (PCA) has identified replicable PE domains congruent with the recognised symptom domains in psychotic disorders. Most often these PE domains have separated the positive, cognitive and negative symptom domains (Chen et al., 1997; Ericson et al., 2011; Fossati et al., 2003). Positive symptoms refer to the gain or addition in perceptions and experiences, and can be more finely separated into paranoia, hallucinations, delusions and grandiosity domains (American Psychiatric Association, 2013b). Cognitive symptoms, also referred to as cognitive disorganisation, often occur in individuals with psychotic disorders (American Psychiatric Association, 2013b) but certain aspects also overlap with other psychiatric disorders, such as attention deficit hyperactivity disorder (ADHD)(American Psychiatric Association, 2013a). Negative symptoms refer to the removal or subtraction of perceptions or experiences. Subdomains of negative symptoms include anhedonia, asociality,

apathy, alogia (poverty of speech), avolition and inattention (American Psychiatric Association, 2013b). Similar to the cognitive symptoms, certain aspects of these negative symptoms are features of psychotic disorders but also other psychiatric disorders. For example, anhedonia and avolition are also symptoms of major depression (American Psychiatric Association, 2013a).

Although at the extreme, PE domains resemble symptom domains in psychotic disorders, PEs are different in a number of ways. Firstly, PEs are common in the general population, particularly adolescence (Fonseca-Pedrero, Paíno-Piñeiro, Lemos-Giráldez, Villazón-García, & Muñiz, 2009; Ronald et al., 2014; Van Os, Linscott, Myin-Germeys, Delespaul, & Krabbendam, 2009; Verdoux & van Os, 2002). Second, as opposed to the categorical outcome of a diagnosis, PEs can be viewed as quantitative traits, which vary in frequency and severity (Ronald et al., 2014; Van Os et al., 2009). Third, PEs are not always associated with distress for the individual and should be viewed as a part of normal variation in the general population (McGrath et al., 2015; Ronald et al., 2014).

There is a closely related group of traits referred to as schizotypal traits or schizotypy. These schizotypal traits differ to PEs in that they focus on differences in personality that reflect liability to psychotic disorders rather than the presentation of subclinical psychotic symptoms (Pedrero & Debbané, 2017). Schizotypal traits are not a focus of this thesis and will not be discussed further.

### 1.2.2 - PEs as risk factors

Although PEs are themselves not pathological, they have been associated with a number of mental health related factors and outcomes in adolescence and adulthood.

One adolescent mental health outcome associated with PEs is suicide. For example, PEs during adolescence have been identified as a risk factor for adolescent suicide attempts and ideation (Cederlöf et al., 2016; Kelleher et al., 2013; Kelleher, Cederlöf, & Lichtenstein, 2014). Suicide is a leading cause of mortality with approximately one million deaths by suicide per year worldwide (World Health Organization, 2012). The most recent study reporting an association between PEs and suicide was based on data collected by the Child and Adolescent Twin Study in Sweden (CATSS)(Cederlöf et al., 2016). This study reported that paranoia, hallucinations (auditory and visual), thought interference, and grandiosity at age 15 or 18 all positively and significantly predicted suicide attempts later in adolescence (hazard ratios of 1.6-2.5). This study also demonstrated a positive and significant association between paranoia and hallucinations separately and subsequent diagnosis of substance abuse disorder during adolescence (hazard ratio of 2.7-3.0). Furthermore, the authors report that PEs (assessed at 15 or 18) preceded the suicide attempt or diagnosis when they included outcomes occurring anytime after 15. Although this study provides robust evidence for an association between certain positive PE domains, suicide attempts, and substance abuse disorder, it did not investigate the effect of cognitive or negative domains of PEs.

There is also evidence for association between PEs and adult psychiatric disorders including psychotic (McGrath et al., 2016; Poulton et al., 2000; Van Os et al., 2009; Welham et al., 2009) and non-psychotic psychiatric disorders (McGrath et al., 2016; S. A. Sullivan et al., 2014). There is evidence of a bidirectional effect in that PEs can both precede and succeed the diagnosis of a psychiatric disorder (McGrath et al., 2016), however the majority of studies have primarily focused on PEs occurring prior to diagnosis (Poulton et al., 2000; S. A. Sullivan et al., 2014; Van Os et al., 2009; Welham et al., 2009). One of the first studies demonstrating a link between PEs prior to the onset of psychotic disorders specifically was based on the Dunedin

Multidisciplinary Health and Development Study (Poulton et al., 2000). This study reported that individuals at age 11 reporting prevalent PEs had a 16-fold increase in risk of developing schizophreniform-disorder by age 26. The most recent study investigating the relationship between PEs and the subsequent onset of a range of psychiatric disorders (not schizophrenia) reported a significant and positive association between PEs and several mood disorders, anxiety disorders, impulse-control disorders, eating disorders and substance-use disorders, with odds ratios between 1.7 and 3.2 (McGrath et al., 2016). These studies have focused on the positive domain of PEs (mainly paranoia, hallucinations, and thought interference), with the effect of cognitive and negative PE domains on subsequent psychiatric outcomes less explored. A meta-analysis showed that unaffected first-degree relatives of schizophrenic individuals had increased cognitive deficits compared to the general population (Sitskoorn, Aleman, Ebisch, Appels, & Kahn, 2004). This provides evidence that cognitive deficits in a non-clinical sample have shared aetiology with psychotic disorders. There is also some recent evidence of an overlap in aetiology of negative symptoms during adolescence and schizophrenia based on a molecular genetic approach called polygenic risk scoring (H. J. Jones et al., 2016). This study will be discussed further in Section 1.6.

As previously mentioned, studies investigating PEs have focused on those occurring during adolescence. This is supported by the fact that adolescence is known to be influential in predicting mental health outcomes in general, but also the previous report that PEs during adolescence are associated with increased negative outcomes relative to PEs in childhood (Kelleher et al., 2012; Trotman et al., 2013). Collectively, the literature highlights the importance of PEs before the onset of psychiatric disorders, primarily in adolescence, suggesting that an understanding of adolescent PEs may provide insight into the aetiology of a range of psychiatric disorders.

### 1.2.3 - Assessment of adolescent PEs

Adolescent PEs are assessed using either interviews or questionnaires. Clinical interviews are time consuming, expensive, and may introduce an inter-rater bias, making questionnaires an appealing alternative. The validity of assessing adolescent PEs using questionnaires has been previously explored by several studies.

Self-report questionnaires of positive PEs lead to higher median estimates of prevalence than from interviews (Linscott & Van Os, 2013). However, when compared to the findings of clinical interviews, self-report questionnaire items assessing experiences of paranoia and hallucinations have good validity (Kelleher, Harley, Murtagh, & Cannon, 2009; Laurens et al., 2007).

Self-report questionnaires are used to assess cognitive disorganisation in non-clinical samples. A commonly used self-report questionnaire for assessing cognitive disorganisation in clinical and non-clinical samples is in the O-LIFE (Oxford-Liverpool Inventory of Feelings and Experiences) (Mason, Claridge, & Jackson, 1995). A previous study, comparing scores from the O-LIFE cognitive disorganisation questionnaire with scores from a clinical interview, reported that although patients with schizophrenia had higher cognitive disorganisation O-LIFE scores than healthy participants, there was no significant correlation between the cognitive disorganisation O-LIFE scores and the cognitive disorganisation scores from clinical interview (Cochrane, Petch, & Pickering, 2010). Although this comparison was based on schizophrenic patients instead of a non-clinical sample, this finding does not support the comparability of self-report measures of cognitive disorganisation to clinical interview based measures.

The positive and cognitive domains during adolescence are typically assessed using self-report measures. However, the negative symptom PE domain during

adolescence is typically assessed using observer- (such as parent-) report interviews or questionnaires. This is due to evidence of a discrepancy between self- and observer-reported negative symptoms in schizophrenic patients, suggesting that self-report of negative symptoms is less reliable than observer-report (Hamera, Schneider, Potocky, & Casebeer, 1996; Selten, Wiersma, & van den Bosch, 2000).

### **1.3 - Molecular genetics**

In this section I will describe the function of deoxyribonucleic acid (DNA), genetic variation, and then the development of methods for identifying genomic regions associated with common given traits or diseases, leading up to the advent of the genome-wide association study (GWAS) era.

#### 1.3.1 - DNA

One approach to understanding the mechanisms underlying a given trait or disease is through identification of associated differences within DNA. The sequence of DNA encodes information for the synthesis and regulation of proteins, the molecular machines that carry out biological processes. Regions of DNA that encode proteins are called protein-coding genes. Regions of DNA that do not encode proteins are called non-coding regions. A large proportion of non-coding DNA is thought to have other functions such as encoding RNA that is not translated into proteins but regulates the expression of other regions of the genome. In humans, DNA is organised into chromosomes of which we have two copies (diploid), one from our mother and one from our father. Twenty-two of the chromosome pairs are called autosomes (non-sex chromosome) and 1 chromosome pair are the sex chromosomes called X and Y. Females have two X-chromosomes. Males have one X-chromosome and one Y-chromosome.

### 1.3.2 – Genetic variation

99.5% of the DNA sequence in humans is identical between all humans (Levy et al., 2007). The other 0.5% of DNA showing difference in sequence between individuals is termed genetic variation. Genetic variation is the result of random changes in the sequence of DNA (genetic mutations) that may or may not be passed on to future offspring dependant on the reproductive fitness of the organism. Broadly speaking *de novo* mutations (occurring within an individual) can either have a positive, neutral or negative effect on an organism's reproductive fitness, affecting the extent to which the mutation is passed on to future generations and the mutation frequency in a population. There are many different types of genetic variation, from single nucleotide changes, referred to as either single nucleotide polymorphism (SNPs) or single nucleotide variations (SNVs), to large chromosomal rearrangements, such as large insertions, deletions and translocations. The larger the genetic variation (in terms of genomic length), the more likely it is to have an effect on the organism. The different versions of a given DNA sequence caused by genetic variation are referred to as alleles. As previously mentioned, humans are diploid, so they have two copies of each chromosome (except the sex chromosomes in males) and therefore, two alleles of genetic locus (region), one maternal and one paternal. The combination of alleles at a given locus is referred to as a genotype. At a given locus where there are only two possible alleles in the population (biallelic), the two alleles are often distinguished by their frequency in a given population. The allele that is more common in the population is called the major allele, and the allele that is less common in the population is called the minor allele. As a result, the frequency of a specific genetic variant is referred to as the minor allele frequency (MAF).

Another important feature of DNA is linkage disequilibrium (LD), the non-random association between alleles at different loci in a given population. If chromosomes were inherited without any recombination, *de novo* mutations, or rearrangements, then all existing genetic variation on a given chromosome would have a perfect pair-wise correlation of 1. However, due to the occurrence of recombination, *de novo* mutations, and large-scale chromosomal rearrangements, the correlation between genetic variants generally decreases as the genomic distance between them increases. Other factors that alter the correlation or LD between genetic variants include the rate of mutation, natural selection, and random genetic drift (Ardlie, Kruglyak, & Seielstad, 2002). There are different metrics used to measure the LD between two variants. The most widely used measure of LD between two variants in population genetics is  $r^2$ , which can vary between 0 and 1 (1 = perfect correlation, 0 = no correlation).

### 1.3.3 - Identification of genetic variants underlying rare and common disease

One aspect of molecular genetics is the identification of genetic variation underlying different outcomes. This process can elucidate the function of genomic regions and the biological processes underlying a given phenotype. With this knowledge it is possible to predict, and if desired, prevent or treat health outcomes.

This is particularly true for Mendelian traits where genetic variation affecting a single gene is responsible for determining the outcome. As a result, genetic variation underlying a Mendelian disease that reduces the organism's reproductive fitness will be under strong negative selective pressure, leading to its reduced frequency in the population, and as a result, Mendelian diseases are very rare. Due to the rarity of Mendelian disease, and unifactoral aetiology, pedigrees (families) that show multiple occurrences of the disorder are used to identify the causal gene via a method called linkage analysis. Although linkage analysis is a very powerful

tool for identifying the causal variant for Mendelian diseases, it is not well suited to the identification of genetic variation associated with common diseases (N. J. Risch, 2000).

According to the common disease-common variant hypothesis (N. Risch & Merikangas, 1996), genetic factors underlying common diseases must themselves be common. However, in contrast to the large effect size variation in Mendelian diseases, common diseases are driven by many genetic (and environmental) factors with individually small effect sizes, allowing them to stably exist in the general population. The large number of small effect size genetic variants (termed polygenicity) underlying common diseases and traits has now been demonstrated repeatedly using both polygenic scoring and mixed linear model methodology (Dudbridge, 2016; Visscher, Brown, McCarthy, & Yang, 2012). As a result, association analysis testing for a correlation between genetic variation and common disease using unrelated individuals, an approach more appropriate for the identification of small effect size genetic variation, is the method of choice (N. J. Risch, 2000). Although the majority of variance in a common trait or disease is attributable to common genetic variation of small effect size, some variance can also be attributed to rarer genetic variation due to the presence of rare variation with small effect size, and very few individuals carrying large effect rare variants that lead to a Mendelian form of the common trait or disease (Manolio et al., 2009).

Until recently, association studies were exclusively candidate gene studies that focused on specific genes based on functional plausibility for a given phenotype. The candidate gene approach has had some success in psychiatry, whereby genes implicated in neurobiological pathways are studied. For example, *5-HTT-LPR*, a genetic variant within the serotonin transporter gene (*5-HTT*) has been associated with a number of psychiatric traits/disorders, first of which were affective

disorders (Collier et al., 1996). However, a key limitation of the candidate gene approach is its dependence on *a priori* hypotheses and our poor ability to identify plausible genes.

## **1.4 – Genome-wide association studies (GWAS)**

In this section I will introduce the GWAS design, its improved ability to control for population stratification, the multiple testing problem of genome-wide analysis, the factors affecting statistical power, and finally the use of consortia to overcome sample size limitations.

### 1.4.1 – The premise of GWAS

Hypothesis-free genome-wide association studies (GWASs) became feasible following the availability of high resolution genetic maps of the human genome, such as from the HapMap project (Gibbs et al., 2003), and the development of high-throughput genotyping technology (Ding & Jin, 2009). Unlike the traditional candidate gene approach where variants within regions of interest are tested for association, GWASs test for associations with genetic variation across the genome. The markers of genetic variation used in GWAS are common variants, typically single nucleotide polymorphisms (SNPs) or insertions/deletions (INDELS). Based on the common disease-common variant hypothesis, common variation is assumed to explain a larger amount of phenotypic variance (larger  $r^2$ ) than rare variation. An increase in  $r^2$  corresponds to increased statistical power (ability to reject a false null hypothesis). With such a limited understanding of the biological mechanisms and genetic architecture underlying psychiatric phenotypes, this hypothesis-free approach is more suitable than candidate gene studies. GWASs have highlighted many novel genomic regions associated with a range of phenotypes, enabling an improved understanding of the aetiology of many complex human diseases, and highlighted prevention and therapeutic strategies (Visscher et al., 2012).

GWASs typically use regression models to test for changes in minor allele copy number that correlate with a given phenotype. Additive effects of each minor allele copy (as opposed to dominant or recessive) have been shown to explain the majority of variance (Hill, Goddard, & Visscher, 2008). Therefore, individuals homozygous for the major allele are coded as 0, heterozygous individuals are coded as 1, and individuals homozygous for the minor allele are coded as 2. Linear regression is used for the analysis of quantitative phenotypes and logistic regression is used for the analysis binary phenotypes. Both these types of regression have underlying assumptions which if violated can result in increased type-I and II error rates. An assumption regarding linear regression is the normality of the residuals (Berry, 1993). This issue will be discussed further in Chapter 3.

#### 1.4.2 – Control of population stratification

One advantage of the genome-wide approach compared to the candidate gene approach is the more sensitive control of population structure. In candidate gene studies population structure was either not controlled for or researchers would perform either genomic control or structured association. Genomic control is where the association statistic, typically the chi-square, is divided by an inflation factor ( $\lambda$ ), the degree to which an implausible genetic variant is associated with the same trait (Bacanu, Devlin, & Roeder, 2000). However, this approach was not totally effective and doesn't stop false negatives (Price et al., 2006). Structured association was an alternative approach that assigns individuals to discrete subpopulations based on allele frequencies at each locus (Pritchard, Stephens, & Donnelly, 2000). The main limitations of this approach were its sensitivity to the number of discrete clusters specified and the computational cost when using large datasets (Price et al., 2006). Another approach for controlling for population

structure is by including measures of ancestry as covariates in the model. The measures of ancestry are most commonly estimated by applying principal components analysis or multidimensional scaling to genome-wide common genetic variation (Price et al., 2006). These methods identify linearly uncorrelated axes of covariance (components) between all genetic variants. These components have been shown to be correlated with other ancestry measures and control for population stratification when included as covariates (Price et al., 2006). A more novel approach to control for population stratification is mixed linear modelling (Yang, Zaitlen, Goddard, Visscher, & Price, 2014). This approach involves creating a genetic relationship matrix (GRM) that describes the genome-wide structure within a sample, using a random effects model to estimate the phenotypic variance attributable to the GRM, and then estimating association statistics for each genetic variant that account for the phenotypic variance explained by the GRM.

#### 1.4.3 – The multiple testing problem and genome-wide significance

Although genome-wide association studies provide some advantages, the hypothesis-free design leads to a large number of partially dependent (due to LD) and independent tests, which if unaccounted for will increase the type-1 error rate. The simple Bonferroni correction of significance is too stringent due to the correlation / LD between genetic variation making them non-independent. Several lines of evidence estimated that genome-wide analyses were performing approximately one million independent test in a European sample, and therefore a genome-wide significance threshold of  $p \leq 5 \times 10^{-8}$  should be adopted (Dudbridge & Gusnanto, 2008; Hoggart, Clark, De Iorio, Whittaker, & Balding, 2008; Pe'er, Yelensky, Altshuler, & Daly, 2008).

#### 1.4.4 – Factors affecting power in GWAS

GWASs, due to the heavy burden of multiple testing and the small effect size of individual genetic variants, are often limited in their ability to reject a false null hypothesis, also referred to as limited statistical power. Other than effect size, major factors affecting statistical power include sample size, total phenotypic variance that can be explained by tagged genetic variation (SNP-heritability) (Purcell, Cherny, & Sham, 2003), genetic and phenotypic heterogeneity (Manchia et al., 2013), and disorder/disease prevalence (Purcell et al., 2003).

Studies often try to overcome sample size limitations by combining samples for mega-analysis, or by meta-analysing test statistics derived from several samples. Indeed, the approach of increasing sample size to improve power is effective. This was nicely demonstrated by GWASs of schizophrenia where each study had an increased sample size and a corresponding increase in the number of schizophrenia loci identified at genome-wide significance (Ripke et al., 2014; Ripke, O’Dushlaine, et al., 2013; Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium, 2011). However, the approach of increasing sample size is not always the most effective approach for increasing statistical power. This was demonstrated by the 2013 mega-GWAS of major depressive disorder, which identified no replicable associated loci despite including over 9,000 major depressive disorder cases in the analysis (Ripke, Wray, et al., 2013). At the time this was the largest GWAS to find no significant locus, highlighting the importance of other factors affecting power.

Major depressive disorder has several differences compared to other psychiatric disorders including higher prevalence, lower heritability, and increased phenotypic heterogeneity (Power et al., 2017). The issue of phenotypic heterogeneity is of particular importance as its effect on statistical power increases with sample size,

and phenotypic heterogeneity is something that can be reduced. One approach to reduce phenotypic heterogeneity is by stratification based on other patient information, such as age of onset. This has been demonstrated to increase statistical power in the face of reduced sample size using both ischaemic stroke phenotypes and major depression (Power et al., 2017; Traylor et al., 2013).

Although diagnostic categories have great utility in a clinical setting, patients within them often have no unifying symptom and differing profiles of symptom severity. This is particularly true among psychiatric diagnoses. Therefore, another approach for the reduction of phenotypic heterogeneity is the use of measures that assess specific symptom domains using quantitative measures (Manchia et al., 2013). The use of measures that assess specific phenotypes on a quantitative scale have been reported to reduce phenotypic heterogeneity and thereby increase statistical power (Kraemer, 2007).

#### 1.4.5 – GWAS consortia

As previously mentioned, a major limitation of statistical power is sample size. The main approach for addressing sample size limitations, without starting a new larger scale study to investigate the phenotype of interest, is collaboration between research groups to pool resources. These collaborations usually occur in one of three ways: 1) Meta-analysis of summary statistics from multiple uncoordinated studies, 2) Meta-analysis of summary statistics from multiple coordinated studies, or 3) Mega-analysis of individual-level data from multiple studies. The main benefit of meta-analysis approaches is ability to include information from samples where the individual-level data is not available. Many samples often have restricting regulations over the sharing of individual-level data for data security purposes. Traditionally, the first approach was most common. However, the lack of a priori agreement of measures, statistical methods and quality control parameters led to

discrepancies between the studies, making it less appropriate to compare test statistics between studies. More recently, the second approach has become more common, whereby summary statistics from studies that have been coordinated are meta-analysed. This second approach, sometimes referred to as a multicentre study, is preferable to the first due to the predefined analysis pipeline improving the comparability between studies. The third approach is generally harder to carry out due to data sharing restrictions and computational requirements to analyse large samples. However, if the individual level data from each sample can be stored in a single location with sufficient computational power, the mega-analysis approach offers two key advantages. Firstly, almost all statistical methods are based on asymptotic (or large sample) theory, and therefore become more robust as sample size increases. However, several studies have demonstrated that a well controlled meta-analysis give almost identical results to mega-analysis. Secondly, the handling and analysis of all the data on one site and by one experimenter enables the experimenter to go back to the raw data to perform more exploratory analyses more easily, and reduces the likelihood of discrepancy in the processing of each sample.

## **1.5 - Genetics of adolescent PEs**

In order for a genetic approach to be appropriate for the investigation of PEs, there must be evidence that PEs are heritable. Furthermore, given adolescent PEs themselves are not pathological, there must be evidence that PEs are genetically associated with clinical phenotypes. In this section I will briefly outline methods used to calculate heritability and co-heritability estimates, and then present the evidence that adolescent PEs are both heritable and genetically associated with clinical outcomes.

### 1.5.1 – Estimating heritability and co-heritability

Heritability is calculated by estimating the relationship between genotypic similarity and phenotypic similarity. This can be done using family-based designs where the average genotypic similarity between different family relationships is used. For example, in the twin design, the within pair differences of monozygotic (MZ) and dizygotic (DZ) twins are compared (Plomin, DeFries, Knopik, & Neiderheiser, 2013). Twin-based heritability estimates typically include the phenotypic variance explained by all additive genetic factors. The classic twin design assumes equal shared environment among both MZ and DZ siblings. The equal shared environment assumption has been questioned on the basis that twins who resemble one another more closely are likely to be treated more similarly. However, studies investigating this assumption have reported that the equal shared environment assumption is accurate for most aspects of psychopathology (Martin, Boomsma, & Machin, 1997).

Another approach is to use unrelated individuals and estimate the genotypic similarity between individuals using genotypic data (Yang, Benyamin, McEvoy, Gordon, Henders, & Others, 2010). This approach traditionally uses common genetic variation from a genotyping array (mainly SNPs), and so these estimates are referred to as the SNP-heritability. Estimating the SNP-heritability requires larger sample sizes than twin-based methods for accurate estimates with similar factors affecting power as mentioned in Section 1.4.4 (Visscher et al., 2014). A popular method for estimating the SNP-heritability is genomic-relatedness-matrix restricted maximum likelihood (GREML) (Yang, Benyamin, McEvoy, Gordon, Henders, & Others, 2010) in Genome-wide Complex Trait Analysis (GCTA) (Yang, Lee, Goddard, & Visscher, 2011), which consists of two analytical steps. First, all SNPs are used to calculate the genetic relationship matrix (GRM) of the sample.

Secondly, the GRM (independent variable) is used in a mixed linear model to estimate SNP-heritability of the phenotype (dependent variable). GCTA estimates the proportion phenotypic variance explained by additive genetic effects.

Another method for estimating the SNP-heritability of the a phenotype is called LD-score regression (Bulik-Sullivan et al., 2015). LD-score regression is a GWAS (genome-wide association study) summary statistic based method for estimating the phenotypic variance explained by additive genetic effects. This method uses linear regression to estimate the relationship between LD-score and the chi-square of genetic variants. The premise behind this approach is that the more genetic variation that an index variant is in LD with, the more likely it is to tag a causal variant, and therefore on average have a higher test statistic. Because LD is not correlated with population structure or other sources of confounding, this method is also able to distinguish polygenicity from confounding.

Similar to methods estimating heritability, the genetic correlation between two phenotypes can be estimated using predicted genotypic similarity within families (Plomin et al., 2013), or estimated genotypic similarity using genotypic data from unrelated individuals (S. H. Lee, Yang, Goddard, Visscher, & Wray, 2012). In contrast to methods estimating heritability, where the phenotypic variance is partitioned into variance components, methods estimating genetic correlation partition the phenotypic *covariance*.

### 1.5.2 – Heritability of PEs

Twin-based analyses have estimated moderate twin heritability estimates for adolescent PEs (Ericson et al., 2011; Hur, Cherny, & Sham, 2012; Zavos et al., 2014). One study using separate measures to assess specific adolescent PE domains reported heritability estimates between 15% and 59% (Zavos et al., 2014). In fact, a meta-analysis of almost all twin studies reported that all reliably measured traits

are significantly heritable (Polderman et al., 2015). The previously mentioned Zavos et al. study also reported a significant genetic correlation of 0.3-0.6 between certain specific PEs (Zavos et al., 2014).

Due to sample size limitations, existing SNP heritability estimates for PEs are less accurate but provide evidence that adolescent PEs are at least in part regulated by common genetic variation (Sieradzka et al., 2015). This study is discussed further in Section 1.6. These findings support the validity of a molecular genetic approach for understanding the factors affecting adolescent PEs.

### 1.5.3 – Genetic association between PEs and clinical outcomes

Several studies investigating this question reported that offspring reported increased adolescent PEs if their parents had been diagnosed with a psychiatric disorder, particularly schizophrenia and affective disorders (Binbay et al., 2012; R. B. Jones et al., 2016; Zavos et al., 2014). As discussed in the systematic review in Section 1.6, molecular genetic studies using polygenic risk score analysis have reported limited and mixed findings regarding the genetic overlap between adolescent PEs and psychiatric disorders (H. J. Jones et al., 2016; Sieradzka et al., 2014; Zammit et al., 2014). Collectively, these findings provide limited evidence that adolescent PEs may share genetic pathways with major psychiatric disorders, but further research is required.

## **1.6 - Genome-wide analysis of adolescent PEs – A systematic review**

Here, all genome-wide studies of adolescent PEs with the following aims will be summarised: 1) to identify genetic variation underlying adolescent PEs, 2) to estimate the variance in adolescent PEs attributable to common genetic variation using measured genotypes, and 3) to evaluate genetic associations between

adolescent PEs and psychiatric disorders. To identify all such papers, a series of search terms were used in PubMed to capture the following three criteria: 1) age group, 2) phenotype, and 3) type of analysis. The search was conducted using the following format: (adolescent OR adolescence OR teenager OR teenagers OR teenage OR teen OR child OR childhood OR children OR young adult) AND (psychotic OR psychosis OR schizotypy OR schizotypal OR psychotic experiences OR psychotic-like-experiences OR prodromal OR psychosis proneness OR paranoia OR hallucinations OR anhedonia OR negative symptoms OR cognitive disorganisation OR cognitive disorganization OR grandiosity OR delusions) AND (gwas OR genome-wide OR polygenic OR gcta OR greml OR snp-heritability OR ldsc OR ld score regression). The resulting publications were then selected for the review based on the following inclusion criteria:

- Studies using genome-wide array or sequence data.
- Studies focusing on at least one dimension of PEs. If other psychological traits are analysed, only the results referring to PEs are discussed.
- Studies using adolescent sample or sample that overlaps with adolescence (10-20 years of age)
- Studies in English.

In addition to the systematic database search using PubMed, bibliographies of relevant research and review papers were investigated by hand to identify further relevant studies.

#### 1.6.1 - Results of systematic review

Collectively, four publications were identified as genome-wide studies of adolescent PEs using the above search terms on 23rd of August 2017 (Table 1.1).

One study performed a genome-wide association analysis of PEs (Zammit et al., 2014). This study used a binary phenotype of no positive symptoms or definite positive symptoms at either 12 or 18 and reported no genome-wide significant variation.

One study estimated the SNP-heritability of PEs (Sieradzka et al., 2015). The study estimated the SNP-heritability of the full range of specific PEs at age 16 and reported evidence for a common genetic aetiology in some types of PEs (Cognitive Disorganisation, Grandiosity, Anhedonia, Paranoia). SNP-heritability was estimated using the GREML method in GCTA software.

Three of these studies have used polygenic risk scoring to test for an association between schizophrenia and PEs assessed during adolescence (H. J. Jones et al., 2016; Sieradzka et al., 2014; Zammit et al., 2014). Positive psychotic experiences have been reported to show no association with schizophrenia genetic risk three times based on two samples (H. J. Jones et al., 2016; Sieradzka et al., 2014; Zammit et al., 2014). The one and only study testing for an association with cognitive disorganisation showed no significant positive association (Sieradzka et al., 2014). Of the two studies testing for an association with negative symptoms, one showed a significant positive association (H. J. Jones et al., 2016), the other did not (Sieradzka et al., 2014).

One study tested for a positive association between the full range of specific PEs and bipolar disorder (Sieradzka et al., 2014) and reported no significant positive associations.

Two of these publications tested whether candidate genetic variation previously associated with schizophrenia plays a role in PEs during adolescence (Sieradzka et al., 2014; Zammit et al., 2014). No candidate variation achieved significance after

accounting for multiple testing. The strongest association was between Paranoia a gene called *TCF4* (transcription factor 4)(Sieradzka et al., 2014).

In summary, there is some evidence that adolescent PEs are in part influenced by common genetic factors. Studies of adolescent PE have reported some suggestive evidence that *TCF4* is involved in the aetiology of adolescent PEs, but no genetic locus has been significantly associated with adolescent PEs. The evidence of a common genetic overlap with related psychotic disorders is inconclusive due to mixed findings across studies. The discrepancy between studies could be partly explained by differences in the measures used including the latent variable assessed and the degree of specificity. To improve the accuracy estimates of SNP-heritability, the ability to identify associated genetic loci, and the robustness of evidence of genetic association between adolescent PEs and psychiatric disorders, greater statistical power is required. As previously discussed in Section 1.4.4 and Section 1.4.5, this could be achieved by pooling information from multiple samples and the use of both quantitative and specific measures.

**Table 1.1. Summary of all publications performing genome-wide analysis of psychotic experiences during adolescence.**

Study	Sample	Measure/s	Method	Results
<p><b>1. Zammit et al., 2014</b></p>	<p>Avon Longitudinal Study of Parents and Children (ALSPAC): <math>N = 3,483</math> unrelated individuals assessed for PEs at 12 and 18 years.</p> <p>Primary analysis: Individuals separated into two groups based on the presence of at least one definite PE at either 12 or 18. The no PE and definite PE groups contained 3,059 and 424 individuals respectively.</p> <p>Secondary analysis: Individuals separated into two groups based on the presence of at least one probably PE at either 12 or 18. No PE and probable PE groups contained 2,588 and 912 individuals respectively.</p>	<p>Psychosis-Like Symptoms interview (PLIKSi): Semi-structured interview based on the Schedule for Clinical Assessment in Psychiatry (SCAN). Includes 11 questions assessing hallucinations, delusions and thought interference.</p>	<p>Candidate variation analysis of 17 SNPs that showed a genome-wide significant association with schizophrenia, and or combined schizophrenia and bipolar phenotype.</p> <p>GWAS using logistic regression to compare allele frequencies between PE groups.</p> <p>Polygenic risk score analysis using logistic regression to compare schizophrenia genetic risk scores between PE groups. Schizophrenia genetic risk was based on schizophrenia GWAS summary statistics from the Psychiatric Genomics Consortium (PGC) 1.</p>	<p>Candidate variation analysis returned no variant achieving significance after correction for multiple testing.</p> <p>GWAS identified no variant achieving <math>p &lt; 5 \times 10^{-8}</math>. 121 variants achieved <math>p &lt; 5 \times 10^{-5}</math> representing 31 independent signals.</p> <p>Polygenic risk scoring: No significant association between schizophrenia polygenic risk scores and definite/none PE groups in primary and secondary analyses.</p>

**Table 1.1 cont.**

Study	Sample	Measure/s	Method	Results
<p><b>2.</b> <b>Sieradzka et al., 2014</b> (see also <b>Krapohl et al., 2016</b>)</p>	<p>Discovery sample: Twins Early Development Study (TEDS): <i>N</i> = 2,152 unrelated individuals assessed at age 16.</p> <p>Replication sample: ALSPAC: <i>N</i> = 3,427 unrelated individuals assessed at age 16.</p>	<p>Discovery sample: Specific Psychotic Experiences Questionnaire (SPEQ): Contains six measures assessing specific PE domains including paranoia, hallucinations, grandiosity, cognitive disorganisation, anhedonia and parent-rated negative symptoms.</p> <p>Replication sample: Psychosis-Like Symptoms Questionnaire (PLIKS-Q). Assesses positive PE domain combining paranoia, hallucinations and delusions.</p>	<p>Polygenic risk scoring using linear regression to compare differences in schizophrenia genetic risk and bipolar disorder genetic risk with differences in specific PE domains. Schizophrenia and bipolar disorder genetic risks were based on schizophrenia (PGC 2) and bipolar disorder GWAS summary statistics respectively. A one-tailed hypothesis that there would be a positive association was used.</p> <p>Candidate variation analysis of 33 SNPs previously associated schizophrenia or a combined schizophrenia bipolar phenotype at genome-wide significance. This was initially carried out in TEDS, with any significant SNPs being replicated in the ALSPAC sample.</p> <p>Linear regression was used to compare differences in schizophrenia genetic risk based on genome-wide significant variation only. Schizophrenia genetic risk was calculated using an effect size weighted and unweighted approach.</p>	<p>Polygenic risk score analysis returned no evidence of a positive association between any PE domain and schizophrenia genetic risk. Contrary to the one-tailed hypothesis, results indicated that schizophrenia genetic risk negatively predicts anhedonia and parent-rated negative symptoms, and bipolar disorder genetic risk negatively predicts anhedonia.</p> <p>Candidate variation analysis of previously implicated variation returned no variant achieving significance after correction for multiple testing. The strongest evidence for association was within <i>TCF2</i>. The association between candidate SNPs within <i>TCF2</i> were not significant in the replication sample. However, there were differences in the phenotype used in the discovery and replications samples.</p> <p>Neither the weighted nor unweighted schizophrenia genetic risk score based on genome-wide significant genetic variation significantly predicted any PE domain.</p>

**Table 1.1 cont.**

<b>Study</b>	<b>Sample</b>	<b>Measure/s</b>	<b>Method</b>	<b>Results</b>
<p><b>3.</b> <b>Sieradzka et al., 2015</b></p>	<p>TEDS: <math>N = 2,152</math> unrelated individuals assessed at 16.</p>	<p>See Discovery sample information for study 2.</p>	<p>SNP-based heritability was estimated using GREML in GCTA. The analysis was performed using both LD-pruning and MAF stratification to account for LD. The analysis was also performed without controlling for LD.</p>	<p>The different approaches used to control for LD lead to different heritability estimates. MAF-stratified results were deemed most accurate. MAF-stratified heritability estimates provided evidence that differences in paranoia (6%), cognitive disorganisation (23%), grandiosity (10%) and anhedonia (32%) are in part attributable to common genetic variation. However, estimates have large standard errors and were non-significant except for anhedonia. Hallucinations and Parent-rated Negative Symptom scales returned 0% heritability estimates.</p>

**Table 1.1 cont.**

Study	Sample	Measure/s	Method	Results
<p><b>4.</b> <b>Jones et al., 2016</b></p>	<p>ALSPAC:   <i>N</i> = 5,444 unrelated individuals assessed for positive symptoms at 12 or 18 years. Definite and no positive symptom groups (<i>N</i>=419 and 5,025 respectively) defined by at least one definite PE at either 12 or 18.   <i>N</i> = 3,673 unrelated individuals assessed for negative symptoms at age 16.5. High/low negative symptom groups (<i>N</i> = 337 and 3,336 respectively) were defined using a CAPE score threshold of 14.   <i>N</i> = 4,106 unrelated individuals assessed for depression outcome likelihood at age 15.5. High/low groups (<i>N</i> = 373 and 3,733 respectively) defined using a 15% likelihood of diagnosis.   <i>N</i> = 4,107 unrelated individuals assessed for anxiety outcome likelihood at age 15.5. High/low groups (<i>N</i> = 444 and 3,663 respectively) defined using a 15% likelihood of diagnosis.</p>	<p>Positive symptoms including (hallucinations, delusions, and thought interference) were assessed using the PLIKSi. (see study 1)</p> <p>Negative psychosis-like symptoms were assessed using 10 questions from the Community Assessment of Psychotic Experiences (CAPE) self-report questionnaire. This measure assesses features such as apathy, anergia and asociality.</p> <p>Depressive and anxiety disorder outcomes were derived from the semi-structured Development and Well-Being Assessment (DAWBA) interview.</p>	<p>Polygenic risk score analysis using logistic regression tested for association between genetic risk for schizophrenia using PGC 2 schizophrenia GWAS summary statistics and positive symptoms, negative symptoms, anxiety and depression disorders.</p> <p>Sensitivity analyses were performed to determine whether effects varied when using different thresholds to distinguish groups or when participants were excluded based on a diagnosis of a psychotic disorder at 18 or a parental diagnosis. Effects of adjusting for parental history of schizophrenia or depression were also tested.</p>	<p>A significant positive association between schizophrenia PRS and negative symptoms and anxiety disorder was reported. Positive symptoms showed a near significant positive association when using more relaxed <i>p</i>-value thresholds for the PRS but a near significant negative association when using a stringent <i>p</i>-value threshold for the PRS.</p> <p>Sensitivity analyses showed results as robust to across different thresholds used to determine groups. Results were also not dependent on any diagnoses of the participants or parental history of schizophrenia or depression.</p>

## **1.7 - Future directions for genetic research of adolescent PEs**

This chapter has highlighted the importance and validity of understanding adolescent PEs using a genetic approach. Previous studies have demonstrated that at least some adolescent PEs are partly influenced by common genetic variation. However, no genetic variant has been associated with adolescent PEs at genome-wide. Furthermore, previous studies have also provided limited and mixed evidence for a genetic overlap between adolescent PEs and schizophrenia and other typically adult-onset disorders, requiring further investigation. One overarching limitation of all previous genome-wide studies of PEs is sample size, leading to limited statistical power to provide a robust genetic characterisation of adolescent PEs. Due to the predicted small effect size of associated genetic variation, larger sample sizes are required to identify specific genetic variation. Many of these studies have also suffered from low statistical power due to the effect of phenotypic heterogeneity by the use of binary and broad/non-specific measures of PEs. Phenotypic heterogeneity must be kept to a minimum to improve statistical power. To take this important area of research forward, further investigation is required using large samples assessed using specific and quantitative measures of PEs.

## **1.8 - Aims**

Based on previous literature regarding the genetic factors affecting adolescent PEs and their relationship with psychiatric disorders, this thesis aims to further characterise genetic variation associated with adolescent PEs and the genetic overlap between adolescent PEs and related adult psychiatric disorders. This will be achieved using larger samples of adolescents assessed using both quantitative and specific measures of psychotic experiences. Specifically, this thesis will: 1) Harmonise genetic data from samples with adolescent PE data to enable combined analysis, 2) Harmonise measures of specific PEs between samples, 3) Estimate SNP-heritability of specific PEs. 4) Perform

GWAS of specific PEs to identify associated genetic variation and biological pathways. 5)

Estimate the genetic correlation between specific PEs and typically adult-onset psychiatric disorders including schizophrenia, bipolar disorder and major depressive disorder.

# **Chapter 2 - GWAS of adolescent psychotic experiences in the Twins Early Development Study**

## **2.1 - Introduction**

As discussed in Section 1.5.1, adolescent psychotic experiences (PEs) are heritable and have been associated with a number of psychiatric disorders. Therefore, investigation of the genetic factors associated with adolescent PEs could be informative about the developmental pathways associated with a range of psychiatric disorders. As discussed in Section 1.4.0, a successful approach for the identification of genetic loci associated with common traits is association testing, particularly on a genome-wide scale.

As discussed in Section 1.6.1, there has been one previous GWAS of adolescent PEs (Zammit et al., 2014). This study used data from the Avon Longitudinal Study of Parents and Children (ALSPAC) study and tested for genetic associations with the presence of a definite PE at age 12 or 18. This study focused on the positive symptom domain by assessing individuals using the PLIKSi, which captures paranoia, hallucinations and delusions. No genome-wide significant ( $p < 5 \times 10^{-8}$ ) genetic association was identified with the dichotomous definite or none PE groups, although 31 linkage disequilibrium (LD) independent loci were reported as showing a suggestive association based on a significance threshold of  $p < 5 \times 10^{-5}$ . Further investigation of genetic associations with the full range of PEs assessed using specific quantitative measures is required.

Here, genome-wide association analysis was performed for six quantitative measures of specific PEs applied in an adolescent general population twin sample called the Twins Early Development Study (TEDS) (Haworth, Davis, & Plomin, 2013).

## **2.2 - Methods**

### 2.2.1 - Participants

TEDS recruited twins born in England and Wales between 1994 and 1996, and assessed these individuals longitudinally (Haworth et al., 2013). TEDS originally recruited 13,488 families who had responded with a written consent form. The Institute of Psychiatry ethics committee approved TEDS and their consent procedure (ref: 05/Q0706/228).

In this study we used data collected as a part of the Longitudinal Experiences And Perceptions (LEAP) project, a study within TEDS investigating the aetiology of adolescent psychotic experiences. Families who had withdrawn from TEDS, had never returned any data or had issues with their known address were not invited to participate in LEAP. Of the 10,874 TEDS families invited to participate, 5,076 parents and 5,059 twin pairs provided data on quantitative dimension specific PEs at age 16 years (mean = 16.32 years; standard deviation = 0.68). Participants were excluded based on lack of consent at first contact or for the present study, presence of severe medical disorder(s) including autism spectrum disorder, lack of zygosity information or experience of severe perinatal complications.

### 2.2.2 - Measure of specific PEs

This study used the Specific Psychotic Experiences Questionnaire (SPEQ) to assess dimension-specific psychotic experiences (Ronald et al., 2014). SPEQ assesses five self-report subscales (Paranoia, Hallucinations, Cognitive Disorganisation, Grandiosity and Anhedonia) and one parent-rated subscale (Parent-rated Negative Symptoms). These subscales have high internal consistency, with Cronbach's  $\alpha$  ranging from .77 to .93, and test re-test reliability with a correlation of .65 to .74 across a 9-month interval ( $p=0.001$ ). The SPEQ has been reported as a valid measure of adolescent PEs based on the opinion of expert clinicians, the correlations between the SPEQ subscales and

measures of anxiety, depression and personality, and a moderate to high correlation between the positive subscales of the SPEQ and a previously validated measure of positive PEs called the psychosis-like symptoms questionnaire (PLIKS-Q) (Ronald et al., 2014).

The validity of the SPEQ' subscales has been confirmed by expert clinicians.

### 2.2.3 - Phenotypic analyses

Descriptive statistics for phenotypic measures were calculated using R.

When performing association analyses it is desirable that the phenotypic values fit a normal distribution with a low amount of skew. Four of the specific PEs (Paranoia, Hallucinations, Grandiosity and Delusions, and Parent-rated Negative Symptoms) had a moderate level of skew ( $>1$ ) and were therefore normalised. Normalisation was carried out using a rank-based inverse normal transformation called Van der Waerden transformation in SPSS. This method ranks the data points and then places them into quantiles of a normal distribution. The correlation of a measure before and after transformation was always  $>.91$ .

### 2.2.4 - DNA collection and genotyping

DNA collection and genotyping procedures were carried out by the TEDS research team prior to the beginning of this project. DNA was extracted from 4,440 TEDS unrelated children using buccal cheek swabs. In total, 3,665 samples were successfully hybridised to the AffymetrixGeneChip 6.0 SNP genotyping platform at the Affymetrix headquarters in Santa Clara, California, USA, as part of the TEDS Wellcome Trust Case Control Consortium 2 (WTCCC2) study of reading and mathematical abilities.

Of these samples, 513 were excluded based on one or more of the following parameters: low call rate or heterozygosity outliers, atypical population ancestry, sample

duplication, relatedness to other sample members based on an identity by descent threshold of <5%, unusual hybridisation intensity, gender mismatches, and having less than 90% of genotypes called identically on the genome-wide array and Sequenom panel. This resulted in 3,152 unrelated individuals successfully genotyped consisting of 1,446 males and 1,706 females.

Of these 3,152 genotyped individuals, data on specific psychotic experiences was only available for 2,179 individuals. Of these 2,179 individuals, 837 had a genetically identical sibling for whom it was possible to infer the genotype of. Individuals with inferred genotypes are termed 'pseudo-genotyped'. Given that both siblings of these genotyped monozygotic twin pairs have provided phenotypic information, they can both be included in the association analysis, providing family structure is accounted for. This gives a total sample of 3,016 genotyped individuals who have provided data on specific psychotic experiences (42.1% Male, 57.9% Female).

It is possible that more distantly related individuals removed during quality control of the genetic data could have also been included in subsequent analyses. However, related individuals were removed before the data was received for this thesis. Nonetheless, the ability of currently available methods to account for the presence of identical siblings and more distant relatives simultaneously has not been validated.

The AffymetrixGeneChip 6.0 SNP genotyping platform captured variation at 1,852,600 sites across the genome. Prior to receiving the genetic data, it had been imputed using the 1000 Genomes Phase 1 V3 reference genome. The imputed genetic data was converted to 'hard-call' format using a certainty threshold of 0.9. Light quality control parameters had also been applied. For the purpose of this project, the following more stringent quality control parameters were applied by me: Individual genotyping rate of >90%, Hardy-Weinberg Equilibrium of  $p > 1 \times 10^{-6}$ , SNP genotyping rate of >98%, minor allele frequency threshold of >1%. This left 4,282,342 genetic variants for analysis.

### 2.2.5 - Association analysis

Sex, age and eight principal components of population structure were used as covariates. All analyses used the phenotype residuals resulting from linear regression of the Van de Waerden transformed subscales on the 10 covariates in R using the 'lm' function. The eight principle components of population structure, previously calculated using principle components analysis by Maciek Trzaskowski (Trzaskowski, Eley, et al., 2013), were included as covariates to control for population stratification.

Genome-wide association analysis of all six specific psychotic experiences using related (i.e. MZ twin pairs) and unrelated individuals was performed in PLINK. Additional covariance arises between related individuals due to shared environmental factors. The non-independence of related individuals must be accounted for to estimate correct standard errors. This study used the generalised estimating equation (GEE) method, which creates a covariance matrix and then uses a Huber Sandwich Estimator equation to estimate the correct standard errors (Minică, Dolan, Kampert, Boomsma, & Vink, 2015). This was implemented using the R package called 'gee'. An exchangeable working correlation structure was used because the data was cross sectional with nothing to distinguish members within clusters, making them exchangeable, as the relationship between them is the same. The 'gee' package was implemented in PLINK using the R-plugin function.

Initially, only observed (i.e. not imputed) genetic variation was tested for an association to reduce the computation time. Regions containing genetic variation achieving or close to achieving genome-wide significance ( $p < 5 \times 10^{-8}$ ) were subsequently analysed using imputed genetic variation. Observed genetic variants achieving or close to achieving genome-wide significance were also analysed using MERLIN (Abecasis, Cherny, Cookson, & Cardon, 2002) to confirm validity of using GEE within PLINK. MERLIN uses a mixed effect model approach to account for related individuals. There is a body of

research comparing the two approaches for the control of non-independent observations, with neither method prevailing as superior (Hubbard et al., 2010; Subramanian & O'Malley, 2010).

## **2.3 - Results**

### 2.3.1 - Descriptive statistics

Descriptive statistics for the SPEQ measure are summarised in Table 2.1.

### 2.3.2 - Genome-wide association analysis

Independent genome-wide association analysis of the six specific psychotic experiences and genotyped SNPs collectively returned two variants (rs7830364 and rs7845752) achieving genome-wide significance ( $p < 5 \times 10^{-8}$ ), representing two independent loci, both for cognitive disorganisation (Figure 2.1). Another two loci were close to genome-wide significance for Parent-rated Negative Symptoms, with top variants (rs7587811 and rs16876921) achieving  $p$ -values of  $7.01 \times 10^{-8}$  and  $1.44 \times 10^{-7}$ . Figures 2.2-2.5 show the phenotypic mean per genotype for these four SNPs achieving or close to genome-wide significance. Supplementary Figures 2.1-2.4 show regional associations for only observed genetic variation at these four loci. Collectively, these four loci were considered as regions of interest for subsequent analysis.

The three observed SNPs that showed the strongest associations with PEs (rs7830364 and rs7845752 for Cognitive Disorganisation, and rs7587811 for Parent-reported Negative Symptoms) were tested for association using MERLIN to validate the approach of using GEE in PLINK. The two genome-wide significant SNPs for Cognitive Disorganisation achieved  $p < 3 \times 10^{-5}$  and the one near genome-wide significant SNP for Parent-reported Negative Symptoms achieved  $p = 1.6 \times 10^{-8}$ . Given the instability of very small  $p$ -values, we consider this evidence that GEE in PLINK is a valid approach for accounting for related individuals.

Across all psychotic experiences, 73 variants achieved suggestive significance ( $p < 1 \times 10^{-5}$ ), representing 47 LD independent loci (at least 250Kb apart). A summary of top SNPs in the independent loci associated at genome-wide and suggestive significance for each of the specific psychotic experiences is available in Table 2.2. Quality control procedures and control of family structure using GEE were deemed sufficient with observed p-values showing a roughly uniform distribution between zero and one equating to inflation factors ( $\lambda$ ) between 1.00 – 1.02.

Analysis of the four independent regions of interest (surrounding rs7830364, rs7845752, rs7587811 and rs16876921) using the 1K Genome imputed TEDS dataset was supportive. The locus with the strongest evidence for association with cognitive disorganisation is on chromosome 8, containing 24 variants achieving genome-wide significance when including imputed variation, with the top variant being rs74921500 at  $p = 8.57 \times 10^{-9}$  (Figure 2.6). rs74921500 is within a protein-coding gene called *CSMD1* (CUB and Sushi multiple domains 1). The other locus achieving genome-wide significance with cognitive disorganisation had 3 variants achieving genome-wide significance when including imputed variation, with the top variant being rs7841444 at  $p = 1.68 \times 10^{-8}$  (Figure 2.7). rs7841444 is within *LOC105375732*, an uncharacterised non-coding RNA, proximal to *HAS2* (hyaluronan synthase 2) and *HAS2-AS1* (hyaluronan synthase 2 – antisense 1). The locus most associated with parent-rated negative symptoms, just below genome-wide significance, had 5 variants in ~4Kb region achieving genome-wide significance when including imputed variation, all of which had a  $p = 2.69 \times 10^{-8}$  (Figure 2.8). This region is within a protein-coding gene called *SPAG16* (sperm associated antigen 16). The other regions of interest (surrounding rs16876921) showing a strong association with parent-rated negative symptoms showed no additional evidence for association when including imputed variation due to a limited number of LD proxies. The top variant in this region was still the genotyped SNP (rs16876921) with  $p = 1.44 \times 10^{-7}$  (Figure 2.9). rs16876921 is within *SEPSECS-AS1* (Sep

(O-phosphoserine) tRNA:Sec (selenocysteine) tRNA synthase- Antisense 1), a non-coding RNA with no known function. The variant is also 45Kb upstream of *PI4K2B* (phosphatidylinositol 4-kinase type 2 beta), a protein-coding gene important in the phosphatidylinositol pathway.

**Table 2.1. Descriptive statistics for TEDS sample of related and unrelated individuals and six dimensions of adolescent psychotic experiences assessed using SPEQ.**

	<i>N</i>	<i>Mean</i>	<i>Median</i>	<i>SD</i>	<i>Variance</i>	<i>Range</i>	<i>Skew</i>
<i>Paranoia</i>	2979	11.96	10	10.34	106.84	0-71	1.51
<i>Hallucinations</i>	2985	4.54	2	5.85	34.17	0-42	2.05
<i>Cognitive Disorganisation</i>	2978	3.83	3	2.83	7.99	0-11	0.50
<i>Grandiosity and Delusions</i>	2980	5.20	4	4.28	18.36	0-24	1.12
<i>Anhedonia</i>	2979	15.93	15	7.71	59.45	0-49	0.53
<i>Negative Symptoms</i>	2997	2.63	1	3.56	12.67	0-28	2.35

*Note.* N, Sample size; SD, Standard Deviation.

Table 2.2. List of independent variants with  $p < 1 \times 10^{-5}$  for each of the specific psychotic experiences.

***Paranoia***

<b><i>Rank</i></b>	<b><i>Index SNP</i></b>	<b><i>Chromosome</i></b>	<b><i>Position (bp)</i></b>	<b><i>A1</i></b>	<b><i>MAF</i></b>	<b><i>Beta</i></b>	<b><i>p-value</i></b>	<b><i>Gene symbols</i></b>
1	rs12682930	9	110379567	T	0.06	-0.29	$6.06 \times 10^{-7}$	LOC105376205
2	rs7582778	2	240350457	A	0.13	0.20	$1.39 \times 10^{-6}$	~30Kb from HDAC4
3	rs4128707	12	58288363	G	0.45	-0.13	$4.13 \times 10^{-6}$	LOC101927608
4	rs12168697	22	41107688	T	0.44	0.13	$6.85 \times 10^{-6}$	~30Kb from LOC105373039
5	rs11638592	15	61468155	T	0.15	-0.17	$9.48 \times 10^{-6}$	RORA
6	rs238215	20	47870506	T	0.24	0.14	$9.52 \times 10^{-6}$	ZNFX1
7	rs7801276	7	148632720	T	0.49	0.12	$9.95 \times 10^{-6}$	~6Kb from RNY5

***Hallucinations***

<b><i>Rank</i></b>	<b><i>Index SNP</i></b>	<b><i>Chromosome</i></b>	<b><i>Position (bp)</i></b>	<b><i>A1</i></b>	<b><i>MAF</i></b>	<b><i>Beta</i></b>	<b><i>p-value</i></b>	<b><i>Gene symbols</i></b>
1	rs662968	15	50008443	G	0.45	-0.13	$5.99 \times 10^{-7}$	~50Kb from DTWD1
2	rs9467476	6	25368978	G	0.06	0.26	$1.39 \times 10^{-6}$	LRRC16A
3	rs10777029	12	87731557	T	0.34	-0.12	$3.04 \times 10^{-6}$	~6Kb from LOC105369879
4	rs17204910	15	61451622	C	0.17	-0.15	$4.66 \times 10^{-6}$	~350Kb from LOC105374394
5	rs2740351	17	643426	G	0.44	0.12	$5.08 \times 10^{-6}$	FAM57A
6	rs1501359	5	45337972	C	0.05	-0.24	$6.93 \times 10^{-6}$	HCN1
7	rs11627856	14	48848243	G	0.11	-0.18	$7.50 \times 10^{-6}$	~20Kb from STT3A pseudo
8	rs2600855	3	62539014	A	0.13	-0.17	$7.58 \times 10^{-6}$	CADPS

Table 2.2 cont.

***Cognitive Disorganisation***

<b><i>Rank</i></b>	<b><i>Index SNP</i></b>	<b><i>Chromosome</i></b>	<b><i>Position (bp)</i></b>	<b><i>A1</i></b>	<b><i>MAF</i></b>	<b><i>Beta</i></b>	<b><i>p-value</i></b>	<b><i>Gene symbols</i></b>
1	rs7830364	8	3682023	G	0.01	-1.22	1.24x10 <sup>-8</sup>	CSMD1
2	rs7845752	8	122695330	T	0.01	-1.65	2.98x10 <sup>-8</sup>	LOC105375732, ~35Kb from HAS2
3	rs553850	9	37377959	A	0.14	-0.56	6.72x10 <sup>-7</sup>	LOC105376035
4	rs9315289	13	34852254	T	0.34	-0.41	1.28x10 <sup>-6</sup>	~70Kb from LOC105370158
5	rs279779	1	16776804	C	0.41	0.38	3.97x10 <sup>-6</sup>	NECAP2
6	rs1946972	9	32874871	C	0.09	-0.60	5.02x10 <sup>-6</sup>	~80Kb from APTX and TMEM215
7	rs17071637	6	110742197	C	0.05	-0.76	6.07x10 <sup>-6</sup>	~1Kb from LOC105377936 and ~4Kb from SLC22A16
8	rs508994	11	116280252	A	0.12	-0.51	8.25x10 <sup>-6</sup>	~250Kb from LOC101929011

***Grandiosity and Delusions***

<b><i>Rank</i></b>	<b><i>Index SNP</i></b>	<b><i>Chromosome</i></b>	<b><i>Position (bp)</i></b>	<b><i>A1</i></b>	<b><i>MAF</i></b>	<b><i>Beta</i></b>	<b><i>p-value</i></b>	<b><i>Gene symbols</i></b>
1	rs12197499	6	36214491	A	0.11	0.20	1.15x10 <sup>-6</sup>	PNPLA1, LOC105375036
2	rs4790637	17	4411505	T	0.26	-0.14	3.20x10 <sup>-6</sup>	SPNS2
3	rs7026159	9	86850104	T	0.09	0.20	4.26x10 <sup>-6</sup>	~30Kb from SLC28A3
4	rs17555239	15	25840403	T	0.41	0.12	5.88x10 <sup>-6</sup>	~15Kb from LOC105370737
5	rs2211442	10	55536004	C	0.10	-0.19	6.91x10 <sup>-6</sup>	~15Kb from PCDH15
6	rs7239816	18	67757790	C	0.07	-0.23	8.23x10 <sup>-6</sup>	RTTN
7	rs9541773	13	35252044	G	0.09	-0.20	8.30x10 <sup>-6</sup>	LOC105370159
8	rs17550688	9	8241736	A	0.10	-0.20	8.37x10 <sup>-6</sup>	~70Kb from PTPRD
9	rs3791964	2	218720394	A	0.26	0.14	9.78x10 <sup>-6</sup>	TNS1

Table 2.2 cont.

***Anhedonia***

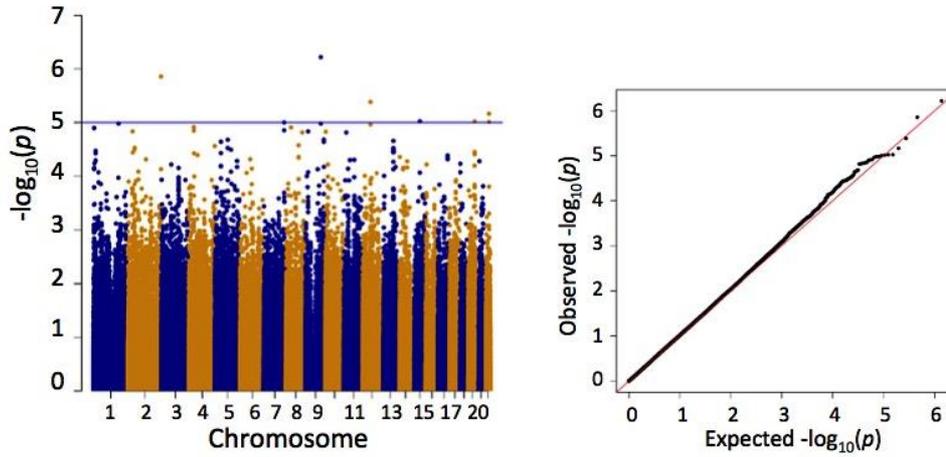
<b><i>Rank</i></b>	<b><i>Index SNP</i></b>	<b><i>Chromosome</i></b>	<b><i>Position (bp)</i></b>	<b><i>A1</i></b>	<b><i>MAF</i></b>	<b><i>Beta</i></b>	<b><i>p-value</i></b>	<b><i>Gene symbols</i></b>
1	rs319027	11	89249010	T	0.01	-3.10	1.33x10 <sup>-6</sup>	NOX4
2	rs2275706	1	220087703	T	0.02	-3.08	1.38x10 <sup>-6</sup>	SLC30A10
3	rs10435834	9	111640631	T	0.33	-1.05	3.49x10 <sup>-6</sup>	IKBKAP
4	rs4843658	16	87690403	C	0.12	1.46	5.45x10 <sup>-6</sup>	JPH3

***Parent-rated Negative Symptoms***

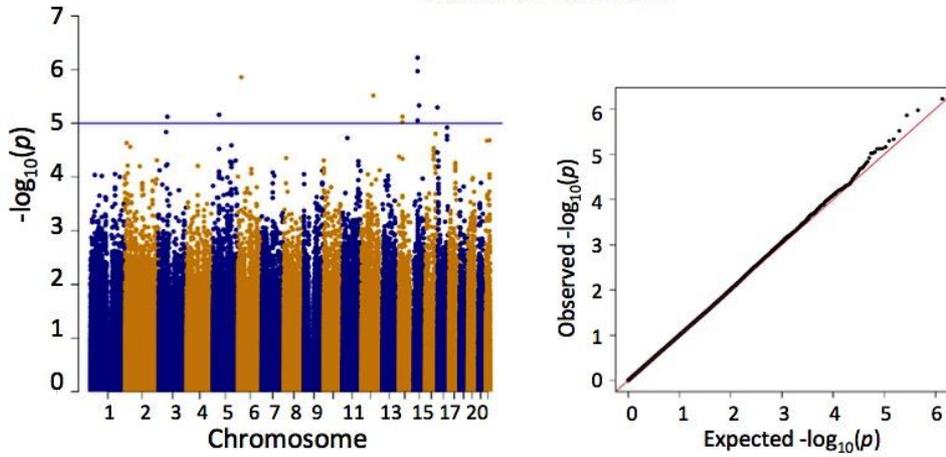
<b><i>Rank</i></b>	<b><i>Index SNP</i></b>	<b><i>Chromosome</i></b>	<b><i>Position (bp)</i></b>	<b><i>A1</i></b>	<b><i>MAF</i></b>	<b><i>Beta</i></b>	<b><i>p-value</i></b>	<b><i>Gene symbols</i></b>
1	rs7587811	2	214457257	C	0.01	0.56	7.01x10 <sup>-8</sup>	SPAG16
2	rs16876921	4	25179865	A	0.01	-0.40	1.44x10 <sup>-7</sup>	SEPSECS antisense RNA 1
3	rs16835045	3	129754605	G	0.05	0.31	1.38x10 <sup>-6</sup>	~0.5Kb from OR7E21P
4	rs2518203	6	102305195	A	0.14	-0.17	1.60x10 <sup>-6</sup>	GRIK2
5	rs10818365	9	122454415	T	0.45	-0.12	1.91x10 <sup>-6</sup>	~150Kb from LOC105376250
6	rs832539	5	56199386	T	0.27	0.14	2.11x10 <sup>-6</sup>	LOC105378980
7	rs1978648	2	43371542	T	0.27	0.14	2.24x10 <sup>-6</sup>	LOC100506047
8	rs4320122	4	120024874	C	0.29	0.14	3.88x10 <sup>-6</sup>	LOC102723967
9	rs12091513	1	115842658	A	0.03	-0.30	6.08x10 <sup>-6</sup>	NGF
10	rs6505386	17	32223728	G	0.15	-0.16	7.43x10 <sup>-6</sup>	ASIC2
11	rs12142944	1	84380880	T	0.03	-0.27	9.31x10 <sup>-6</sup>	TTLL7

Note. Each independent variant is labelled with the gene in which it is present or most proximal. A1, test allele; MAF, minor allele frequency; Beta, unstandardized effect size.

### Paranoia



### Hallucinations



### Cognitive Disorganisation

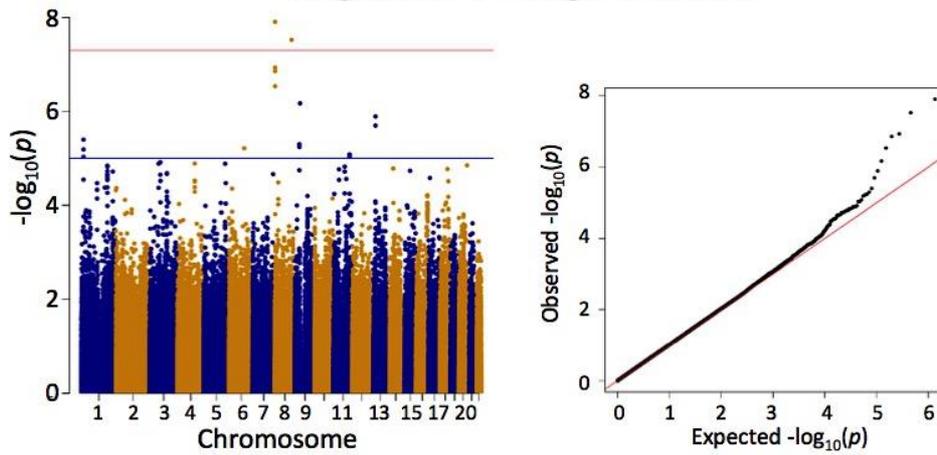
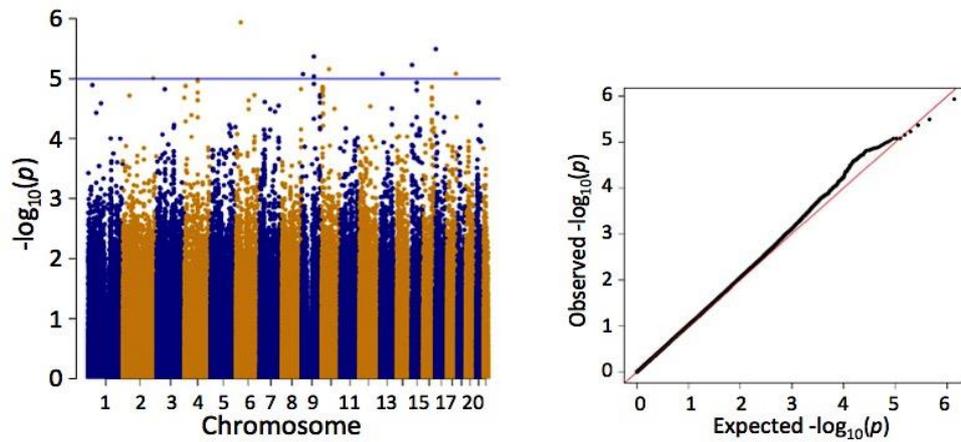
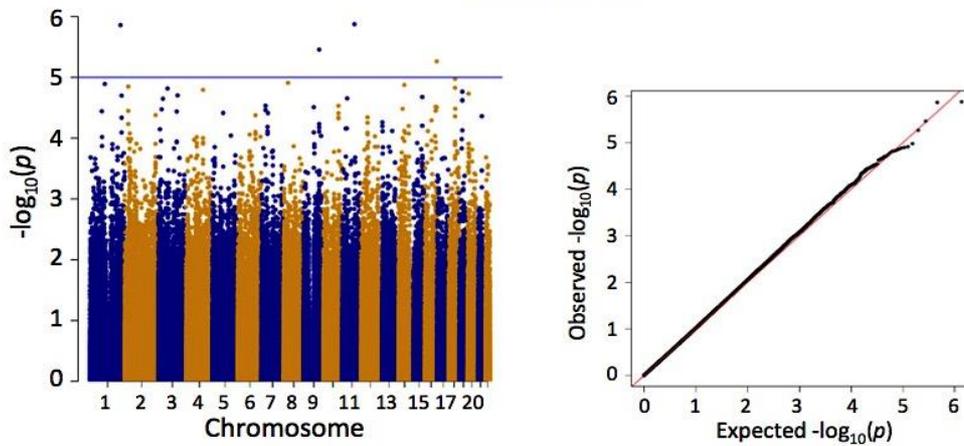


Figure 2.1. Manhattan plot and QQ-plot of specific psychotic experiences in adolescence in TEDS.

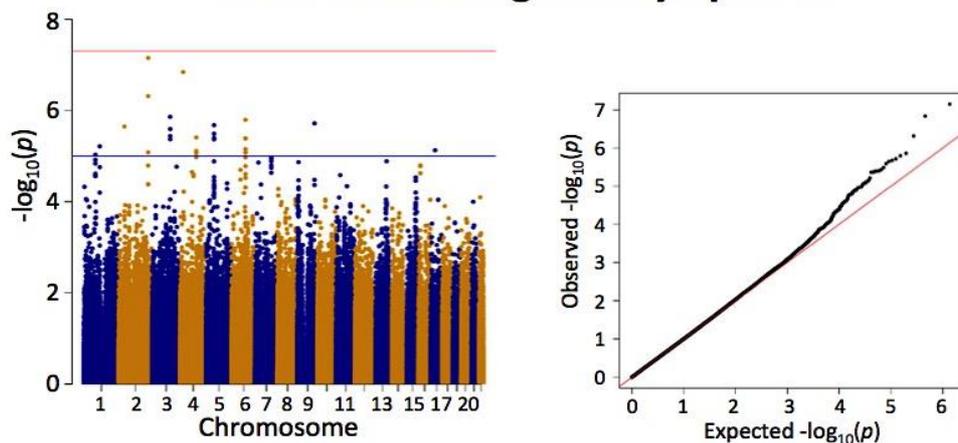
## Grandiosity and Delusions



## Anhedonia

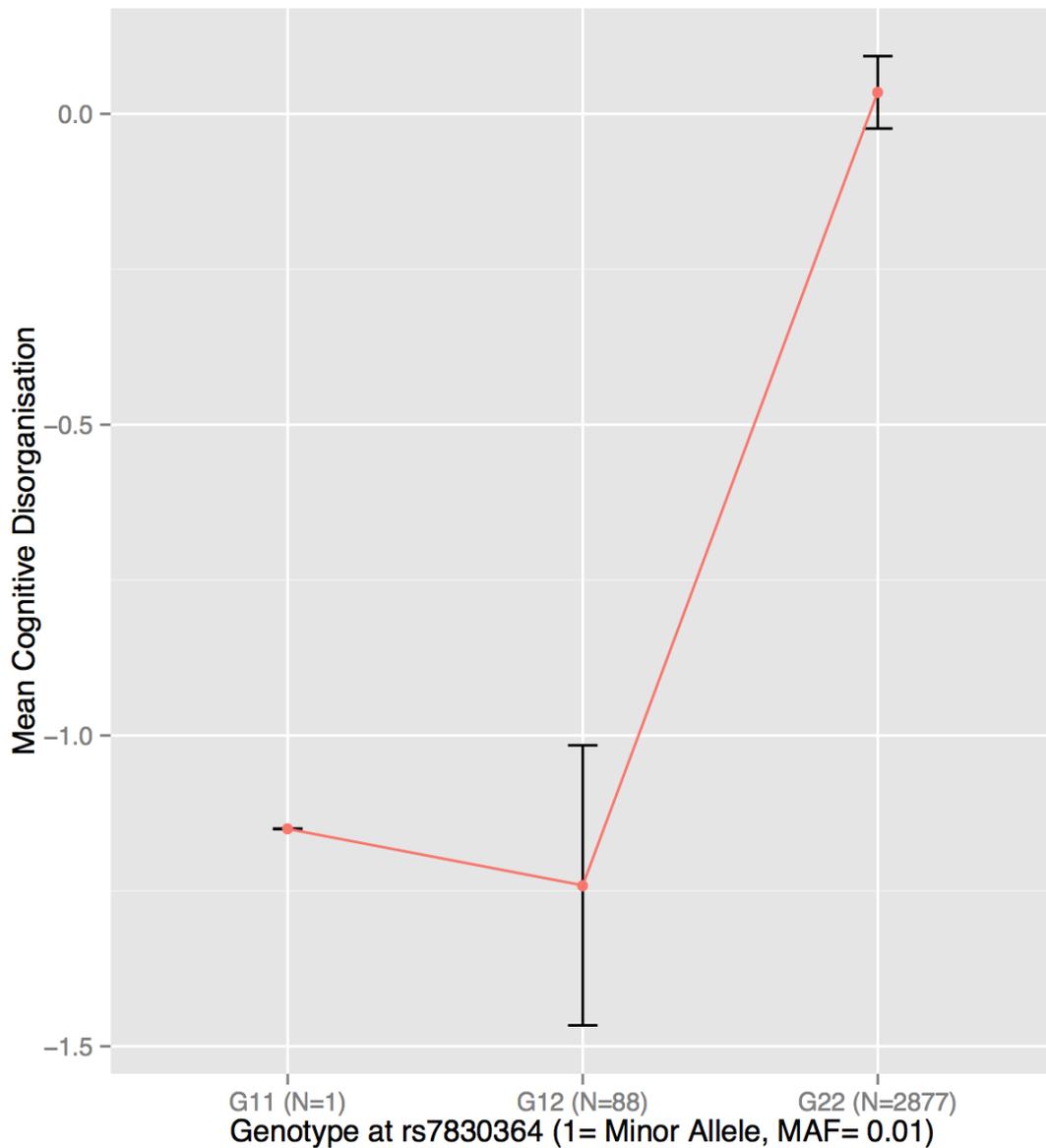


## Parent-rated Negative Symptoms



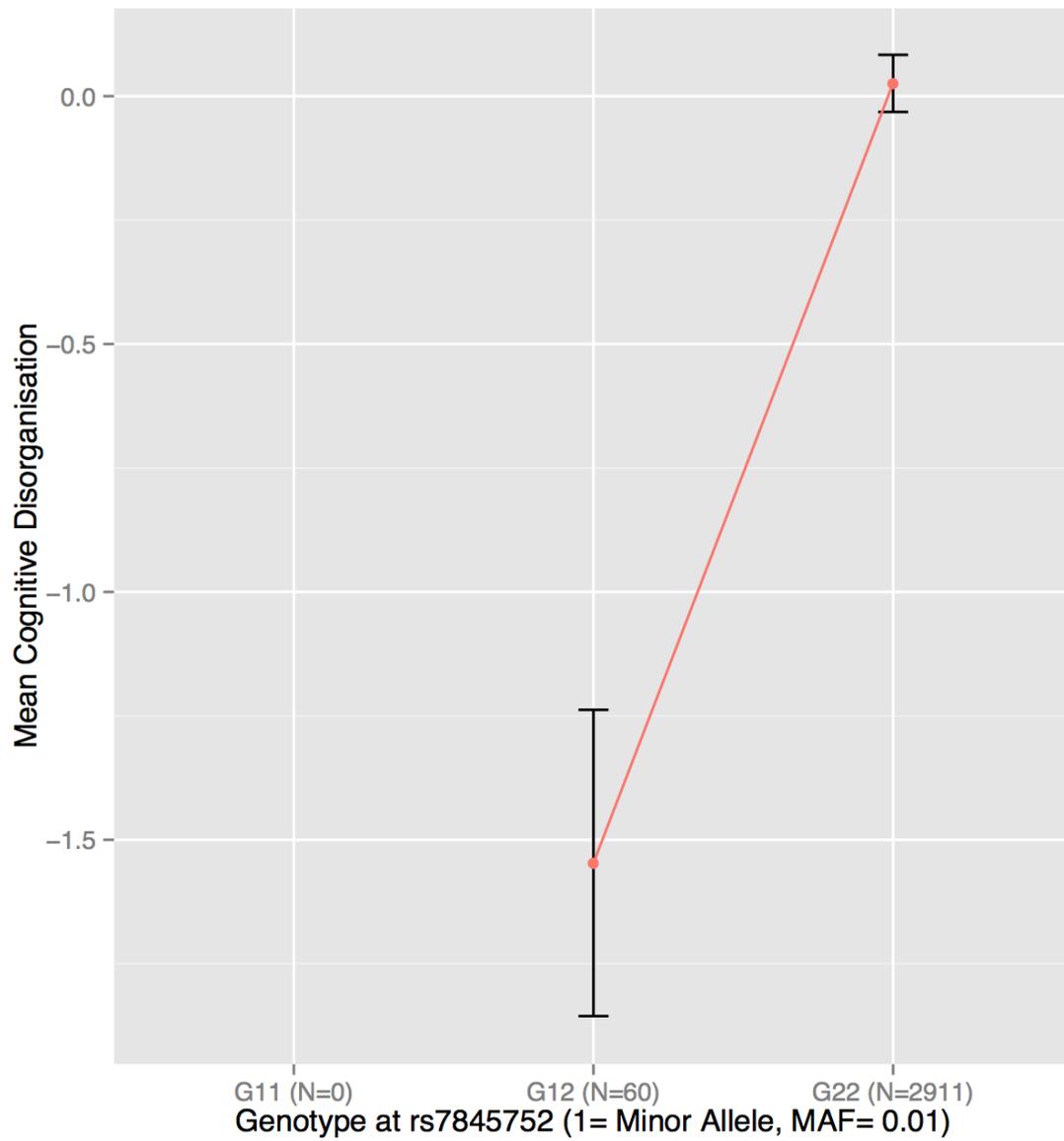
**Figure 2.1 cont. Manhattan plot and QQ-plot of specific psychotic experiences in adolescence in TEDS.**

*Note.* Plots on the left show  $-\log_{10}(p)$  values for genotyped variation across the genome. Plots on the right show observed  $-\log_{10}(p)$  values against expected  $-\log_{10}(p)$  values based on a chi-squared distribution.



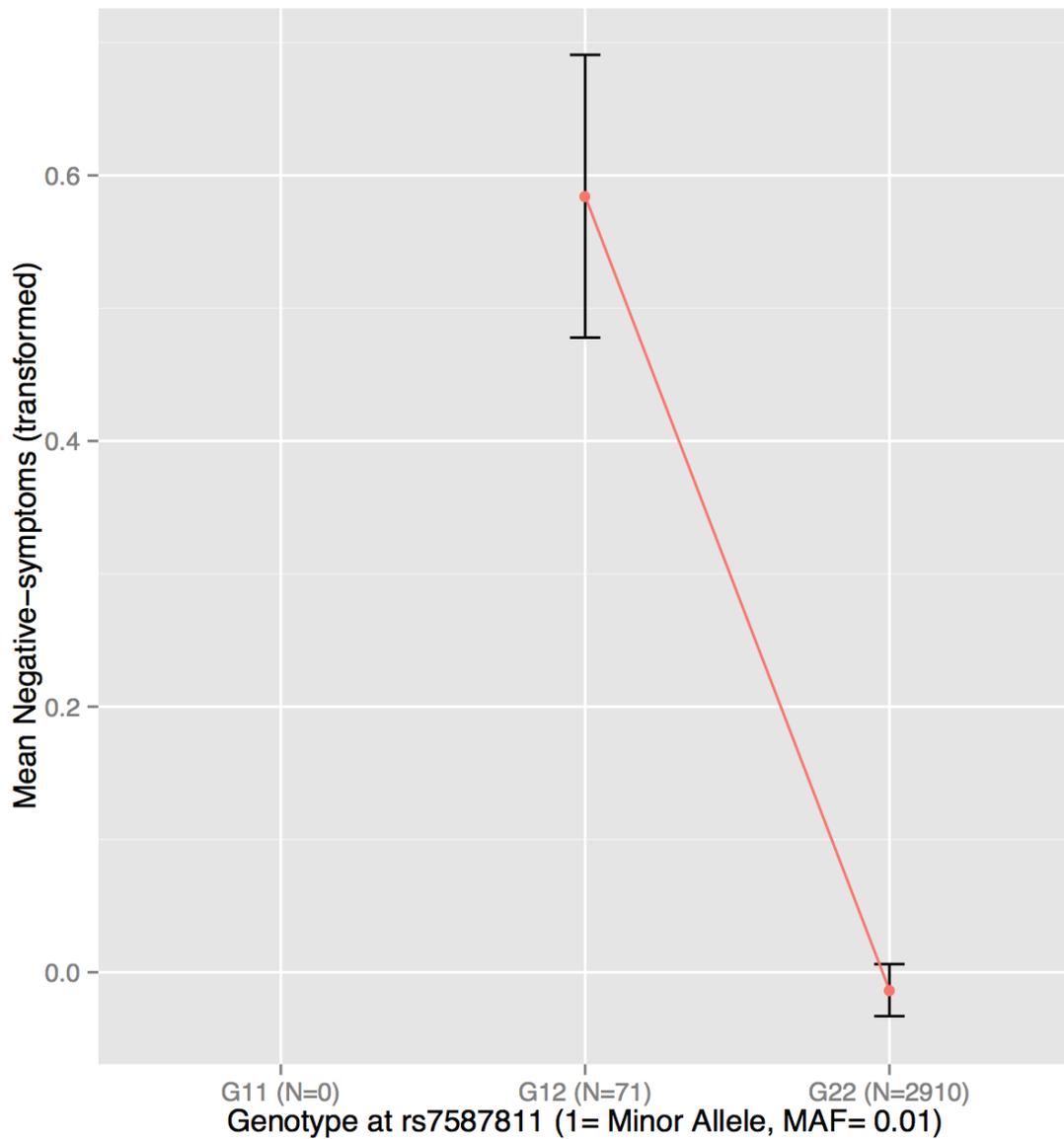
**Figure 2.2. Mean Cognitive Disorganisation scores by genotype at rs7830364.**

*Note.* Standard error bars are sandwich estimator corrected. G11, homozygous minor allele genotype; G12, heterozygous genotype; G22, homozygous major allele genotype; MAF, minor allele frequency.



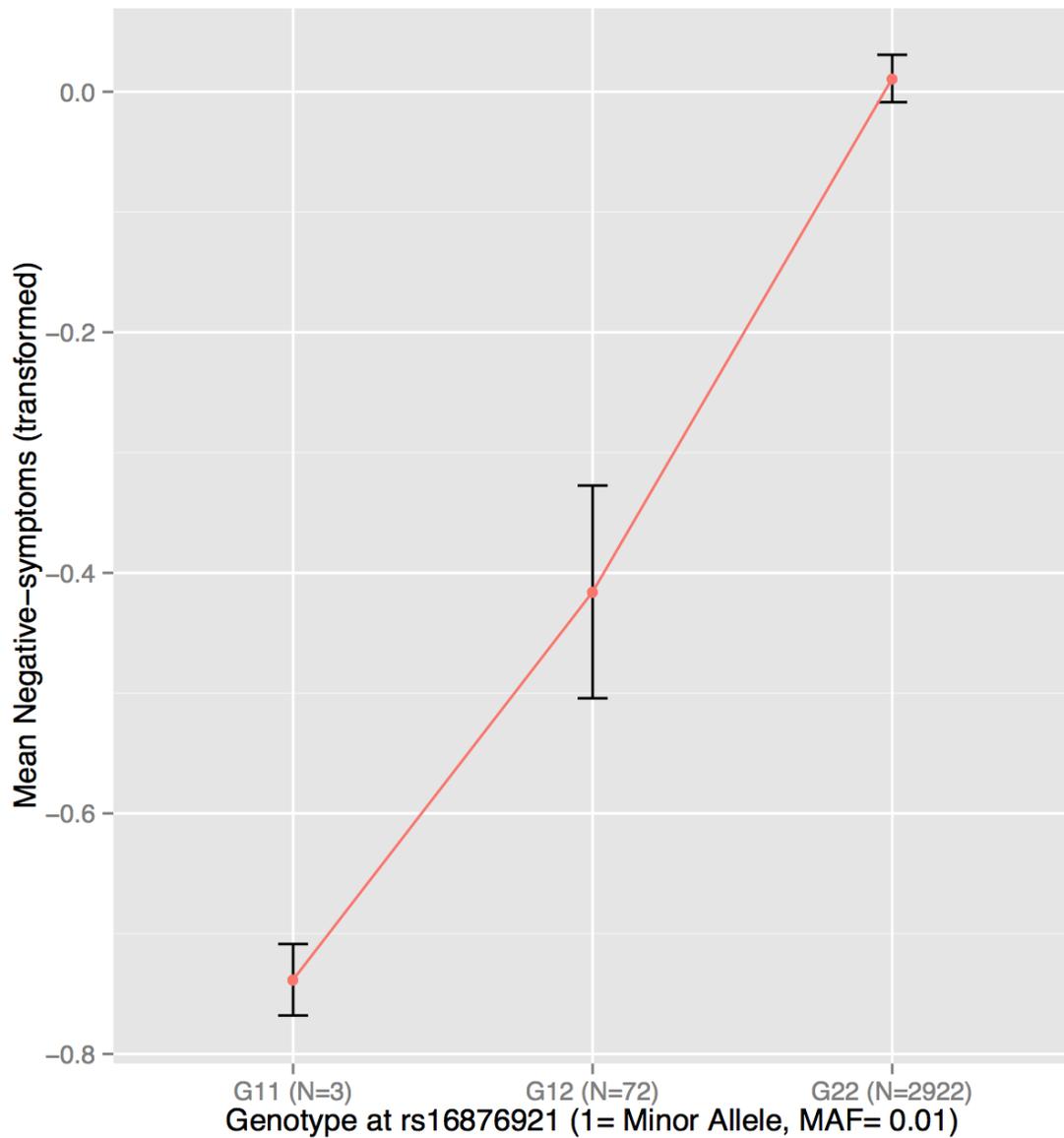
**Figure 2.3. Mean Cognitive Disorganisation scores by genotype at rs7845752.**

*Note.* Standard error bars are sandwich estimator corrected. G11, homozygous minor allele genotype; G12, heterozygous genotype; G22, homozygous major allele genotype; MAF, minor allele frequency.



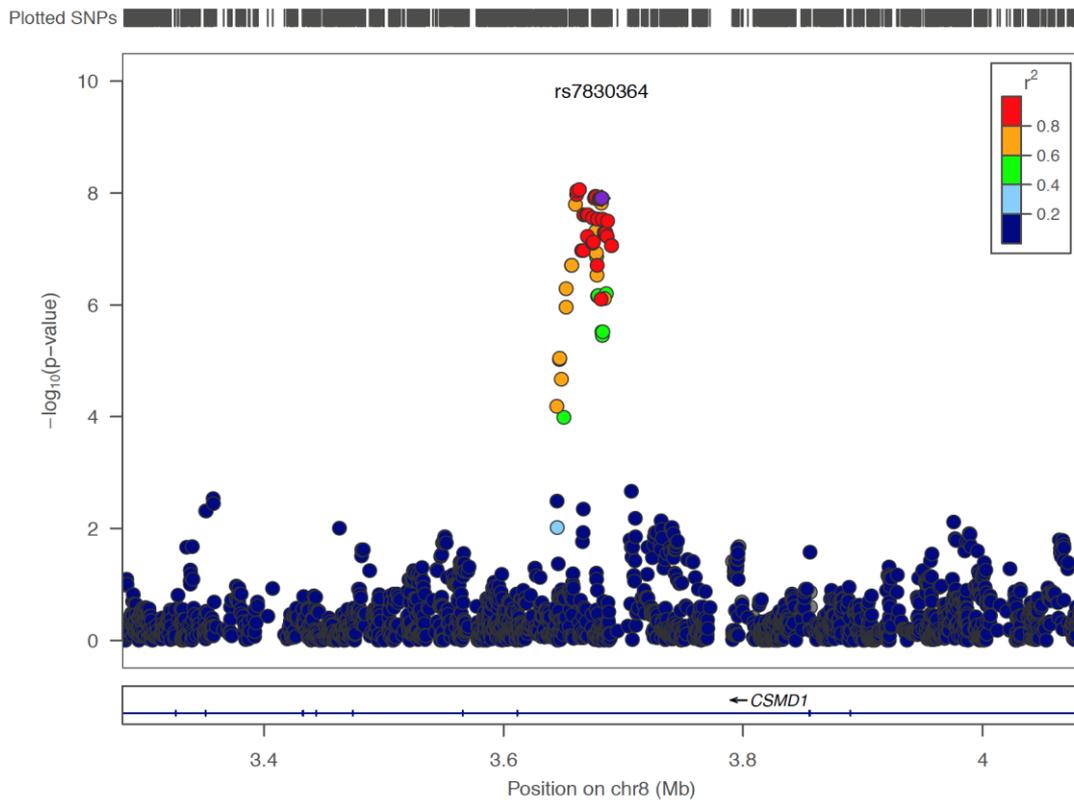
**Figure 2.4. Mean Negative Symptoms scores by genotype at rs7587811.**

*Note.* Standard error bars are sandwich estimator corrected. Negative symptoms scores have been transformed using Van de Waerden transformation. G11, homozygous minor allele genotype; G12, heterozygous genotype; G22, homozygous major allele genotype; MAF, minor allele frequency.



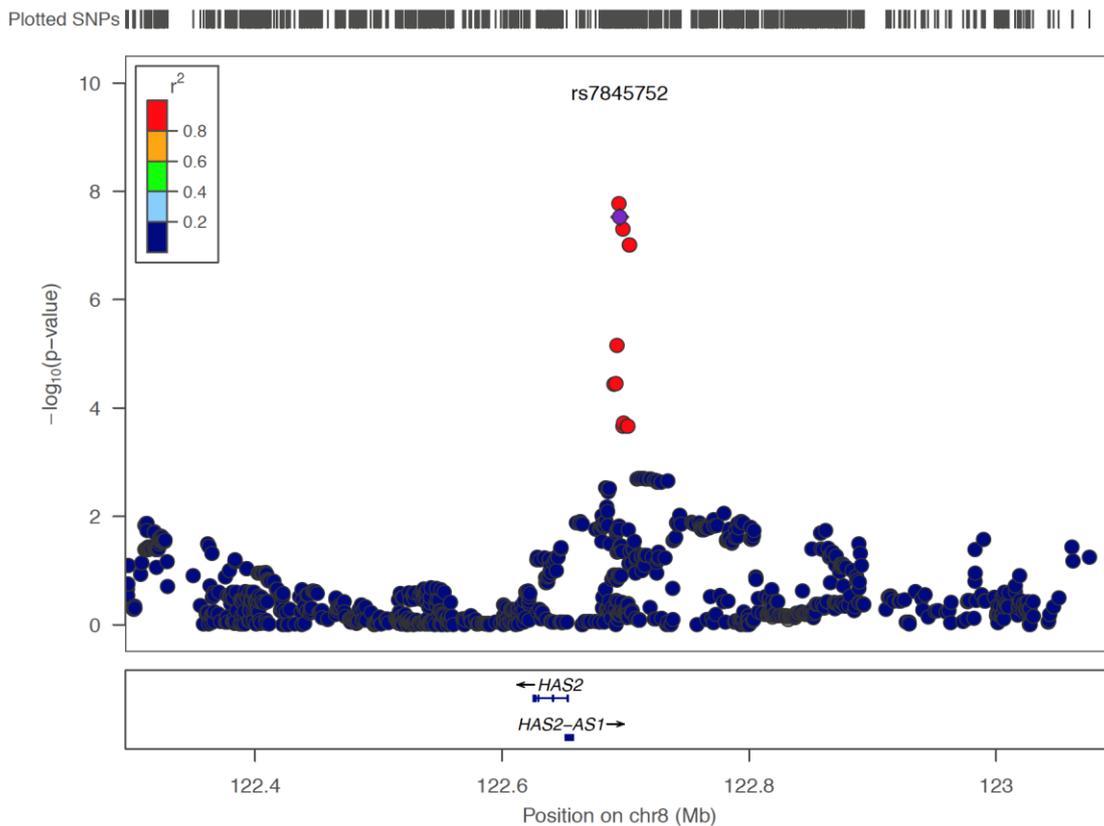
**Figure 2.5. Mean Negative Symptoms scores by genotype at rs16876921.**

*Note.* Standard error bars are sandwich estimator corrected. Negative symptoms scores have been transformed using Van de Waerden transformation. G11, homozygous minor allele genotype; G12, heterozygous genotype; G22, homozygous major allele genotype; MAF, minor allele frequency.



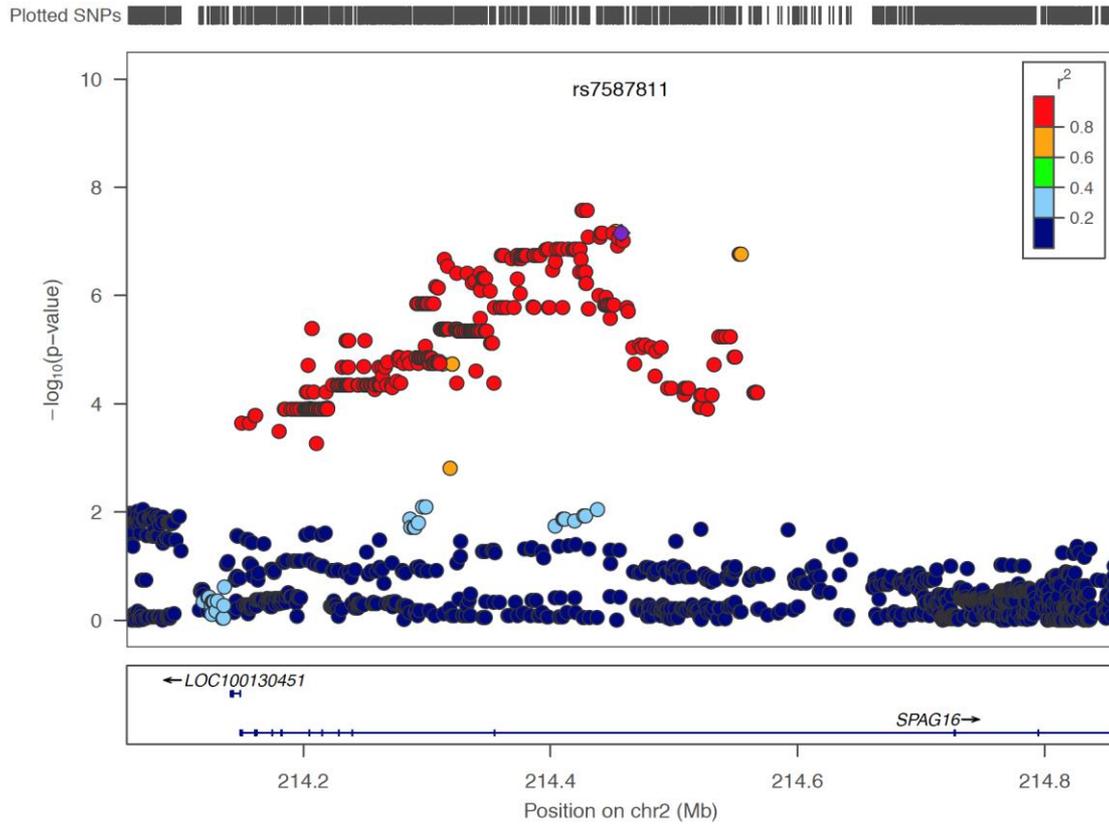
**Figure 2.6. Regional association plot of variation surrounding rs7830364 for Cognitive Disorganisation during adolescence.**

*Note.* The most strongly associated genotyped variant with Cognitive Disorganisation in this region is highlighted in purple (rs7830364). The colour of other points indicates the correlation ( $r^2$ ) of variants with rs7830364. This peak is within the protein-coding gene *CSMD1*. The 'plotted SNPs' bar at the top indicates the density of tagged SNPs in this region. This figure includes genotyped and 1K genome imputed variation. The arrow next to the gene name indicates the direction of transcription. The short vertical lines along the gene represent splice sites.



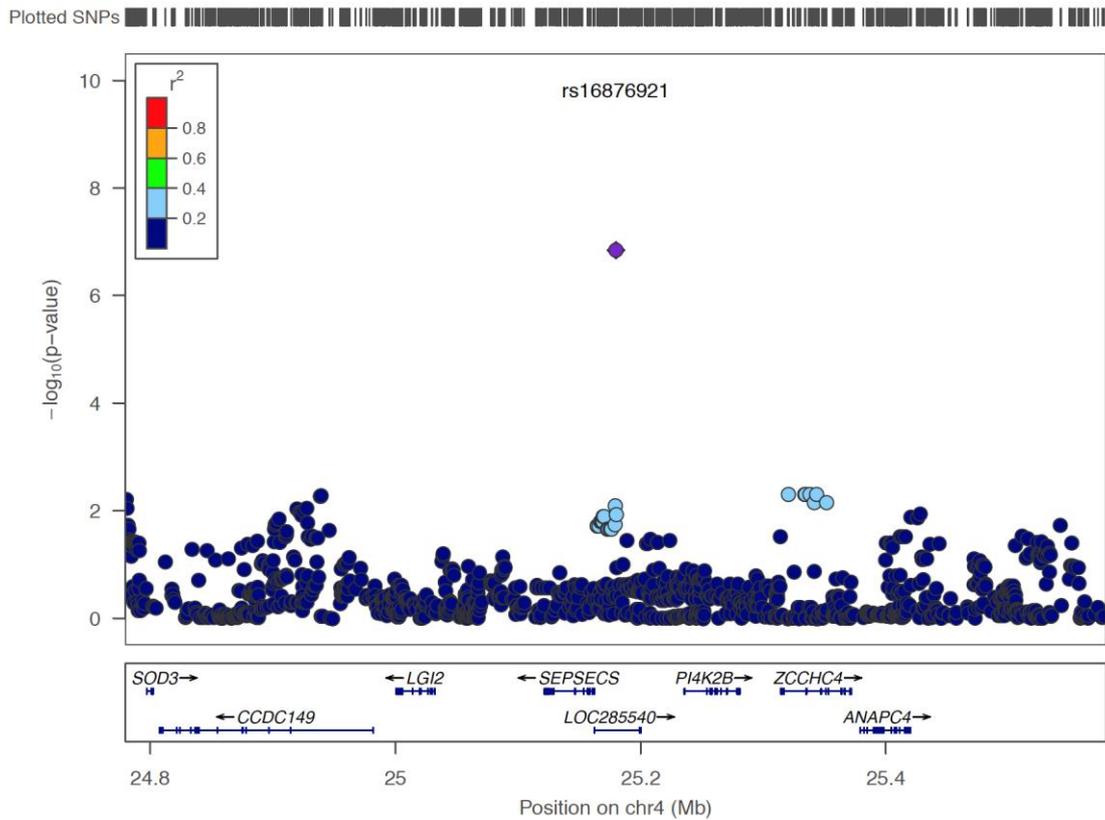
**Figure 2.7. Regional association plot of variation surrounding rs7845752 for Cognitive Disorganisation during adolescence.**

*Note.* The most strongly associated genotyped variant with Cognitive Disorganisation in this region is highlighted in purple (rs7845752). The colour of other points indicates the correlation of variants with rs7845752. This peak is within an uncharacterised non-coding RNA proximal to *HAS2* and *HAS2-AS1*. The 'plotted SNPs' bar at the top indicates the density of tagged SNPs in this region. This figure includes genotyped and 1K genome imputed variation. The arrow next to the gene name indicates the direction of transcription. The short vertical lines along the gene represent splice sites.



**Figure 2.8. Regional association plot of variation surrounding rs7587811 for Parent-reported Negative Symptoms during adolescence.**

*Note.* The most strongly associated genotyped variant with Parent-reported Negative Symptoms in this region is highlighted in purple (rs7587811). The colour of other points indicates the correlation of variants with rs7587811. This peak is within a protein-coding gene called *SPAG16*. This figure includes genotyped and 1K genome imputed variation. The arrow next to the gene name indicates the direction of transcription. The short vertical lines along the gene represent splice sites.



**Figure 2.9. Regional association plot of variation surrounding rs16876921 for Parent-reported Negative Symptoms during adolescence.**

*Note.* The most strongly associated genotyped variant with Parent-reported Negative Symptoms in this region is highlighted in purple (rs16876921). The colour of other points indicates the correlation of variants with rs16876921. This variant is within a non-coding RNA called *SEPSECS-AS1*. This figure includes genotyped and 1K genome imputed variation. The arrow next to the gene name indicates the direction of transcription. The short vertical lines along the gene represent splice sites.

## 2.4 - Discussion

This is the first GWAS of specific adolescent PEs assessed quantitatively. This study has identified the first genome-wide significant variation for Parent-reported Negative Symptoms and two genome-wide significant associations for Cognitive Disorganisation. There were an additional 47 independent loci achieving suggestive significance ( $p < 1 \times 10^{-5}$ ).

The strongest association with Cognitive Disorganisation was within *CSMD1*, a protein-coding gene thought to regulate complement activation and inflammation expressed throughout the central nervous system (Kraus et al., 2006). *CSMD1* has been previously associated with cognitive ability and executive function in healthy males (Koiliari et al., 2014), and psychotic disorders (Ripke et al., 2014; Xu et al., 2014), supporting its role in Cognitive Disorganization in adolescence and the overlapping aetiology of adolescent PEs and psychotic disorders in adulthood. Interestingly, *CSMD1* has also been implicated in the autoimmune disorder multiple sclerosis (Baranzini et al., 2009) and the overlap between multiple sclerosis and schizophrenia (Andreassen et al., 2015).

The other genome-wide significant locus for Cognitive Disorganisation was proximal to *HAS2* and its antisense (*HAS2-AS1*). The *HAS2* protein synthesises hyaluronan and has been implicated in breast cancer (P. Li et al., 2015; Okuda et al., 2012) and both osteo- and rheumatoid- arthritis (Chang, Yamada, & Yamamoto, 2005; Yoshida et al., 2004) but not with any neurodevelopmental phenotypes.

The genome-wide significant locus for Parent-reported Negative Symptoms was within *SPAG16*, a gene encoding two proteins important for the microtubular backbone (axoneme) of tail of sperm and postmeiotic germ cells (Zhang et al., 2007). There is no evidence of *SPAG16* function in neuropsychiatric traits/disorders except one study reporting a suggestive association with a visual endophenotype for schizophrenia and autism (Goodbourn et al., 2014). Otherwise, *SPAG16* has been mainly implicated in

multiple sclerosis and rheumatoid arthritis (de Bock et al., 2014; Knevel et al., 2013), again consistent with the hypothesis of genetic overlap between adolescent PEs and autoimmune disorders.

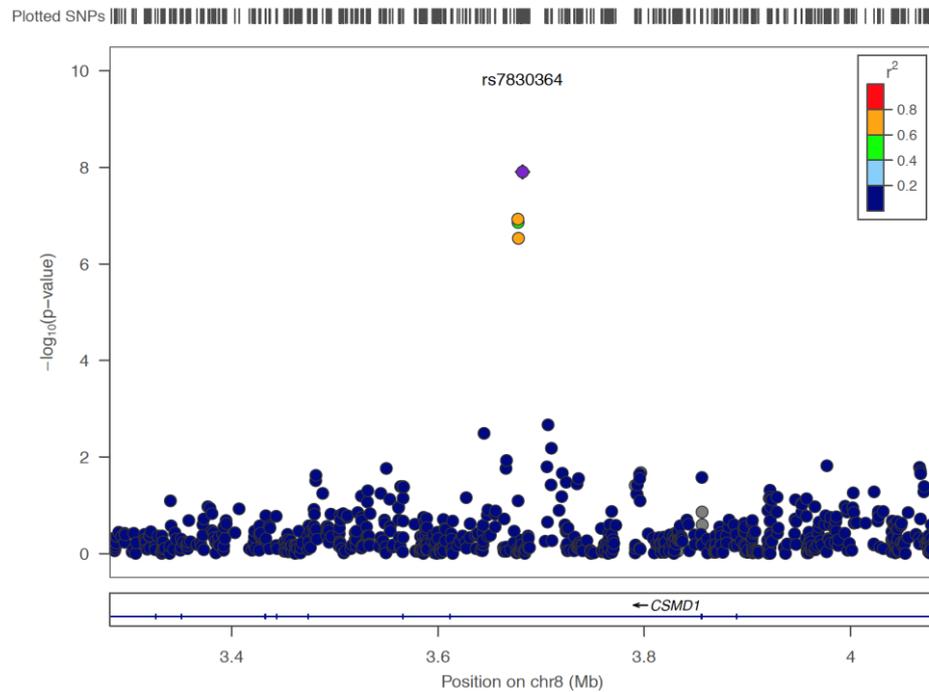
A role for immune-related pathways in the aetiology of psychiatric disorders, particularly schizophrenia, has been suggested by epidemiological studies for many years (Miller & Raison, 2016; Muller & J Schwarz, 2010) and has been more recently supported via molecular genetics (The Pathway Analysis Subgroup of the Psychiatric Genomics Network Consortium, 2015). Given the phenotypic evidence of association of psychiatric disorders with both adolescent PEs (see Section 1.2.2) and immune-related pathways, adolescent PEs and immune pathways may also be associated with one another.

In order to increase sample size this study included the siblings of MZ individuals in the analyses using a generalised estimating equation to account for the additional covariance between related individuals (Minica, Boomsma, Vink, & Dolan, 2014; Minică et al., 2015). This method for accounting for related individuals was supported in this study as the inflation factor in all analyses was 1.00 - 1.02 and top associations were broadly replicated when using a mixed linear model approach applied in MERLIN (Abecasis et al., 2002). Although related individuals were included, the major limitation of this study is sample size. Integration of other adolescent samples both with genome-wide variation and specific PEs phenotypes is required to improve statistical power and identify further genetic variation and gene pathways associated with adolescent PEs.

In conclusion, the use of specific quantitative measures of adolescent PEs has successfully identified three genetic loci significantly associated with adolescent PEs. However, larger sample sizes are required to improve statistical power to detect further genome-wide significant variation, and as such efforts were not made to replicate these

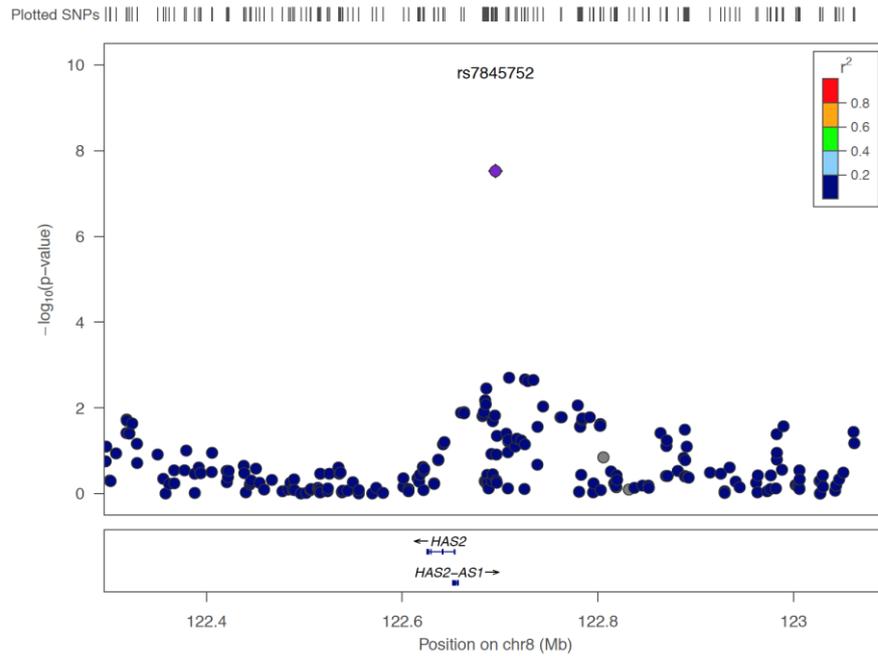
associations at this stage. This study has also provided some indication that adolescent PEs and immune-related pathways may overlap in aetiology to some degree.

## 2.5 – Appendix



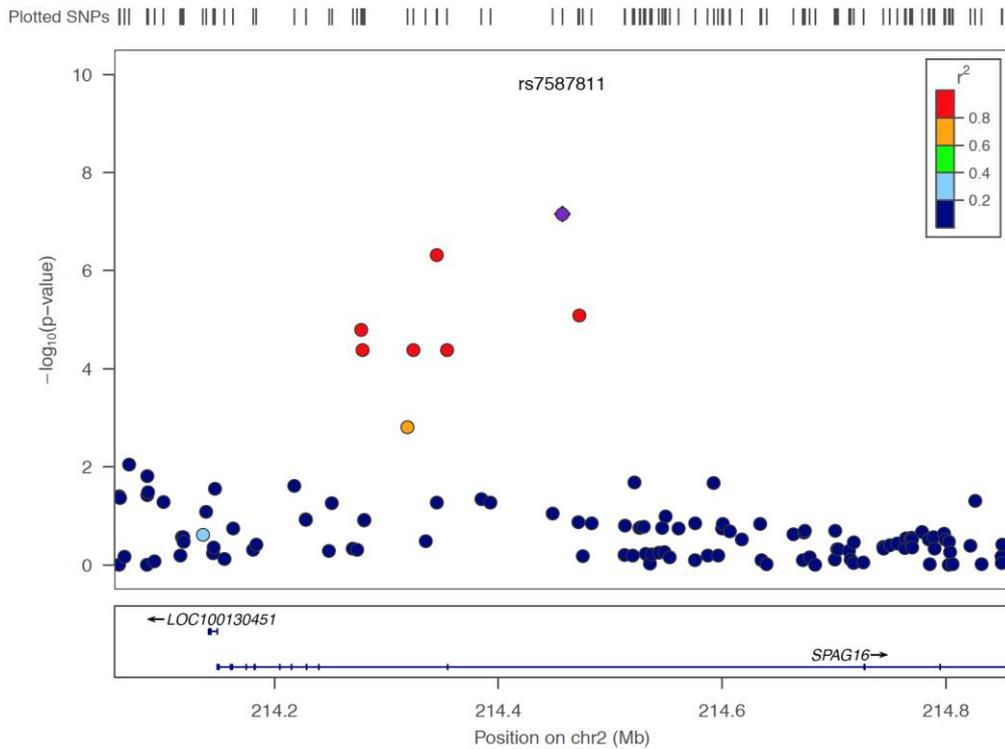
**Supplementary Figure 2.1. Regional association plot for observed variation only surrounding rs7830364 for Cognitive Disorganisation during adolescence.**

*Note.* The most strongly associated genotyped variant with Cognitive Disorganisation in this region is highlighted in purple (rs7830364). The colour of other points indicates the correlation of variants with rs7830364. This peak is within the protein-coding gene *CSMD1* (CUB and Sushi multiple domains 1). This figure includes only genotyped variation. The arrow next to the gene name indicates the direction of transcription. The short vertical lines along the gene represent splice sites.



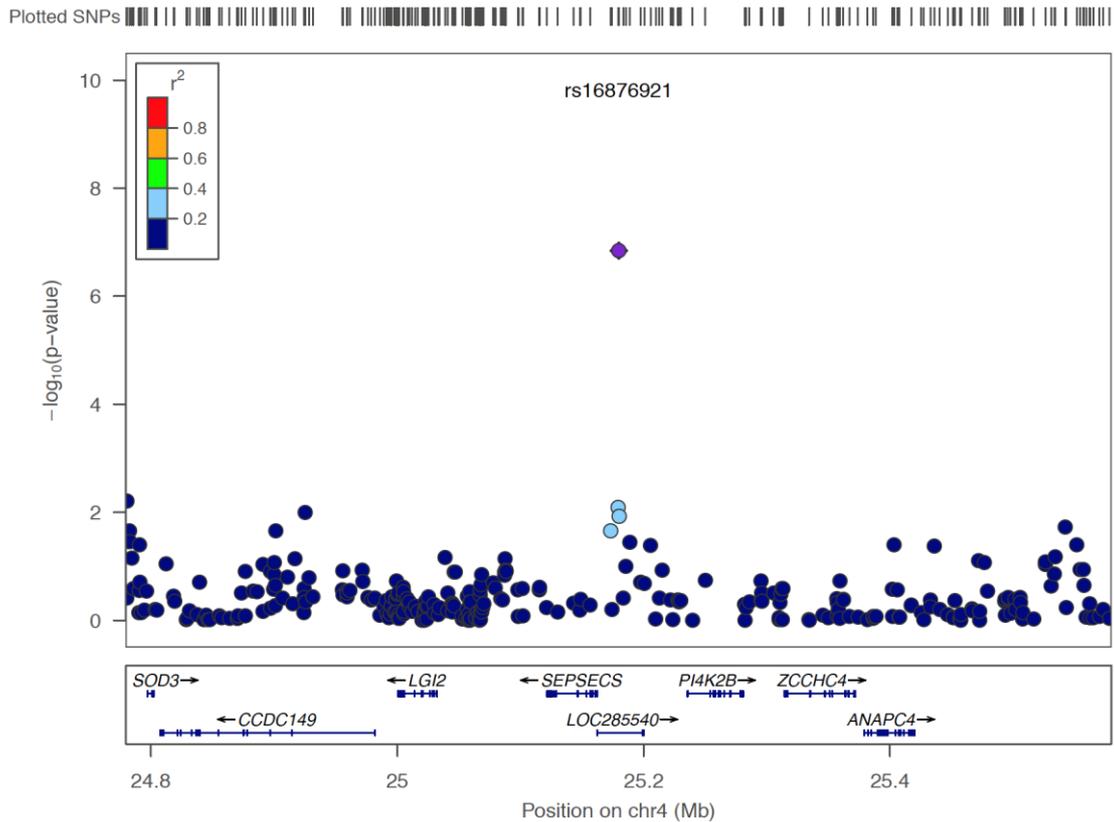
**Supplementary Figure 2.2. Regional association plot for observed variation only surrounding rs7845752 for Cognitive Disorganisation during adolescence.**

*Note.* The most strongly associated genotyped variant with Cognitive Disorganisation in this region is highlighted in purple (rs7845752). The colour of other points indicates the correlation of variants with rs7845752. This peak is within an uncharacterised non-coding RNA proximal to *HAS2* (Hyaluronan Synthase 2) and *HAS2-AS1*. This figure includes only genotyped variation. The arrow next to the gene name indicates the direction of transcription. The short vertical lines along the gene represent splice sites.



**Supplementary Figure 2.3. Regional association plot for observed variation only surrounding rs7587811 for Parent-reported Negative Symptoms during adolescence.**

*Note.* The most strongly associated genotyped variant with Parent-reported Negative Symptoms in this region is highlighted in purple (rs7587811). The colour of other points indicates the correlation of variants with rs7587811. This peak is within a protein-coding gene called *SPAG16* (sperm associated antigen 16). This figure includes only genotyped variation. The arrow next to the gene name indicates the direction of transcription. The short vertical lines along the gene represent splice sites.



**Supplementary Figure 2.4. Regional association plot for observed variation only surrounding rs16876921 for Parent-reported Negative Symptoms in adolescence.**

*Note.* The most strongly associated genotyped variant with Parent-reported Negative Symptoms in this region is highlighted in purple (rs16876921). The colour of other points indicates the correlation of variants with rs16876921. This variant is within a non-coding RNA called *SEPSECS-AS1*. This variant is also proximal to *PI4K2B*. This figure includes only genotyped variation. The arrow next to the gene name indicates the direction of transcription. The short vertical lines along the gene represent splice sites.

# **Chapter 3 - Investigating the effect of the procedures used when controlling for the normality assumption and covariates**

## **3.1 – Introduction**

As briefly mentioned in Section 1.4.1, the most common statistical method for assessing the association between a specific genetic variant and a given phenotype is either logistic or linear regression. This thesis aims to use quantitative measures in all primary analyses and linear regression will be used in Chapters 5-7. Therefore linear regression will be the focus of this chapter. This chapter represents a departure from phenotypic and genetic analyses of psychotic experience data (which will continue in chapters 4 onwards) to explore a methodological issue.

Linear regression, as well as many other statistical methods, has the underlying assumption that the residuals of the model are normally distributed (Berry, 1993). This assumption is important for the accurate approximation of significance. Violation of the normality assumption can lead to heteroskedasticity, the comparison of variables with unequal variance, which potentially increases both type-I error rates and reduced power (Feingold, 2002). In genetic studies, the normality of residuals is largely dictated by the distribution of the dependent (phenotypic) variable due to the very small effect size of individual genetic variants (Servin & Stephens, 2007).

The normality assumption can either be satisfied by the transformation of the phenotypic variable (normalisation), or controlled for using heteroskedasticity robust methods such as generalized estimating equations (GEEs). One of the most popular approaches is the normalisation of the dependent variable. There are several transformations that can be used for this purpose, the most popular being log, power or

Box-Cox transformations, and rank-based inverse normal transformations (INTs), also referred to as quantile normalisations, such as the Van de Waerden transformation (Beasley, Erickson, & Allison, 2009). In many cases the use of log transformation is insufficient for normalising data. Conversely, rank-based INTs always create a perfect normal distribution when there are no tied observations. Previous studies have reported that although rank-based INTs can lead to loss of information, this approach controls power and type-I error rate (Peng, Robert, DeHoff, & Amos, 2007; K. Wang & Huang, 2002). However, a comprehensive review of rank-based INTs demonstrated that in certain scenarios, rank-based INTs do not control type-I error, although they remain useful in large samples where alternative methods, such as resampling, are less practical (Beasley et al., 2009).

It is often desirable to adjust for covariates in analysis. In genetic studies, principal components of ancestry are commonly included to reduce confounding by population structure. When a transformation to normality is used, the covariates may be included in the analysis model after transformation, or alternatively they may be regressed against the response prior to the residuals being transformed to normality. The latter approach has been used in a number of recent high profile studies (S. E. Jones et al., 2016; Locke et al., 2015; Wain et al., 2015) and is also automated in the 'rntransform' function within GenABEL, a popular R package (Aulchenko, Ripke, Isaacs, & Van Duijn, 2007). One reason is that in collaborative consortia, it is more convenient to perform in-house adjustments for covariates and transformations prior to data sharing. Another reason is that confounders may be considered to have their effects on the untransformed, rather than the normalised, variable. Finally, pre-adjustment for covariates will break many of the ties that are present in data derived from questionnaires or other rating scales that are usually represented by a small number of discrete values.

Although the approach of regressing covariate effects from the dependent variable prior to normalisation has some practical advantages, to the best of our knowledge, after extensive searches and reading of the surrounding literature, the effect of this procedure has not been documented. This chapter investigates whether applying rank-based INT to residuals reintroduces a linear correlation with covariates. Three factors predicted to effect the outcome of this process are evaluated: original skew of the dependent variable, proportion of tied observations in the dependent variable, and the original correlation between the dependent variable and covariates. This chapter also investigates the consequence of an alternative procedure where the dependent variable is normalised (randomly splitting tied observations) prior to regressing out covariate effects. Both simulated and real data are used.

## **3.2 – Methods**

### 3.2.1 - Simulation of phenotypic data

Two types of phenotypic data were simulated: quantitative variables containing no tied observations (herein referred to as continuous variables) and quantitative variables containing tied observations (herein referred to as questionnaire-type variables). These variables were simulated to exhibit different degrees of skew ranging from -2 to 2.

Skewed variables were created using the R 'rbeta' function, which randomly generates numbers following a beta distribution with two shape parameters to control the degree of skew. Each simulated variable contained 10,000 observations. To create tied observations in the questionnaire-type variables, the initially continuous data were collapsed into evenly distributed and discrete response bins. The number of response bins, determining the proportion of tied observations, was varied between 5 and 160 to capture the typical ranges of questionnaire-type data.

The R functions used to create continuous and questionnaire-type variables, called 'SimCont' and 'SimQuest' respectively, are available in the appendix (Supplementary Notes 3.1 and 3.2).

A normal distribution has skew = 0 but also kurtosis = 0. Given that the simulated variables were generated to follow a beta distribution, variables with a skew equal to zero may not have a kurtosis equal to zero. To ensure that the correction of kurtosis was not driving effects seen when skew is equal to zero, continuous and questionnaire-type variables were also generated using the 'rnorm' function in R to exhibit both a skew and kurtosis of zero. The functions used to create continuous and questionnaire-type with skew and kurtosis fixed to zero, called 'SimContNorm' and 'SimQuestNorm' respectively, are available in the appendix (Supplementary Notes 3.3 and 3.4).

### 3.2.2 - Simulation of covariate data

To create correlated covariate data, noise was added to each simulated phenotypic variable until the desired phenotype-covariate correlation was achieved. Phenotype-covariate correlations (Pearsons) were varied between -0.5 and 0.5 to investigate the full range of possible dependent variable-covariate relationships. Noise was added to the questionnaire variables using the 'jitter' function in R.

The R function used to create covariates for each phenotypic variable, called 'CovarCreator', is available in the appendix (Supplementary Note 3.5).

### 3.2.3 - Testing the effect of rank-based inverse normal transformation after regressing out covariate effects

Linear regression of each covariate against the corresponding phenotypic variable was used to calculate phenotypic residuals, which are linearly uncorrelated with the covariates. The Spearman's rank-based correlation between the phenotypic residuals and covariates was measured. Phenotypic residuals were then normalised using the

'rntransform' from the GenABEL package in R, which applies a rank-based INT similar to van de Waerden transformation. To determine whether the transformed residuals were still linearly uncorrelated with covariates, the Pearson correlation between the transformed residuals and covariates was calculated.

#### 3.2.4 - Testing the effect of applying rank-based inverse normal transformation (randomly splitting ties) before regressing out covariate effects

This was carried out using the same simulated questionnaire-type and continuous variables and covariates. The raw questionnaire-type and continuous variables underwent rank-based INT using a modified version of the 'rntransform' function from GENABEL that randomly ranks any tied observations. The modified version of 'rntransform', called 'rntransform\_random', is available in the appendix (Supplementary Note 3.6). Linear regression of each covariate against the corresponding normalised questionnaire-type and continuous variables was used to calculate phenotypic residuals, which are linearly uncorrelated with the covariates.

One concern with rank-based INT, particularly when randomly splitting ties, is that the linear relationship between the phenotypic variable and independent variables (including covariates) may be severely distorted. To determine the extent to which rank-based INT when randomly splitting ties distorts phenotypic variables, the Pearson correlations between the untransformed and transformed phenotypic variables were calculated. To determine the extent to which rank-based INT when randomly splitting ties distorts the relationship between the phenotypic variables and covariates, the Pearson correlation between the transformed phenotypic variables and covariates was calculated.

Another concern with normalising the phenotypic variable before regressing out covariates is that the process of regressing out covariates may re-introduce skew in the

residuals. To determine the extent to which regressing covariates from normalised phenotypic variables re-introduced skew, the skew of the residuals was assessed.

### 3.2.5 - Demonstration using real data

To determine whether the predicted effects (when using simulated data) of performing rank-based INT before or after regressing out covariate effects are accurate, the same procedure was applied to real questionnaire data provided by the Twins Early Development Study (TEDS)(Haworth et al., 2013). TEDS data was acquired via the standard data request form procedure. The TEDS study has ethical approval from the Institute of Psychiatry ethics committee (ref: 05/Q0706/228). Data from two questionnaires were used measuring Paranoia and Anhedonia. Both of these measures are part of the SPEQ (Specific Psychotic Experiences Questionnaire)(Ronald et al., 2014). Unrelated individuals were included in subsequent analyses. Individuals with missing phenotypic data were excluded from all analyses. Sum scores of unrelated individuals were calculated by summing the response of each item. Each item of both the Paranoia and Anhedonia scales were coded as values from 0-5, with the total ranges of the Paranoia and Anhedonia scales being 0-75 and 0-50 respectively. Sum scores were calculated using different numbers of items (1, 2, 4, 8) to create different numbers of response bins (5, 10, 20, 40) as in the simulation study. The covariates used were age (continuous variable skew of -0.32) and sex (binary variable with skew of 0.22). Table 3.1 shows the skew, number of response bins (proportion of ties) and correlation with covariates for each of dependent variable. The TEDS data were analysed using the same procedure as the simulated data.

**Table 3.1. Skew, range, and correlation with covariates for dependent variables derived from TEDS sample.**

Dependent variable	Range	Skew	Pearson correlation with age	Pearson correlation with sex
Paranoia	5	1.357	0.055	0.018
Paranoia	10	1.195	0.043	-0.026
Paranoia	20	1.095	0.030	-0.022
Paranoia	40	1.296	0.022	-0.059
Anhedonia	5	1.868	-0.006	0.177
Anhedonia	10	0.858	-0.025	0.127
Anhedonia	20	0.651	-0.020	0.135
Anhedonia	40	0.537	-0.013	0.205

### 3.3 – Results

#### 3.3.1 - Simulated data

##### 3.3.1.1 - The effect of rank-based inverse normal transformation after regressing out covariate effects using simulated data

As expected, regressing covariates against phenotypic variables created phenotypic residuals that were linearly uncorrelated with covariates. Although there was no linear correlation, in almost all simulations a rank-based correlation existed between the residuals and covariates (Supplementary Figures 3.1-3.7). As a consequence, rank-based INT of residuals re-introduced a linear correlation between the phenotypic variables and covariates (Supplementary Figures 3.8-3.14). Three factors predicted to affect the extent to which rank-based INT of residuals re-introduced a correlation between the phenotypic variables and covariates were tested. These factors were the original skew of the phenotypic variable, the original correlation between the phenotypic variable and covariate, and the proportion of tied observations in the original phenotypic data.

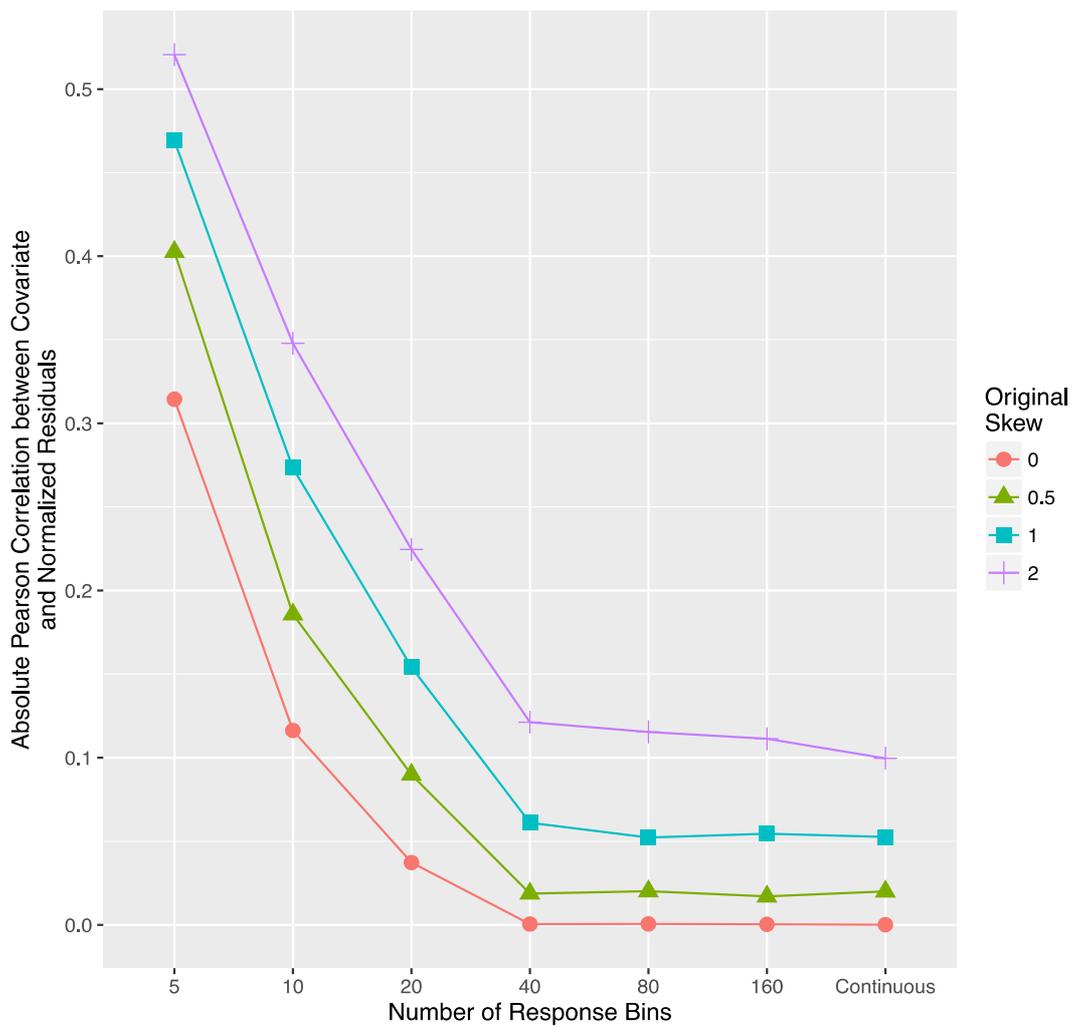
First, in terms of skew, greater skew of the phenotypic variable was associated with a higher correlation between the normalised phenotypic residuals and the covariate data (Figure 3.1, Supplementary Figures 3.8-3.14). The direction of skew had no effect on the correlation between the normalised residuals and the covariate data. The effect of normalising residuals when skew was equal to zero remained when kurtosis was also fixed to zero (Supplementary Figure 3.15).

Second, the direction of the original correlation between the original phenotypic variable and the covariates was reversed after rank-based INT of residuals. In questionnaire-type data, when the proportion of tied observations was high, the magnitude of correlation between the original questionnaire data and covariates had a negative relationship with the degree to which normalisation re-introduced the correlation with covariates (Supplementary Figure 3.8). However, this negative relationship reversed as the proportion of ties decreased (Supplementary Figure 3.16). This means when the proportion of tied observations was low (or in continuous data), the magnitude of correlation between the original questionnaire data and covariates had a positive relationship with the degree to which normalisation re-introduced the correlation with covariates.

Third, in terms of the proportion of ties in the phenotypic variable, a decreased number of response bins in the questionnaire-type data (i.e. smaller range and more tied observations) resulted in an increased correlation between covariates and normalised residuals (Figure 3.1). However, even when there were 160 response bins, or the data were continuous, rank-based INT still re-introduced a correlation with covariates when the data had an original skew  $>0.5$  (Supplementary Figures 3.13-3.14).

As previously mentioned, although there is no linear correlation between phenotypic residuals and covariates, a rank-based correlation between the phenotypic residuals and covariates remained in almost all simulations. The factors affecting the magnitude

of rank-based correlation between phenotypic residuals and covariates are the same as those influencing the effect of rank-based INT of residuals (Supplementary Figures 3.17-3.18).



**Figure 3.1.** The relationship between the number of available responses (x-axis) and correlation between normalised residuals and covariate (y-axis) for different values of the skew in the raw phenotypic data.

*Note.* Within this figure, the correlation between the untransformed phenotypic data and covariate data is at 0.06.

3.3.1.2 - The effect of rank-based inverse normal transformation (randomly splitting ties) before regressing out covariates using simulated data

Rank-based INT of phenotypic variables, randomly splitting ties, before subsequent regression of covariates against the normalised phenotypic data, always resulted in phenotypic residuals with no linear correlation with covariates, and in the majority of simulations, skew less than 0.05.

The process of rank-based INT whilst randomly splitting ties decreased the correlation with covariates by a small amount (median change of 5%)(Supplementary Table 3.1). The extent to which the covariate correlation decreased was dependent on the original correlation between the covariate and the dependent variable, the skew of the dependent variable, and the proportion of tied responses in the dependent variable (Supplementary Figures 3.19-3.24).

The Pearson correlations between dependent variables before and after rank-based INT (randomly splitting tied observations) were between 0.77 and 1.00. An increased proportion of tied observations and increased skew led to a decreased correlation after rank-based INT (Supplementary Figure 3.25).

Regressing covariates after normalising the dependent variables introduced a smaller degree of skew when covariates had either a low skew themselves or a low correlation with the dependent variable. The degree to which regressing covariate effects introduced skew was not dependent on the proportion of tied observations. Overall, regressing covariates introduced a small amount of skew to the dependent variable (0.00 – 0.11) unless the covariate had a correlation with the dependent variable over 0.25 and a skew greater than 0.05 (Supplementary Figure 3.26). However, highly skewed covariates may introduce larger amounts of skew even when exhibiting a low correlation with the dependent variable.

### 3.3.2 - Real data

#### 3.3.2.1 - Effect of rank-based inverse normal transformation of residuals when using real data

The observed effect of applying rank-based procedures to residuals within simulated questionnaire-type data was validated using real questionnaire data from TEDS. When using the age covariate (continuous) the magnitude and direction of effect of applying rank-based procedure to residuals were similar to those of simulated questionnaire-type data (Supplementary Tables 3.2-3.3). The effect of rank-based procedures on residuals when using real questionnaire data was slightly reduced in comparison to effects observed when using simulated questionnaire-type data.

When the sex covariate (binary) was used, the magnitude, and in some cases the direction, of the effect of rank-based procedures varied from effects observed in simulated data. Although regressing the effect of a binary covariate altered the outcome of rank-based procedures, application of rank-based procedures to residuals still re-introduced a correlation with covariates (Supplementary Tables 3.4-3.5). Importantly, when a dichotomous variable was used, a large number of ties in the data still existed reducing the efficacy of rank-based INT.

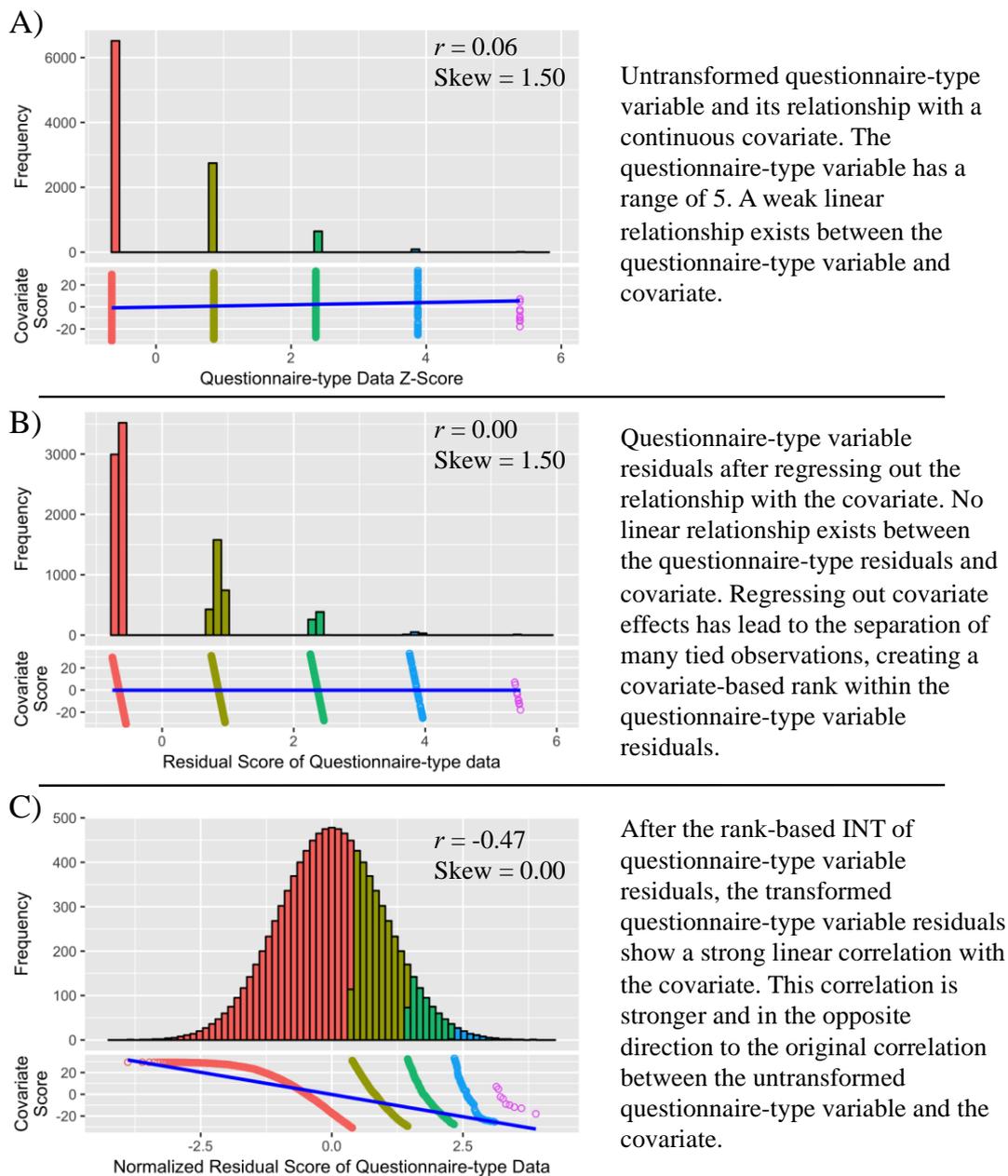
#### 3.3.2.2 - Effect of rank-based inverse normal transformation before regressing out covariates when using real data

The effect of rank-based INT (randomly splitting tied observations) before regressing out covariate effects in real questionnaire data was comparable to the effects observed when using simulated data. Rank-based INT, randomly splitting ties, and subsequent regression of covariates created residuals that were linearly uncorrelated with covariates and normally distributed (Supplementary Tables 3.6-3.7). The correlation between the dependent variable and covariate did vary slightly before and after rank-

based INT (Supplementary Tables 3.6-3.7). Contrary to the observed effects when using simulated data, the correlation between the dependent variable and the covariate did not always decrease. The Pearson correlation between raw and normalised questionnaire data varied between 0.83 and 0.99 dependent on the skew of the raw data and the number of response bins (Supplementary Table 3.8). Similar to the results of based on simulated data, the effect of regressing covariates out of the normalised variables did not re-introduce skew greater than 0.02 in any situations (Supplementary Tables 3.6-3.7).

### **3.4 – Discussion**

This study has demonstrated that regressing covariates against the dependent (phenotypic) variable and then transforming the resulting residuals to normality re-introduces a correlation between the covariates and the normalised dependent variable. This effect occurs because the process of regressing covariates against the response variable leads to a covariate-based rank in the residuals, which is then used to redistribute the data (Figure 3.2). This effect of regressing covariates against response variables occurs when the response variable is continuous (contains no tied observations) or questionnaire-type (contains tied observations), however the effect increases as the proportion of tied observations increases. The degree to which the covariate correlation is re-introduced during rank-based INT is dependent on the original skew of the response variable, although when the data contain a large proportion of tied observations, a correlation with covariates is re-introduced even when there is no skew.

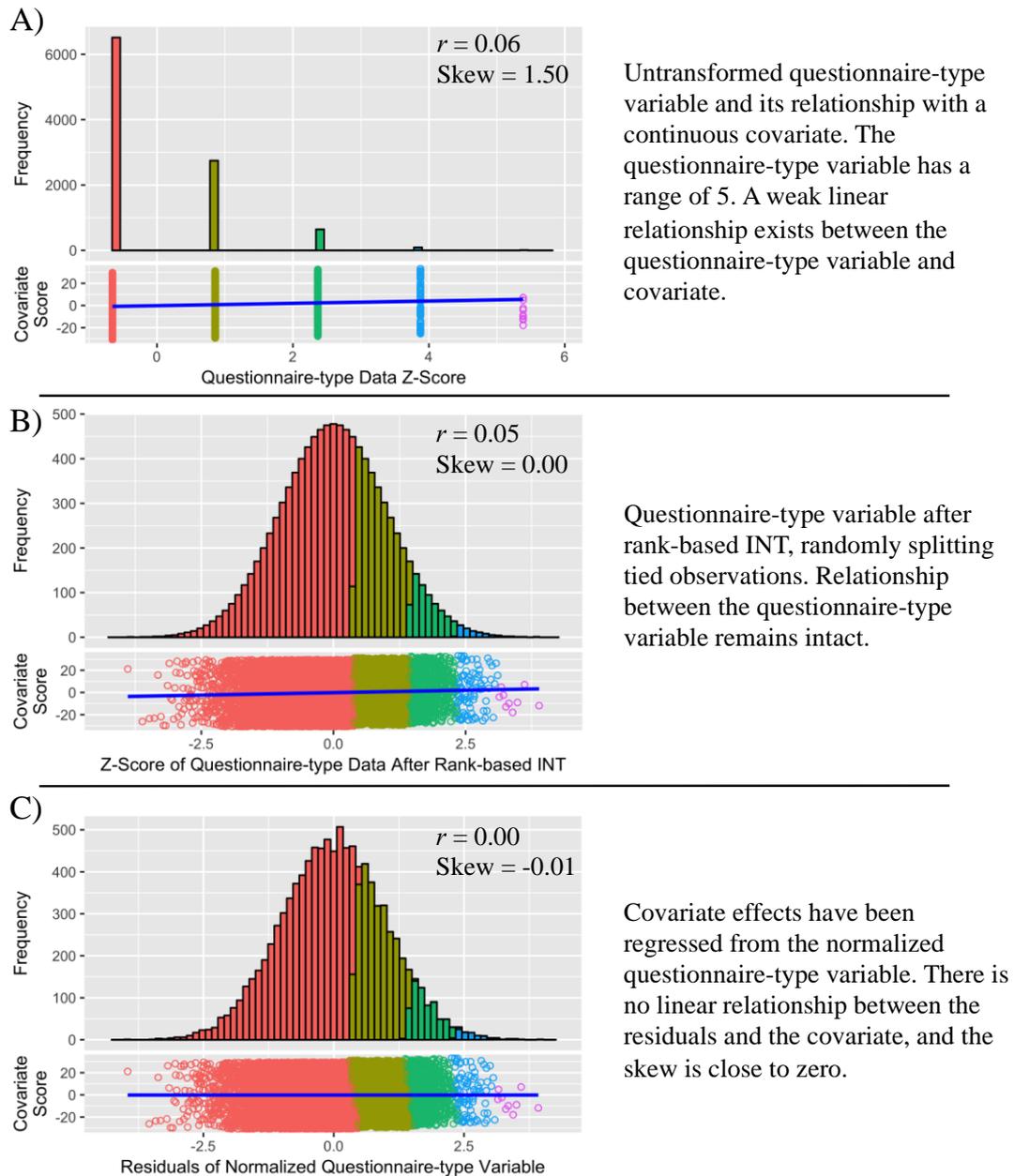


**Figure 3.2. The effect of applying a rank-based INT to residuals of questionnaire-type data, i.e. after regressing out covariates.**

*Note.* All correlations referred to in this figure are Pearson (linear) correlations.

This study has also evaluated an alternative procedure for preparing data for parametric analyses, whereby the response variable undergoes rank-based INT, randomly separating ties, before regressing out covariate effects. The findings demonstrate that this alternative approach is preferable as it creates a normally distributed response variable with no correlation with covariates (Figure 3.3). The notion of normalising the

response variables before estimating its relationship with covariates may seem counterintuitive as the process of normalisation may disrupt the true relationship between variables. Although this may be true in some scenarios, when the variables are skewed and/or contain tied observations, the change in relationship between variables due to normalisation (Supplementary Tables 3.6-3.7) is small relative to the change in relationship when normalising residuals (Supplementary Tables 3.2-3.4). In contrast, regressing covariates after normalisation will leave no correlation between the variables, meaning that any confounding by those covariates will be eliminated.



**Figure 3.3. The effect of applying a rank-based INT to questionnaire-type data before regressing out covariates.**

*Note.* All correlations referred to in this figure are Pearson (linear) correlations.

Given the importance of phenotypic transformations, authors must describe the details of this process. Many studies do not clearly describe the details in which the data is processed, however there are some studies that have clearly applied rank-based INT to residuals (S. E. Jones et al., 2016; Locke et al., 2015; Wain et al., 2015). It is not thought that the results of these studies are seriously in error as they have either dealt with

traits that have a very low skew and/or are continuous, or they have replicated their findings using binary outcomes based on untransformed data. However, it is thought that the potential problems with rank-based INT of residuals after correcting for covariates are not well known, and that researchers should be aware of these issues before applying such a procedure. It is suggested that, when possible, researchers adjust for covariates after rather than before applying a normalising transformation, or employ other methods that do not assume normality.

Although, this chapter concludes that normalisation of the dependent variable should be performed prior to the regression of covariates, regressing out covariates that are either highly skewed or highly correlated with the dependent variable, may introduce substantial skew to the residuals. However, this scenario is considered unlikely.

One of the findings of this study was that rank-based INT of residuals often leads to an increased magnitude of correlation between the dependent variable and covariate than there was between the raw variables. This indicates that rank-based INT of residuals introduces an artificial correlation between the dependent variable and covariate.

Another form of bias caused by conditioning variables on a covariate is called collider bias. A collider is a variable that is causally associated with two or more variables. If a variable is conditioned on an associated collider variable, it can introduce an artificial causal association with another variable associated with the collider variable. Real data examples of collider bias have been previously described (Cole et al., 2009).

This study has demonstrated that rank-based INT of phenotypic residuals after adjusting for covariates can lead to an overcorrection of covariate effects leading to a correlation in the opposite direction between the normalised phenotypic residuals and covariates, and in questionnaire-type data, often of a greater magnitude. This finding has implications for all rank-based procedures and highlights the importance of clearly documenting how the raw data is handled. Normalisation of phenotypic data before

regressing out covariates has been explored as an alternative procedure and has been shown to produce normally distributed phenotypic residuals that are uncorrelated with covariates. Based on these results, subsequent parametric analysis in this thesis will use phenotypic data that has undergone rank-based INT (randomly splitting ties) prior to regressing out covariate effects.

## 3.5 – Appendix

### Supplementary Note 3.1. 'SimCont' – Function to simulate continuous variables.

```
SimCont<-function(DesiredSkew, Seed=10101, NumOfSamp, StartingValue){
  set.seed(Seed)

  if(DesiredSkew >= 0){
    y<-StartingValue
    d<-1
    while(d){
      sim <- scale(rbeta(NumOfSamp, y, 10))

      if(skewness(sim) > DesiredSkew-0.0001 & skewness(sim) < DesiredSkew+0.0001) break

      if(skewness(sim) > DesiredSkew) {y<-y+0.001}
      if(skewness(sim) < DesiredSkew) {y<-y-0.001}
    }

    if(DesiredSkew < 0){
      y<-StartingValue
      d<-1
      while(d){
        sim <- scale(rbeta(NumOfSamp, 10, y))

        if(skewness(sim) > DesiredSkew-0.0001 & skewness(sim) < DesiredSkew+0.0001) break

        if(skewness(sim) > DesiredSkew) {y<-y-0.001}
        if(skewness(sim) < DesiredSkew) {y<-y+0.001}
      }

      cat('StartingValue:',StartingValue,'\n')
      cat('Skew:',skewness(sim),'\n')
      cat('N:',length(sim),'\n')
      cat('FinalValue=',y,'\n')

      return(as.numeric(scale(sim)))
    }
  }
}
```

### Supplementary Note 3.2. 'SimQuest' – Function to simulate questionnaire-type variables.

```
SimQuest<-function(DesiredSkew, NumOfResponse, Seed=10101, NumOfSamp,
StartingValue){
  set.seed(Seed)

  Breaker<-function(x,n){
    a<-NULL
    for(y in seq(1:(n-1))){
      a[y]<-(abs((max(x)-min(x)))/n)*y
    }
    return(c(min(x), min(x)+a, max(x)))
  }

  TieCreator<-function(x,y){
    n<-length(y)

    binned<- .bincode(round(x,2), round(y,2), right = TRUE, include.lowest = T)
    return(binned)
  }

  if(DesiredSkew>=0){
    y<-StartingValue
    d<-1
    while(d){
      sim <- scale(rbeta(NumOfSamp, y, 10))

      sim_breaks<-Breaker(sim,NumOfResponse)
      tied_sim<-TieCreator(sim,sim_breaks)

      if(skewness(tied_sim) > DesiredSkew-0.0001 & skewness(tied_sim) <
DesiredSkew+0.0001) break

      if(skewness(tied_sim) > DesiredSkew) {y<-y+0.001}
      if(skewness(tied_sim) < DesiredSkew) {y<-y-0.001}
    }
  }

  if(DesiredSkew<0){
    y<-StartingValue
    d<-1
    while(d){
      sim <- scale(rbeta(NumOfSamp, 10, y))

      sim_breaks<-Breaker(sim,NumOfResponse)
      tied_sim<-TieCreator(sim,sim_breaks)

      if(skewness(tied_sim) > DesiredSkew-0.0001 & skewness(tied_sim) <
DesiredSkew+0.0001) break

      if(skewness(tied_sim) > DesiredSkew) {y<-y-0.001}
      if(skewness(tied_sim) < DesiredSkew) {y<-y+0.001}
    }
  }

  skewness(tied_sim)-DesiredSkew

  cat('Number of available responses:',length(unique(tied_sim)),'\n')
  cat('Skew:',skewness(tied_sim),'\n')
  cat('N:',length(tied_sim),'\n')
  cat('FinalValue=',y,'\n')

  return(as.numeric(scale(tied_sim)))
}
```

### Supplementary Note 3.3. 'SimContNorm' – Function to simulate continuous variables with skew and kurtosis equal to zero.

```
SimContNorm<-function(NumOfSamp, Seed=10101, StartingValue=10){
  set.seed(Seed)

  d<-1
  while(d){
    sim <- scale(rnorm(NumOfSamp))

    if(skewness(sim) > -0.0001 & skewness(sim) < 0.0001 & kurtosis(sim) < 0.01 &
      kurtosis(sim) > -0.01) break

    cat('Skew:',skewness(sim),'\n')
    cat('Kurtosis:',skewness(sim),'\n')
  }
  cat('Skew:',skewness(sim),'\n')
  cat('Kurtosis:',skewness(sim),'\n')
  cat('N:',length(sim),'\n')

  return(as.numeric(scale(sim)))
}
```

### Supplementary Note 3.4. 'SimQuestNorm' – Function to simulate questionnaire-type variables with skew and kurtosis equal to zero.

```
SimQuestNorm<-function(NumOfResponse, Seed=10101, NumOfSamp){
  set.seed(Seed)

  Breaker<-function(x,n){
    a<-NULL
    for(y in seq(1:(n-1))){
      a[y]<-(abs((max(x)-min(x)))/n)*y
    }
    return(c(min(x), min(x)+a, max(x)))
  }

  TieCreator<-function(x,y){
    n<-length(y)

    binned<- .bincode(round(x,2), round(y,2), right = TRUE, include.lowest = T)

    return(binned)
  }

  d<-1
  while(d){
    sim <- scale(rnorm(NumOfSamp))

    sim_breaks<-Breaker(sim,NumOfResponse)
    tied_sim<-TieCreator(sim,sim_breaks)

    if(skewness(tied_sim) > -0.0001 & skewness(tied_sim) < 0.0001 & kurtosis(tied_sim)
    < 0.01 & kurtosis(tied_sim) > -0.01) break

  }

  cat('Number of available responses:',length(unique(tied_sim)),'\n')
  cat('Skew:',skewness(tied_sim),'\n')
  cat('Kurtosis:',kurtosis(tied_sim),'\n')
  cat('N:',length(tied_sim),'\n')

  return(as.numeric(scale(tied_sim)))
}
```

### Supplementary Note 3.5. 'CovarCreator' – Function to create correlated covariates for continuous and questionnaire-type variables.

```
CovarCreator<-function(x,dis.cor,dir,start=1){
y<-start
while(TRUE){
  cov<-jitter(x, factor = y, amount = NULL)
  j<-cor(cov,x,use='complete.obs')
  if(j < dis.cor-0.0001) {y<-y-1}
  if(j > dis.cor+0.0001) {y<-y+1}
  if(j <= dis.cor+0.0001 & j >= dis.cor-0.0001) break()
}

print(y)

if(dir == 'neg') {cov<--cov}

j<-cor(cov,x,use='complete.obs')
print(j)

return(cov)
}
```

### Supplementary Note 3.6. 'rntransform\_random' – Function to perform rank-based INT whilst randomly splitting tied observations.

```
rntransform_random<-function (formula, data, family = gaussian)
{
  if (is(try(formula, silent = TRUE), "try-error")) {
    if (is(data, "gwa.data"))
      data1 <- phdata(data)
    else if (is(data, "data.frame"))
      data1 <- data
    else stop("'data' must have 'gwa.data' or 'data.frame' class")
    formula <- data1[[as(match.call()[["formula"]], "character")]]
  }
  var <- ztransform(formula, data, family)
  out <- rank(var, ties.method='random') - 0.5
  out[is.na(var)] <- NA
  mP <- 0.5/max(out, na.rm = T)
  out <- out/(max(out, na.rm = T) + 0.5)
  out <- qnorm(out)
  out
}
```

**Supplementary Table 3.1. Difference in covariate correlation with the dependent variable before and after rank-based INT when splitting tied observations randomly. This is based on simulated data.**

<b>Original Covariate Correlation</b>	<b>Mean covariate correlation after normalisation</b>	<b>Covariate correlation % difference after normalisation</b>
0.5	0.476	4.88%
0.25	0.239	4.61%
0.12	0.114	5.07%
0.06	0.057	5.04%
0.03	0.029	7.27%
0.01	0.010	21.30%

Supplementary Table 3.2. Outcome of rank-based INT after regressing effect of a continuous covariate (age) from real questionnaire data.

Phenotype	Range	Original skew	Pearson correlation between phenotype and covariate	Pearson correlation between phenotype residuals and covariate	Pearson correlation between normalised phenotype residuals and covariate	Final skew
<b>Paranoia</b>	5	1.357	0.055	$-1.55 \times 10^{-15}$	-0.275	$8.91 \times 10^{-6}$
<b>Paranoia</b>	10	1.195	0.043	$-1.89 \times 10^{-15}$	-0.140	$5.40 \times 10^{-6}$
<b>Paranoia</b>	20	1.095	0.030	$-1.59 \times 10^{-15}$	-0.079	$1.59 \times 10^{-5}$
<b>Paranoia</b>	40	1.296	0.022	$-1.42 \times 10^{-15}$	-0.045	$1.47 \times 10^{-5}$
<b>Anhedonia</b>	5	1.868	$-5.73 \times 10^{-3}$	$5.81 \times 10^{-16}$	0.462	$8.91 \times 10^{-6}$
<b>Anhedonia</b>	10	0.858	-0.025	$1.07 \times 10^{-15}$	0.172	$5.40 \times 10^{-6}$
<b>Anhedonia</b>	20	0.651	-0.020	$1.06 \times 10^{-15}$	0.081	$1.59 \times 10^{-5}$
<b>Anhedonia</b>	40	0.537	-0.013	$7.74 \times 10^{-16}$	0.037	$1.47 \times 10^{-5}$

**Supplementary Table 3.3. Outcome of regressing effect of a continuous covariate (age) from real questionnaire data on Spearman correlation.**

<b>Phenotype</b>	<b>Range</b>	<b>Original skew</b>	<b>Spearman correlation between phenotype and covariate</b>	<b>Spearman correlation between phenotype residuals and covariate</b>
<b>Paranoia</b>	5	1.357	0.063	-0.266
<b>Paranoia</b>	10	1.195	0.049	-0.123
<b>Paranoia</b>	20	1.095	0.031	-0.061
<b>Paranoia</b>	40	1.296	0.027	-0.025
<b>Anhedonia</b>	5	1.868	-1.01x10 <sup>-3</sup>	0.436
<b>Anhedonia</b>	10	0.858	-0.027	0.146
<b>Anhedonia</b>	20	0.651	-0.027	0.062
<b>Anhedonia</b>	40	0.537	-0.016	0.031

Supplementary Table 3.4. Outcome of rank-based INT after regressing effect of a dichotomous covariate (sex) from real questionnaire data.

Phenotype	Range	Original skew	Pearson correlation between phenotype and covariate	Pearson correlation between phenotype residuals and covariate	Pearson correlation between normalised phenotype residuals and covariate	Final skew
<b>Paranoia</b>	5	1.357	0.018	$-1.00 \times 10^{-16}$	-0.264	0.624
<b>Paranoia</b>	10	1.195	-0.026	$-8.74 \times 10^{-17}$	0.125	0.209
<b>Paranoia</b>	20	1.095	-0.022	$-2.93 \times 10^{-16}$	0.065	0.097
<b>Paranoia</b>	40	1.296	-0.059	$-5.60 \times 10^{-16}$	$-7.96 \times 10^{-3}$	0.077
<b>Anhedonia</b>	5	1.868	0.177	$4.78 \times 10^{-16}$	-0.212	0.624
<b>Anhedonia</b>	10	0.858	0.127	$4.07 \times 10^{-16}$	-0.040	0.209
<b>Anhedonia</b>	20	0.651	0.135	$4.09 \times 10^{-16}$	0.049	0.097
<b>Anhedonia</b>	40	0.537	0.205	$-1.77 \times 10^{-16}$	$-5.51 \times 10^{-3}$	0.077

**Supplementary Table 3.5. Outcome of regressing effect of a dichotomous covariate (sex) from real questionnaire data on Spearman correlation.**

Phenotype	Range	Original skew	Spearman correlation between phenotype and covariate	Spearman correlation between phenotype residuals and covariate
Paranoia	5	1.357	0.018	-0.269
Paranoia	10	1.195	-0.031	0.118
Paranoia	20	1.095	-0.026	0.054
Paranoia	40	1.296	-0.066	-0.020
Anhedonia	5	1.868	0.200	-0.187
Anhedonia	10	0.858	0.124	-0.025
Anhedonia	20	0.651	0.140	0.065
Anhedonia	40	0.537	0.211	$9.05 \times 10^{-3}$

**Supplementary Table 3.6. Effect of rank-based INT (randomly ranking tied observations) on the relationship between real questionnaire variable and continuous covariate (age). This table also shows to what extent regressing covariate effects reintroduces skew.**

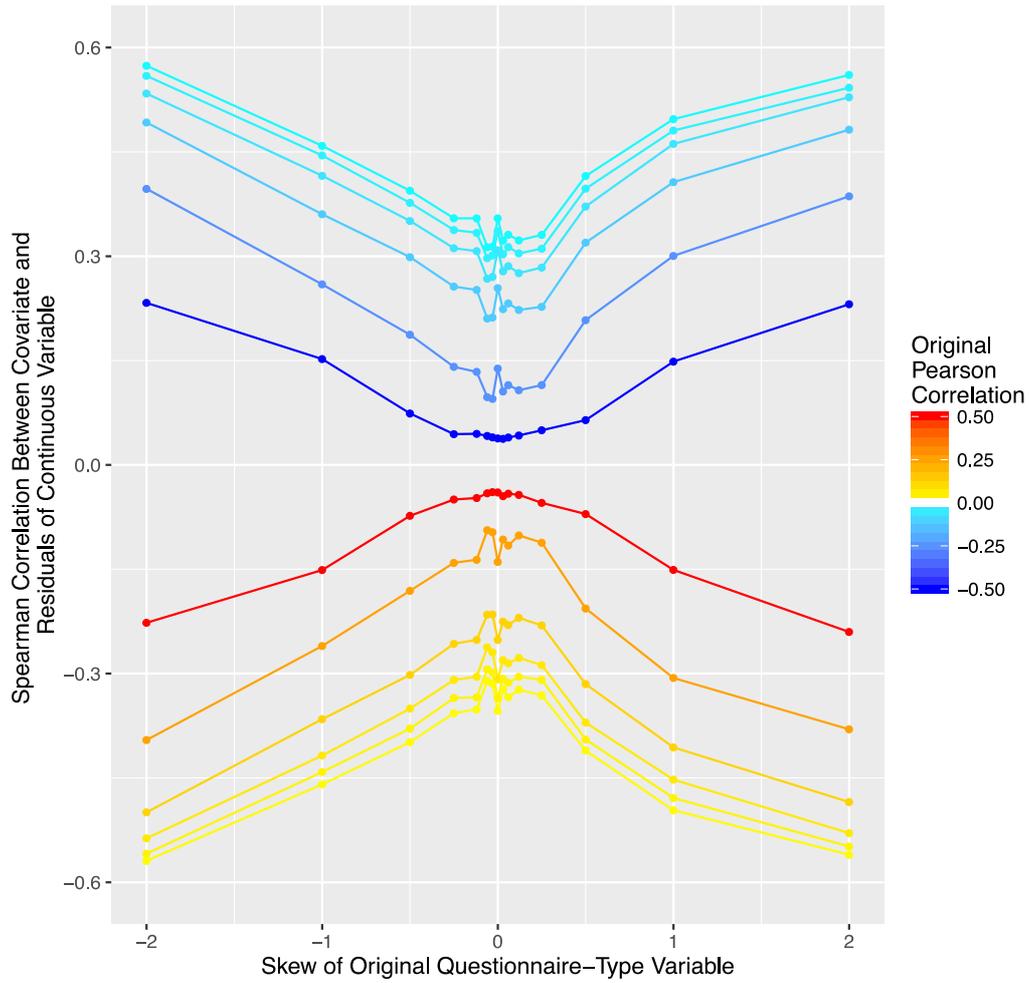
<b>Phenotype</b>	<b>Range</b>	<b>Original Skew of Dependent Variable</b>	<b>Original Correlation between Dependent Variable and Age</b>	<b>Correlation between Normalised Dependant Variable and Age</b>	<b>Correlation between Residuals of Normalised Dependant Variable and Age</b>	<b>Skew of Residuals of Normalised Dependant</b>
<b>Paranoia</b>	5	1.357	0.055	0.054	$-2.52 \times 10^{-15}$	$-3.87 \times 10^{-3}$
<b>Paranoia</b>	10	1.195	0.043	0.043	$-2.99 \times 10^{-15}$	$-2.56 \times 10^{-3}$
<b>Paranoia</b>	20	1.095	0.030	0.029	$-1.96 \times 10^{-15}$	$-1.67 \times 10^{-3}$
<b>Paranoia</b>	40	1.296	0.022	0.023	$-1.66 \times 10^{-15}$	$-3.67 \times 10^{-4}$
<b>Anhedonia</b>	5	1.868	-0.006	-0.013	$1.01 \times 10^{-15}$	$5.34 \times 10^{-4}$
<b>Anhedonia</b>	10	0.858	-0.025	-0.028	$1.56 \times 10^{-15}$	$3.27 \times 10^{-4}$
<b>Anhedonia</b>	20	0.651	-0.020	-0.024	$1.24 \times 10^{-15}$	$1.67 \times 10^{-3}$
<b>Anhedonia</b>	40	0.537	-0.013	-0.014	$8.05 \times 10^{-16}$	$1.02 \times 10^{-4}$

**Supplementary Table 3.7. Effect of rank-based INT (randomly ranking tied observations) on the relationship between real questionnaire variable and binary covariate (sex). This table also shows to what extent regressing covariate effects reintroduces skew.**

Phenotype	Range	Original Skew of Dependent Variable	Original Correlation between Dependent Variable and Sex	Correlation between Normalised Dependant Variable and Sex	Correlation between Residuals of Normalised Dependant Variable and Sex	Skew of Residuals of Normalised Dependant
Paranoia	5	1.357	0.018	0.012	$7.14 \times 10^{-16}$	$-2.81 \times 10^{-4}$
Paranoia	10	1.195	-0.026	-0.033	$5.17 \times 10^{-16}$	$1.54 \times 10^{-3}$
Paranoia	20	1.095	-0.022	-0.026	$5.96 \times 10^{-17}$	$1.36 \times 10^{-3}$
Paranoia	40	1.296	-0.059	-0.065	$-4.01 \times 10^{-16}$	$1.18 \times 10^{-3}$
Anhedonia	5	1.868	0.177	0.175	$4.80 \times 10^{-16}$	-0.017
Anhedonia	10	0.858	0.127	0.127	$5.92 \times 10^{-16}$	-0.010
Anhedonia	20	0.651	0.135	0.136	$4.94 \times 10^{-16}$	$-1.35 \times 10^{-3}$
Anhedonia	40	0.537	0.205	0.206	$-7.20 \times 10^{-17}$	$-3.32 \times 10^{-3}$

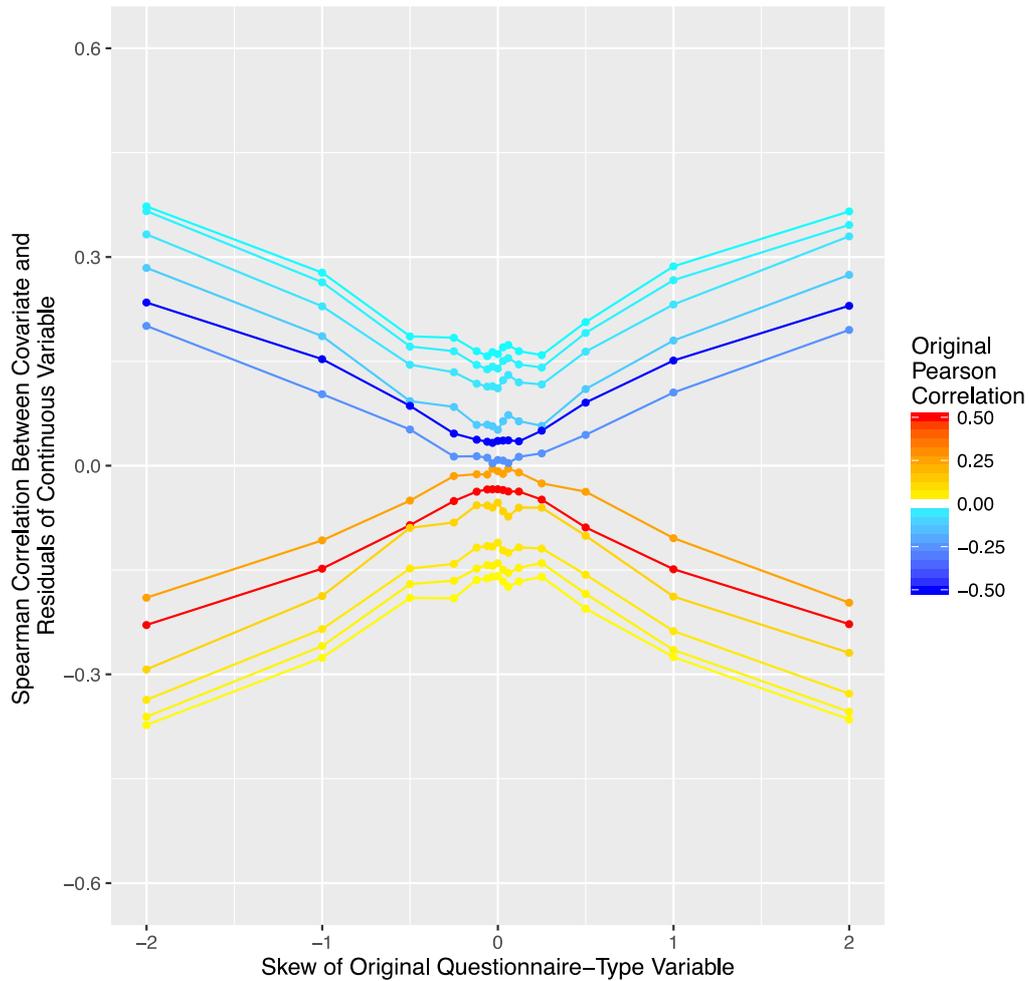
**Supplementary Table 3.8. Effect of rank-based INT of real questionnaire variable when randomly ranking tied observations. Shows Pearson correlation between dependent variable before after normalisation.**

<b>Phenotype</b>	<b>Range</b>	<b>Original Skew</b>	<b>Correlation after rank-based INT</b>
<b>Paranoia</b>	5	1.357	0.889
<b>Paranoia</b>	10	1.195	0.939
<b>Paranoia</b>	20	1.095	0.957
<b>Paranoia</b>	40	1.296	0.949
<b>Anhedonia</b>	5	1.868	0.833
<b>Anhedonia</b>	10	0.858	0.958
<b>Anhedonia</b>	20	0.651	0.980
<b>Anhedonia</b>	40	0.537	0.990



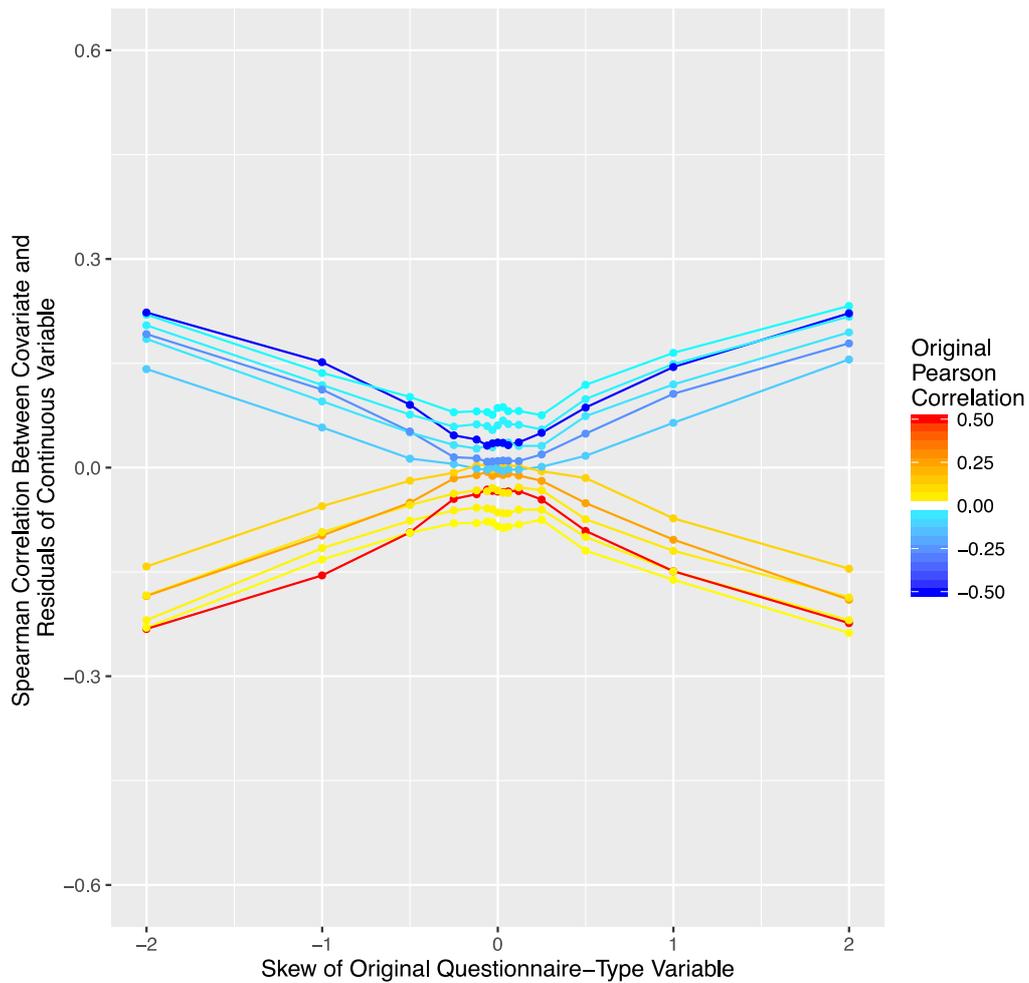
**Supplementary Figure 3.1. Effect of regressing out covariate effects from questionnaire-type data with a range of 5 before rank-based non-parametric analyses.**

*Note.* X-axis shows the skew of the original phenotypic data. Y-axis shows the Spearman rank-based correlation between residuals and covariates. Colours indicate the original Pearson correlation between the questionnaire-type data and covariate.



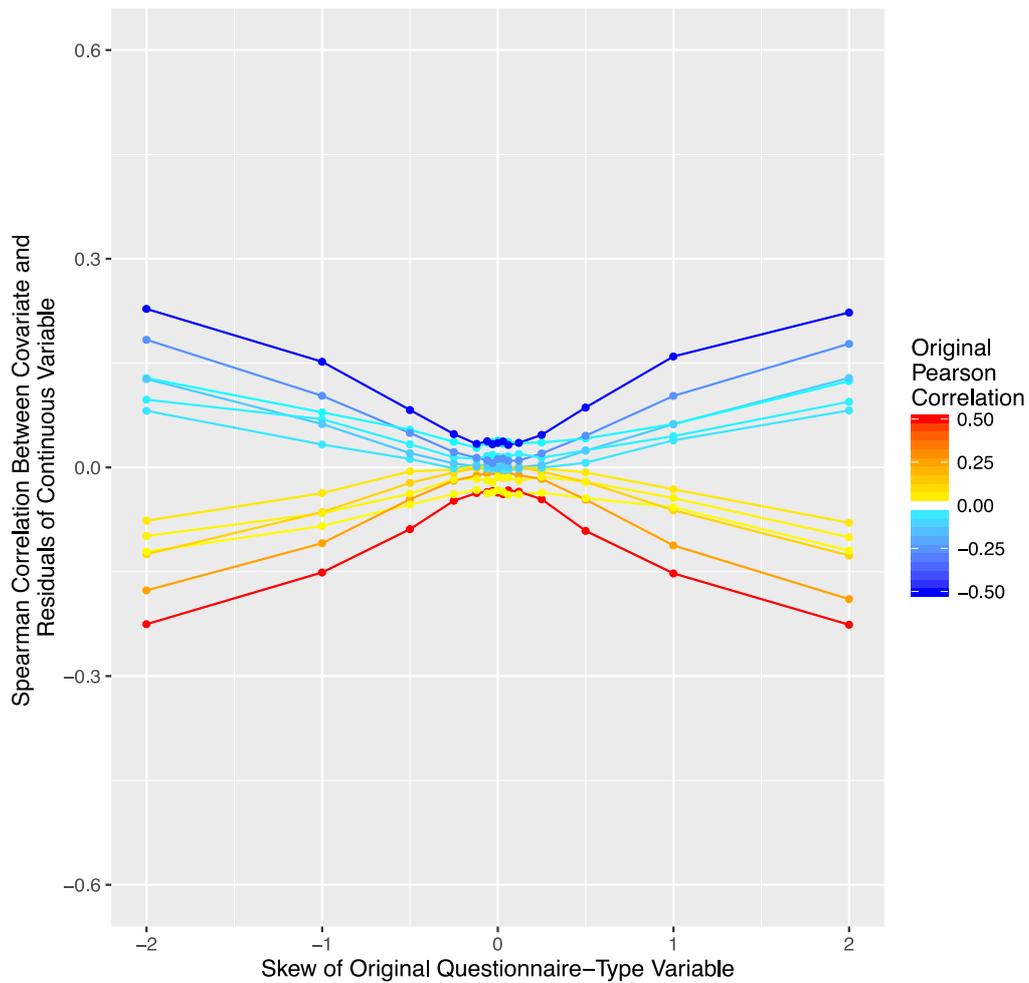
**Supplementary Figure 3.2. Effect of regressing out covariate effects from questionnaire-type data with a range of 10 before rank-based non-parametric analyses.**

*Note.* X-axis shows the skew of the original phenotypic data. Y-axis shows the Spearman rank-based correlation between residuals and covariates. Colours indicate the original Pearson correlation between the questionnaire-type data and covariate.



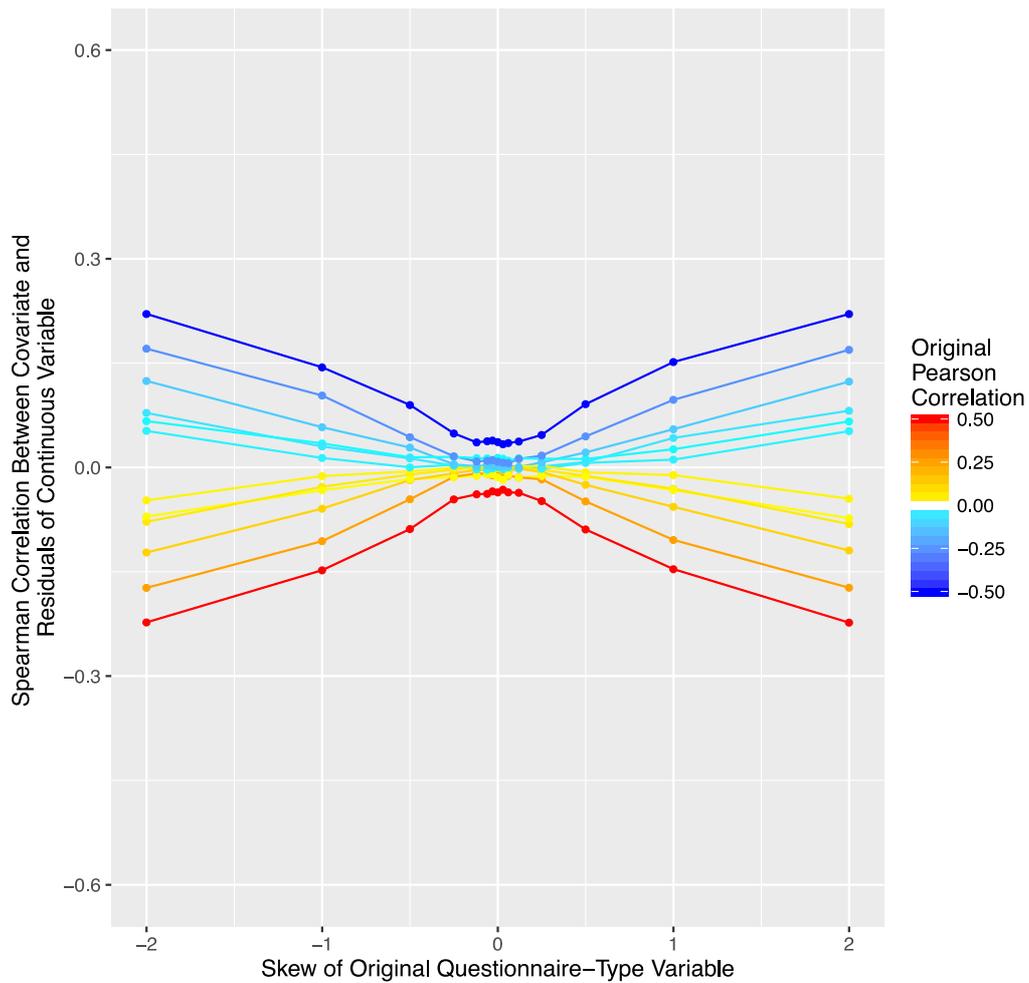
**Supplementary Figure 3.3. Effect of regressing out covariate effects from questionnaire-type data with a range of 20 before rank-based non-parametric analyses.**

*Note.* X-axis shows the skew of the original phenotypic data. Y-axis shows the Spearman rank-based correlation between residuals and covariates. Colours indicate the original Pearson correlation between the questionnaire-type data and covariate.



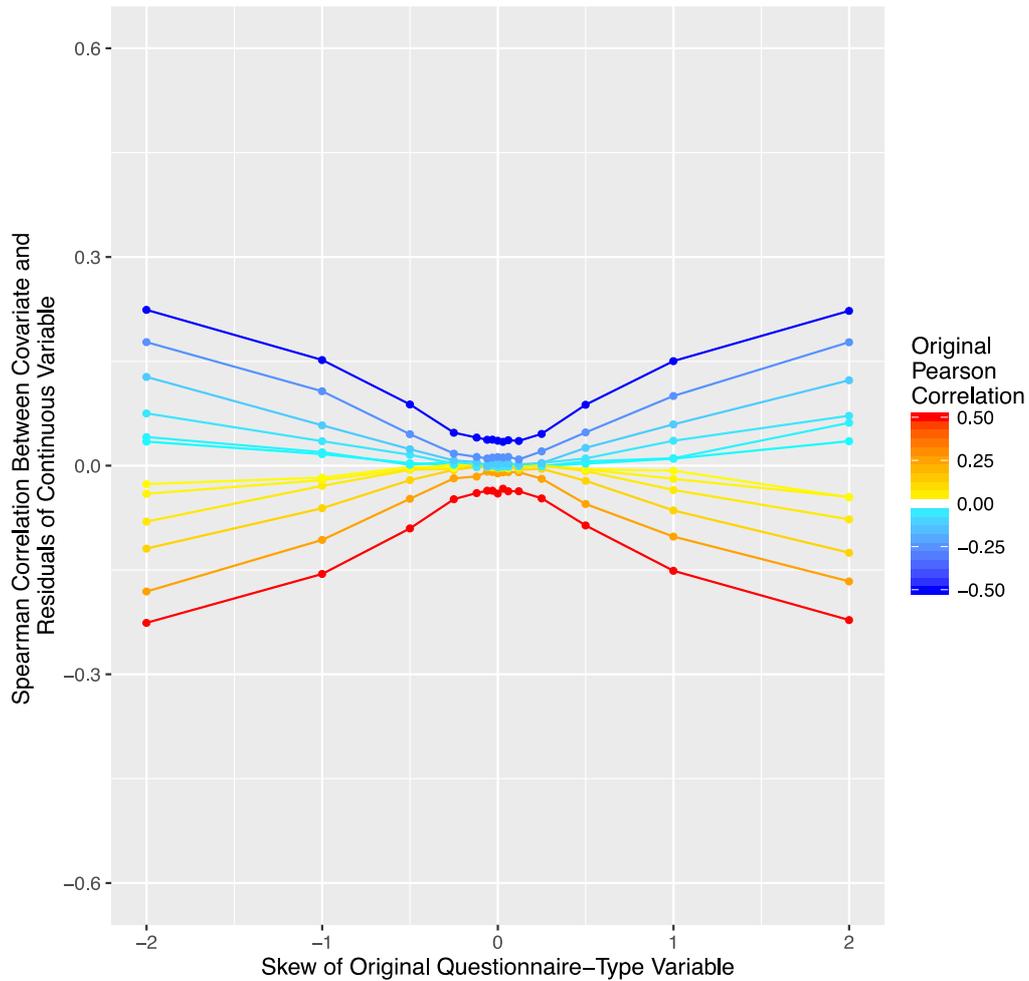
**Supplementary Figure 3.4. Effect of regressing out covariate effects from questionnaire-type data with a range of 40 before rank-based non-parametric analyses.**

*Note.* X-axis shows the skew of the original phenotypic data. Y-axis shows the Spearman rank-based correlation between residuals and covariates. Colours indicate the original Pearson correlation between the questionnaire-type data and covariate.



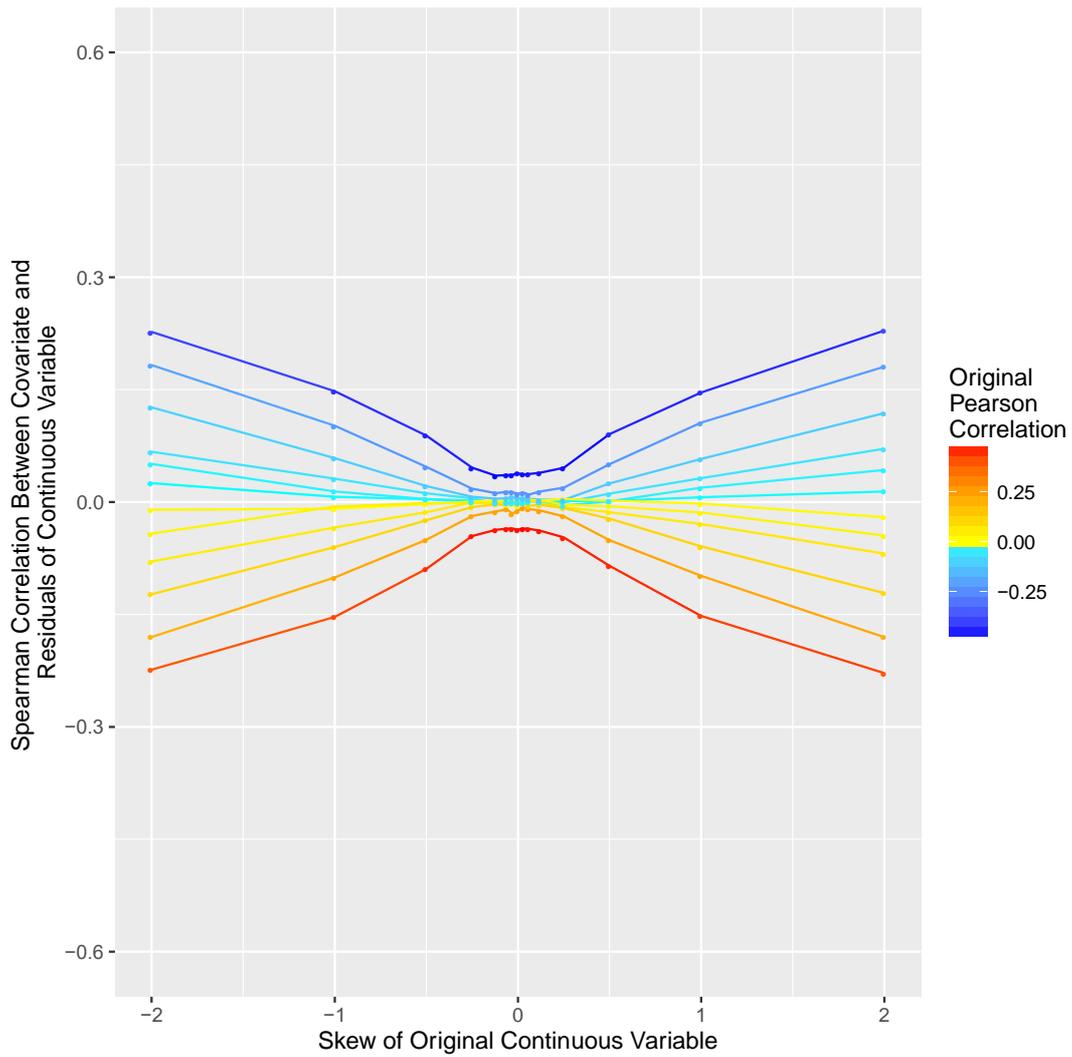
**Supplementary Figure 3.5. Effect of regressing out covariate effects from questionnaire-type data with a range of 80 before rank-based non-parametric analyses.**

*Note.* X-axis shows the skew of the original phenotypic data. Y-axis shows the Spearman rank-based correlation between residuals and covariates. Colours indicate the original Pearson correlation between the questionnaire-type data and covariate.



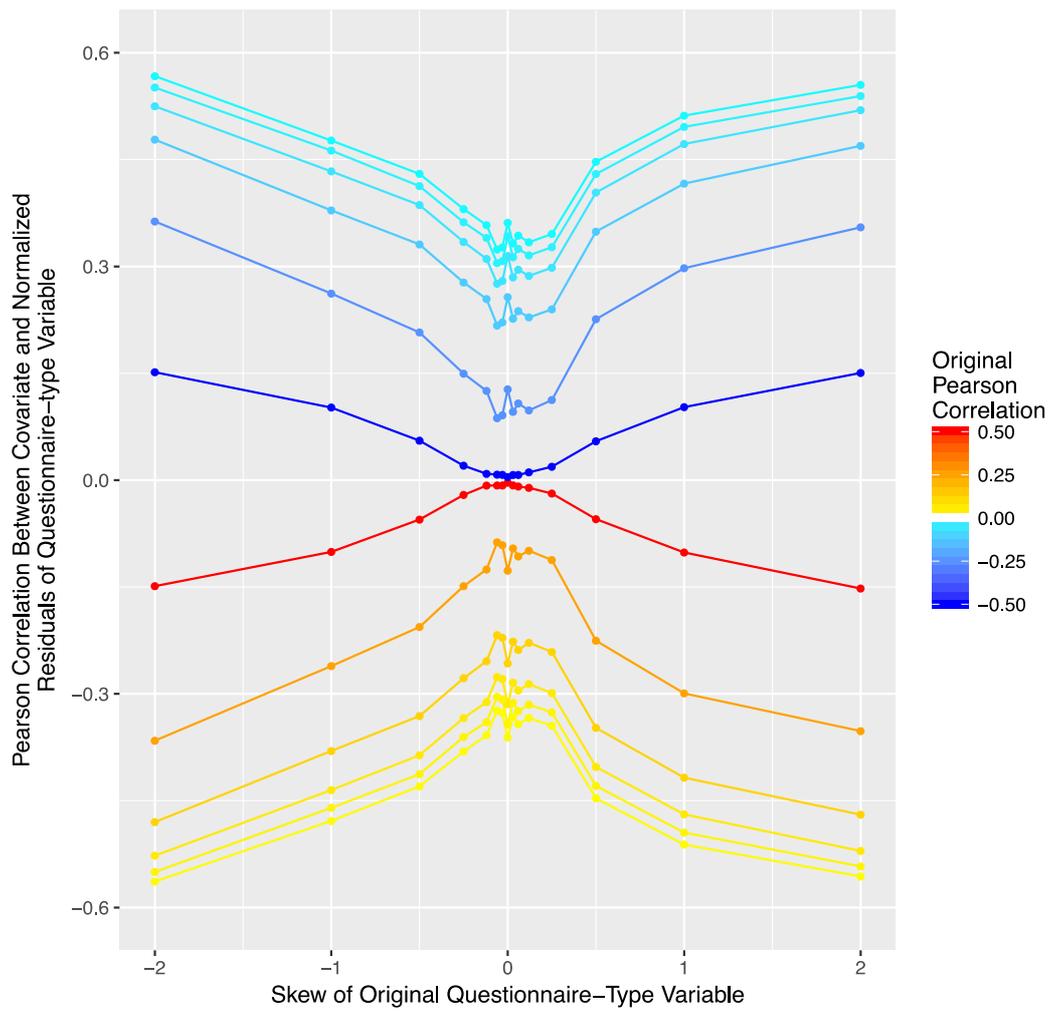
**Supplementary Figure 3.6. Effect of regressing out covariate effects from questionnaire-type data with a range of 160 before rank-based non-parametric analyses.**

*Note.* X-axis shows the skew of the original phenotypic data. Y-axis shows the Spearman rank-based correlation between residuals and covariates. Colours indicate the original Pearson correlation between the questionnaire-type data and covariate.



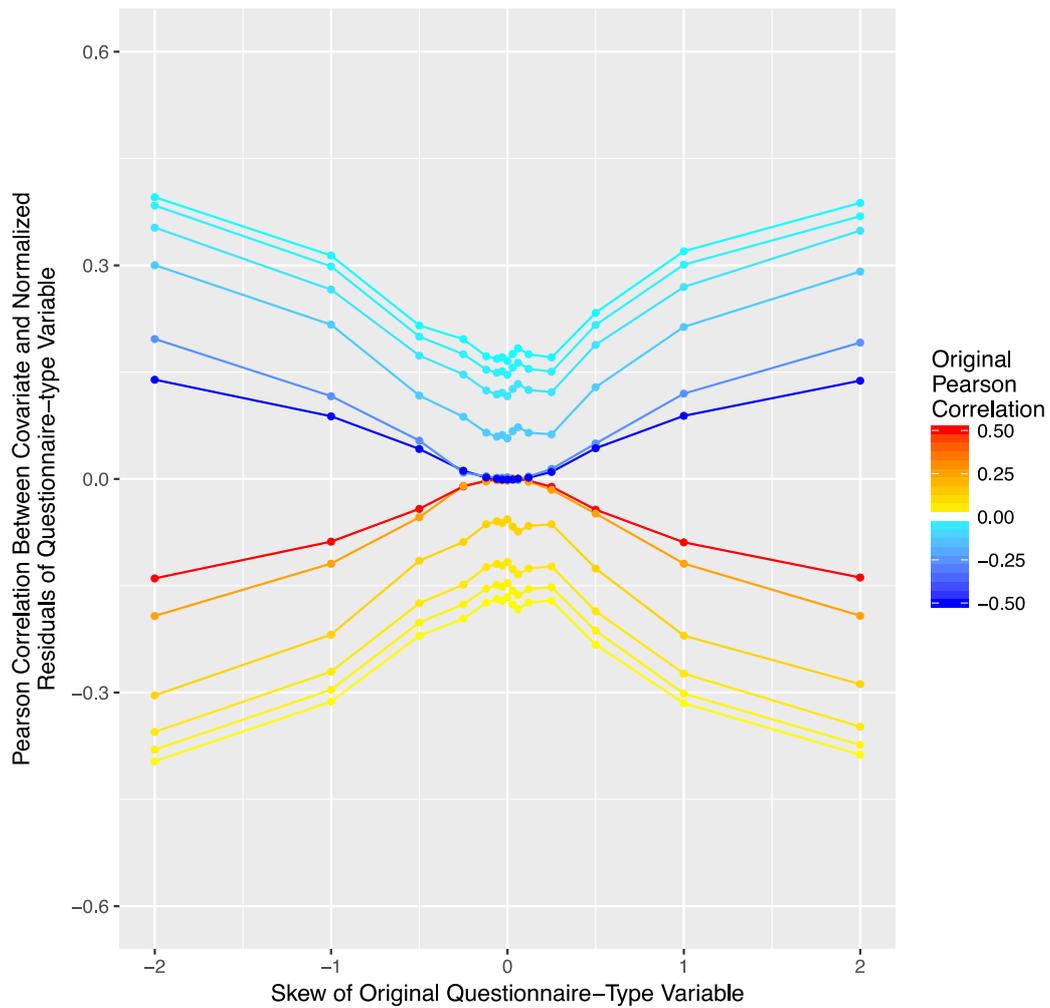
**Supplementary Figure 3.7. Effect of regressing out covariate effects from continuous data before rank-based non-parametric analyses.**

*Note.* X-axis shows the skew of the original phenotypic data. Y-axis shows the Spearman rank-based correlation between residuals and covariates. Colours indicate the original Pearson correlation between the continuous data and covariate.



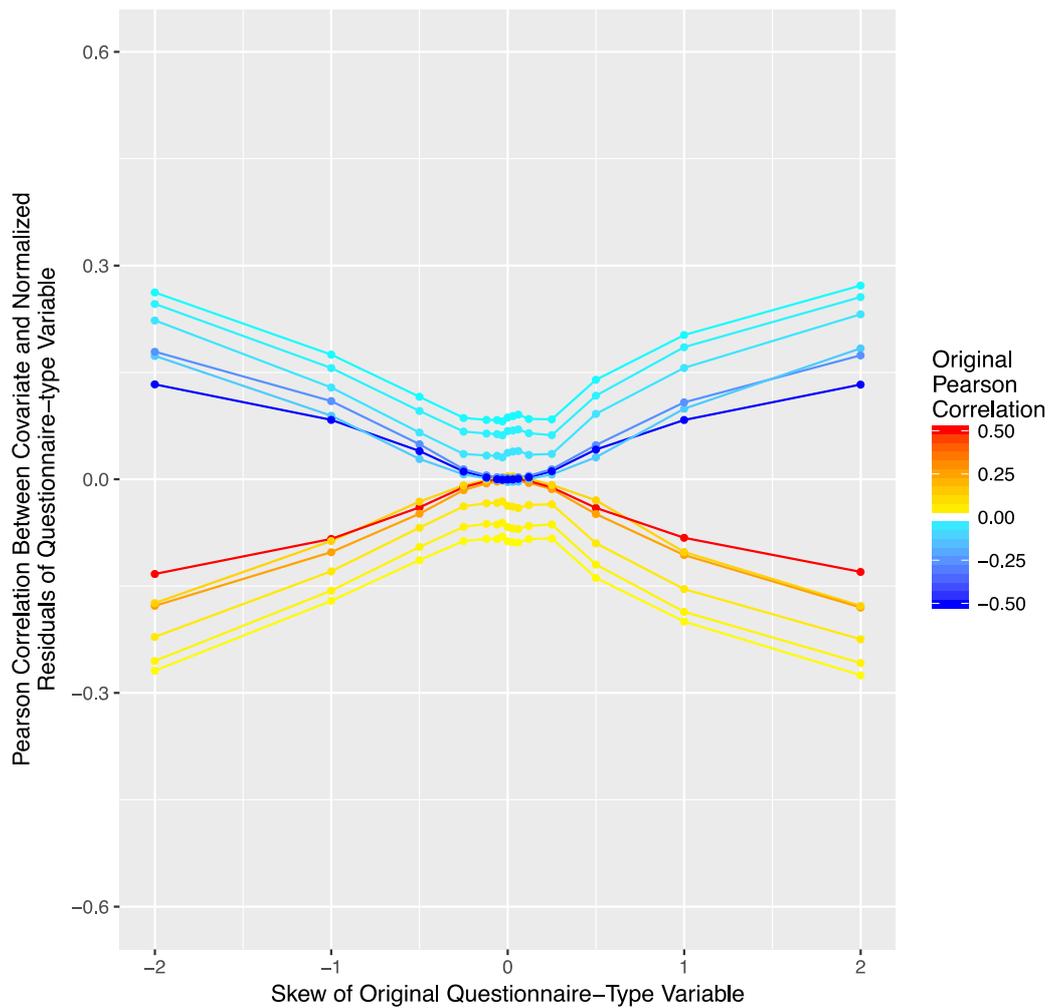
**Supplementary Figure 3.8. Effect of rank-based INT of questionnaire-type data residuals (after regressing out covariates) with range of 5.**

*Note.* The relationship between the original skew of the raw questionnaire-type data (x-axis), the original correlation between the raw questionnaire-type data and covariate data (colour coded), and the correlation between normalised questionnaire-type residuals (y-axis). This figure is based on simulated data.



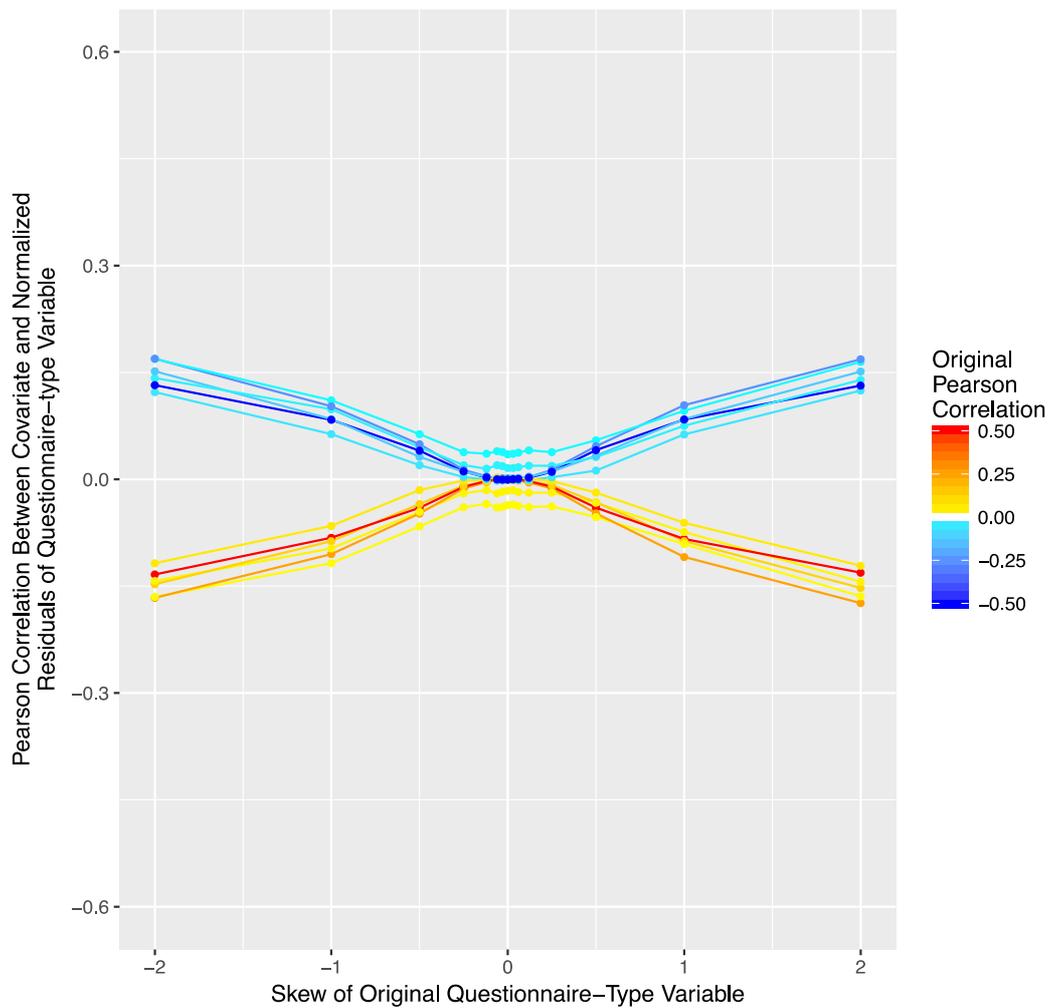
**Supplementary Figure 3.9. Effect of rank-based INT of questionnaire-type data residuals (after regressing out covariates) with range of 10.**

*Note.* The relationship between the original skew of the raw questionnaire-type data (x-axis), the original correlation between the raw questionnaire-type data and covariate data (colour coded), and the correlation between normalised questionnaire-type residuals (y-axis). This figure is based on simulated data.



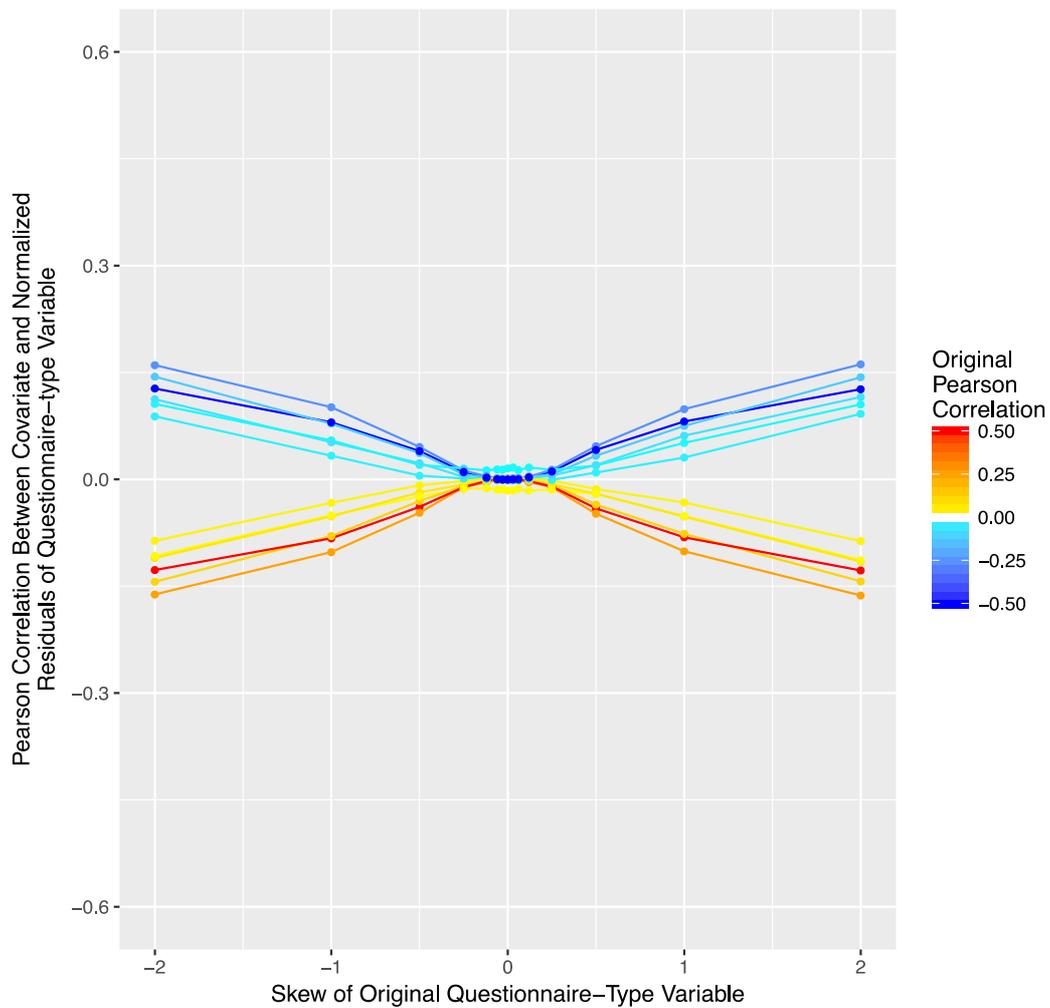
**Supplementary Figure 3.10. Effect of rank-based INT of questionnaire-type data residuals (after regressing out covariates) with range of 20.**

*Note.* The relationship between the original skew of the raw questionnaire-type data (x-axis), the original correlation between the raw questionnaire-type data and covariate data (colour coded), and the correlation between normalised questionnaire-type residuals (y-axis). This figure is based on simulated data.



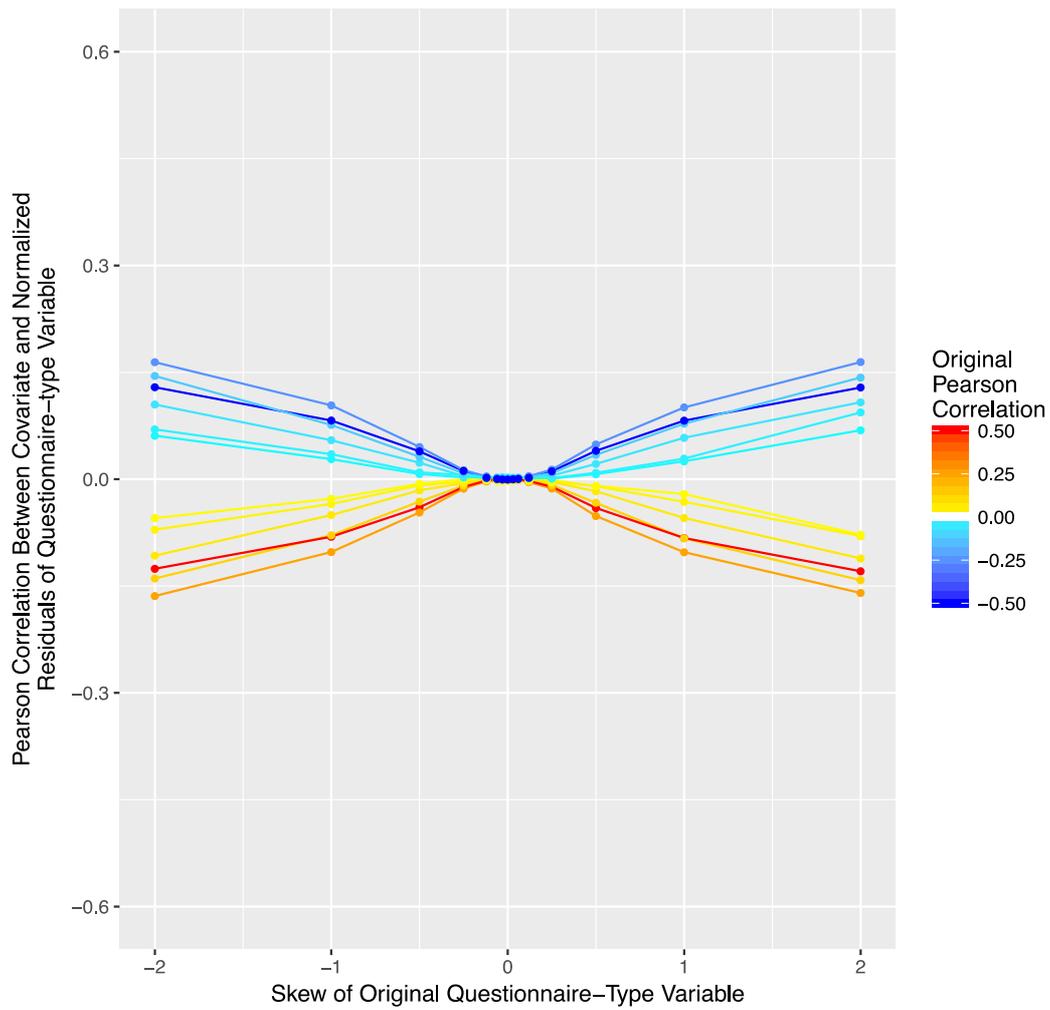
**Supplementary Figure 3.11. Effect of rank-based INT of questionnaire-type data residuals (after regressing out covariates) with range of 40.**

*Note.* The relationship between the original skew of the raw questionnaire-type data (x-axis), the original correlation between the raw questionnaire-type data and covariate data (colour coded), and the correlation between normalised questionnaire-type residuals (y-axis). This figure is based on simulated data.



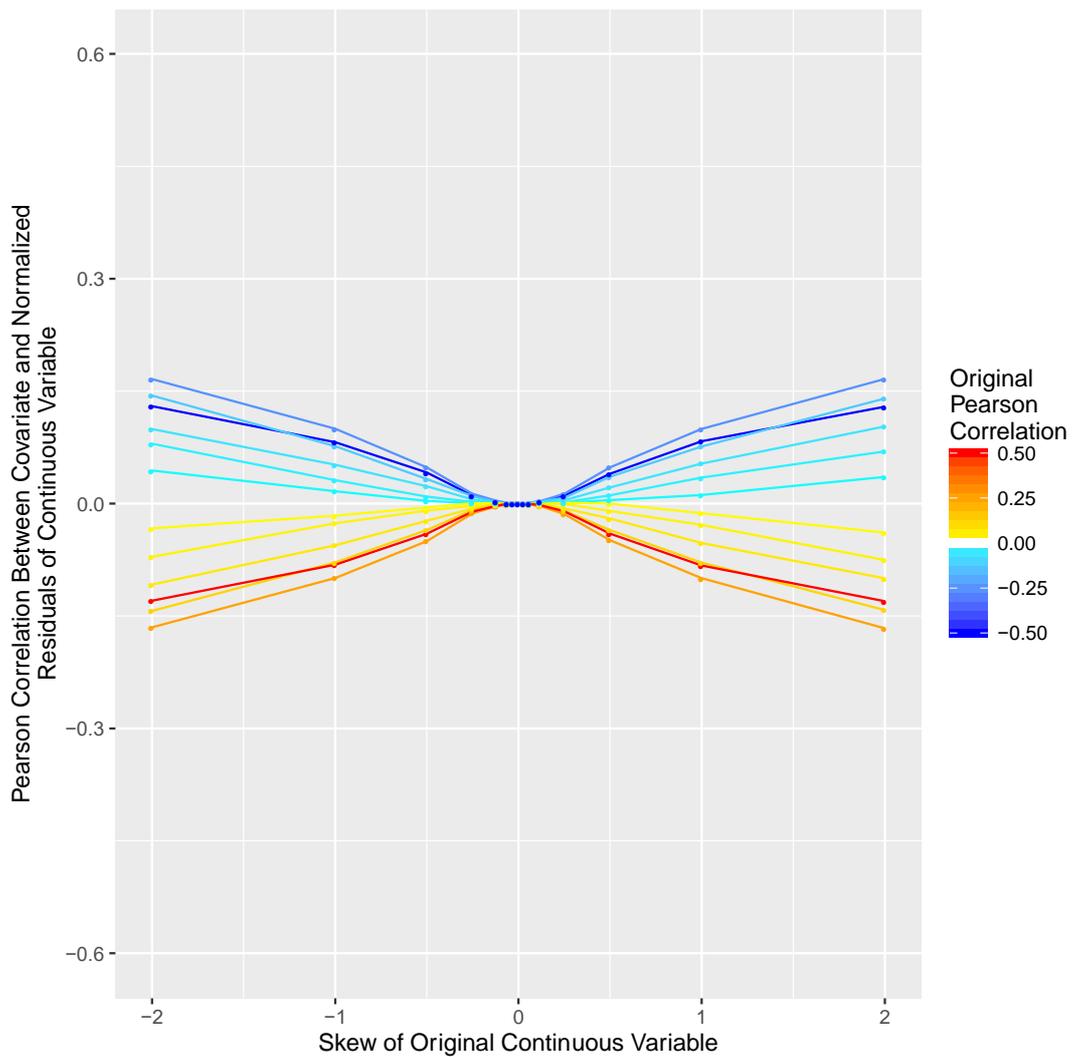
**Supplementary Figure 3.12. Effect of rank-based INT of questionnaire-type data residuals (after regressing out covariates) with range of 80.**

*Note.* The relationship between the original skew of the raw questionnaire-type data (x-axis), the original correlation between the raw questionnaire-type data and covariate data (colour coded), and the correlation between normalised questionnaire-type residuals (y-axis). This figure is based on simulated data.



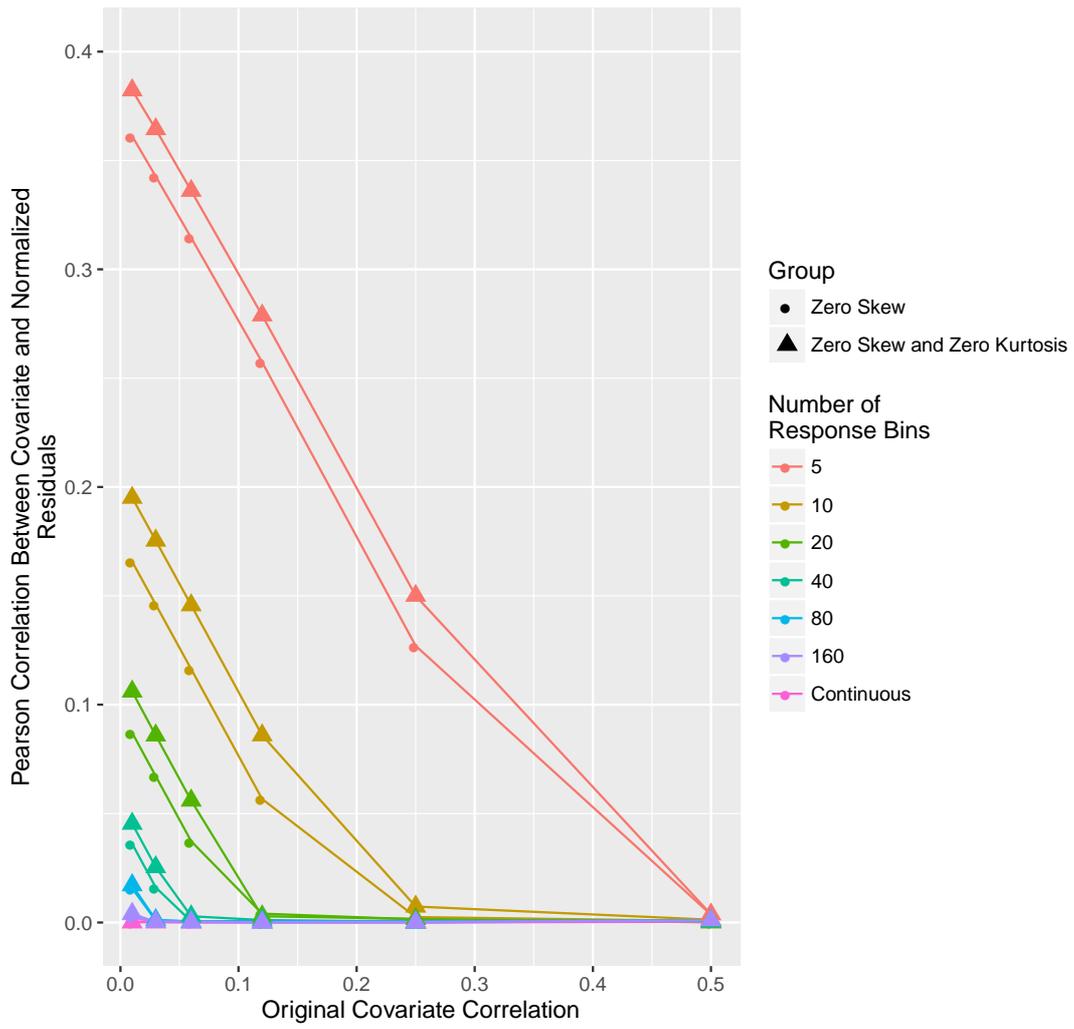
**Supplementary Figure 3.13. Effect of rank-based INT of questionnaire-type data residuals (after regressing out covariates) with range of 160.**

*Note.* The relationship between the original skew of the raw questionnaire-type data (x-axis), the original correlation between the raw questionnaire-type data and covariate data (colour coded), and the correlation between normalised questionnaire-type residuals (y-axis). This figure is based on simulated data.

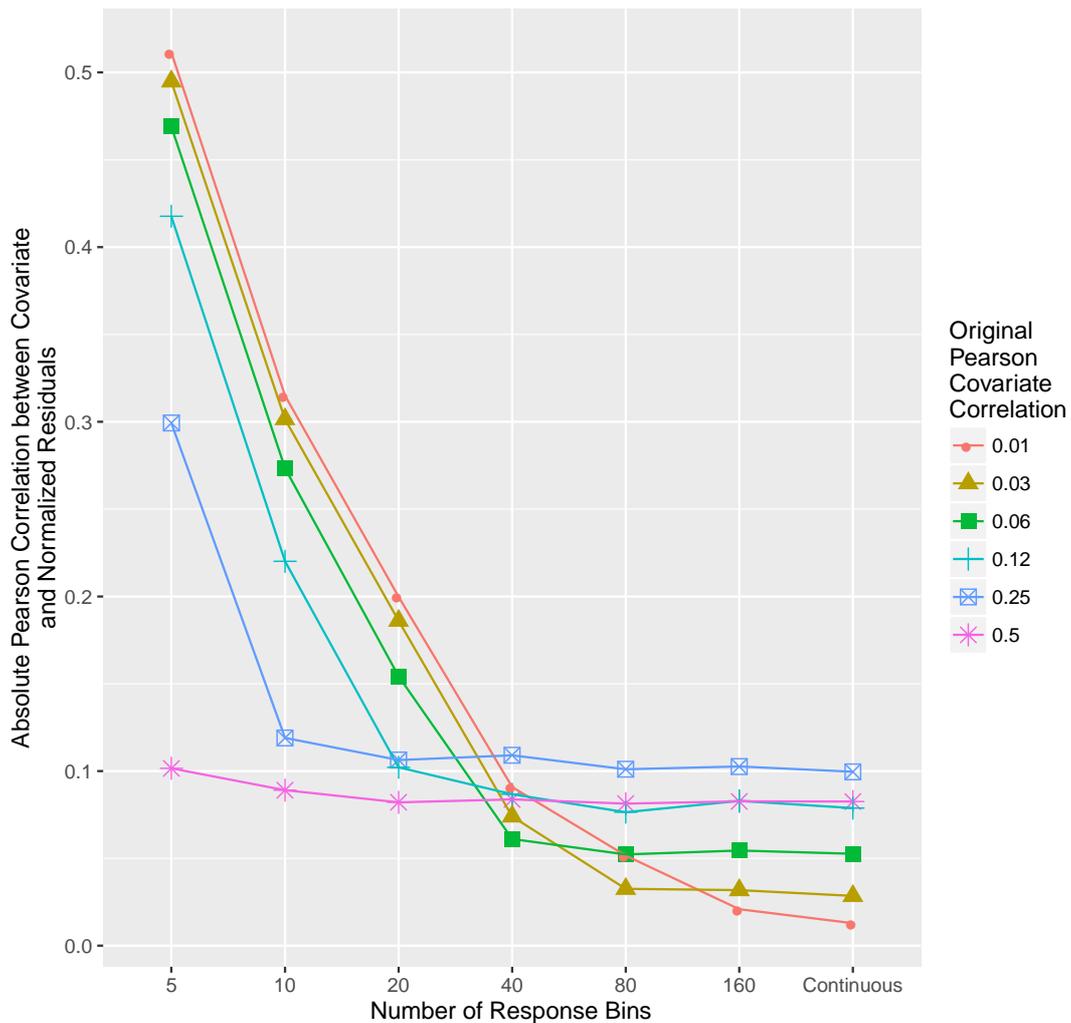


**Supplementary Figure 3.14. Effect of rank-based INT of continuous data residuals (after regressing out covariates).**

*Note.* The relationship between the original skew of the raw continuous data (x-axis), the original correlation between the raw continuous data and covariate data (colour coded), and the correlation between normalised continuous residuals (y-axis). This figure is based on simulated data.

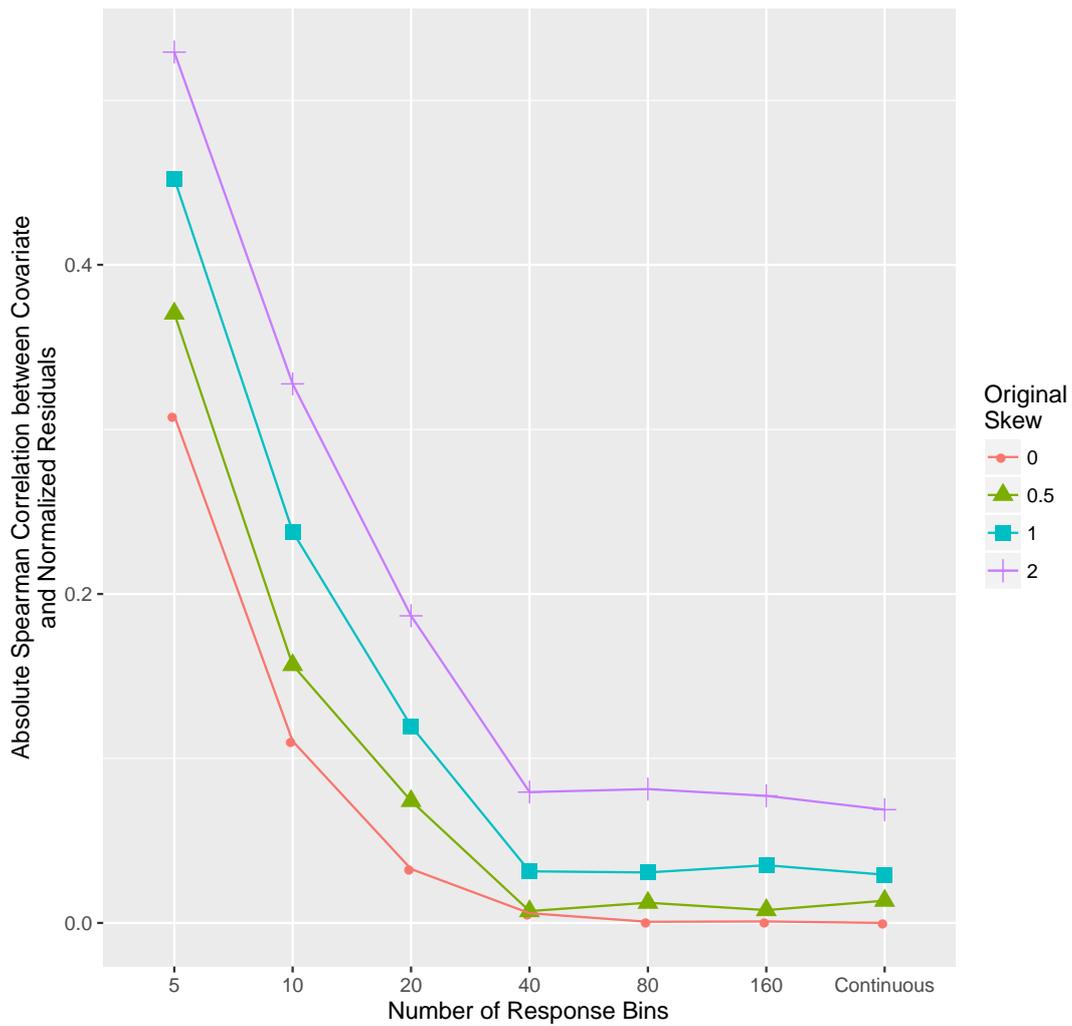


**Supplementary Figure 3.15. Effect of rank-based INT when kurtosis and skew are equal to zero.**



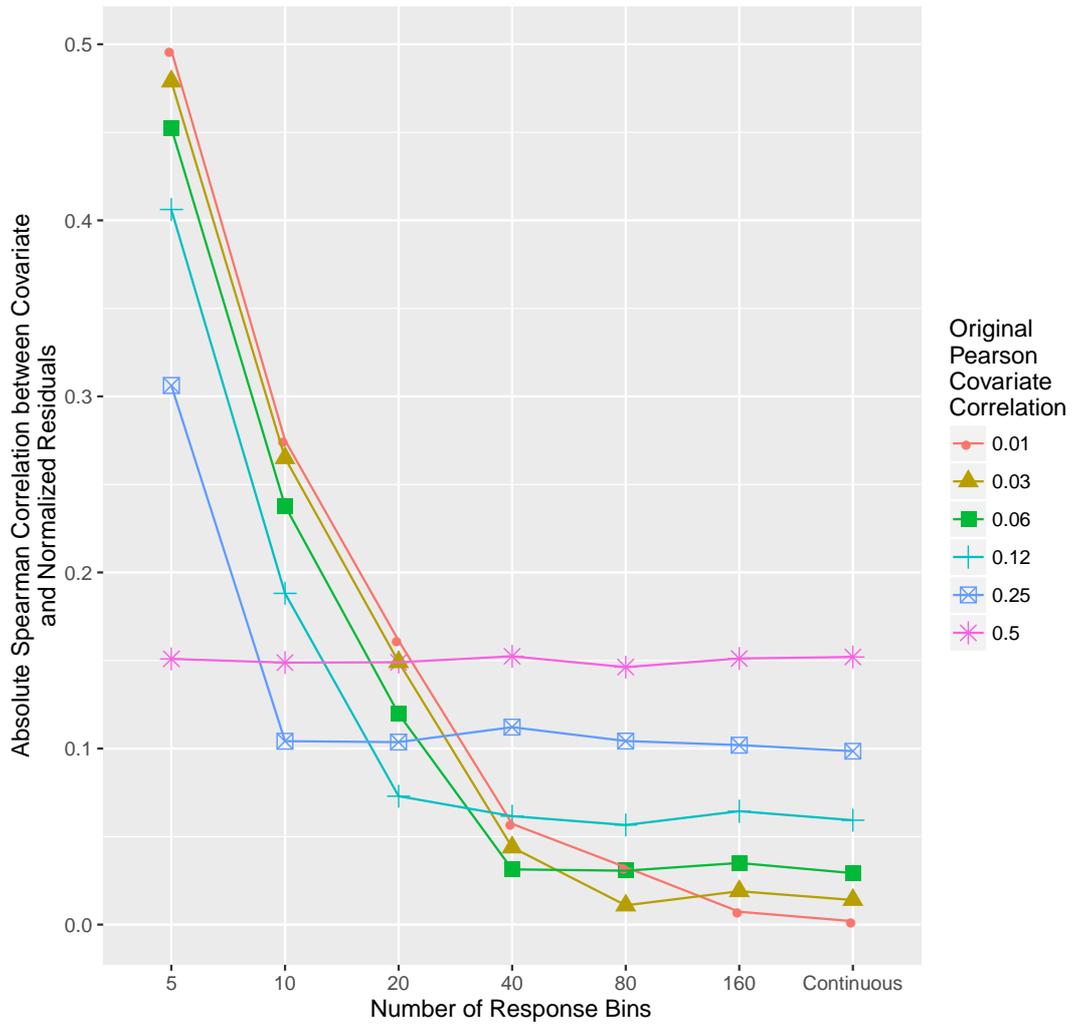
**Supplementary Figure 3.16. Effect of proportion of ties and magnitude of original covariate correlation on the confounding effect of normalising the dependent variable residuals.**

*Note.* The number of response bins ( $x$ -axis) is a measure of the proportion of tied observations. As the number of available responses increases, the proportion of tied observations decreases. The  $y$ -axis is the absolute correlation between normalised residuals and the covariate, and therefore indicates the degree to which normalisation reintroduces the covariate correlation with residuals. Colour indicates the original correlation between the covariate and simulated variables. This figure is based on simulated variables with a skew of 1.



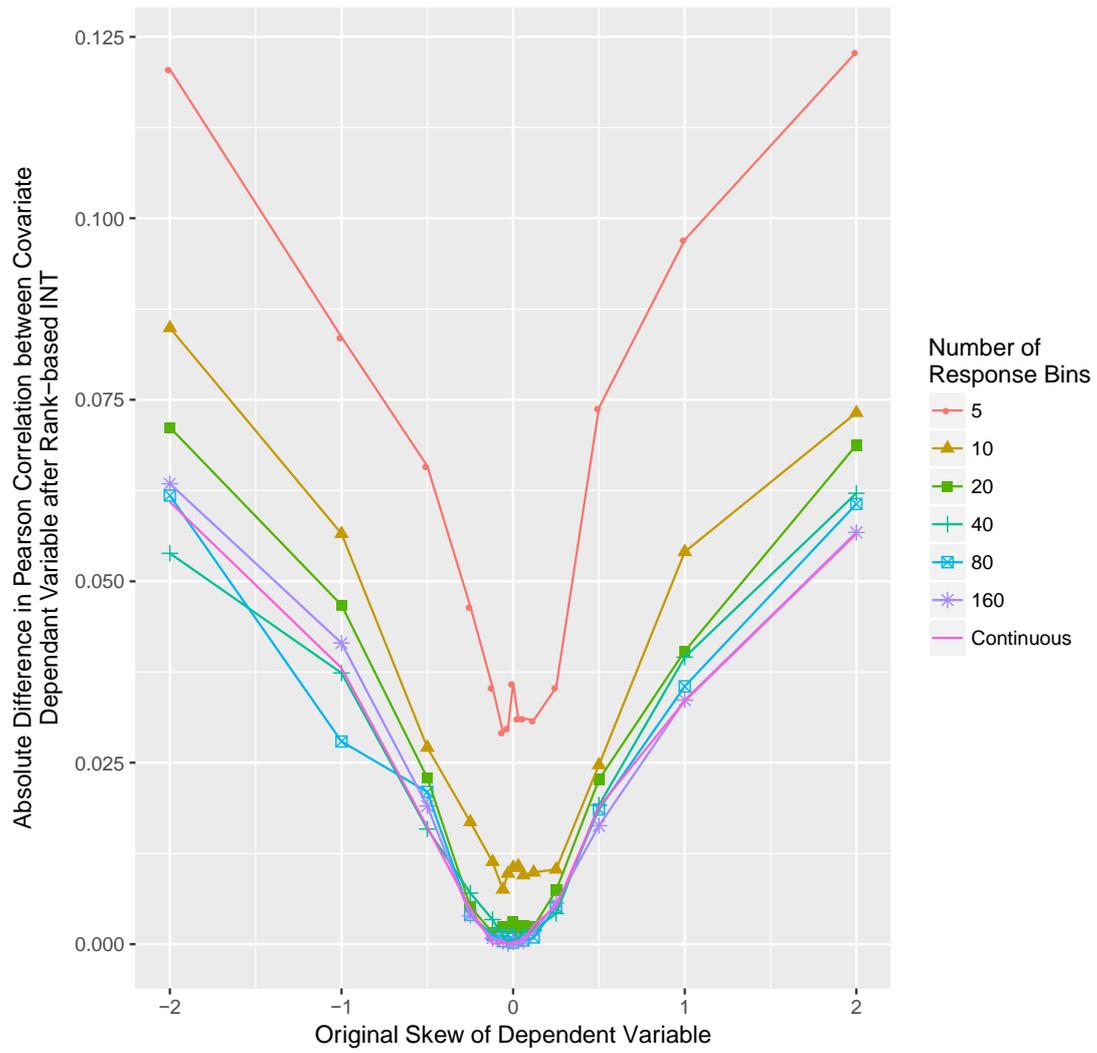
**Supplementary Figure 3.17. The relationship between the number of available responses in the dependent variable (x-axis) and the absolute Spearman rank-based correlation between normalised residuals and covariate (y-axis) for different values of the skew.**

*Note.* Within this figure, the Pearson correlation between the raw dependent variable and covariate data is at 0.06.

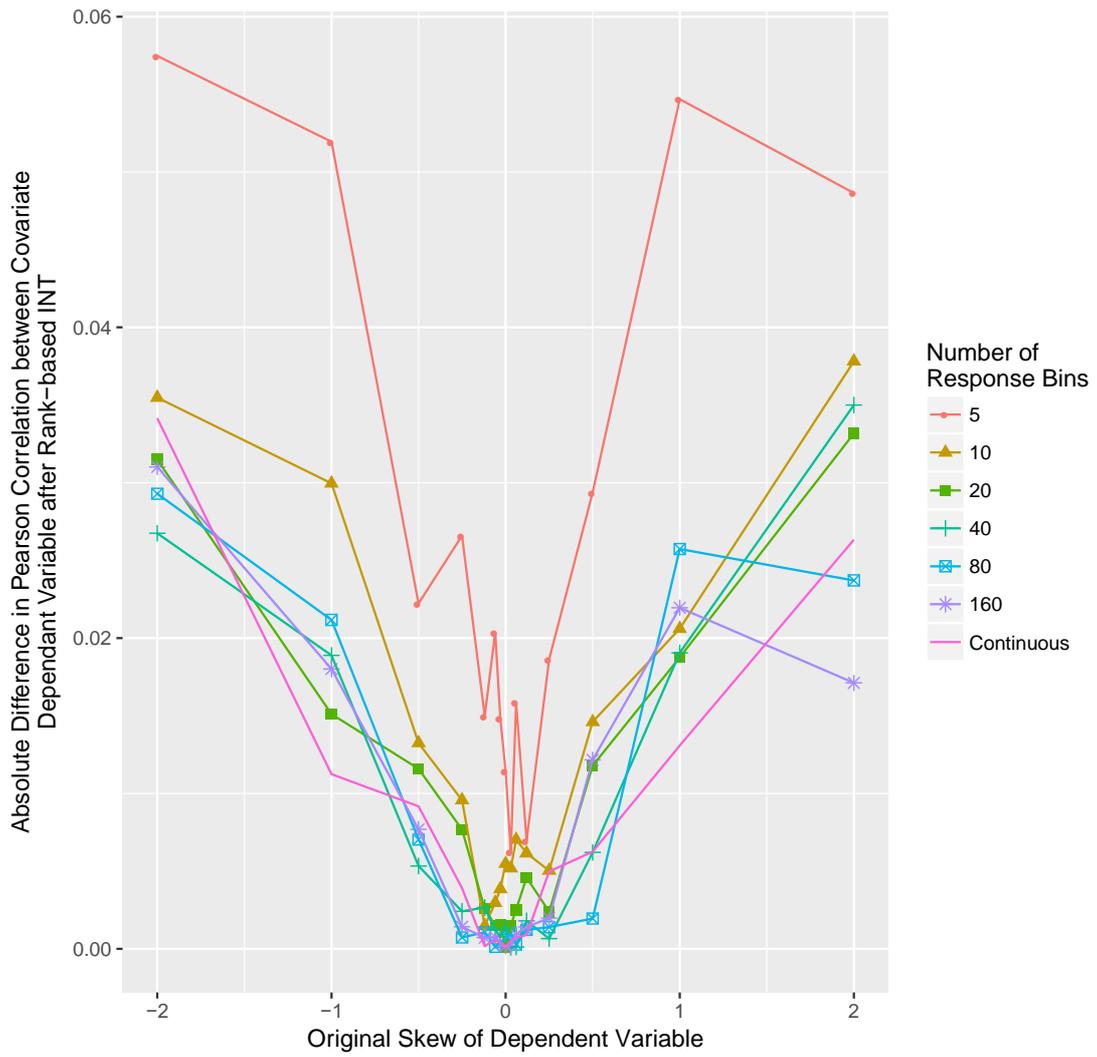


**Supplementary Figure 3.18. Effect of proportion of ties and magnitude of original covariate correlation on the Spearman correlation between dependent variable residuals and the covariate.**

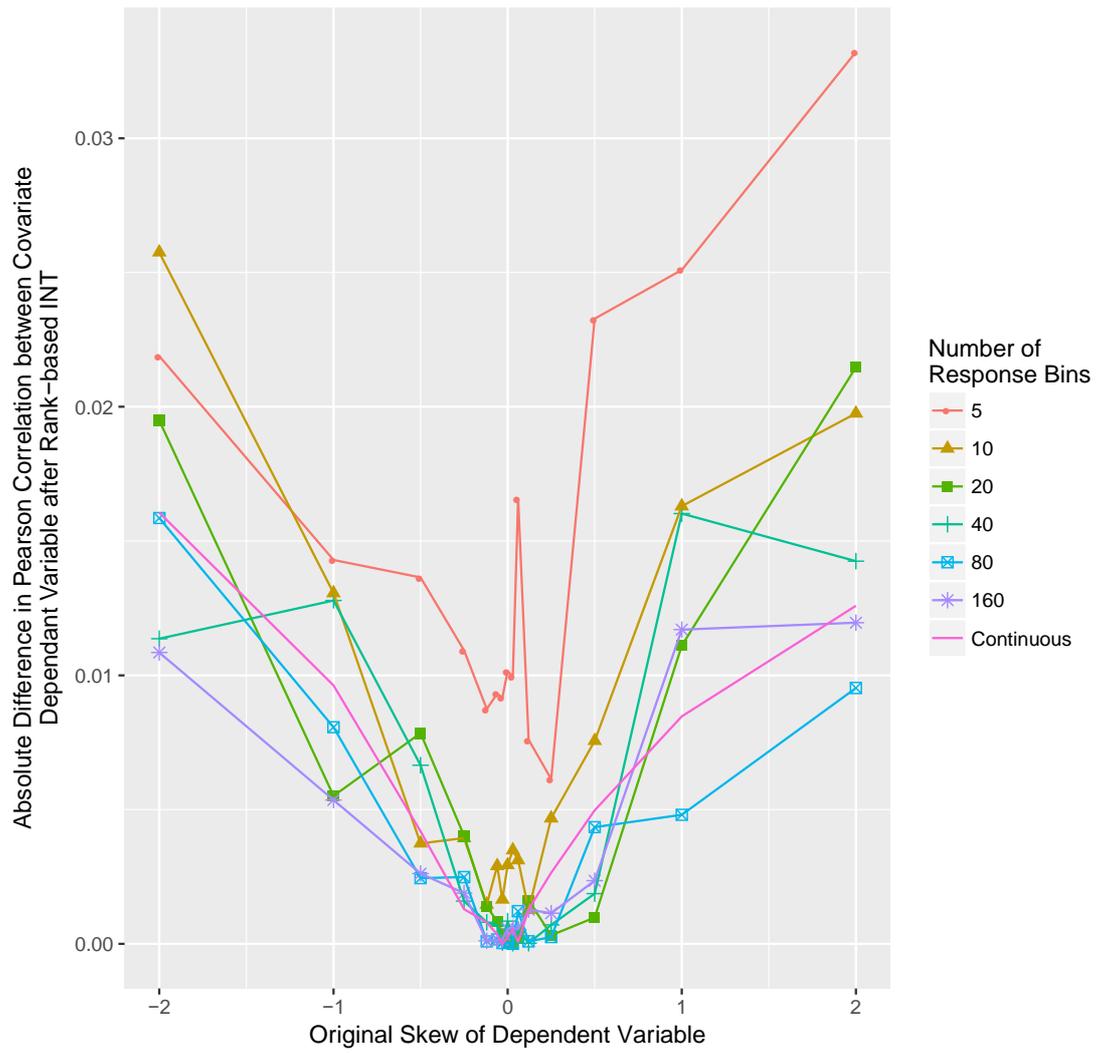
*Note.* The number of response bins ( $x$ -axis) is a measure of the proportion of tied observations. The  $y$ -axis is the absolute Spearman rank-based correlation between residuals and covariates. Colour indicates the original Pearson correlation between the covariates and raw simulated variables. This figure is based on simulated variables with a skew of 1.



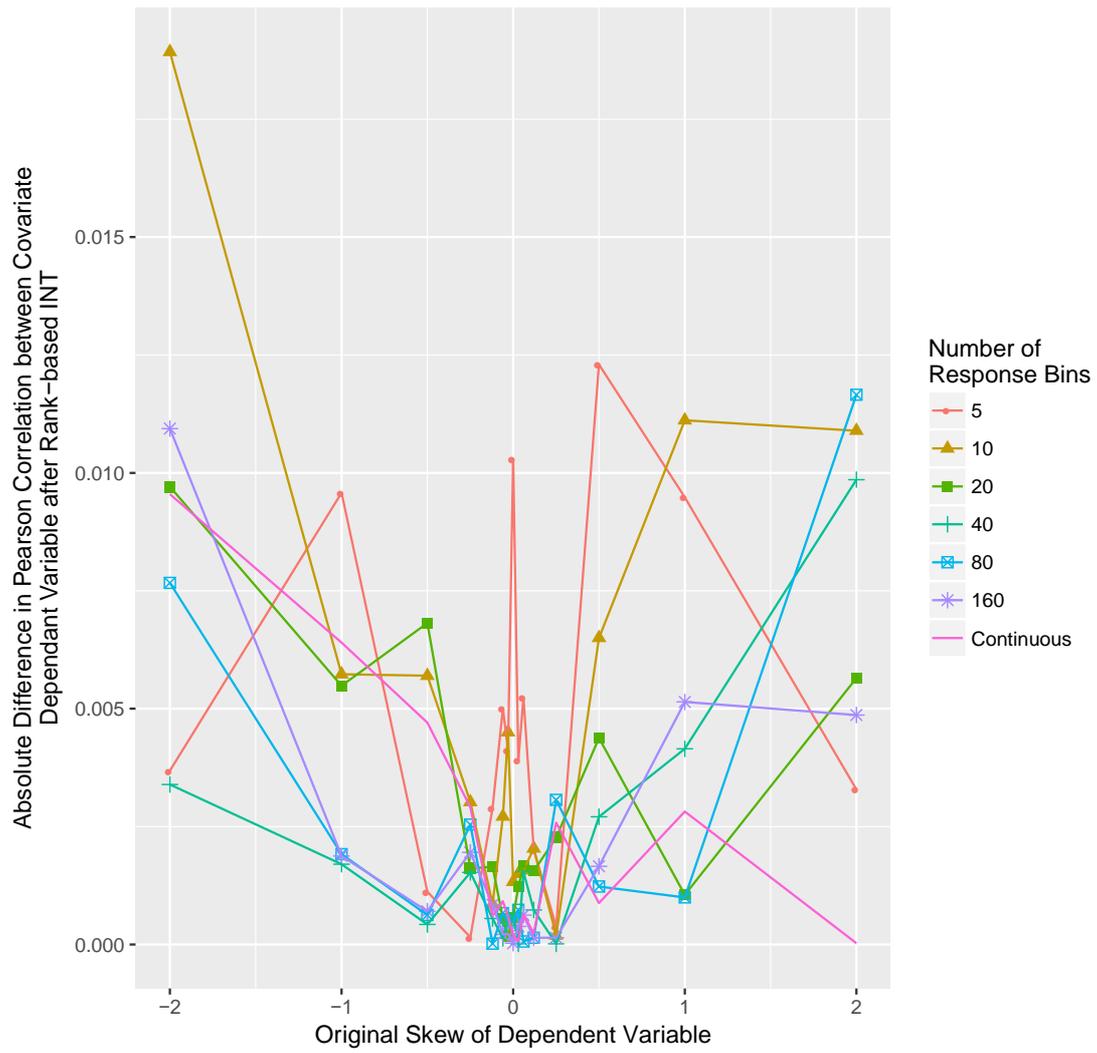
**Supplementary Figure 3.19. Difference in covariate correlation with dependent variable after rank-based INT when original dependent-covariate correlation is 0.5.**



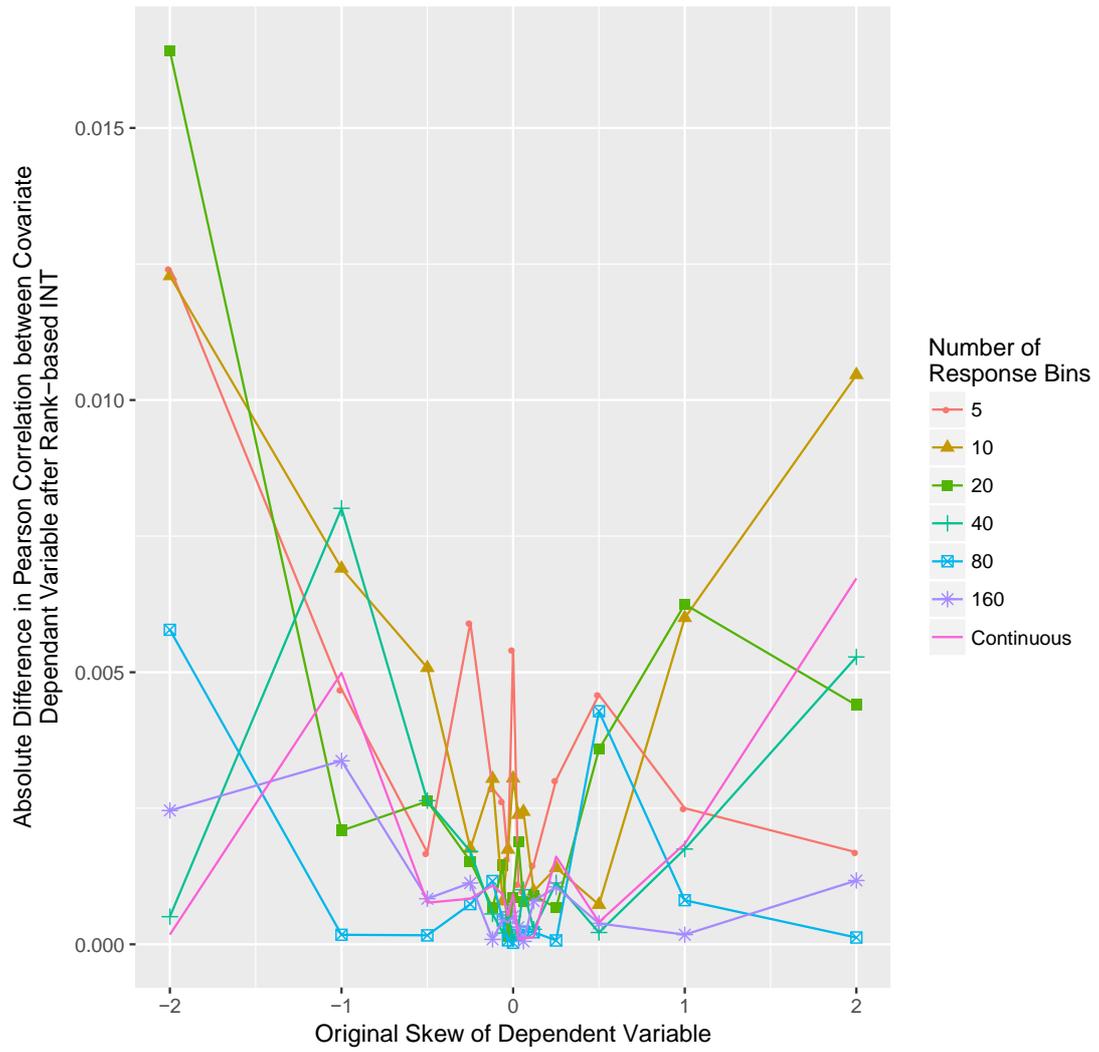
**Supplementary Figure 3.20. Difference in covariate correlation with dependent variable after rank-based INT when original dependent-covariate correlation is 0.25.**



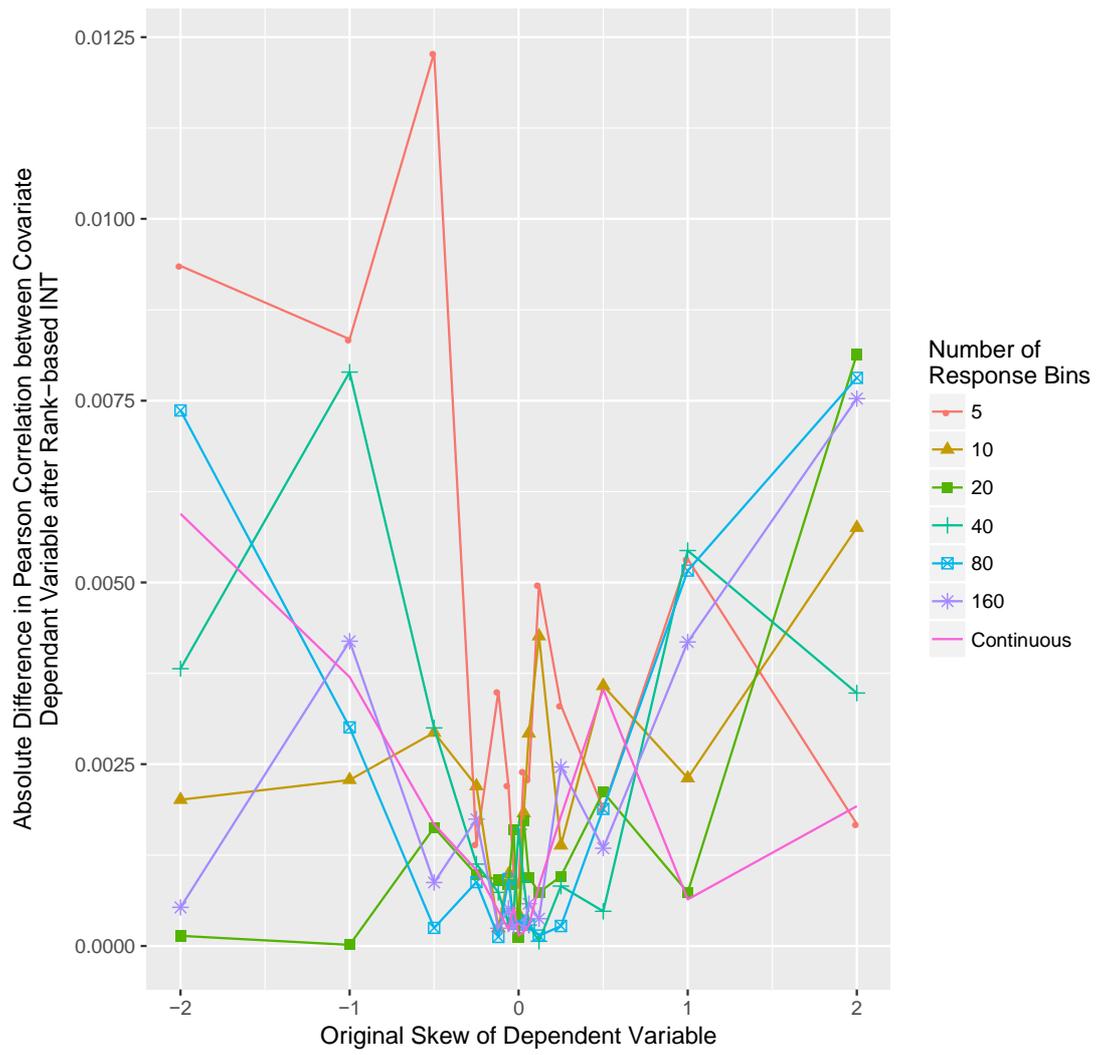
**Supplementary Figure 3.21. Difference in covariate correlation with dependent variable after rank-based INT when original dependent-covariate correlation is 0.12.**



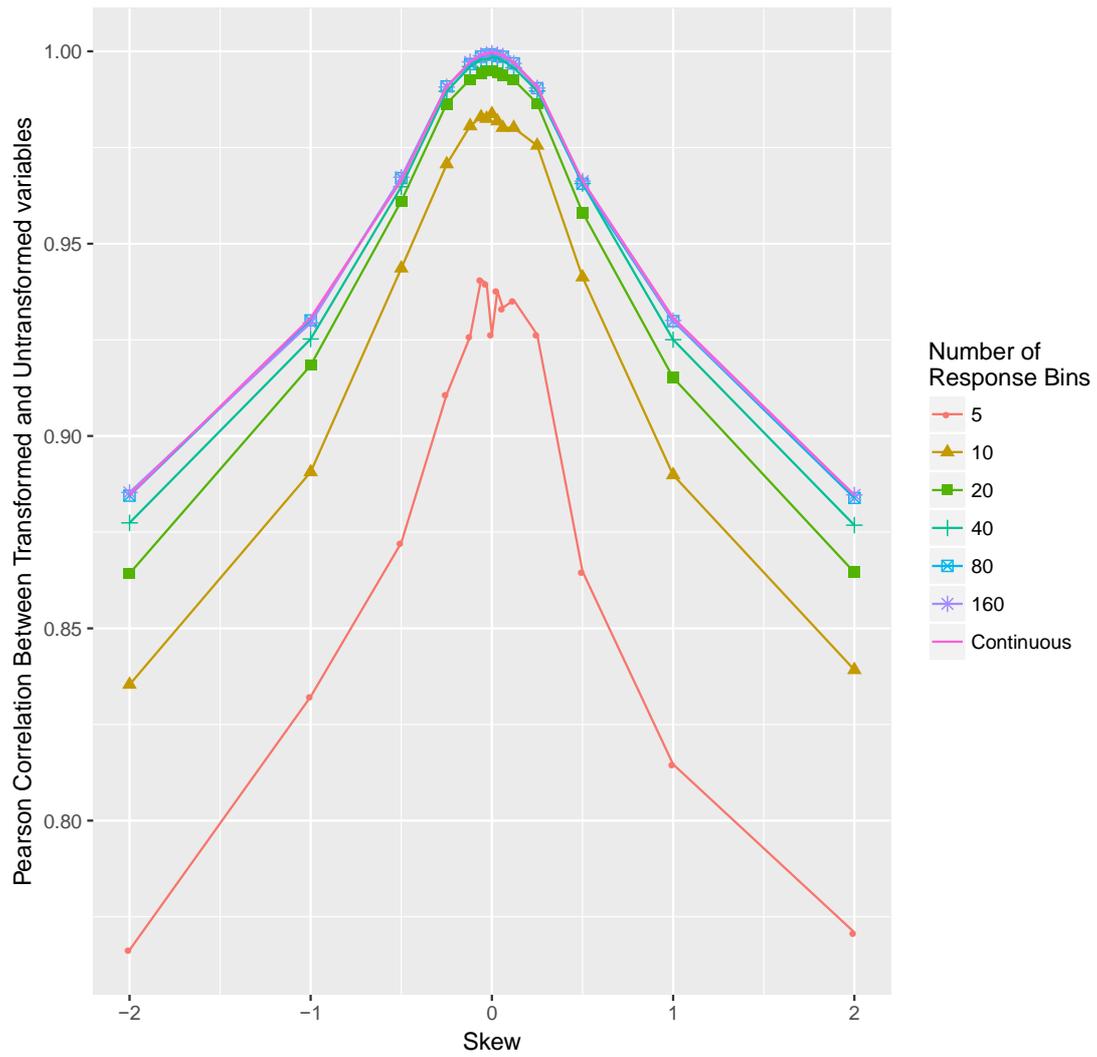
**Supplementary Figure 3.22. Difference in covariate correlation with dependent variable after rank-based INT when original dependent-covariate correlation is 0.06.**



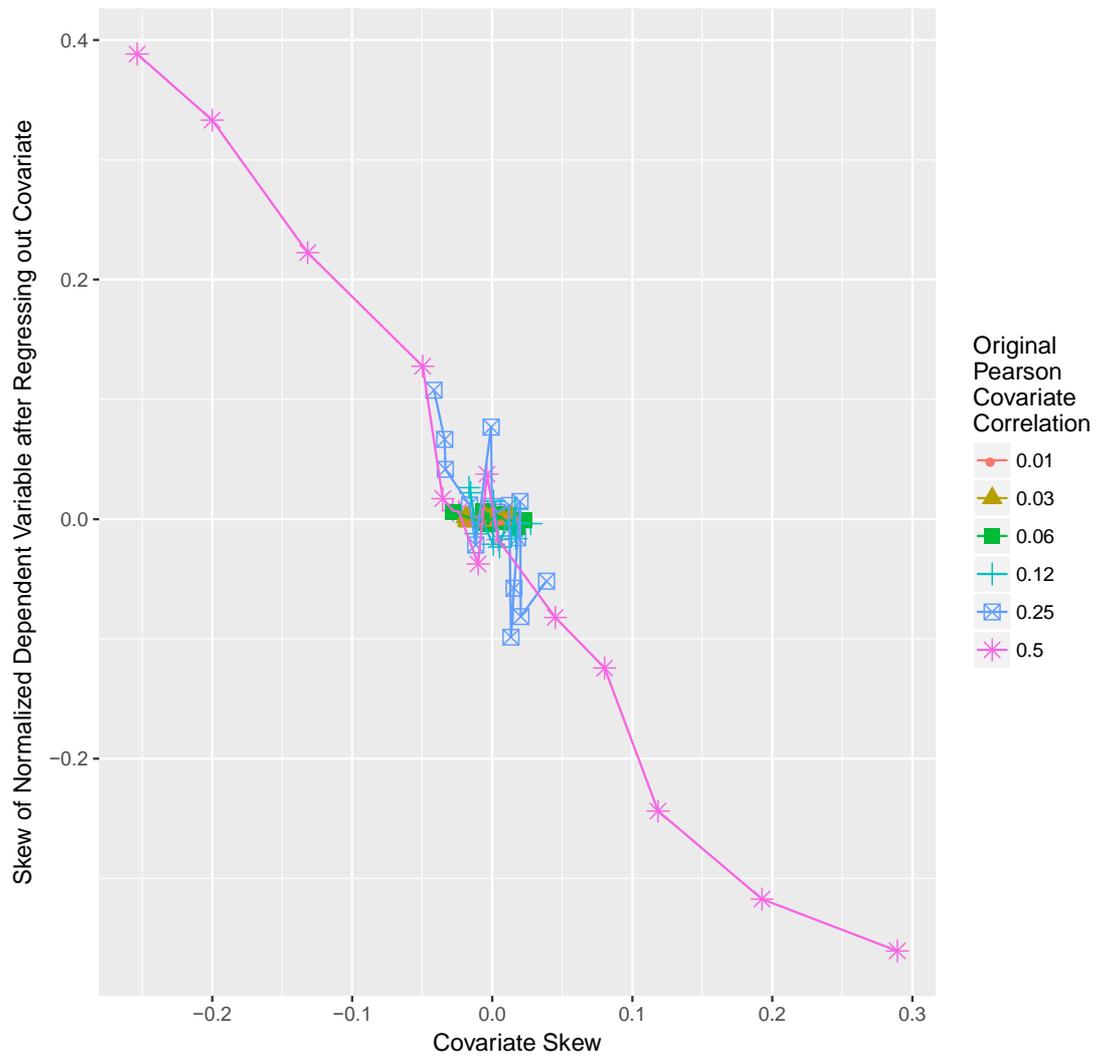
**Supplementary Figure 3.23. Difference in covariate correlation with dependent variable after rank-based INT when original dependent-covariate correlation is 0.03.**



**Supplementary Figure 3.24. Difference in covariate correlation with dependent variable after rank-based INT when original dependent-covariate correlation is 0.01.**



**Supplementary Figure 3.25. Correlation between the dependent variable before and after rank-based INT (randomly splitting tied observations).**



**Supplementary Figure 3.26. Magnitude of skew reintroduced when regressing out covariate effects from normalised dependent variables.**

*Note.* Proportion of ties did not affect this relationship. This figure is based on simulated continuous variables.

# **Chapter 4 - Harmonising Subscales of Specific Psychotic Experiences Across TEDS, ALSPAC and CATSS**

## **4.1 – Introduction**

As discussed in Section 1.6.1, the main limitations of previous molecular genetic studies of adolescent psychotic experiences (PEs) are sample size and the lack of specific and quantitative measures, both of which increase statistical power. As described in Section 1.4.5, the best approach for increasing sample size, apart from starting a new larger scale study to investigate the phenotype of interest, is collaboration between research groups to pool data. Particularly in behavioural or psychiatric research, a likely consequence of combining multiple samples is an increase in phenotypic heterogeneity due to a number of factors including (but not limited to) the use of different measures, environmental differences, and cultural differences. The complex interplay of environmental and cultural differences between samples is difficult to control for and is therefore best avoided by selecting highly comparable samples, based on features such as geographical region, culture and age. The use of different measures between samples is potentially less of an issue when using widely used diagnostic protocols carried out by health care professionals. However, as previously discussed in Section 1.4.4 these diagnostic protocols often lead to binary case-control categories, often containing large degrees of phenotypic heterogeneity. In order to gain the relative improvements in statistical power and interpretation of specific and quantitative measures, careful consideration of the measures used in each sample must occur. Some measures of specific behaviours have been employed in multiple samples enabling the direct comparison of individuals from different samples. For example, a recent meta-analysis of several personality traits was able to combine several samples that had been assessed

via several overlapping questionnaires (Van den Berg et al., 2014). To account for differences between samples for the same questionnaires, this study used item response theory. However, the direct comparison of behavioural traits assessed in different samples often isn't possible due to samples regularly employing bespoke questionnaire measures. An alternative approach is to look within the relevant questionnaires from each sample to identify individual items that are comparable across samples.

A key aim of this thesis is to pool data from samples with sufficient PE data to enable combined analysis. Three European samples with adolescent PE data available have been identified: the Twins Early Development Study (TEDS), the Avon Longitudinal Study of Parents and Children (ALSPAC), and the Child and Adolescent Twin Study in Sweden (CATSS). However, different measures have been used in each sample. This chapter describes the process of harmonising the measures in the three samples onto common scales of specific PEs. Figure 4.1 is a schematic representation of the phenotypic harmonisation procedure. The procedure can be broken down into three stages: 1) Using the TEDS' Specific Psychotic Experiences Questionnaire (SPEQ) as a template, identify items in ALSPAC and CATSS samples that assess specific PEs. 2) Perform psychometric analysis of PE items within ALSPAC and CATSS to derive measures of specific PEs and assess their reliability. 3) Adapt the TEDS measures of specific PEs to improve comparability with measures in ALSPAC and CATSS, and then assess them psychometrically.

## **4.2 - Methods**

### ***4.2.1 - Samples***

Adolescent PEs had been previously assessed in three general population samples called TEDS (Haworth et al., 2013), ALSPAC (Boyd et al., 2012) and CATSS (Anckarsäter et al., 2011). Each of these samples used different measures of adolescent PEs. Below is a

description of the three studies. The available items assessing adolescent PEs are described in Section 4.2.2.

**TEDS** – This sample recruited twins born in England and Wales between 1994 and 1996 and assessed these individuals longitudinally. TEDS originally recruited 13,488 families, who responded with a written consent form. The Institute of Psychiatry ethics committee approved TEDS and their consent procedure (ref: 05/Q0706/228). More information can be found on the following website: <http://www.teds.ac.uk>.

**ALSPAC** – This sample includes 14,062 children born to pregnant residents of the former Avon region in South West England who had an expected date of delivery between the 1<sup>st</sup> of April 1991, and the 31<sup>st</sup> of December 1992. Informed consent has been received from all participants within this study. Ethical approval for this study was obtained from the ALSPAC Law and Ethics Committee and the Local Research Ethics Committees. More information can be found on the following website: <http://www.bristol.ac.uk/alspac>.

**CATSS** – The CATSS sample includes all twins born in Sweden since the 1<sup>st</sup> of July 1992. Currently, data is available for ~5,350 twins at age 18 from the ‘CATSS-18’ subset. All participants are informed of the information being collected and are repeatedly given the opportunity to withdraw. The study has ethical approval from the Karolinska Institute Ethical Review Board. More information can be found on the following website: <http://www.ki.se/en/meb/the-child-and-adolescent-twin-study-in-sweden-catss>.

#### *4.2.2 - Exclusion criteria*

Exclusion criteria were chosen based on those used in previous studies of behavioural traits (Docherty et al., 2010). Each samples’ data dictionary was searched to identify variables relating to the exclusion criteria. The relevant variables from each sample were then matched across samples where possible (Table 4.1). After exclusion criteria

were applied, TEDS, ALSPAC and CATSS respectively provided 4,869, 14,177, and 5,407 unrelated individuals (Table 4.2). These figures do not account for the availability of adolescent PE data, as the PE items used in this study are yet to be identified.

**Table 4.1. Exclusion variables applied in each sample. These exclusion criteria are standard practice for genome-wide association studies of behavioural and cognitive traits (Docherty et al., 2010).**

	TEDS	ALSPAC	CATSS
Exclusion criteria	<ul style="list-style-type: none"> <li>•Unknown zygosity</li> <li>•Unknown sex at age 16</li> <li>•Low birth weight</li> <li>•Short gestational age</li> <li>•Maternal drinking during pregnancy</li> <li>•Long stay in hospital after birth</li> <li>•Diagnosis of autism</li> <li>•Cerebral palsy</li> <li>•Genetic, chromosomal or inherited disorders</li> <li>•Brain damage or disorders affecting the brain</li> <li>•Severely deaf</li> <li>•Developmental delay</li> <li>•Complete blindness</li> <li>•Death of either twin</li> </ul>	<ul style="list-style-type: none"> <li>•Unknown sex</li> <li>•Low birth weight</li> <li>•Short gestational age</li> <li>•Maternal drinking during pregnancy</li> <li>•Long period in hospital after birth</li> <li>•Diagnosis of autism</li> <li>•Cerebral palsy</li>   <li>•Severely deaf</li> <li>•Developmental delay</li> </ul>	<ul style="list-style-type: none"> <li>•Unknown zygosity</li> <li>•Unknown sex</li> <li>•Low birth weight</li> <li>•Birth trauma</li>   <li>•Diagnosis of autism</li> <li>•Cerebral palsy</li> <li>•Chromosomal abnormalities</li> <li>•Brain damage</li> <li>•Deafness</li>   <li>•Complete blindness</li> </ul>

**Table 4.2. Number of individuals from TEDS, ALSPAC and CATSS. Figures shown before and after exclusion criteria have been applied.**

	TEDS	ALSPAC	CATSS
<i>Related individuals before exclusions</i>	10324	15445	10742
<i>Unrelated individuals before exclusions</i>	5162	15243	5669
<i>Unrelated individuals after exclusions</i>	4869	14177	5407

### 4.2.3 - Measures

**TEDS** - The SPEQ consists of six subscales including Paranoia, Hallucinations, Grandiosity and Delusions, Cognitive Disorganisation, Anhedonia, and Parent-rated Negative Symptoms. Details of the SPEQ subscales are available at the following references (Ronald et al., 2013, Sieradzka et al., 2014). Items within the SPEQ are listed and annotated in Supplementary Table 4.1.

**ALSPAC** - This study used the Psychotic Like Experiences Questionnaire (PLIKS-Q) at age 16 to assess adolescent PEs. The PLIKS-Q only assesses some positive PE symptoms including paranoia, hallucinations, and thought insertion. The PLIKS-Q was applied within the 'Life of a 16+ Teenager' questionnaire. This questionnaire included a number of other scales assessing behaviour, all of which were interrogated to identify any items relevant for the assessment of other specific PEs. Items assessing parent-reported negative symptoms were searched for in the corresponding parent questionnaire called 'Your Son/Daughter 16+ Years On'.

Items within the PLIKS-Q consisted of an initial question, and then several follow up questions regarding the individuals' experience (example in Supplementary Note 4.1). It was decided to include frequency information provided by the first follow up question (part a). Including the frequency of experience information provided several advantages including an improved ability to separate the individuals based on the prevalence of the experience, but also matching the frequency type information collected by both TEDS and CATSS positive symptom items.

**CATSS** - Adolescent PEs were assessed at age 18 using the APSS (Adolescent Psychotic-like Symptom Screener). Similar to PLIKS-Q, APSS only captures information on some positive PE symptoms (paranoia, hallucinations, and grandiosity and delusions). All scales applied to these individuals were interrogated to identify items assessing other adolescent PEs. The more relevant scales include Child Mania Rating Scale (CMRS),

Adult ADHD Self-Report Scale (ASRS), and Centre of Epidemiologic Studies Depression Scale (CES-D). Parent-reported negative symptoms were looked for within the parent report Adult Behaviour Checklist (ABCL), and the Autism – Tics, ADHD and other Comorbidities Inventory (A-TAC).

#### 4.2.4 - Analyses

Derivation of comparable subscales of specific PEs across the three studies was achieved via three analytical stages (Figure 4.1).

##### **Stage 1: Identification of items assessing specific adolescent PEs in ALSPAC and CATSS samples using the TEDS' SPEQ measure as a template.**

The data dictionaries for ALSPAC and CATSS were manually searched for all items assessing latent PE traits in common with those captured by the TEDS' SPEQ. The relevant items were then labelled according to the specific PE domain that it assessed. This process was carried out under the guidance of two expert clinicians (Dr Alastair Cardno and Professor Daniel Freeman), both with prior experience of creating measures for specific PEs during adolescence.

##### **Stage 2: Psychometric analysis of PE items in ALSPAC and CATSS to create and assess measures of specific PEs.**

This stage was performed separately for ALSPAC and CATSS. Principal components analysis (PCA) was used to investigate the correlation structure between PE items, highlighting how items cluster into dimensions of PEs.

PCA is a method that uses the correlation between variables to identify axes of variance within the data (Jolliffe, 2002). This dimension reduction of data enables the construction of scales to measure underlying latent variables. PCA was performed using the R function 'prcomp' from the 'stats' package with the 'center' and 'scale' options set

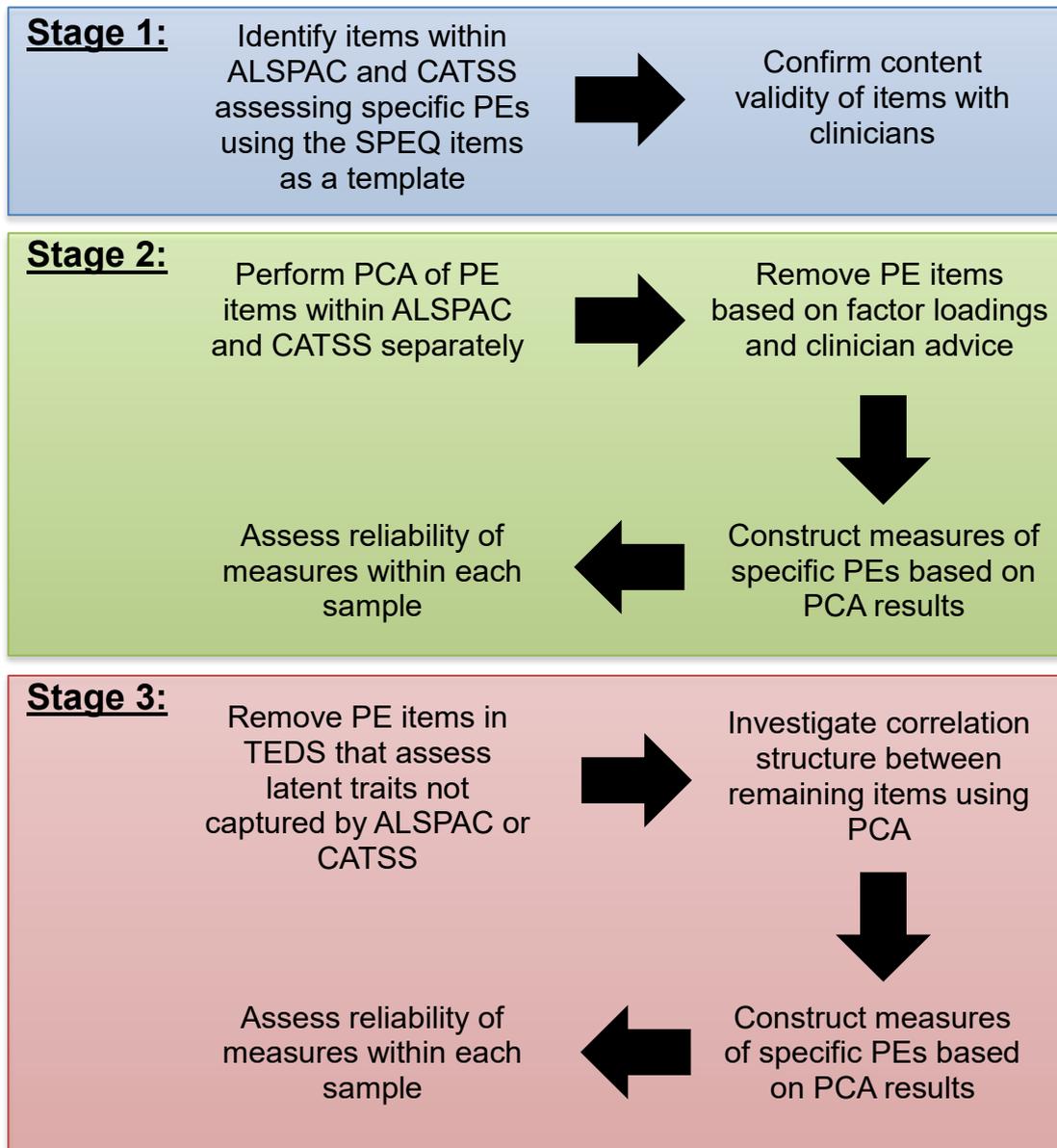
to 'TRUE' (R Core Team, 2015). Based on evidence from scree plots of variance explained by principal components, components were extracted and rotated to improve interpretation of the data. Oblique rotation was used in light of previous evidence of some correlation between specific PEs (Ronald et al., 2014). Extraction and rotation of components was performed using the R function 'principal' from the 'stats' package (R Core Team, 2015).

Measures of specific PEs were based on the PE dimensions identified by PCA and the clinicians' advice. Some specific PE domains can be broken down further into subdomains, e.g. negative symptoms consist of five distinct behaviours including alogia (poverty of speech), avolition, asociality, inattention, and blunted affect. If there was an excess of items assessing one specific aspect of a PE domain, the appropriate number of items would be removed based on clinicians' advice and the factor loadings from PCA.

The Cronbach's alphas of these specific measures were then calculated to assess reliability. This was performed using the R function 'alpha' from the 'psych' package (Revelle, 2015) with the following parameters: 'cumulative = FALSE, max = 10, na.rm = TRUE, check.keys = TRUE, n.iter = 1, delete = TRUE'.

### **Stage 3: Adaptation of TEDS' specific PE measures to improve comparability with ALSPAC and CATSS measures.**

A number of items were removed from the TEDS PE measures to ensure that the measures in TEDS assessed latent traits in common with those captured by the measures in ALSPAC and CATSS. PCA of the remaining PE items in TEDS was then performed to highlight specific PE domains. The number of items assessing each aspect of each PE domain was adjusted to ensure comparability with the measures in ALSPAC and CATSS samples. Calculating the Cronbach's alphas then assessed the reliability of the derived measures.



**Figure 4.1. Schematic representation of the phenotypic harmonisation process.**

## 4.3 - Results

### 4.3.1 – Stage 1: Identifying PE items in ALSPAC and CATSS that matched SPEQ items

**ALSPAC:** Items assessing paranoia and hallucinations matched well with items within the TEDS' SPEQ. Grandiosity was not sufficiently assessed in ALSPAC to create a scale corresponding the TEDS' grandiosity and delusions scale. Although there were no available measures that were designed to specifically assess anhedonia or parent-rated negative symptoms, items from other psychopathology-relevant measures that captured

latent traits resembling these specific PE domains could be used. This is possible due to the wide range of measures administered to the ALSPAC participants, and the overlap between psychological traits, such as social problems. However, cognitive disorganisation was not sufficiently captured within this cohort at age 16. In total, 19 items corresponded to SPEQ items with their content validity confirmed by clinicians. See Table 4.3 for selected items.

**CATSS:** Similar to ALSPAC, items assessing paranoia and hallucinations matched well with items within the TEDS' SPEQ, however, grandiosity was not assessed sufficiently to create a scale corresponding the TEDS' grandiosity and delusions scale. Similar to the situation in ALSPAC, although there were no available measures in CATSS that were designed to specifically assess cognitive disorganisation or parent-rated negative symptoms, items from other psychopathology-relevant measures that captured latent traits resembling these specific PE domains could be used. However, anhedonia was not sufficiently captured within CATSS. In total, 25 items corresponded well to SPEQ items with their content validity confirmed by clinicians. See Table 4.4 for selected items.

**Table 4.3. List of items assessing adolescent psychotic experiences in the ALSPAC sample.**

<b><i>ALSPAC Life of a 16+ Teenager Questionnaire - Section D: Your Current Feelings</i></b>	
<b>Part A</b>	
<b>response options:</b>	0 = No, never, 1 = Yes, maybe, 2 = Yes, definitely
<b>Part B</b>	
<b>response options:</b>	0 = Not at all, 1 = Once or twice, 2 = Less than once a month, 3 = More than once a month, 4 = Nearly every day
<b>para1 a:</b>	Some people believe that other people can read their thoughts. Have other people ever read your thoughts?
<b>para1 b:</b>	How often have other people read your thoughts since your 15th birthday?
<b>para2 a:</b>	Have you ever thought you were being followed or spied on?
<b>para2 b:</b>	How often has this happened since your 15th birthday?
<b>para3 a:</b>	Have you ever believed that you were being sent special messages through the television or the radio, or that a programme had been arranged just for you alone?
<b>para3 b:</b>	How often has this happened since your 15th birthday?
<b>halluc1 a:</b>	Have you ever heard voices that other people couldn't hear
<b>halluc1 b:</b>	How often have you heard voices that other people couldn't hear since your 15th birthday?
<b>halluc2 a:</b>	Have you ever seen something or someone that other people could not see?
<b>halluc2 b:</b>	How often have you seen something or someone that other people could not see since your 15th birthday?
<b>Note:</b>	Composite scores for each item were created by adding the Part A and Part B responses.

Table 4.3 cont.

<b><i>ALSPAC Life of a 16+ Teenager Questionnaire - Section D: Your Current Feelings</i></b>	
<b>Leading statement:</b>	For each of the following questions, please mark the box that best describes the way you have felt over the past month:
<b>Response options:</b>	0 = No, never, 1 = Yes, sometimes, 2 = Yes, often, 3 = Yes, nearly always
<b>anhed1:</b>	Have you felt that you experience few or no emotions at important events, such as on your birthday?
<b>anhed2:</b>	Have you felt that you are lacking 'get up and go'?
<b>anhed3:</b>	Have you felt that you have only a few hobbies or interests?
<b><i>ALSPAC Life of a 16+ Teenager Questionnaire - Section H: Your Current Feelings</i></b>	
<b>Leading statement:</b>	In the past two weeks...
<b>Response options:</b>	0 = Not true, 1 = Sometimes true, 2 = True
<b>anhed4:</b>	I have been having fun. (Reversed)
<b>anhed5:</b>	I didn't enjoy anything at all.
<b>anhed6:</b>	I felt so tired that I just sat around and did nothing.
<b>anhed7:</b>	I have had a good time. (Reversed)
<b><i>ALSPAC Your Daughter 16+ Years On' Questionnaire - Section A: Your Study Teenager</i></b>	
<b>Response options:</b>	0 = Often, 1 = Sometimes, 2 = Hardly ever, 3 = Never
<b>negsym1:</b>	How often does he/she tell you about things that happen at school/college/work?
<b>negsym2:</b>	How often does he/she tell you about things that happen while he's/she's been out?

Table 4.3 cont.

<b><i>ALSPAC Your Daughter 16+ Years On' Questionnaire - Section D: Your Teenager's Feelings</i></b>	
<b>Leading statement:</b>	In the past 6 months...
<b>Response options:</b>	0 = Not true, 1 = Somewhat true, 2 = Certainly true, NA = Don't know
<b>negsym3:</b>	He/She has at least one good friend.
<b>negsym4:</b>	He/She is easily distracted, his/her concentration wanders.
<b>negsym5:</b>	He/She sees tasks through to the end has good attention span.
<b>negsym6:</b>	He/She did not respond when told to do something.
<b><i>ALSPAC Your Daughter 16+ Years On' Questionnaire - Section F: Your Teenager's Health</i></b>	
<b>Response options:</b>	0 = No, 1 = Yes
<b>negsym7:</b>	Thinking back over the last month, has she been feeling tired or felt she had no energy?

*Note.* A few items in this list may not be retained in the final scale due to results of stage 2 of analysis. Items are labelled based on content validity against the SPEQ subscales. para = paranoia, halluc = hallucinations, anhed = anhedonia, negsym = parent-rated negative symptoms.

**Table 4.4. List of items assessing adolescent psychotic experiences in the CATSS sample.**

<b>Adolescent Psychotic-like Symptom Screener (APSS)</b>	
<b>Leading statement:</b>	Have you ever...
<b>Response options:</b>	0 = Never or rarely, 1 = Sometimes, 2 = Often, 3 = Very Often
<b>para1:</b>	Thought you were being followed or spied on?
<b>para2:</b>	Thought you were being sent special messages through the television?
<b>para3:</b>	Thought other people could read your thoughts?
<b>halluc1:</b>	Seen things other people cannot see?
<b>halluc2:</b>	Heard voices that nobody else can hear?
<b>Adult ADHD Self-Report Scale (ASRS)</b>	
<b>Response options:</b>	0 = Never, 1 = Rarely, 2 = Sometimes, 3 = Often, 4 = Very often
<b>cogdis1:</b>	How often do you have trouble wrapping up the fine details of a project, once the challenging parts have been done?
<b>cogdis2:</b>	When you have a task that requires a lot of thought, how often do you avoid or delay getting started?
<b>cogdis3:</b>	How often do you have difficulty keeping your attention when you are doing boring or repetitive work?
<b>cogdis4:</b>	How often do you have difficulty concentrating on what people say to you, even when they are speaking to you directly?
<b>cogdis5:</b>	How often are you distracted by activity or noise around you?

Table 4.4 cont.

<b>Adult Behaviour Checklist (ABCL) - Parental Report</b>	
<b>Leading statement:</b>	How accurate are the following statements for your child in the past six months?
<b>Response options:</b>	0 = Not true, 1 = Somewhat true, 2 = Very or often true
<b>negsym1:</b>	Fails to finish things he/she should do
<b>negsym2:</b>	Underactive, slow moving, or lacks energy
<b>negsym3:</b>	Doesn't get along with other people
<b>negsym4:</b>	Would rather be alone than with others
<b>negsym5:</b>	Not liked by others
<b>negsym6:</b>	Refuses to talk
<b>negsym7:</b>	Has trouble making or keeping friends
<b>negsym8:</b>	Secretive, keeps things to self
<b>negsym9:</b>	Withdrawn, doesn't get involved with others
<b>negsym10:</b>	Stares blankly
<b>negsym11:</b>	Feels tired without good reason
<b>negsym12:</b>	Enjoys being with people
<b>negsym15:</b>	Can't concentrate, can't pay attention for long
<b>Autism - Tics, ADHD and other Comorbidities (A-TAC) -Parental Report</b>	
<b>Response options:</b>	0 = No, 1 = Yes, to a certain degree, 2 = Yes
<b>negsym13:</b>	Does the twin have difficulties expressing emotions and reactions with facial gestures, prosody, or body language?
<b>negsym14:</b>	Does the twin have difficulties to make and keep friends?

*Note.* A few items in this list may not be retained in the final scale due to results of stage 2 of analysis. Items are labelled based on content validity against SPEQ subscales. para = paranoia, halluc = hallucinations, cogdis = cognitive disorganisation, negsym = parent-rated negative symptoms.

#### 4.3.2 - Stage 2: Psychometric analysis of identified PE items within ALSPAC and CATSS

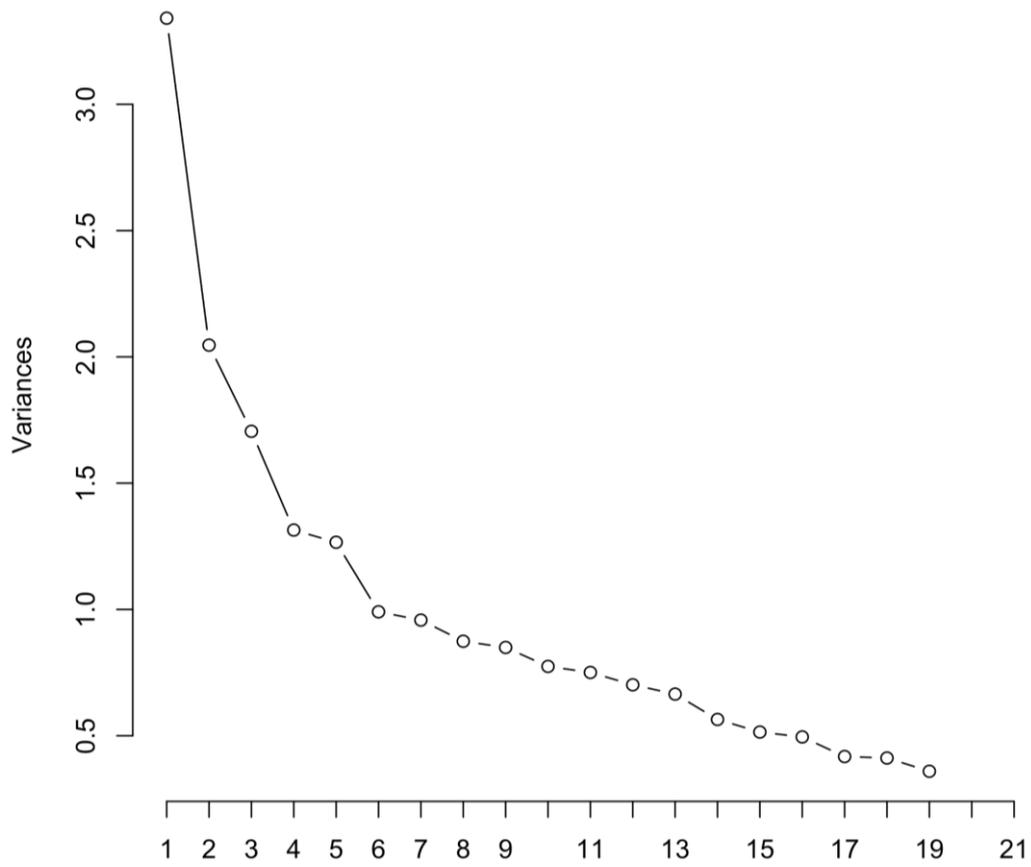
**ALSPAC:** Using a scree plot to interpret PCA results, the majority of variance in these items was explained by three components (Figure 4.2). Therefore, extraction of three principal components was performed to improve interpretability. The three principal component model fit the data well with a root mean square of the residuals (RMSR) of 0.08. Using a cut-off of  $\pm 0.3$  to interpret loadings, all paranoia and hallucination items loaded onto a single component, anhedonia items all loaded onto a single component and five out of seven parent rated negative symptom items loaded onto single component. Based on the items loading within the three components (Table 4.5), the components (or latent variables) were named paranoia and hallucinations, anhedonia and parent-rated negative symptoms. It was agreed in consultation with clinicians that the two poorly loading parent-rated negative symptom items, negsym3 and negsym7 ('at least one good friend' and 'feeling tired or has no energy') were measuring key aspects of the construct and should be kept despite poor loadings.

Calculation of Cronbach's alpha for the subscales showed that paranoia and hallucinations, and anhedonia subscales had good reliability, with standardised alphas of 0.69 and 0.73 respectively. As a result of the two poorly loading items for parent-rated negative symptoms (negsym3 and negsym7), this subscale only achieved an alpha of 0.60 (Table 4.6). If one of the poorly loading items was removed from the subscale then the standardised alpha increased to  $\geq 0.63$ . However, with the understanding that these behavioural traits do not necessarily behave well psychometrically, all items were retained due to good content validity.

**CATSS:** PCA showed that the majority of variance among these 25 items was explained by three principal components (Figure 4.3). Therefore, extraction of three principal components was performed with oblique rotation to improve interpretability (Table 4.7). The three principal component model fit the data well with an RMSR of 0.06.

Similar to ALSPAC, using a loading cut-off of  $\pm 0.3$ , all paranoia and hallucination items loaded onto a single component. This latent variable captured by this component was therefore named paranoia and hallucinations. All cognitive disorganisation items loaded at  $\pm 0.3$  on a single component. Of the 15 parent-rated negative-symptom items, 13 loaded onto a single component using a cut-off of  $\pm 0.3$ . The other two had loadings of 0.24 and 0.20 with fairly high cross loading onto the cognitive disorganisation component. The poorly loading parent-rated negative symptom items were retained in the scale due to good content validity. Of the 15 parent-rated negative symptoms, seven assessed asociality, whereas only two assessed each of the other aspects of negative symptoms (poverty of speech, inattention, avolition, and blunted affect). Although PCA identified these items as assessing a single latent factor, based on previous literature and the advice of clinicians, it was deemed important that each specific aspect of negative symptoms have an equal contribution to an individuals' overall score. Therefore, 5 asociality items were removed from the scale leaving negsym7 and negsym9, the two best-matched items across samples that also had the highest loading on the negative symptom component in PCA. The results of PCA changed minimally when re-run after removal of the excess asociality negative symptom items (see Supplementary Figure 4.1 and Supplementary Table 4.2).

Calculation of Cronbach's alphas for the subscales showed that the reliability of the positive symptom, cognitive disorganisation subscale and parent-rated negative symptom subscales was good with standardised alphas of 0.73, 0.79 and 0.76 respectively (Table 4.8).



**Figure 4.2. Scree plot of ALSPAC psychotic experience items.**

**Table 4.5. Principal component loadings of psychotic experience items in ALSPAC.**

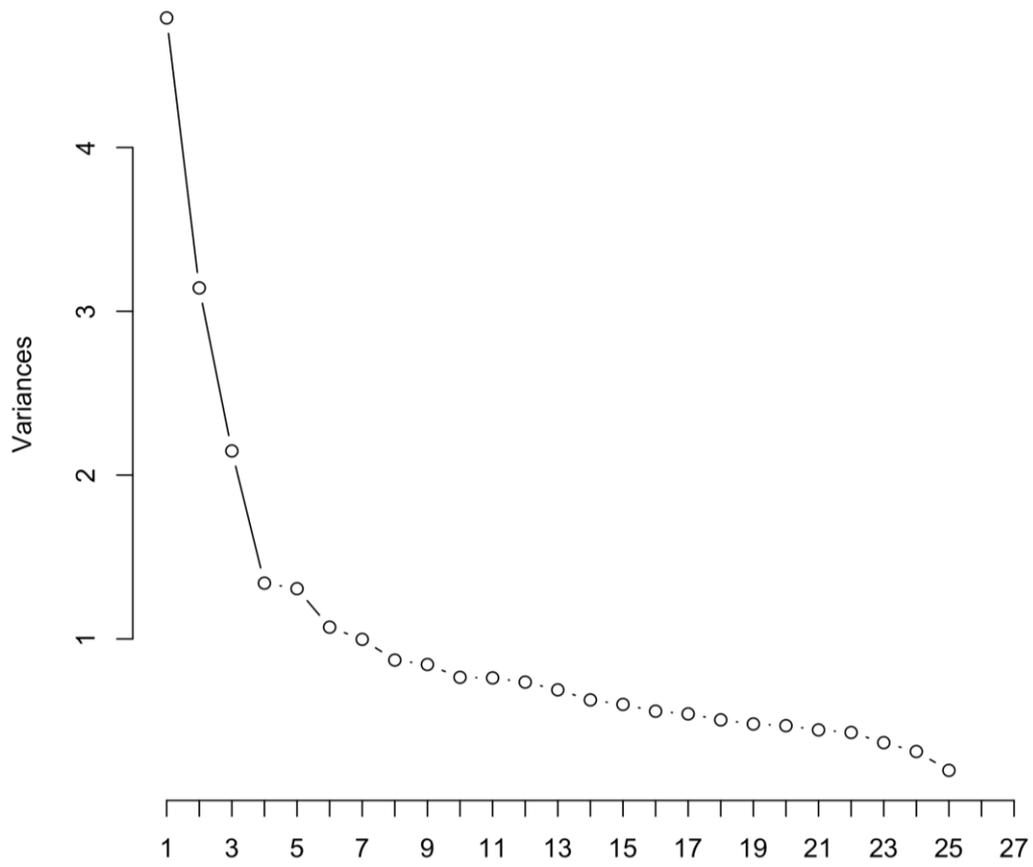
	<b>Component 1</b>	<b>Component 2</b>	<b>Component 3</b>
para1.comp	-0.02	<b>0.61</b>	-0.02
para2.comp	0.09	<b>0.63</b>	0.02
para3.comp	-0.03	<b>0.57</b>	0.00
halluc1.comp	0.02	<b>0.67</b>	-0.02
halluc2.comp	0.00	<b>0.74</b>	-0.01
anhed1	<b>0.50</b>	0.18	0.05
anhed2	<b>0.61</b>	0.17	0.05
anhed3	<b>0.57</b>	0.15	0.03
anhed4	<b>0.65</b>	-0.18	-0.04
anhed5	<b>0.57</b>	0.10	0.03
anhed6	<b>0.47</b>	0.16	0.03
anhed7	<b>0.72</b>	-0.14	-0.04
negsym1	0.05	-0.13	0.67
negsym2	0.03	-0.11	<b>0.68</b>
negsym3	0.09	-0.03	0.14
negsym4	-0.07	0.12	<b>0.64</b>
negsym5	-0.01	0.06	<b>0.69</b>
negsym6	0.00	0.05	<b>0.56</b>
negsym7	0.18	0.02	0.12

*Note.* Shows the presence of three underlying latent variables that have been labelled as paranoia and hallucinations (Component 2), anhedonia (Component 1), and parent rated negative symptoms (Component 3). Numbers in bold and italic highlight a principal component loading of  $\pm 0.3$ . The RMSR is 0.08 indicates this model provides a good fit.

**Table 4.6. Cronbach's alpha for the paranoia and hallucinations, anhedonia and parent-rated negative symptoms subscales identified within ALSPAC.**

<b>Paranoia and Hallucinations</b>	
Full	0.69
excl. para1.comp	0.66
excl. para2.comp	0.64
excl. para3.comp	0.66
excl. halluc1.comp	0.64
excl. halluc2.comp	0.61
<b>Anhedonia</b>	
Full	0.73
excl. anhed1	0.70
excl. anhed2	0.67
excl. anhed3	0.69
excl. anhed4	0.71
excl. anhed5	0.69
excl. anhed6	0.71
excl. anhed7	0.69
<b>Parent-rated Negative Symptoms</b>	
Full	0.60
excl. negsym1	0.54
excl. negsym2	0.54
excl. negsym3	0.63
excl. negsym4	0.54
excl. negsym5	0.51
excl. negsym6	0.54
excl. negsym7	0.64

*Note.* Table shows alpha when excluding each item of the scale. Parent-rated negative symptom scale has a poor alpha due to negsym3 and negsym7. These items have good content validity and will therefore be retained.



**Figure 4.3. Scree plot of CATSS psychotic experience items.**

**Table 4.7. Principal component loadings of psychotic experience items in CATSS.**

	<b>Component 1</b>	<b>Component 2</b>	<b>Component 3</b>
para1	-0.10	0.12	<b>0.70</b>
para2	-0.03	-0.16	<b>0.64</b>
para3	0.01	0.00	<b>0.78</b>
halluc1	0.03	0.02	<b>0.72</b>
halluc2	0.09	0.02	<b>0.73</b>
cogdis1	-0.03	<b>0.68</b>	0.01
cogdis2	-0.01	<b>0.72</b>	-0.10
cogdis3	-0.02	<b>0.71</b>	0.05
cogdis4	-0.01	<b>0.70</b>	0.06
cogdis5	-0.04	<b>0.72</b>	0.03
negsym1	0.24	<b>0.43</b>	-0.07
negsym2	<b>0.55</b>	0.18	-0.03
negsym3	<b>0.35</b>	0.11	0.04
negsym4	<b>0.72</b>	-0.05	0.00
negsym5	<b>0.48</b>	0.00	-0.01
negsym6	<b>0.52</b>	0.05	0.08
negsym7	<b>0.81</b>	-0.07	0.01
negsym8	<b>0.62</b>	0.04	0.03
negsym9	<b>0.74</b>	0.01	0.01
negsym10	<b>0.33</b>	0.05	0.12
negsym11	<b>0.32</b>	<b>0.34</b>	0.10
negsym12	<b>0.39</b>	0.02	-0.02
negsym13	<b>0.45</b>	0.08	-0.15
negsym14	<b>0.76</b>	-0.06	-0.02
negsym15	0.20	<b>0.39</b>	0.05

*Note.* Highlights the presence of three underlying latent variables that have been labelled called positive symptoms (Component 3), cognitive disorganisation (Component 2), and parent rated negative symptoms (Component 1).

**Table 4.8. Cronbach's alpha for the paranoia and hallucinations, cognitive disorganisation and parent-rated negative symptoms subscale identified within CATSS.**

<b>Paranoia and Hallucinations</b>	
Full	0.73
excl. para1	0.67
excl. para2	0.73
excl. para3	0.67
excl. halluc1	0.67
excl. halluc2	0.67
<b>Cognitive Disorganisation</b>	
Full	0.79
excl. cogdis1	0.77
excl. cogdis2	0.75
excl. cogdis3	0.74
excl. cogdis4	0.75
excl. cogdis5	0.76
<b>Parent-rated Negative Symptoms</b>	
Full	0.76
excl. negsym1	0.74
excl. negsym2	0.72
excl. negsym6	0.74
excl. negsym7	0.73
excl. negsym8	0.72
excl. negsym9	0.72
excl. negsym10	0.74
excl. negsym11	0.74
excl. negsym13	0.75
excl. negsym15	0.75

*Note.* Table shows alpha when excluding each item of the scale.

#### 4.3.3 - Stage 3: Adaptation and psychometric analysis of TEDS' SPEQ scale based on the subscales available in ALSPAC and CATSS

The SPEQ has many more items assessing paranoia and hallucinations than in ALSPAC and CATSS. This means that some aspects of the SPEQ paranoia and hallucinations subscales were not assessed in ALSPAC and CATSS. For example, ALSPAC and CATSS only assessed visual and auditory hallucinations, whereas in TEDS other types of hallucinations were measured, including olfactory and tactile hallucinations. To overcome this potential issue, TEDS' SPEQ items were removed if they assessed aspects of subscales that were not assessed in ALSPAC and CATSS subscales. Adaptations to the TEDS paranoia and hallucinations scales resulted in three items assessing paranoia and five items assessing hallucinations. No items had to be excluded from the anhedonia, cognitive disorganisation or parent-rated negative symptoms scales, leaving ten items assessing each of these subscales (Table 4.9). These adaptations to the SPEQ were carried out in consultation with expert clinicians who were involved in the original development of the SPEQ.

PCA showed that the majority of variance in these remaining items was explained by four components (Figure 4.4). Extraction and rotation of these four components showed that the paranoia and hallucinations items fell into a single component (Table 4.10) and the other three subscales (cognitive disorganisation, parent rated-negative symptoms and anhedonia) remained distinct corresponding to the results of ALSPAC and CATSS psychometric analysis. The composition of these PCA results resembled ALSPAC and CATSS PCA results closely. Paranoia and hallucinations items were therefore considered as measuring a single latent variable.

Although PCA identifies items within a component as assessing a single latent variable, given the prior knowledge that paranoia and hallucinations are separable latent traits, the numbers of items assessing these two domains were balanced across samples to

ensure that when the individual phenotypic scores are calculated, these two domains would be weighted equally across samples. In CATSS and ALSPAC there are three items assessing paranoia and two items assessing hallucinations. In each of these studies one hallucination item assesses auditory hallucinations, and the other assesses visual hallucinations. In the TEDS paranoia and hallucinations scale, there are three items assessing paranoia, but five items assessing hallucinations. Given the prior knowledge that paranoia and hallucinations are separable traits, this imbalance in the number of paranoia items and hallucination items may result in an over-representation of hallucinations in TEDS individuals compared to CATSS and ALSPAC individuals. In order to promote phenotypic homogeneity across the samples the TEDS hallucination items were condensed into two items, one assessing auditory (labelled as Aud\_Hall) and the other assessing visual (labelled as Vis\_Hall). In order to preserve the valuable information provided by multiple items assessing both visual and auditory hallucinations, the mean across auditory hallucination items will act as a single item assessing auditory hallucinations. Similarly, the mean of the two visual hallucination items will be used as a single item assessing visual hallucinations. This procedure was not carried out for anhedonia, cognitive disorganisation or parent-rated negative symptom scales, as the items within them are measuring a more unidimensional trait meaning the number of items available in each sample will introduce less heterogeneity. After converting hallucinations items into two items, the Cronbach's alpha of the paranoia and hallucinations scale was 0.72 (Table 4.11).

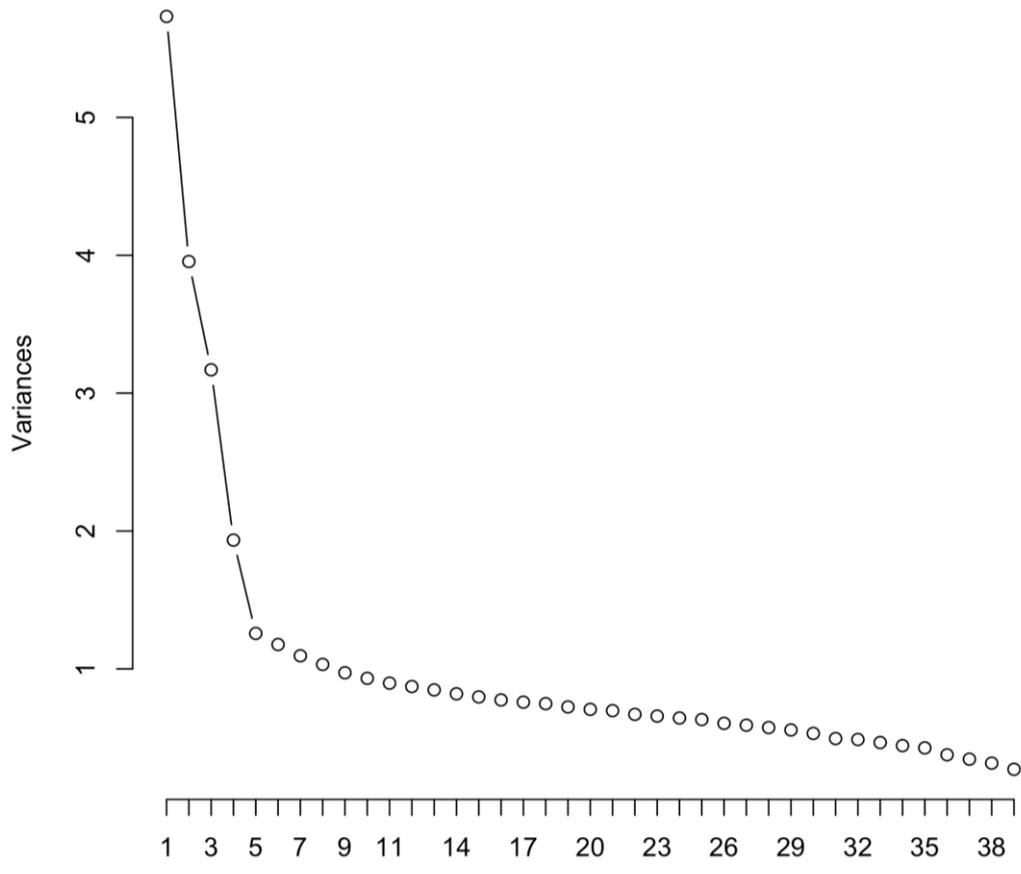
**Table 4.9. List of SPEQ items that match items available in ALSPAC and CATSS.**

<b><i>Paranoia and Hallucinations</i></b>	
<b>Leading statement:</b>	How often have you thought...
<b>Response options:</b>	0 = Not at all, 1 = Rarely, 2 = Once a month, 3 = Once a week, 4 = Several times a week, 5 = Daily
<b>para4</b>	I might be being observed or followed
<b>para12</b>	People might be conspiring against me
<b>para15</b>	I can detect coded messages about me in the press/TV/internet
<b>halluc1</b>	Hear noises or sounds when there is nothing about to explain them?
<b>halluc3</b>	Hear sounds or music that people near you don't hear?
<b>halluc5</b>	See things that other people cannot
<b>halluc7</b>	Sees shapes, lights, or colours even though there is nothing really there?
<b>halluc8</b>	Hear voices commenting on what you're thinking or doing?
<b><i>Anhedonia</i></b>	
<b>Response options:</b>	0 = Very false for me, 1 = Moderately false for me, 2 = Slightly false for me, 3 = Slightly true for me, 4 = Moderately true for me, 6 = Very true for me
<b>Item 1:</b>	When something exciting is coming up in my life, I really look forward to it.
<b>Item 2:</b>	When I'm on my way to an amusement park, I can hardly wait to ride the rollercoasters.
<b>Item 3:</b>	When I think about eating my favourite food, I can almost taste how good it is.
<b>Item 4:</b>	I don't look forward to things like eating out at restaurants.
<b>Item 5:</b>	I get so excited the night before a major holiday I can hardly sleep.
<b>Item 6:</b>	When I think of something tasty, like chocolate biscuit, I have to have one.
<b>Item 7:</b>	Looking forward to a pleasurable experience is in itself pleasurable.
<b>Item 8:</b>	I look forward to a lot of things in my life.
<b>Item 9:</b>	When ordering something off a menu, I imagine how good it will taste.
<b>Item 10:</b>	When I hear about a new movie starring my favourite actor, I can't wait to see it.

Table 4.9 cont.

<b><u>Cognitive Disorganisation</u></b>	
<b>Response options:</b>	0 = Yes, 1 = No
<b>Item 1:</b>	Are you easily confused if too much happens at the same time?
<b>Item 2:</b>	Do you frequently have difficulty in starting to do things?
<b>Item 3:</b>	Are you a person whose mood goes up and down easily?
<b>Item 4:</b>	Do you dread going into a room by yourself where other people have already gathered and are talking?
<b>Item 5:</b>	Do you find it difficult to keep interested in the same thing for a long time?
<b>Item 6:</b>	Do you find it difficult in controlling your thoughts?
<b>Item 7:</b>	Are you easily distracted from work by daydreams?
<b>Item 8:</b>	Do you ever feel that your speech is difficult to understand because the words are all mixed up and don't make sense?
<b>Item 9:</b>	Are you easily distracted when you talk read or talk to someone?
<b>Item 10:</b>	Is it hard for you to make decisions?
<b>Item 11:</b>	When in a crowded room, do you often have difficulty in following a conversation?
<b><u>Parent-rated Negative Symptoms</u></b>	
<b>Leading statement:</b>	My child...
<b>Response options:</b>	0 = Not at all true, 1 = Somewhat true, 2 = Mainly true, 3 = Definitely true
<b>Item 1:</b>	Usually gives brief, one word replies to questions, even if encouraged to say more.
<b>Item 2:</b>	Often does not have much to say for him/herself.
<b>Item 3:</b>	Has few or no friends.
<b>Item 4:</b>	Is often inattentive and appears distracted.
<b>Item 5:</b>	Often does not pay attention when being spoken to.
<b>Item 6:</b>	Often sits around for a long time doing nothing.
<b>Item 7:</b>	Has a lack of energy and motivation.
<b>Item 8:</b>	Has very few interests or hobbies.
<b>Item 9:</b>	Often fails to smile or laugh at things others would find funny.
<b>Item 10:</b>	Seems emotionally "flat", for example, rarely changes the emotions he/she shows.

*Note.* Cognitive Disorganisation, Anhedonia, and Parent-rated Negative Symptoms scales have not been altered from the original scales in the specific psychotic experiences questionnaire (SPEQ).



**Figure 4.4. Scree plot of SPEQ items in TEDS that match psychotic experience items available in the ALSPAC and CATSS samples.**

**Table 4.10. Principal component loadings of psychotic experience items in TEDS.**

	Component 1	Component 2	Component 3	Component 4
para4	0.01	<b>0.52</b>	-0.01	0.09
para12	0.08	<b>0.50</b>	0.03	0.05
para15	0.05	<b>0.40</b>	-0.05	0.00
halluc1	-0.02	<b>0.67</b>	-0.01	0.10
halluc3	-0.01	<b>0.78</b>	-0.01	-0.02
halluc5	0.00	<b>0.79</b>	-0.01	-0.07
halluc7	-0.01	<b>0.74</b>	0.00	0.01
halluc8	0.01	<b>0.67</b>	0.04	0.00
cogdis1	-0.01	-0.03	-0.06	<b>0.63</b>
cogdis2	0.04	-0.02	0.07	<b>0.60</b>
cogdis3	0.01	0.06	0.00	<b>0.48</b>
cogdis4	0.06	0.06	0.01	<b>0.44</b>
cogdis5	0.00	-0.07	0.04	<b>0.54</b>
cogdis6	-0.07	0.18	0.03	<b>0.51</b>
cogdis7	-0.01	0.05	-0.05	<b>0.54</b>
cogdis8	-0.01	0.15	0.03	<b>0.45</b>
cogdis9	0.02	0.01	-0.05	<b>0.61</b>
cogdis10	-0.03	-0.03	-0.01	<b>0.54</b>
cogdis11	0.04	0.03	0.06	<b>0.51</b>
anhed1	0.05	0.04	<b>0.60</b>	0.08
anhed2	-0.01	-0.06	<b>0.64</b>	-0.01
anhed3	0.06	0.05	<b>0.35</b>	0.21
anhed4	0.02	0.01	<b>0.45</b>	0.02
anhed5	-0.01	0.03	<b>0.63</b>	-0.10
anhed6	-0.04	0.02	<b>0.54</b>	-0.23
anhed7	0.00	-0.05	<b>0.67</b>	0.01
anhed8	0.05	0.06	<b>0.61</b>	0.21
anhed9	-0.02	-0.02	<b>0.69</b>	-0.01
anhed10	-0.02	0.00	<b>0.56</b>	-0.06
negsym1	<b>0.54</b>	0.03	0.00	-0.02
negsym2	<b>0.69</b>	0.05	0.02	-0.10
negsym3	<b>0.74</b>	-0.01	0.02	-0.06
negsym4	<b>0.71</b>	-0.01	0.07	-0.04
negsym5	<b>0.66</b>	-0.02	-0.10	0.04
negsym6	<b>0.72</b>	0.00	-0.01	0.07
negsym7	<b>0.65</b>	-0.03	0.02	0.05
negsym8	<b>0.47</b>	0.06	0.09	-0.04
negsym9	<b>0.64</b>	0.01	-0.04	0.04
negsym10	<b>0.65</b>	0.00	-0.04	0.07

*Note.* The correlation between principal components supports the use of oblique rotation. The RMSR is <0.08 indicating this model provides good fit.

**Table 4.11. Cronbach's alpha of paranoia and hallucinations (after converting hallucinations into two items), cognitive disorganisation, anhedonia and parent-rated negative symptom scales in the TEDS sample using matched items.**

<b>Paranoia and Hallucinations</b>	
Full	0.72
excl. para4	0.68
excl. para12	0.68
excl. para15	0.73
excl. Aud_Hall	0.62
excl. Vis_Hall	0.65
<b>Cognitive Disorganisation</b>	
Full	0.77
excl. cogdis1	0.75
excl. cogdis2	0.75
excl. cogdis3	0.76
excl. cogdis4	0.76
excl. cogdis5	0.76
excl. cogdis6	0.75
excl. cogdis7	0.76
excl. cogdis8	0.76
excl. cogdis9	0.75
excl. cogdis10	0.76
excl. cogdis11	0.76

**Table 4.11 cont.**

<b>Anhedonia</b>	
Full	0.78
excl. anhed1	0.76
excl. anhed2	0.75
excl. anhed3	0.78
excl. anhed4	0.77
excl. anhed5	0.75
excl. anhed6	0.77
excl. anhed7	0.75
excl. anhed8	0.76
excl. anhed9	0.75
excl. anhed10	0.76
<b>Parent-rated Negative Symptoms</b>	
Full	0.85
excl. negsym1	0.84
excl. negsym2	0.83
excl. negsym3	0.83
excl. negsym4	0.83
excl. negsym5	0.83
excl. negsym6	0.83
excl. negsym7	0.83
excl. negsym8	0.85
excl. negsym9	0.83
excl. negsym10	0.83

*Note.* Table shows alpha when excluding each item of the scale.

#### **4.4 - Discussion**

This chapter has identified items within three non-clinical adolescent samples assessing a broad range of specific PEs. These PE items have been demonstrated to fall into distinct PE domains, comparable to those previously reported (Ronald et al., 2014; Wigman et al., 2012), and comparable across the samples within this study. The expert clinicians consulted here advised that the PE scales derived within each sample are capturing comparable latent traits across the samples in spite of the PE items within each sample not being identical. The outcome of this phenotypic harmonisation is the derivation of reliable, comparable, quantitative, and specific PE measures within three large-scale non-clinical adolescent samples. These specific PE measures enable the combined analysis of Paranoia and Hallucinations, and Parent-rated Negative Symptoms

across three samples, and Cognitive Disorganisation and Anhedonia across two samples. This provides an opportunity for combined analysis of multiple samples and improved statistical power. Therefore these measures will be used in all subsequent chapters.

The phenotypic harmonisation process raised two important points for consideration.

Firstly, the ALSPAC and CATSS samples were not directly assessed for certain specific PE domains (cognitive disorganisation, anhedonia, and negative symptoms). However, the traits/symptoms relating to different areas of psychopathology frequently overlap, and therefore, items within measures from other areas of psychopathology could be used. For example, the cognitive disorganisation domain of PEs was not directly assessed in CATSS. However, the CATSS sample had administered the Adult Self-report ADHD Scale (ASAS) containing several items that assessed problems in attention and concentration that resembled items within the Cognitive Disorganisation measure in the SPEQ (Ronald et al., 2014). This approach was facilitated by the large range of measures administered to participants of ALSPAC and CATSS, capturing a broad range of traits relating to psychopathology. Consultation with expert clinicians was crucial for validating the derived measures as both content valid and well matched across samples.

A second but related issue is determining the number of items within each measure assessing specific aspects of the trait. This is important for both content validity within samples and for ensuring the comparability of the measures across samples. For example, the negative symptoms PE domain contains several aspects of behaviour including alogia (poverty of speech), asociality, inattention, avolition, and blunted affect. Although PCA identified these items as assessing a single latent factor, based on previous literature and the advice of clinicians, it was deemed important that each specific aspect of negative symptoms have an equal contribution to an individuals' overall score. Therefore there must be a similar proportion of items assessing each specific aspect of the negative symptoms domain. As previously described, this issue

was particularly important for the Parent-rated Negative Symptoms measures in the CATSS sample. Of the 15 items assessing parent-rated negative symptoms, seven assessed asociality, leading to an increased weighting of asocial behaviour relative to other aspects of the domain. Therefore, five asociality items were removed to ensure a more equal contribution of the different aspects of the negative symptoms PE domain, and ensuring comparability across samples.

Although this approach has provided opportunities for the analysis of several specific PEs across several samples, it has not been possible to create Grandiosity measures for ALSPAC and CATSS samples. As a result, this domain of PEs will not be included in subsequent analyses. Furthermore, Cognitive Disorganisation was not assessed in ALSPAC, and Anhedonia was not assessed in CATSS, meaning these PE domains will only be analysed across the two samples with available data. This study has been able to identify a sufficient number of items to highlight the dimensional structure of PEs. However, due to the high correlation between paranoia and hallucinations, and the small number of items for these traits in ALSPAC and CATSS, PCA was unable to separate these specific domains, and it was decided to study these two domains as one.

A consideration is the inability to assess the phenotypic correlation of PE measures between samples. Phenotypic heterogeneity is likely to exist given that these samples were assessed separately using different items (although highly similar in some situations). Molecular genetic methods for estimating the genetic correlation between traits could be used to estimate the extent to which these phenotypes have a genetic overlap. However, methods for the estimation of genetic correlation are underpowered when looking between these samples. Another approach to indicate sample heterogeneity is to compare SNP-heritability estimates for a trait within versus across samples. This will be investigated further in Chapter 6.

In conclusion, this chapter has derived scales within each sample assessing four common latent traits: Paranoia and Hallucinations (TEDS, ALSPAC and CATSS), Anhedonia (TEDS and ALSPAC), Cognitive Disorganisation (TEDS, CATSS), and Parent-rated Negative Symptoms (TEDS, ALSPAC and CATSS). The results of PCA, the approval of expert clinicians, and the similarity of items across samples, collectively support the validity of these scales as assessing adolescent PEs that are both dimension specific and in common across samples. These scales will be used to study specific PEs across TEDS, ALSPAC and CATSS in subsequent chapters.

## 4.5 – Appendix

**Supplementary Table 4.1. List of SPEQ items capturing adolescent PEs.**

<b><u>Paranoia</u></b>	
<b>Leading statement:</b>	How often have you thought...
<b>Response options:</b>	0 = Not at all, 1 = Rarely, 2 = Once a month, 3 = Once a week, 4 = Several times a week, 5 = Daily
<b>para1</b>	I need to be on my guard against others
<b>para2</b>	There might be negative comments being spread about me
<b>para3</b>	People are deliberately trying to irritate me
<b>para4</b>	I might be being observed or followed
<b>para5</b>	People are trying to upset
<b>para6</b>	People are looking at me in an unfriendly way
<b>para7</b>	People are being hostile towards me
<b>para8</b>	Bad things are being said about me behind my back
<b>para9</b>	Someone has bad intentions towards me
<b>para10</b>	Someone has it in for me
<b>para11</b>	People would harm me if given the opportunity
<b>para12</b>	People might be conspiring against me
<b>para13</b>	People are laughing at me
<b>para14</b>	I am under threat from others
<b>para15</b>	I can detect coded messages about me in the press/TV/internet
<b><u>Hallucination</u></b>	
<b>Leading statement:</b>	How often have you thought...
<b>Response options:</b>	0 = Not at all, 1 = Rarely, 2 = Once a month, 3 = Once a week, 4 = Several times a week, 5 = Daily
<b>halluc1</b>	Hear noises or sounds when there is nothing about to explain them?
<b>halluc2</b>	Feel that someone is touching you but when you look nobody is there?
<b>halluc3</b>	Hear sounds or music that people near you don't hear?
<b>halluc4</b>	Detects smells which don't seem to come from your surroundings?
<b>halluc5</b>	See things that other people cannot
<b>halluc6</b>	Experience unusual burning sensations or other strange feeling in or on your body that can't be explained?
<b>halluc7</b>	Sees shapes, lights, or colours even though there is nothing really there?
<b>halluc8</b>	Hear voices commenting on what you're thinking or doing?
<b>halluc9</b>	Notice smells or odours that people next to you seem unaware of?

Supplementary table 4.1 cont.

<b><u>Cognitive Disorganisation</u></b>	
<b>Response options:</b>	0 = Yes, 1 = No
<b>Item 1:</b>	Are you easily confused if too much happens at the same time?
<b>Item 2:</b>	Do you frequently have difficulty in starting to do things?
<b>Item 3:</b>	Are you a person whose mood goes up and down easily?
<b>Item 4:</b>	Do you dread going into a room by yourself where other people have already gathered and are talking?
<b>Item 5:</b>	Do you find it difficult to keep interested in the same thing for a long time?
<b>Item 6:</b>	Do you find it difficult in controlling your thoughts?
<b>Item 7:</b>	Are you easily distracted from work by daydreams?
<b>Item 8:</b>	Do you ever feel that your speech is difficult to understand because the words are all mixed up and don't make sense?
<b>Item 9:</b>	Are you easily distracted when you talk read or talk to someone?
<b>Item 10:</b>	Is it hard for you to make decisions?
<b>Item 11:</b>	When in a crowded room, do you often have difficulty in following a conversation?
<b><u>Anhedonia</u></b>	
<b>Response options:</b>	0 = Very false for me, 1 = Moderately false for me, 2 = Slightly false for me, 3 = Slightly true for me, 4 = Moderately true for me, 6 = Very true for me
<b>Item 1:</b>	When something exciting is coming up in my life, I really look forward to it.
<b>Item 2:</b>	When I'm on my way to an amusement park, I can hardly wait to ride the rollercoasters.
<b>Item 3:</b>	When it think about eating my favourite food, I can almost taste how good it is.
<b>Item 4:</b>	I don't look forward to things like eating out at restaurants.
<b>Item 5:</b>	I get so excited the night before a major holiday I can hardly sleep.
<b>Item 6:</b>	When I think of something tasty, like chocolate biscuit, I have to have one.
<b>Item 7:</b>	Looking forward to a pleasurable experience is in itself pleasurable.
<b>Item 8:</b>	I look forward to a lot of things in my life.
<b>Item 9:</b>	When ordering something off a menu, I imagine how good it will taste.
<b>Item 10:</b>	When I hear about a new movie starring my favourite actor, I can't wait to see it.

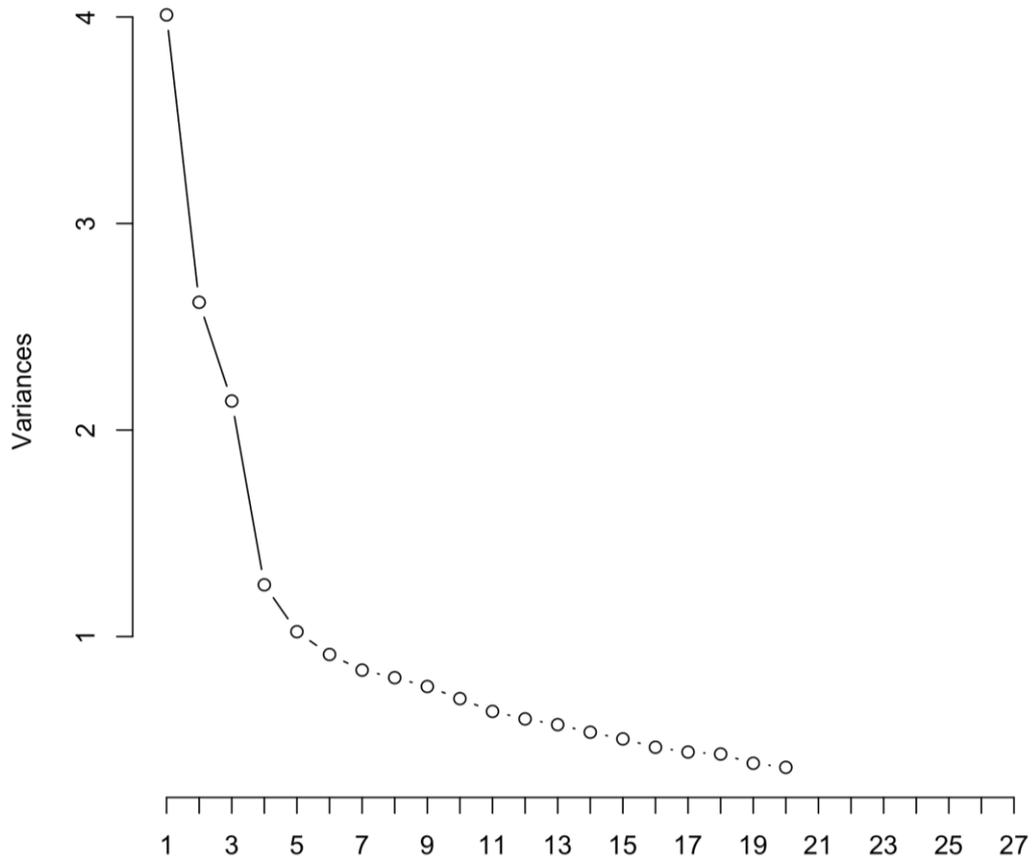
**Supplementary table 4.1 cont.**

<b><i>Parent-rated Negative Symptoms</i></b>	
<b>Leading statement:</b>	My child...
<b>Response options:</b>	0 = Not at all true, 1 = Somewhat true, 2 = Mainly true, 3 = Definitely true
<b>Item 1:</b>	Usually gives brief, one word replies to questions, even if encouraged to say more.
<b>Item 2:</b>	Often does not have much to say for him/herself.
<b>Item 3:</b>	Has few or no friends.
<b>Item 4:</b>	Is often inattentive and appears distracted.
<b>Item 5:</b>	Often does not pay attention when being spoken to.
<b>Item 6:</b>	Often sits around for a long time doing nothing.
<b>Item 7:</b>	Has a lack of energy and motivation.
<b>Item 8:</b>	Has very few interests or hobbies.
<b>Item 9:</b>	Often fails to smile or laugh at things others would find funny.
<b>Item 10:</b>	Seems emotionally "flat", for example, rarely changes the emotions he/she shows .

*Note.* Items are labelled according to specific PE subscale. para = paranoia, halluc = hallucinations, grandi = grandiosity and delusions, cogdis = cognitive disorganisation, anhed = anhedonia, negsym = parent-rated negative symptoms.

**Supplementary Table 4.2. Loadings of CATSS PE items from principal components analysis after removal of excess parent-reported asociality items.**

	<b>Component 1</b>	<b>Component 2</b>	<b>Component 3</b>
para1	-0.13	0.14	<b>0.70</b>
para2	-0.04	-0.16	<b>0.64</b>
para3	0.01	0.00	<b>0.78</b>
halluc1	0.04	0.01	<b>0.72</b>
halluc2	0.13	0.01	<b>0.72</b>
cogdis1	-0.04	<b>0.69</b>	0.02
cogdis2	0.00	<b>0.73</b>	-0.10
cogdis3	-0.02	<b>0.72</b>	0.05
cogdis4	0.01	<b>0.70</b>	0.06
cogdis5	-0.03	<b>0.73</b>	0.03
negsym1	<b>0.31</b>	<b>0.39</b>	-0.09
negsym2	<b>0.63</b>	0.12	-0.04
negsym6	<b>0.61</b>	-0.02	0.06
negsym7	<b>0.69</b>	-0.08	0.02
negsym8	<b>0.70</b>	-0.03	0.02
negsym9	<b>0.75</b>	-0.03	0.00
negsym10	<b>0.46</b>	-0.01	0.10
negsym11	<b>0.37</b>	<b>0.31</b>	0.08
negsym13	<b>0.48</b>	0.04	-0.16
negsym15	0.25	<b>0.36</b>	0.04



**Supplementary Figure 4.1. Scree plot of CATSS PE items after removal of excess parent-reported asociality items.**



# **Chapter 5 - Genome-wide association study of specific psychotic experiences using TEDS, ALSPAC and CATSS samples**

## **5.1 – Introduction**

As discussed in Section 1.6, prior to this thesis there was only one genome-wide association study (GWAS) of adolescent psychotic experiences (PEs)(Zammit et al., 2014). This previous GWAS, based on data collected in the ALSPAC study, reported no variant significantly associated with a binary outcome on a broad positive symptom measure. The statistical power of this study was restricted by its use of a binary and non-specific measure, and a limited sample size (N=3,483). In Chapter 2, a second GWAS of adolescent PEs was performed, which used quantitative measures of specific PEs assessed in the TEDS sample (N=2,978 – 2,997). Although this second GWAS also had limited statistical power due to a small sample size, the use of quantitative and specific PEs could potentially reduce phenotypic heterogeneity compared to the previous study. This second GWAS reported three genome-wide significant loci for specific PEs. To gain further statistical power, larger sample sizes are required. As described in Section 4.1, the main approach to increasing sample size is the pooling of resources across samples with the relevant phenotypic data. In Chapter 4, data from three European general population samples (TEDS, ALSPAC and CATSS) were analysed to create specific measures of PEs that were comparable across samples. This led to the creation of four specific PE measures that were assessed in at least two of the three samples: Paranoia and Hallucinations, Cognitive Disorganisation, Anhedonia, and Parent-rated Negative Symptoms. TEDS, ALSPAC and CATSS have collected genome-wide genotypic data for a subset of their participants, providing an opportunity to perform the largest GWAS of specific adolescent PEs to date.

This chapter aims to identify genetic loci associated with specific adolescent PEs by performing genome-wide association analysis of the four specific PEs assessed across TEDS, ALSPAC, and CATSS. Initially, the genotypic data from each study was harmonised via genotypic imputation to a common reference panel and the application of quality control parameters. Subsequently, the identification of associated genetic variation was performed across all three samples simultaneously (mega-GWAS). In addition, this chapter aimed to identify genes associated with specific adolescent PEs using two approaches (MAGMA and PrediXcan).

## **5.2 – Methods**

### 5.2.1 – Samples

The three samples used in this chapter have been previously described in Section 4.2.1. The exclusion criteria applied within each sample have been detailed in Section 4.2.2.

### 5.2.2 – Measures

The measures used in this chapter are based on the results of Chapter 4. Lists of items for each measure within each sample are in Supplementary Tables 5.1-5.4.

To understand the relationship between the PE domains used in this study and other related behavioural and cognitive domains, the correlations between PEs and anxiety symptoms, depressive symptoms and cognitive ability were assessed. These correlations were assessed primarily using the TEDS sample as only this sample assessed anxiety symptoms, depressive symptoms and cognitive ability at the same time point as PEs. Only the correlation between self-report depressive symptoms and PEs was calculated in all three samples, as this was the only trait of interest that was measured in all three samples at the same time point as PEs.

TEDS assessed anxiety symptoms using the Childhood Anxiety Sensitivity Index (self-report) (Silverman, Fleisig, Rabian, & Peterson, 1991) and Anxiety-Related Behaviours Questionnaire (parent-report)(Hallett, Ronald, Rijdsdijk, & Eley, 2009), depressive symptoms using the Mood and Feelings Questionnaire (self- and parent-report)(Angold, Costello, Messer, & Pickles, 1995), and general cognitive ability using the Mill Hill Vocabulary score and Raven's Progressive Matrices (Raven, Raven, & Court, 1996, 1989). Self-report depressive symptoms were assessed using the Short Mood and Feelings Questionnaire (Angold et al., 1995) in the ALSPAC sample, and the Center for Epidemiological Studies-Depression measure in the CATSS sample (Radloff, 1977).

### 5.2.3 – Handling missing phenotypic data

Individuals with >50% missingness for a given measure were excluded from subsequent analysis of that measure. The remaining missing phenotypic data for each measure was imputed using multiple imputation in R (Buuren & Groothuis-Oudshoorn, 2011). The specified method of imputation was predictive mean matching ('meth = pmm'). Each missing value was replaced with the average of 10 imputations ('m=10') using 50 iterations ('maxit = 50').

### 5.2.4 – Calculation of phenotypic sum scores

Individual scores for each measure were calculated using sum scores. To ensure the equal contribution of each item to the sum score, item response values were rescaled to values between 0 and 1.

### 5.2.5 – Normalisation of phenotypic data and controlling for covariate effects

Sum scores for each psychotic experience subscale were normalised using inverse rank-based normalisation (data ties ranked randomly) and then standardised. This approach of normalisation before controlling for covariate effects was investigated in Chapter 3. The R function used to normalise the phenotypic data (called `rntransform_random`,

Supplementary Note 3.6) was an adaptation of the `rntransform` function from GenABEL that randomly ranks tied observations. The normalised scores were then regressed against the following covariates: sex, age, age<sup>2</sup>, sex\*age, sex\*age<sup>2</sup>, study, and the top 8 principal components of ancestry. Principal components of ancestry were jointly calculated across all three samples using PLINK1.9 (<https://www.cog-genomics.org/plink2>) based only on observed genetic variation (number of linkage disequilibrium (LD) independent genetic markers = 49,596).

All stages of this phenotypic preparation procedure and calculation of ancestry covariates (Section 5.2.3 – 5.2.5) were carried out for the purposes of this thesis.

#### *5.2.6 – DNA collection and genotyping*

**TEDS:** DNA was extracted using buccal cheek swabs. Genotyping was performed using Affymetrix GeneChip 6.0 SNP genotyping platform at the Affymetrix Santa Clara, California, USA, as part of the TEDS Wellcome Trust Case Control Consortium 2 (WTCCC2) study of reading and mathematical abilities. Samples were excluded based on one or more of the following parameters: low call rate or heterozygosity outliers, atypical population ancestry, sample duplication or relatedness to other sample members, unusual hybridisation intensity, gender mismatches, and having less than 90% of genotypes called identically on the genome-wide array and Sequenom panel. This resulted in 3,152 unrelated individuals being successfully genotyped consisting of 1,446 males and 1,706 females. Individuals from monozygotic (MZ) twin pairs could be used to impute the genotype of their sibling based on the assumption that MZ twins are genetically identical. Given that both siblings of these genotyped MZ twin pairs provided phenotypic information, they were both included in subsequent analyses, accounting for family structure (Minică et al., 2015).

**ALSPAC:** DNA was extracted from umbilical cord blood. Genotyping was performed using the Illumina HumanHap550 quad genome-wide SNP genotyping platform by

23andMe subcontracting the Wellcome Trust Sanger Institute, Cambridge, UK and the Laboratory Corporation of America, Burlington NC, USA. Samples were excluded based on the following criteria: incorrect gender assignments, heterozygosity outliers, low call rate, cryptic relatedness and being of non-European ancestry as detected by a multidimensional scaling analysis seeded with HapMap 2 individuals. Subsequently, 8,365 unrelated individuals survived quality control consisting of 4,285 males and 4,080 females.

**CATSS:** DNA was extracted from saliva. Genotyping was performed using the Illumina Infinium PsychArray-24 BeadChip and carried out by SNP&SEQ Technologies in Uppsala, Sweden. Samples were excluded based on one or more of the following parameters: Low call rate, excess heterozygosity, sample duplication, erroneous within family relatedness, cryptic relatedness, gender mismatches, being non-European as detected by principal components analysis seeded by the 1KG EUR (1000 genomes European) reference. This resulted in 17,898 individuals successfully genotyped. This number includes dizygotic (DZ) twin pairs. As in TEDS, the genotypes of MZ siblings were imputed if their co-twin was genotyped. If both siblings of these genotyped twin pairs provided phenotypic information, they were both included in subsequent analyses, while accounting for family structure.

All DNA collection and genotyping was carried out prior to this project. MZ sibs had not been inserted into the genotypic data prior to this study.

#### 5.2.7 – Genotype imputation and quality control procedure

This genotype imputation and quality control process was carried out for this study.

Genotype imputation and harmonisation was performed using only autosomes. The reference panel chosen for imputation was the 1KG Phase 3 version 5 dataset. Stringent quality control and strand alignment (ambiguous SNPs excluded) was performed prior

to imputation. The data was phased using ShapeIt V2 and subsequently imputed in 5Mb sections using Impute2 (B. Howie, Marchini, & Stephens, 2011; B. N. Howie, Donnelly, & Marchini, 2009). In all cases the call accuracy of imputation was between .90 and .99. SNPs with poor INFO scores were removed (INFO < .3). For ease of subsequent analysis, the dosage levels of imputed SNPs were converted to 'hard calls' using a threshold of >0.9 with the intention of following up associations of interest accounting for imputation probabilities.

The 'hard call' genotype data from each sample were merged per chromosome using PLINK1.9. A light SNP-missingness threshold (>20%) was applied to remove SNPs that were neither imputed nor observed in the samples. A second round of stringent SNP- and individual-level QC was then applied: SNP missingness > 2%, individual missingness >5%, minor allele frequency (MAF) < 1%, Hardy-Weinberg equilibrium  $p < 1 \times 10^{-6}$ . After QC 4,487,870 common variants were captured in each of the samples.

#### 5.2.8 – Mega-genome-wide association analysis

After phenotypic and genotypic harmonisation, the three samples were combined to enable genome-wide association mega-analysis of four specific PEs with the following sample sizes (including siblings): Paranoia and Hallucinations = 8,665, Anhedonia = 6,579, Cognitive Disorganisation = 6,297, and Parent-rated Negative Symptoms = 10,098. Genome-wide association analysis of all four PE domains using related (i.e. monozygotic and dizygotic twin pairs) and unrelated individuals was performed in PLINK (<http://pngu.mgh.harvard.edu/purcell/plink/>) (Purcell et al., 2007). As described in Chapter 2, additional covariance arises from related individuals; this was accounted for using the method of generalised estimating equations (GEE) (Minica et al., 2014; Minică et al., 2015) in R specifying an exchangeable correlation matrix.

Power calculations showed that the mega-GWASs of PEs had a sufficient sample sizes to detect reasonable effect sizes (Table 5.1). Power calculations show that the  $r^2$  (effect

size) that could be detected with 80% power is 0.50% for Paranoia and Hallucinations, 0.65% for Anhedonia, 0.78% for Cognitive Disorganisation, and 0.45% for Parent-rated Negative Symptoms.

#### 5.2.9 – Meta-genome-wide association study

To evaluate the effect of mega-analysis rather than meta-analysis, the four PEs were also analysed post hoc within each sample and then meta-analysed using inverse-variance weighting in METAL (Willer, Li, & Abecasis, 2010). To evaluate the effect of randomly splitting ties when normalising the phenotypic data, additional post hoc meta-GWASs of normalised specific PEs, but averaging tied observations, were also carried out using METAL. The results of this analysis are considered less robust due to the remaining skew of the dependent variables, particularly for Paranoia and Hallucinations, and Parent rated Negative Symptoms.

**Table 5.1. Power to detect association at genome-wide significance for mega-genome-wide association analyses of psychotic experience traits.**

<b>Psychotic Experience</b>	<b><i>N</i></b>	<b><i>r</i><sup>2</sup></b>	<b>Power</b>
Paranoia and Hallucinations	7970	0.005	0.81
Paranoia and Hallucinations	7970	0.01	1.00
Anhedonia	6068	0.005	0.52
Anhedonia	6068	0.01	0.99
Cognitive Disorganisation	5083	0.005	0.34
Cognitive Disorganisation	5083	0.01	0.95
Parent-rated Negative Symptoms	8763	0.005	0.88
Parent-rated Negative Symptoms	8763	0.01	1.00

*Note.* *N* = effective sample size after accounting for related individuals (Minica et al., 2014); *r*<sup>2</sup>, variance explained in the outcome by a tagged SNP.

#### 5.2.10 – Replication analysis of genetic associations

After the TEDS+CATSS+ALSPAC sample had been analysed, genotypic data for additional TEDS participants became available as a result of a second wave of genotyping of further individuals in 2016. These individuals were used as a replication sample for any variants or genes associated with adolescent PEs at significance. This replication sample was imputed using the haplotype reference consortium data via the Sanger imputation server (McCarthy et al., 2016). Eva Krapohl, a member of the TEDS team, led the imputation process. The imputed genotypic data was converted to hard-call format (certainty threshold of 0.9) in PLINK. Replication analyses were performed using the same procedure as the discovery analyses.

#### 5.2.11 – Gene-region association analysis

MAGMA (de Leeuw, Mooij, Heskes, & Posthuma, 2015), using the mean of the  $X^2$ , aggregated the association of SNPs within specified gene regions based on the summary statistics of the PE GWASs. This summary statistic approach was used as it is computationally efficient and the multiple linear principal component regression model doesn't account for the presence of related individuals. The mean  $X^2$  approach is equivalent to those employed in VEGAS, PLINK's --set model, and SKAT when using inverse-variance weights. Genetic variants were assigned to genes based on the NCBI 37.3 build with a 10kb annotation window used around genes, resulting in 17,226 genes being tested. LD was calculated using the combined TEDS, ALSPAC, and CATSS sample. To account for multiple testing,  $p$ -values were Bonferroni corrected using the number of genes tested.

#### 5.2.12 – Predicted differential gene expression analysis

Like complex phenotypes, the expression (transcription) of a given gene is regulated by the interplay of genetic and environmental factors. Genetic variants associated with

differential gene expression are called eQTLs (expression quantitative trait loci). Depending on the variance in the expression of a gene explained by tagged genetic variation, it is possible to predict the expression of a gene based on the genotype of individuals at eQTL sites. The gene expression predicted by eQTL variation is called 'genetically regulated expression', and is distinct from gene expression differences that are a consequence of the phenotype of interest, and distinct from gene expression differences due to other factors (including environmental).

PrediXcan (Gamazon et al., 2015) was used to predict genetically regulated expression levels in the prefrontal cortex of individuals in TEDS, ALSPAC and CATSS. This was achieved using a tissue-specific additive gene-expression prediction models that have been trained using the reference transcriptomic dataset called GTEX Tissue Expression, (<http://www.gtexportal.org/home/>) project. Linear regression was then used to test for an association between predicted gene-expression levels and individual PEs, using GEE to control for related individuals. Due to the small sample size used to create the prediction models for the frontal cortex (N = 92), the expression levels of many genes could not be reliably predicted. After removal of genes with expression levels showing no variance, the total number of genes with predicted expression levels was 2,769. To account for multiple testing, *p*-values were Bonferroni corrected using the number of genes tested.

Given the relatively poor ability to predict gene expression in the frontal cortex, and the 19.2% overlap between eQTLs for the blood and brain (McKenzie, Henders, Caracella, Wray, & Powell, 2014), it was of interest to test for differential gene expression in whole blood. Gene expression prediction models for whole blood were from the DGN (Depression Genes and Networks) reference (N=922). This analysis enabled the prediction of 10,699 genes. This analysis of differential gene expression in blood was

post-hoc and for comparison with the frontal cortex results. To help interpret the results,  $p$ -values were Bonferroni corrected to account for multiple testing.

## **5.3 – Results**

### 5.3.1 – Descriptive of individual psychotic experience scores

Descriptive statistics for the raw PE sum scores within each sample are shown in Table 5.2. The relationships between age, sex and the raw PE measures in each sample are presented in Tables 5.3 and 5.4. The correlation between adolescent PEs and anxiety symptoms, depressive symptoms and cognitive ability in the TEDS sample are presented in Table 5.5. The correlations between self-reported depressive symptoms and PEs in TEDS, ALSPAC and CATSS are presented in Table 5.6. The median correlation between a given scale before and after normalisation when randomly splitting ties was 0.92 (Table 5.7), dependent on skew of the original scale.

**Table 5.2. Descriptive statistics for raw psychotic experience domain sum scores in each sample.**

<b>TEDS</b>							
<b>Specific PE</b>	<b><math>\mu</math></b>	<b><math>\sigma</math></b>	<b>Range</b>	<b>Skew</b>	<b><math>\alpha</math></b>	<b><i>N</i></b>	<b><i>N</i> sibs</b>
Paranoia and Hallucinations	0.43	0.51	0 - 3.90	2.03	0.73	2994	827
Anhedonia	3.19	1.54	0 - 9.80	0.53	0.79	2988	821
Cognitive Disorganization	3.83	2.82	0 - 11.00	0.51	0.77	2987	823
Parent-rated Negative Symptoms	0.88	1.19	0 - 9.33	2.35	0.84	2995	833
<b>ALSPAC</b>							
<b>Specific PE</b>	<b><math>\mu</math></b>	<b><math>\sigma</math></b>	<b>Range</b>	<b>Skew</b>	<b><math>\alpha</math></b>	<b><i>N</i></b>	<b><i>N</i> sibs</b>
Paranoia and Hallucinations	0.27	0.51	0 - 4.17	2.78	0.67	3591	0
Anhedonia	1.42	1.19	0 - 7.00	1.09	0.73	3591	0
Parent-rated Negative Symptoms	1.57	1.11	0 - 7.00	0.61	0.61	4019	0
<b>CATSS</b>							
<b>Specific PE</b>	<b><math>\mu</math></b>	<b><math>\sigma</math></b>	<b>Range</b>	<b>Skew</b>	<b><math>\alpha</math></b>	<b><i>N</i></b>	<b><i>N</i> sibs</b>
Paranoia and Hallucinations	0.20	0.43	0 - 5.00	3.92	0.73	2080	849
Cognitive Disorganization	1.95	0.98	0 - 5.00	0.35	0.79	3310	1335
Parent-rated Negative Symptoms	0.57	0.95	0 - 8.50	2.64	0.75	3084	1395

*Note.* These figures are based on the individuals remaining after quality control and were used in all subsequent genetic analyses.  $\mu$ , mean;  $\sigma$ , standard deviation; *N*, Number of genotyped individuals after exclusions; *N* sibs, number of siblings pairs within *N*; TEDS, Twins Early Development Study; ALSPAC, Avon Longitudinal Study of Parents and Children; CATSS, Child and Adolescent Twin Study in Sweden.

**Table 5.3. Pearson’s correlation between raw PE sum scores and age.**

	<b>TEDS</b>	<b>ALSPAC</b>	<b>CATSS</b>
<b>Paranoia and Hallucinations</b>	0.030	0.004	-0.038
<b>Anhedonia</b>	-0.018	-0.025	NA
<b>Cognitive Disorganisation</b>	0.023	NA	0.020
<b>Parent-rated Negative Symptoms</b>	-0.043	0.001	-0.042

*Note.* These figures are based on the individuals remaining after quality control and were used in all subsequent genetic analyses. TEDS, Twins Early Development Study; ALSPAC, Avon Longitudinal Study of Parents and Children; CATSS, Child and Adolescent Twin Study in Sweden.

**Table 5.4. Mean sex differences for untransformed psychotic experience sum scores.**

**TEDS**

	Males		Females		<i>p</i>
	$\mu$	SD	$\mu$	SD	
<b>Paranoia and Hallucinations</b>	0.408	0.497	0.441	0.527	0.081
<b>Anhedonia</b>	3.630	1.544	2.866	1.459	4.85x10 <sup>-41</sup>
<b>Cognitive Disorganisation</b>	3.237	2.622	4.254	2.888	2.78x10 <sup>-23</sup>
<b>Parent-rated Negative Symptoms</b>	0.990	1.246	0.794	1.134	1.11x10 <sup>-5</sup>

**ALSPAC**

	Males		Females		<i>p</i>
	$\mu$	SD	$\mu$	SD	
<b>Paranoia and Hallucinations</b>	0.180	0.418	0.327	0.559	3.34x10 <sup>-19</sup>
<b>Anhedonia</b>	1.259	1.116	1.537	1.218	1.87x10 <sup>-12</sup>
<b>Parent-rated Negative Symptoms</b>	1.681	1.122	1.467	1.083	7.77x10 <sup>-10</sup>

**CATSS**

	Males		Females		<i>p</i>
	$\mu$	SD	$\mu$	SD	
<b>Paranoia and Hallucinations</b>	0.154	0.381	0.233	0.461	1.94x10 <sup>-5</sup>
<b>Cognitive Disorganization</b>	1.824	0.939	2.048	0.998	4.24x10 <sup>-11</sup>
<b>Parent-rated Negative Symptoms</b>	0.628	0.981	0.525	0.909	2.64x10 <sup>-3</sup>

*Note.* Mean difference *p*-values were estimated using the two sample t-test. These figures are based on the individuals remaining after quality control and were used in all subsequent genetic analyses. TEDS, Twins Early Development Study; ALSPAC, Avon Longitudinal Study of Parents and Children; CATSS, Child and Adolescent Twin Study in Sweden.  $\mu$ , mean.

**Table 5.5. Pearson correlation between PEs and anxiety symptoms, depressive symptoms and cognitive ability in TEDS.**

	Depression <sup>a</sup>	Anxiety <sup>a</sup>	Depression(P) <sup>a</sup>	Anxiety(P) <sup>a</sup>	Cognitive Ability <sup>b</sup>
<b>Paranoia and Hallucinations</b>	0.457**	0.396**	0.199**	0.152**	-0.071*
<b>Anhedonia</b>	0.106**	-0.109**	0.110**	0.070**	0.085*
<b>Cognitive Disorganisation</b>	0.525**	0.493**	0.252**	0.243**	-0.091*
<b>Parent-rated Negative Symptoms</b>	0.195**	0.052*	0.498**	0.491**	-0.079*

*Note.* \* =  $p \leq 0.05$ , \*\* =  $p \leq 0.001$ . <sup>a</sup> = Between 2146 – 2162 individuals in analyses. <sup>b</sup> = Between 925 – 928 individuals in analyses. (P) = Parent-report. These figures are based on unrelated TEDS individuals with genetic data.

**Table 5.6. Pearson correlation between PEs and self-reported depressive symptoms in TEDS, ALSPAC and CATSS.**

	TEDS <sup>a</sup>	ALSPAC <sup>b</sup>	CATSS <sup>c</sup>
<b>Paranoia and Hallucinations</b>	0.457**	0.358**	0.293**
<b>Anhedonia</b>	0.106**	0.801**	NA
<b>Cognitive Disorganisation</b>	0.525**	NA	0.454**
<b>Parent-rated Negative Symptoms</b>	0.195**	0.211**	0.271**

*Note.* \* =  $p \leq 0.05$ , \*\* =  $p \leq 0.001$ . <sup>a</sup> = Between 2147 – 2162 individuals in analyses. <sup>b</sup> = Between 2958 – 3505 individuals in analyses. <sup>c</sup> = Between 1214 – 1777 individuals in analyses. These figures are based on unrelated individuals with genetic data.

**Table 5.7. Pearson’s correlations between raw sum scores and scores after inverse-rank based normalisation splitting ties randomly.**

Sample	Psychotic Experience	Correlation
TEDS	Paranoia and Hallucinations	0.887
TEDS	Anhedonia	0.990
TEDS	Cognitive Disorganisation	0.971
TEDS	Parent-rated Negative Symptoms	0.858
ALSPAC	Paranoia and Hallucinations	0.771
ALSPAC	Anhedonia	0.955
ALSPAC	Parent-rated Negative Symptoms	0.979
CATSS	Paranoia and Hallucinations	0.723
CATSS	Cognitive Disorganisation	0.992
CATSS	Parent-rated Negative Symptoms	0.810

5.3.2 – Mega-genome-wide association study of specific adolescent psychotic experiences

The mega-GWAS identified no genome-wide significant variation for the Paranoia and Hallucinations, Cognitive Disorganisation, or parent-rated Negative Symptoms domains. The mega-GWAS of Anhedonia identified one SNP (rs149957215) associated at genome-wide significance ( $p = 3.76 \times 10^{-8}$ ) within a gene called indoleamine 2,3-dioxygenase 2 (*IDO2*) (Figure 5.1). This SNP was imputed in all three samples with an average imputation quality (INFO) of 0.89 and minor allele frequency of 0.013. Examination of rs149957215 in the gnomAD database (Lek et al., 2016) reported a similar minor allele frequency among European individuals (with and without Finnish individuals) and good genotype quality, supporting that imputation of this variant was accurate (results available here <http://gnomad.broadinstitute.org/variant/8-39872495-C-A>). Due to limited LD with rs149957215, neighbouring genetic variation showed no significant evidence of association. However, the association between rs149957215 and Anhedonia did not replicate in the independent TEDS replication sample ( $N = 2,359$  incl. 635 MZ pairs), showing a non-significant association ( $p=0.81$ ) in the opposite direction. The replication sample provided a power of 0.86 to detect an association of the same

magnitude ( $r^2 = 0.47\%$ ) at nominal significance. rs149957215 was well imputed in the replication sample with a MACH  $r^2 = 0.93$ .

The absence of more genome-wide significant associations indicates that the variance explained by individual genetic variants are substantially smaller than the  $r^2$  values that could have been detected with 80% power ( $r^2$  of 0.45% - 0.78%).

The mega-GWASs of the four psychotic experience domains returned several loci achieving a suggestive significance of  $p < 1 \times 10^{-5}$  (Table 5.8, Figures 5.2-5.5). There was no evidence of confounding with lambdas of 0.99–1.01 and LD-score regression intercept of 1.00 in all analyses (Figure 5.6).

The association statistics of genome-wide significant genetic markers in the previous TEDS only GWAS of Chapter 2 are in Table 5.9. The direction of effect remained consistent in the mega-GWASs. However, no variant remained genome-wide significant. rs7830364 within CSMD1 remained nominally significant.

The direction and  $p$ -value of suggestive loci from the PEs GWASs are compared to those in the most recent schizophrenia, bipolar disorder, and major depressive disorder GWASs in Tables 5.10-5.12 respectively. Of the 29 independent suggestive loci, two achieved nominal significance in the schizophrenia GWAS (both with same direction of effect), three achieved nominal significance in the bipolar disorder GWAS (two with same direction of effect), and one achieved nominal significance in the major depression GWAS (with same direction of effect).

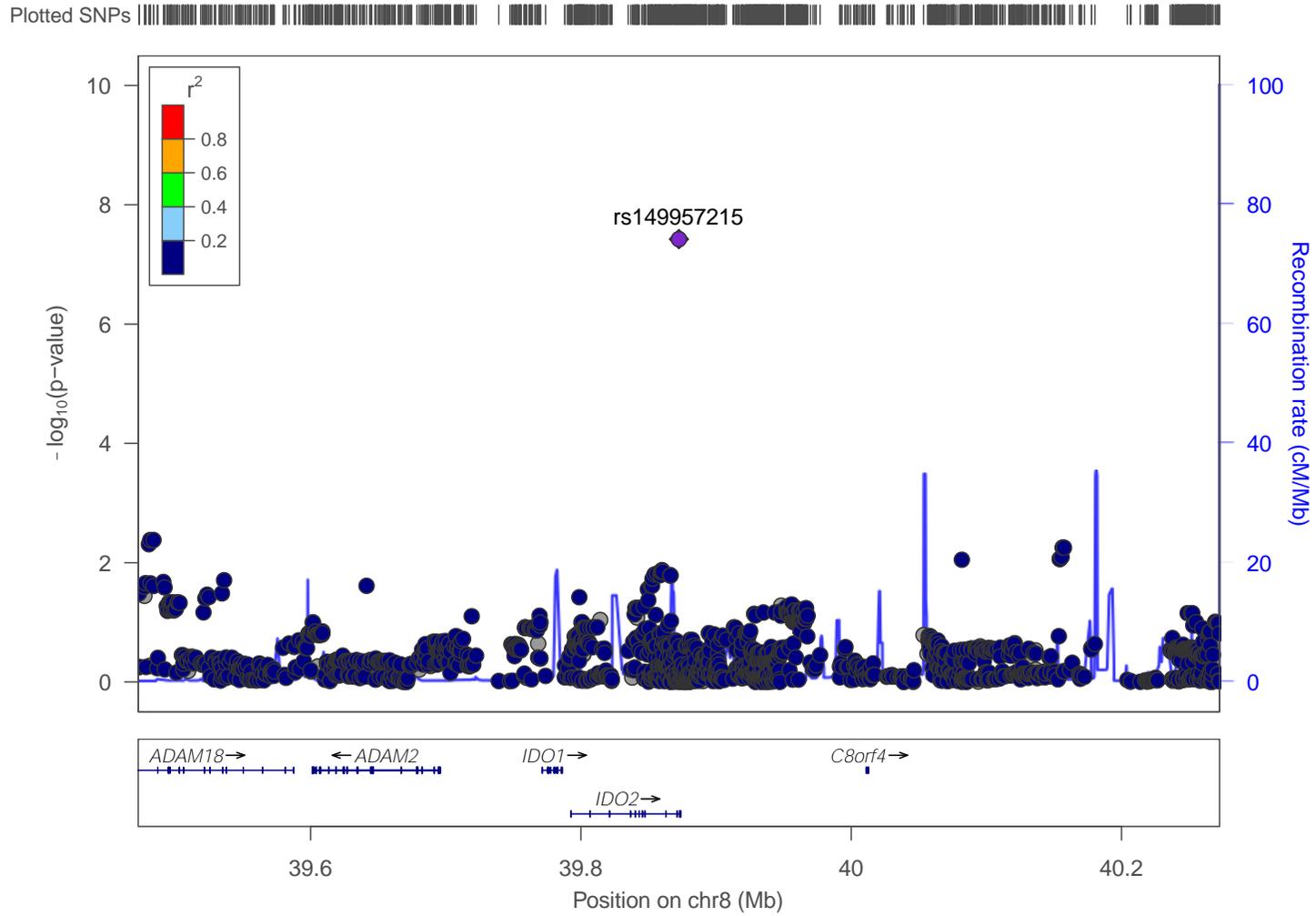


Figure 5.1. LocusZoom plot of genome-wide significant SNP for Anhedonia.

**Table 5.8. Independent loci achieving suggestive significance ( $p < 1 \times 10^{-5}$ ) in mega-genome-wide association study of psychotic experience domains.**

***Paranoia and Hallucinations***

CHR	SNP	BP	A1	A2	MAF	BETA	SE	<i>p</i>	Nearest Gene
16	rs8064063	7214065	C	T	0.290	0.086	0.017	$7.18 \times 10^{-7}$	RBFOX1
2	rs7584721	230401082	A	G	0.078	0.144	0.030	$1.23 \times 10^{-6}$	DNER
1	rs113892704	26675261	A	G	0.198	-0.093	0.020	$3.82 \times 10^{-6}$	AIM1L
5	rs1392391	173683914	A	G	0.122	0.111	0.024	$4.80 \times 10^{-6}$	HMP19
8	rs17749393	96553216	G	A	0.071	0.138	0.031	$6.18 \times 10^{-6}$	C8orf37-AS1
2	rs72919716	78897267	G	A	0.207	-0.089	0.020	$7.72 \times 10^{-6}$	REG3G
8	rs77291092	67335395	A	C	0.064	-0.140	0.031	$8.49 \times 10^{-6}$	RRS1-AS1

***Anhedonia***

CHR	SNP	BP	A1	A2	MAF	BETA	SE	<i>p</i>	Nearest Gene
8	rs149957215	39872495	A	C	0.013	-0.417	0.076	$3.76 \times 10^{-8}$	IDO2
13	rs78013746	61682703	A	C	0.027	0.255	0.054	$2.37 \times 10^{-6}$	MIR3169
6	rs200488	27795109	T	C	0.018	0.297	0.063	$2.89 \times 10^{-6}$	HIST1H4K
11	rs117907077	11033989	A	G	0.024	-0.286	0.062	$3.27 \times 10^{-6}$	ZBED5-AS1
6	rs2531815	28436060	T	C	0.287	0.092	0.020	$4.36 \times 10^{-6}$	ZSCAN23
20	rs6033026	11059873	G	A	0.240	-0.095	0.021	$5.39 \times 10^{-6}$	LOC339593
15	rs7164838	34967574	A	G	0.317	0.088	0.019	$5.55 \times 10^{-6}$	GJD2
11	rs2169485	41079587	G	A	0.178	-0.109	0.024	$6.02 \times 10^{-6}$	LRRC4C
10	rs11195810	113835240	A	G	0.017	0.297	0.066	$6.72 \times 10^{-6}$	GPAM
14	rs12897386	72471862	C	T	0.256	0.093	0.021	$7.62 \times 10^{-6}$	RGS6
9	rs62545506	73241253	T	G	0.104	-0.132	0.030	$8.15 \times 10^{-6}$	TRPM3
14	rs34420225	94290014	A	C	0.212	-0.100	0.022	$9.38 \times 10^{-6}$	PRIMA1
15	rs74519172	55010305	T	C	0.065	0.159	0.036	$9.76 \times 10^{-6}$	UNC13C

Table 5.8 cont.

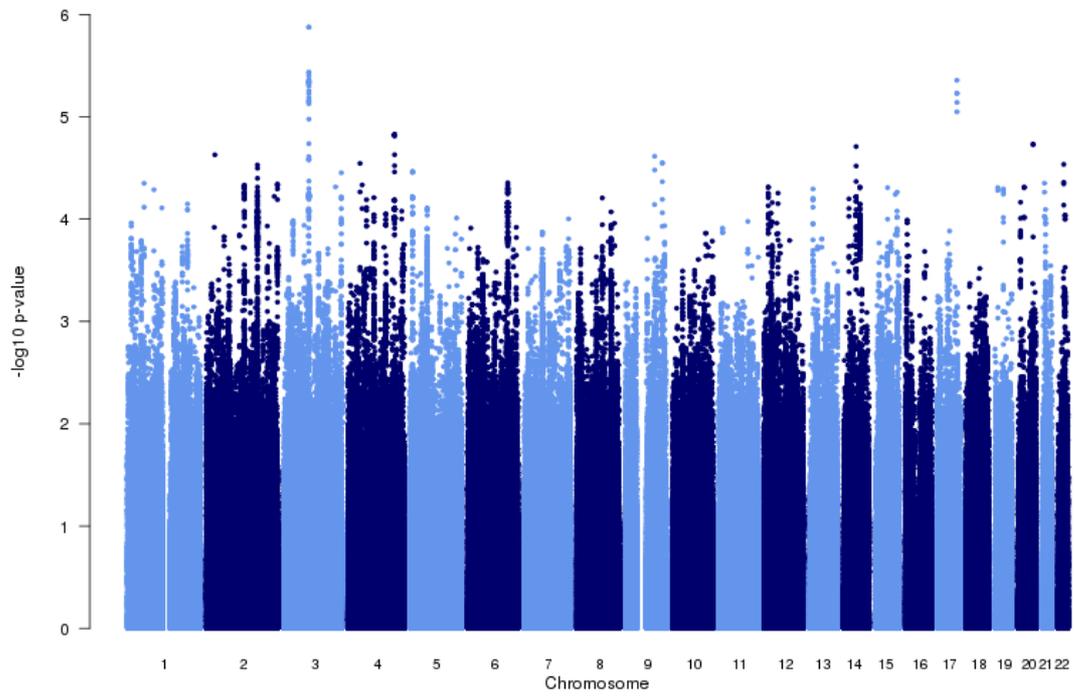
***Cognitive Disorganisation***

CHR	SNP	BP	A1	A2	MAF	BETA	SE	<i>p</i>	Nearest Gene
13	rs1961120	28833372	C	G	0.386	-0.099	0.020	9.00x10 <sup>-7</sup>	PAN3
2	rs200022365	186855226	T	TTTA	0.432	0.097	0.021	2.30x10 <sup>-6</sup>	LOC101927217
2	rs80033666	170682319	C	T	0.089	-0.156	0.034	3.35x10 <sup>-6</sup>	METTL5
2	rs1517844	192405134	C	A	0.400	0.091	0.020	3.95x10 <sup>-6</sup>	NABP1
1	rs6665300	65429558	C	T	0.014	0.376	0.082	4.16x10 <sup>-6</sup>	JAK1
4	rs1911103	126449558	T	C	0.077	-0.164	0.036	5.58x10 <sup>-6</sup>	MIR2054
2	rs7588854	80339218	A	G	0.169	0.119	0.026	6.67x10 <sup>-6</sup>	CTNNA2
3	rs185642755	85127281	C	T	0.012	-0.361	0.081	8.39x10 <sup>-6</sup>	CADM2

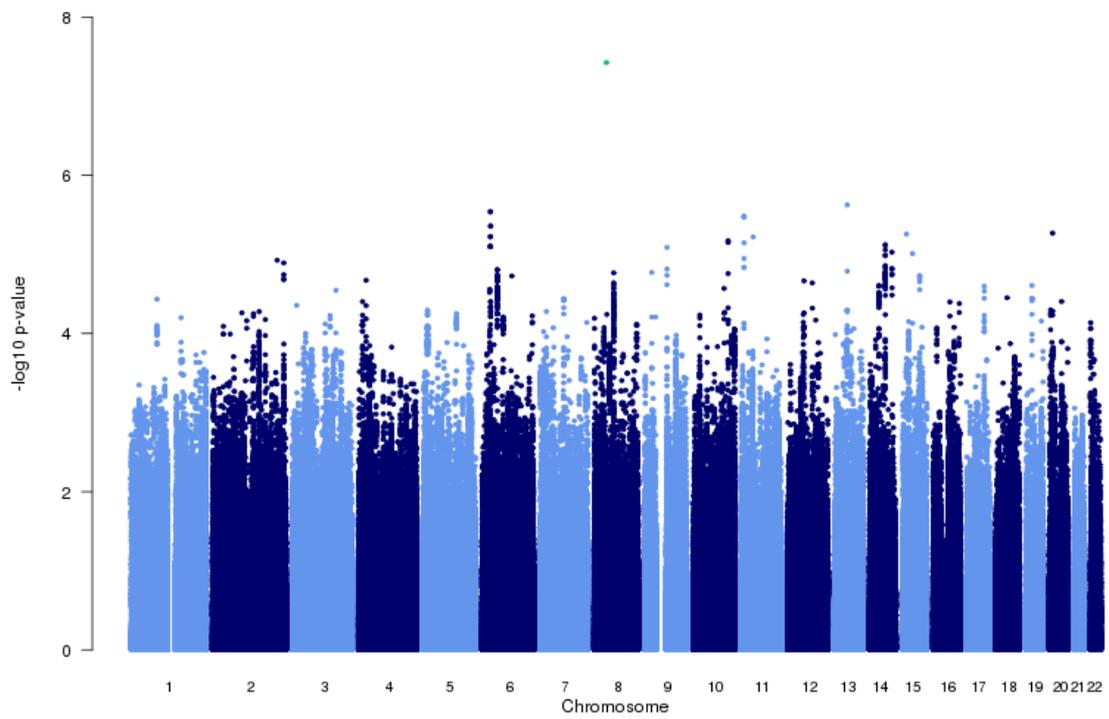
***Parent-rated Negative Symptoms***

CHR	SNP	BP	A1	A2	MAF	BETA	SE	<i>p</i>	Nearest Gene
4	rs4400001	38212771	A	C	0.391	0.080	0.015	2.26x10 <sup>-7</sup>	TBC1D1
8	rs72334712	108862133	CT	C	0.046	0.165	0.034	1.34x10 <sup>-6</sup>	RSPO2
11	rs530272	30519602	C	T	0.017	0.267	0.056	2.20x10 <sup>-6</sup>	MPPED2
2	2:212086966:C:A	212086966	A	C	0.037	0.181	0.039	4.26x10 <sup>-6</sup>	ERBB4
16	rs10500326	4918326	T	G	0.226	-0.082	0.018	5.42x10 <sup>-6</sup>	UBN1
12	rs11063280	4760229	G	A	0.267	0.078	0.017	5.75x10 <sup>-6</sup>	LOC101928989
11	rs12418804	81598184	G	C	0.064	0.140	0.031	6.11x10 <sup>-6</sup>	LOC101928989
6	rs77105684	71903122	G	A	0.026	-0.199	0.044	6.58x10 <sup>-6</sup>	OGFRL1
2	rs114733161	211978580	A	C	0.035	0.181	0.041	8.34x10 <sup>-6</sup>	ERBB4

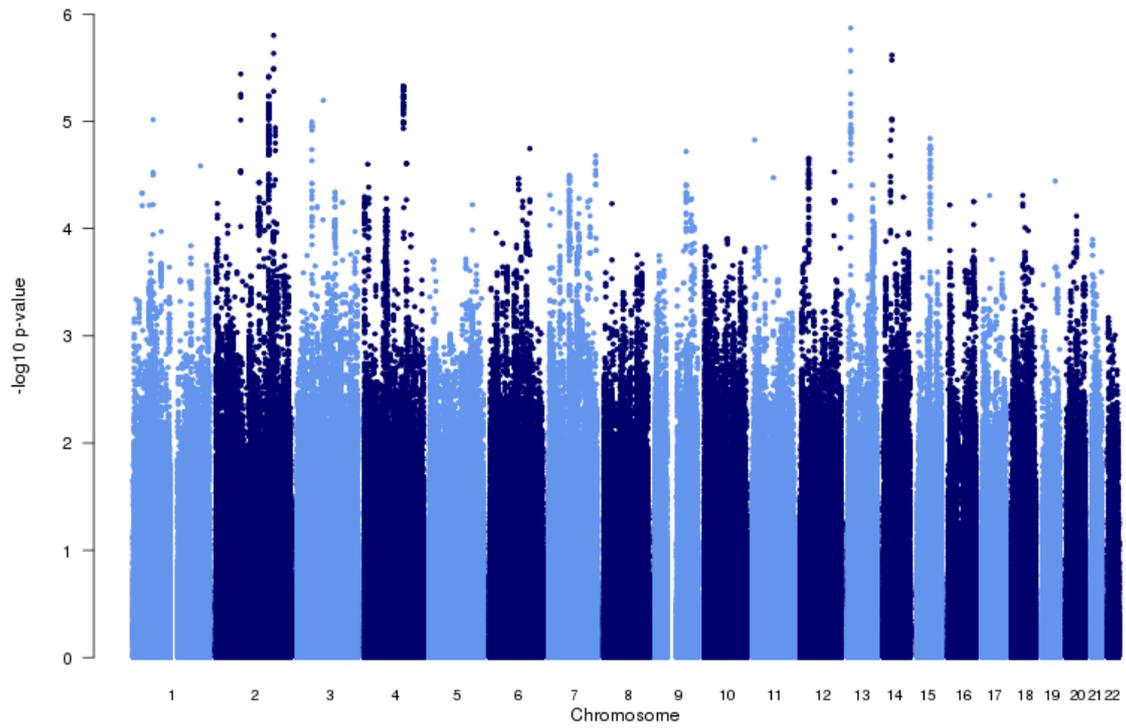
Note. CHR, chromosome number; BP, base-pair position; A1, allele 1 (test allele); A2, allele 2; MAF, minor allele frequency; BETA, unstandardized effect size; Nearest Gene, gene symbol for nearest gene to the lead SNP of each LD independent locus.



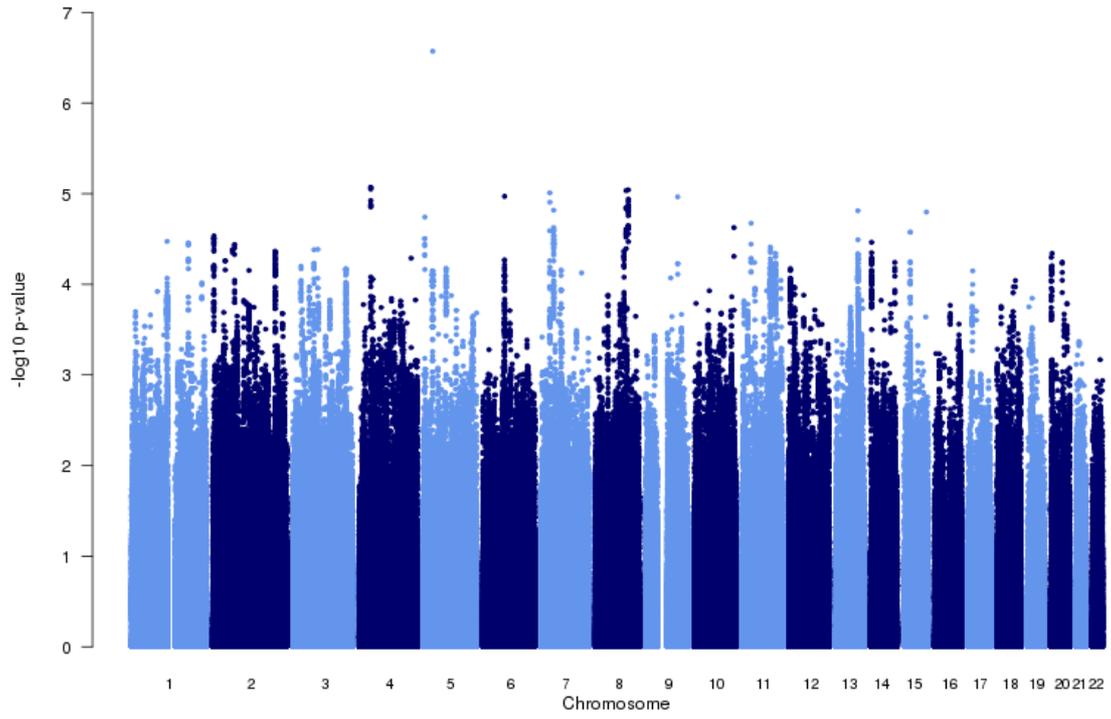
**Figure 5.2. Manhattan plot of Paranoia and Hallucinations mega-GWAS.**



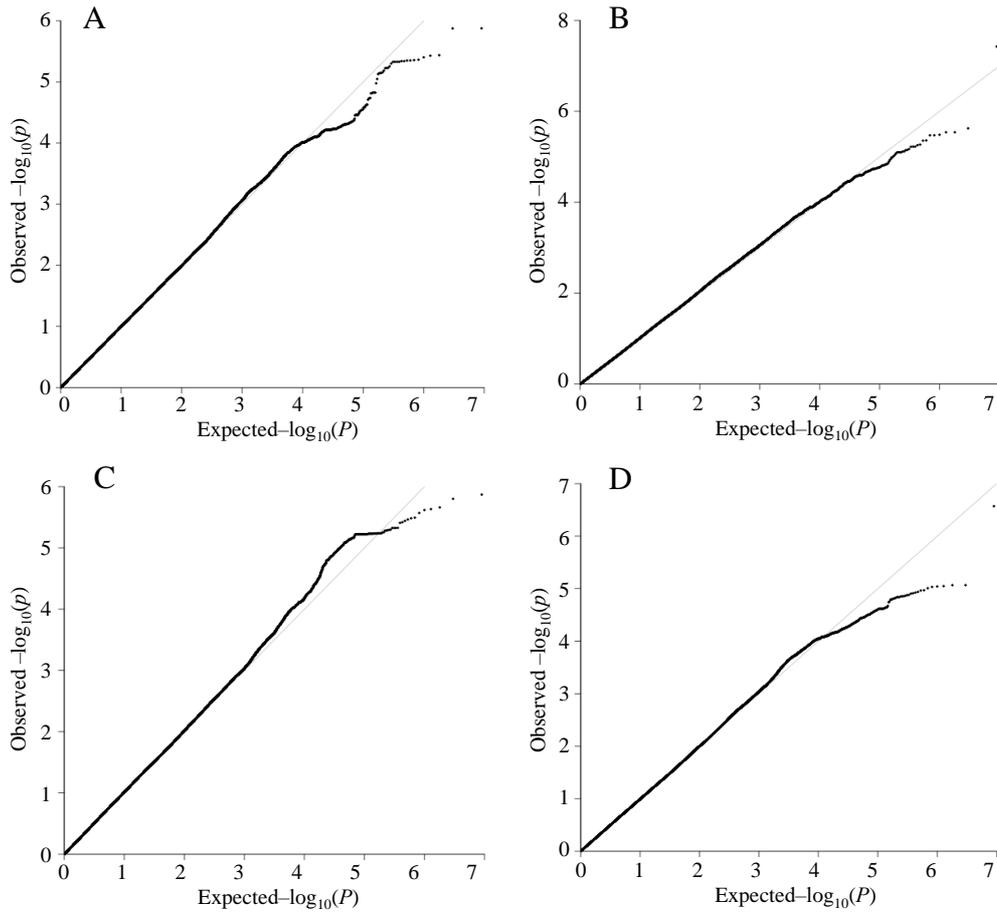
**Figure 5.3. Manhattan plot of Anhedonia mega-GWAS.**



**Figure 5.4. Manhattan plot of Cognitive Disorganisation mega-GWAS.**



**Figure 5.5. Manhattan plot of Parent-rated Negative Symptoms mega-GWAS.**



**Figure 5.6. Quantile-quantile plot of psychotic experience domain mega-GWASs.**

*Note.* A, Paranoia and Hallucinations; B, Anhedonia; C, Cognitive Disorganisation; D, Parent-rated Negative Symptoms.

**Table 5.9. Combined sample (TEDS, ALSPAC and CATSS) association results for genome-wide significant SNPs in TEDS only GWAS of Chapter 2.**

Psychotic Experience	Variant ID	TEDS only GWAS				TEDS+ALSPAC+CATSS GWAS			
		A1	BETA	SE	<i>p</i>	A1	BETA	SE	<i>p</i>
Cognitive Disorganisation	rs7830364	G	-1.22	0.21	1.24x10 <sup>-8</sup>	G	-0.15	0.07	0.03
Cognitive Disorganisation	rs7845752	T	-1.65	0.30	2.98x10 <sup>-8</sup>	C	-0.11	0.07	0.11
Parent-rated Negative Symptoms	rs7587811	C	0.56	0.10	7.01x10 <sup>-8</sup>	C	0.08	0.06	0.23

Note. A1, Allele 1 (test allele); BETA, unstandardized effect size; SE, standard error; TEDS, Twins Early Development Study; ALSPAC, Avon Longitudinal Study of Parents and Children; CATSS, Child and Adolescent Twin Study in Sweden; GWAS, genome-wide association study.

**Table 5.10. Suggestive loci from psychotic experience GWASs in the latest schizophrenia GWAS.**

<i>Paranoia and Hallucinations</i>						
CHR	Variant ID	BP	PSNP	$r^2$	P_SCZ	Direction (PE/SCZ)
3	rs73135634	84961810	rs73135634	*	<b>1.17x10<sup>-3</sup></b>	-/-
17	rs1008621	70362731	rs1008621	*	0.650	+/+
<i>Anhedonia</i>						
CHR	Variant ID	BP	PSNP	$r^2$	P_SCZ	Direction (PE/SCZ)
8	rs149957215	39872495	rs149957215	*	0.339	-/+
13	rs78013746	61682703	rs78013746	*	0.229	+/-
6	rs200488	27795109	rs200488	*	0.159	+/+
11	rs117907077	11033989	rs117907077	*	0.982	-/+
6	rs2531815	28436060	rs2531815	*	0.950	+/+
20	rs6033026	11059873	rs6033026	*	0.952	-/+
15	rs7164838	34967574	rs7164838	*	0.668	+/+
11	rs2169485	41079587	rs2169485	*	0.960	-/+
10	rs11195810	113835240	rs11195810	*	0.796	+/-
14	rs12897386	72471862	rs35727014	1	0.180	+/+
9	rs62545506	73241253	rs62545506	*	0.341	-/-
14	rs34420225	94290014	rs34420225	*	0.270	-/+
15	rs74519172	55010305	rs74519172	*	0.952	+/-
<i>Cognitive Disorganisation</i>						
CHR	Variant ID	BP	PSNP	$r^2$	P_SCZ	Direction (PE/SCZ)
13	rs1961120	28833372	rs1961120	*	0.973	-/-
2	rs200022365	186855226	rs12622553	0.999	0.748	+/-
14	rs7147064	47560742	rs7147064	*	<b>0.028</b>	-/-
2	rs7588854	80339218	rs7588854	*	0.008	+/-
2	rs80033666	170682319	rs80033666	*	0.851	-/-
4	rs1506348	126450002	rs1506348	*	0.403	-/+
3	rs185642755	85127281	rs185642755	*	0.225	-/-
1	rs6665300	65429558	rs79912581	1	0.930	+/+
<i>Parent-rated Negative Symptoms</i>						
CHR	Variant ID	BP	PSNP	$r^2$	P_SCZ	Direction (PE/SCZ)
5	rs147205145	36033829	rs147205145	*	<b>0.023</b>	+/+
4	rs4400001	38212771	rs4583770	1	0.663	+/+
8	rs72334712	108862133	rs10112933	1	0.546	+/-
8	rs35428606	101649797	rs35428606	*	0.338	-/-
7	rs62457829	29549919	rs62457829	*	0.613	+/+

Note: CHR, chromosome number; BP, base-pair position; PSNP, proxy variant in schizophrenia dataset; strength of linkage disequilibrium between index SNP in PE study and proxy in schizophrenia dataset; P\_SCZ,  $p$ -value of proxy SNP in schizophrenia dataset; Direction (PE/SCZ), direction of effect for index and proxy variants in the PE GWASs and schizophrenia GWAS respectively.

**Table 5.11. Suggestive loci from psychotic experience GWASs in the latest bipolar disorder GWAS.**

<i>Paranoia and Hallucinations</i>						
CHR	Variant ID	BP	PSNP	$r^2$	P_BIP	Direction (PE/BIP)
3	rs73135634	84961810	rs9836020	0.944	0.111	-/-
17	rs1008621	70362731	rs1008621	*	<b>0.040</b>	+/+
<i>Anhedonia</i>						
CHR	Variant ID	BP	PSNP	$r^2$	P_BIP	Direction (PE/BIP)
8	rs149957215	39872495	NA	NA	NA	-/NA
13	rs78013746	61682703	rs12583135	0.522	<b>0.004</b>	+/+
6	rs200488	27795109	rs1010261	0.900	0.848	+/-
11	rs117907077	11033989	rs6484605	0.989	<b>0.006</b>	-/+
6	rs2531815	28436060	rs2531815	*	0.195	+/-
20	rs6033026	11059873	rs6033026	*	0.968	-/-
15	rs7164838	34967574	rs7164838	*	0.132	+/+
11	rs2169485	41079587	rs2169485	*	0.810	-/-
10	rs11195810	113835240	rs11195810	*	0.895	+/+
14	rs12897386	72471862	rs12897386	*	0.629	+/-
9	rs62545506	73241253	rs12001853	0.991	0.709	-/+
14	rs34420225	94290014	rs11624417	0.979	0.463	-/+
15	rs74519172	55010305	rs8027123	0.221	0.188	+/+
<i>Cognitive Disorganisation</i>						
CHR	Variant ID	BP	PSNP	$r^2$	P_BIP	Direction (PE/BIP)
13	rs1961120	28833372	rs9319424	0.997	0.204	-/-
2	rs200022365	186855226	rs12622553	0.999	0.059	+/-
14	rs7147064	47560742	rs7147064	*	0.263	-/-
2	rs7588854	80339218	rs1319228	0.975	0.171	+/-
2	rs80033666	170682319	rs3754913	0.909	0.337	-/+
4	rs1506348	126450002	rs1506348	*	0.067	-/+
3	rs185642755	85127281	NA	NA	NA	-/NA
1	rs6665300	65429558	rs17127174	1	0.841	+/+
<i>Parent-rated Negative Symptoms</i>						
CHR	Variant ID	BP	PSNP	$r^2$	P_BIP	Direction (PE/BIP)
5	rs147205145	36033829	NA	NA	NA	+/NA
4	rs4400001	38212771	rs4554086	1	0.907	+/+
8	rs72334712	108862133	rs10112933	1	0.873	+/+
8	rs35428606	101649797	rs2155882	0.997	0.361	-/+
7	rs62457829	29549919	rs1059182	0.896	0.593	+/-

Note: CHR, chromosome number; BP, base-pair position; PSNP, proxy variant in bipolar disorder dataset; strength of linkage disequilibrium between index SNP in PE study and proxy in bipolar disorder dataset; P\_BIP,  $p$ -value of proxy SNP in bipolar disorder dataset; Direction (PE/BIP), direction of effect for index and proxy variants in the PE GWASs and bipolar disorder GWAS respectively.

**Table 5.12. Suggestive loci from psychotic experience GWASs in the latest major depression GWAS.**

<i>Paranoia and Hallucinations</i>						
CHR	Variant ID	BP	PSNP	$r^2$	P_MDD	Direction (PE/MDD)
3	rs73135634	84961810	rs9847448	0.944	0.491	-/+
17	rs1008621	70362731	rs1008622	0.999	0.859	+/+
<i>Anhedonia</i>						
CHR	Variant ID	BP	PSNP	$r^2$	P_MDD	Direction (PE/MDD)
8	rs149957215	39872495	NA	NA	NA	-/NA
13	rs78013746	61682703	rs7991819	0.521	0.962	+/-
6	rs200488	27795109	rs1010261	0.900	<b>0.034</b>	+/-
11	rs117907077	11033989	rs16908726	0.963	0.251	-/+
6	rs2531815	28436060	rs2531815	*	0.082	+/-
20	rs6033026	11059873	rs6033026	*	0.130	-/-
15	rs7164838	34967574	rs7164838	*	0.218	+/+
11	rs2169485	41079587	rs2169485	*	0.057	-/-
10	rs11195810	113835240	rs12254157	0.929	0.493	+/+
14	rs12897386	72471862	rs12883063	0.999	0.308	+/+
9	rs62545506	73241253	rs12001853	0.991	<b>0.011</b>	-/-
14	rs34420225	94290014	rs11624417	0.979	0.242	-/-
15	rs74519172	55010305	NA	NA	NA	+/NA
<i>Cognitive Disorganisation</i>						
CHR	Variant ID	BP	PSNP	$r^2$	P_MDD	Direction (PE/MDD)
13	rs1961120	28833372	rs9319424	0.997	0.686	-/-
2	rs200022365	186855226	rs12618471	0.998	0.142	+/-
14	rs7147064	47560742	rs7147064	*	0.287	-/+
2	rs7588854	80339218	rs1017632	0.905	0.352	+/+
2	rs80033666	170682319	rs3754913	0.909	0.133	-/-
4	rs1506348	126450002	rs1506348	*	0.889	-/-
3	rs185642755	85127281	NA	NA	NA	-/NA
1	rs6665300	65429558	rs17127171	0.994	0.835	+/+
<i>Parent-rated Negative Symptoms</i>						
CHR	Variant ID	BP	PSNP	$r^2$	P_MDD	Direction (PE/MDD)
5	rs147205145	36033829	NA	NA	NA	+/NA
4	rs4400001	38212771	rs6835249	0.747	0.443	+/+
8	rs72334712	108862133	rs11987403	1	0.952	+/-
8	rs35428606	101649797	rs2155882	0.997	0.201	-/+
7	rs62457829	29549919	rs1059182	0.896	0.980	+/+

Note: CHR, chromosome number; BP, base-pair position; PSNP, proxy variant in major depression dataset; strength of linkage disequilibrium between index SNP in PE study and proxy in major depression dataset; P\_MDD,  $p$ -value of proxy SNP in major depression dataset; Direction (PE/ MDD), direction of effect for index and proxy variants in the PE GWASs and major depression GWAS respectively.

### 5.3.3 – Results from mega-analysis and meta-analysis GWAS

Meta-analysis results when using the same PEs data as in the mega-analysis (i.e. ties randomly split during normalisation) were almost identical to those of the mega-analyses. The p-value correlations between all SNPs in the four mega- and meta- GWASs were between 0.972 and 0.996, with slightly lower correlations for PE traits that were originally more skewed. The genome-wide significant SNP in the Anhedonia mega-GWAS (rs149957215) remained significant ( $1.611 \times 10^{-8}$ ) in the meta-analysis, with *p*-values of  $1.01 \times 10^{-5}$  and  $4.247 \times 10^{-4}$  in the TEDS and ALSPAC samples alone.

### 5.3.4 – Evaluating the effect of randomly splitting ties over averaging ties

Averaging tied observations during normalisation was only partially effective, with some PE scales showing a remaining skew greater than 1 (Table 5.13). Meta-analysis results when using PE measures that were normalised with ties averaged contained many highly significant associations. Phenotypes that were more skewed showed more significant associations, particularly in the smaller samples. The large number of highly genome-wide significant associations, and their relationship with increased skew and decreased sample size suggest that many of the significant associations were spurious. For Anhedonia, which had the lowest average skew, the only genome-wide significant variant was rs149957215 ( $p = 1.64 \times 10^{-8}$ ), the same genome-wide significant variant identified in the Anhedonia mega-GWAS.

**Table 5.13. Skew of psychotic experiences after normalisation when averaging tied observations.**

Psychotic Experience	Sample	Skew
Paranoia and Hallucinations	TEDS	0.375
Paranoia and Hallucinations	ALSPAC	1.145
Paranoia and Hallucinations	CATSS	1.334
Anhedonia	TEDS	0.030
Anhedonia	ALSPAC	0.219
Anhedonia	CATSS	NA
Cognitive Disorganisation	TEDS	0.431
Cognitive Disorganisation	ALSPAC	NA
Cognitive Disorganisation	CATSS	0.157
Parent-rated Negative Symptoms	TEDS	0.556
Parent-rated Negative Symptoms	ALSPAC	0.207
Parent-rated Negative Symptoms	CATSS	0.912

#### 5.3.5 – Gene region association from MAGMA

Regional gene-based analysis using MAGMA identified no gene that was significantly associated with any PE domain after Bonferroni correction for multiple testing. The top ten most associated genes for each PE domain are listed in Table 5.14.

#### 5.3.6 – Differential predicted gene expression from PrediXcan

Analysis of predicted frontal cortex gene expression associated with PE domains showed *HACD2* as significantly differentially expressed for Cognitive Disorganisation (Bonferroni corrected  $p = 6.83 \times 10^{-4}$ ). However, the association between *HACD2* and Cognitive Disorganisation was not replicated in the independent replication sample, showing a non-significant association in the opposite direction. The top ten genes showing differential expression for each psychotic experience domain are listed in Table 5.15.

Analysis of predicted gene expression in blood did not identify any significant association with PEs after Bonferroni correction of multiple testing. However, *HACD2* and *CIB2* were the most significantly differentially expressed genes for Cognitive

Disorganisation and Anhedonia respectively. Overall the correlation between association statistics from blood and brain was  $\sim 0$ , even when only comparing genes that were nominally significant genes in the brain (or vice versa).

This indicates that the predicted gene expression levels in the blood and brain are substantially different for the majority of genes. Indeed, when using just overlapping genes on chromosome 1 ( $n=157$ ), the mean correlation was 0.23 and ranged from -0.92 - 0.99. These results demonstrate the large range in eQTL overlap for a given gene between the blood and brain.

The fact that the top associated genes showed similar results in both the blood and brain could indicate that genes associated with adolescent PEs have similar eQTL-mediated regulation in the blood and brain. Alternatively, these genes might just be more reliably imputed in both the brain and blood due to a high heritability of their expression.

**Table 5.14. Top ten genes associated with psychotic experience domains using MAGMA.**

***Paranoia and Hallucinations***

CHR	Gene Symbol	NSNPS	<i>z</i>	<i>p</i>	Corrected <i>p</i>
5	SGCD	1574	4.105	2.018x10 <sup>-5</sup>	0.348
18	MC2R	109	3.676	1.184x10 <sup>-4</sup>	1.000
12	MGAT4C	1987	3.645	1.336x10 <sup>-4</sup>	1.000
1	OR10K2	57	3.640	1.365x10 <sup>-4</sup>	1.000
20	PROCR	86	3.434	2.975x10 <sup>-4</sup>	1.000
16	CALB2	66	3.414	3.202x10 <sup>-4</sup>	1.000
11	RSF1	298	3.395	3.429x10 <sup>-4</sup>	1.000
1	OR10K1	48	3.344	4.136x10 <sup>-4</sup>	1.000
14	MLH3	63	3.311	4.652x10 <sup>-4</sup>	1.000
14	ACYP1	46	3.296	4.913x10 <sup>-4</sup>	1.000

***Anhedonia***

CHR	Gene Symbol	NSNPS	<i>z</i>	<i>p</i>	Corrected <i>p</i>
14	ABHD12B	36	4.045	2.619x10 <sup>-5</sup>	0.451
15	MAPKBP1	91	3.957	3.789x10 <sup>-5</sup>	0.653
10	BNIP3	37	3.657	1.275x10 <sup>-4</sup>	1.000
22	CECR5	43	3.634	1.397x10 <sup>-4</sup>	1.000
14	PYGL	126	3.615	1.501x10 <sup>-4</sup>	1.000
10	PPP2R2D	73	3.607	1.547x10 <sup>-4</sup>	1.000
15	C15orf27	199	3.592	1.644x10 <sup>-4</sup>	1.000
12	IGFBP6	40	3.588	1.664x10 <sup>-4</sup>	1.000
19	ZNF43	143	3.551	1.920x10 <sup>-4</sup>	1.000
15	JMJD7	76	3.530	2.080x10 <sup>-4</sup>	1.000

**Table 5.14 cont.**

***Cognitive Disorganisation***

<b>CHR</b>	<b>Gene Symbol</b>	<b>NSNPS</b>	<b>z</b>	<b>p</b>	<b>Corrected p</b>
13	CLYBL	482	4.112	1.958x10 <sup>-5</sup>	0.337
4	RUFY3	248	3.856	5.773x10 <sup>-5</sup>	0.994
12	ASUN	100	3.822	6.613x10 <sup>-5</sup>	1.000
19	PSG4	1	3.792	7.480x10 <sup>-5</sup>	1.000
9	NR5A1	8	3.625	1.443x10 <sup>-4</sup>	1.000
11	CEP295	31	3.588	1.667x10 <sup>-4</sup>	1.000
7	PEX1	57	3.567	1.807x10 <sup>-4</sup>	1.000
3	TFDP2	551	3.545	1.964x10 <sup>-4</sup>	1.000
9	ADGRD2	10	3.527	2.100x10 <sup>-4</sup>	1.000
4	GRSF1	112	3.468	2.626x10 <sup>-4</sup>	1.000

***Parent-rated Negative Symptoms***

<b>CHR</b>	<b>Gene Symbol</b>	<b>NSNPS</b>	<b>z</b>	<b>p</b>	<b>Corrected p</b>
12	NDUFA9	134	4.192	1.380x10 <sup>-5</sup>	0.238
20	SDCBP2	30	4.036	2.722x10 <sup>-5</sup>	0.469
2	CD28	76	3.793	7.430x10 <sup>-5</sup>	1.000
11	FUT4	23	3.619	1.479x10 <sup>-4</sup>	1.000
5	FST	53	3.545	1.966x10 <sup>-4</sup>	1.000
1	SPAG17	351	3.488	2.433x10 <sup>-4</sup>	1.000
11	DNHD1	138	3.309	4.681x10 <sup>-4</sup>	1.000
19	SLC8A2	30	3.229	6.213x10 <sup>-4</sup>	1.000
16	UBN1	101	3.207	6.717x10 <sup>-4</sup>	1.000
7	GHRHR	33	3.165	7.762x10 <sup>-4</sup>	1.000

*Note:* CHR, chromosome number; NSNPS, Number of genetic variants within the gene region; Corrected p, Bonferroni corrected p-value.

**Table 5.15. Top ten differentially-expressed genes for psychotic experience domains based on predicted gene expression levels.**

***Paranoia and Hallucinations***

<b>Gene Symbol</b>	<b>Beta</b>	<b>SE</b>	<b><i>p</i></b>	<b>Corrected <i>p</i></b>
LY6D	-0.342	0.097	4.15x10 <sup>-4</sup>	1.000
LRR1Q1	-0.430	0.132	1.12x10 <sup>-3</sup>	1.000
FAHD1	0.113	0.036	1.62x10 <sup>-3</sup>	1.000
PLPPR2	-0.315	0.104	2.42x10 <sup>-3</sup>	1.000
STMND1	1.111	0.375	3.08x10 <sup>-3</sup>	1.000
KRT81	0.222	0.076	3.70x10 <sup>-3</sup>	1.000
COMMD2	0.209	0.073	4.26x10 <sup>-3</sup>	1.000
SPATA13	0.248	0.088	4.73x10 <sup>-3</sup>	1.000
ZNHIT6	0.234	0.083	4.84x10 <sup>-3</sup>	1.000
PSMB8	0.131	0.048	5.96x10 <sup>-3</sup>	1.000

***Anhedonia***

<b>Gene Symbol</b>	<b>Beta</b>	<b>SE</b>	<b><i>p</i></b>	<b>Corrected <i>p</i></b>
CIB2	-1.207	0.317	1.42x10 <sup>-4</sup>	0.392
INTS1	-0.192	0.054	3.58x10 <sup>-4</sup>	0.991
FAM198A	0.247	0.073	7.63x10 <sup>-4</sup>	1.000
ACKR2	-0.483	0.146	9.69x10 <sup>-4</sup>	1.000
STRN	-2.089	0.641	1.11x10 <sup>-3</sup>	1.000
AGO2	0.252	0.082	2.16x10 <sup>-3</sup>	1.000
TEAD2	-0.258	0.087	3.07x10 <sup>-3</sup>	1.000
CBLN4	-0.715	0.248	3.94x10 <sup>-3</sup>	1.000
NA	0.643	0.225	4.34x10 <sup>-3</sup>	1.000
ZNF514	-0.243	0.086	4.47x10 <sup>-3</sup>	1.000

**Table 5.15 cont.**

***Cognitive Disorganisation***

<b>Gene Symbol</b>	<b>Beta</b>	<b>SE</b>	<b><i>p</i></b>	<b>Corrected <i>p</i></b>
HACD2	-0.154	0.029	1.06x10 <sup>-7</sup>	2.93x10 <sup>-4</sup>
RASAL2	0.193	0.054	3.87x10 <sup>-4</sup>	1.000
AP4S1	-0.162	0.048	8.25x10 <sup>-4</sup>	1.000
LRBA	-0.420	0.128	9.76x10 <sup>-4</sup>	1.000
NOL6	0.348	0.108	1.29x10 <sup>-3</sup>	1.000
RAD51C	0.077	0.025	2.42x10 <sup>-3</sup>	1.000
SLTM	0.241	0.080	2.51x10 <sup>-3</sup>	1.000
RAB43	0.210	0.070	2.56x10 <sup>-3</sup>	1.000
NFS1	-0.210	0.070	2.60x10 <sup>-3</sup>	1.000
BRIX1	-0.277	0.092	2.70x10 <sup>-3</sup>	1.000

***Parent-rated Negative Symptoms***

<b>Gene Symbol</b>	<b>Beta</b>	<b>SE</b>	<b><i>p</i></b>	<b>Corrected <i>p</i></b>
RNASEL	-1.093	0.305	3.45x10 <sup>-4</sup>	0.956
AKAP3	-0.119	0.033	3.67x10 <sup>-4</sup>	1.000
STXBP5L	-0.138	0.040	4.92x10 <sup>-4</sup>	1.000
EIF5A	-0.580	0.176	9.76x10 <sup>-4</sup>	1.000
MAPK8IP1	2.362	0.725	1.11x10 <sup>-3</sup>	1.000
ADSS	-0.166	0.054	2.18x10 <sup>-3</sup>	1.000
HNRNPC	-1.618	0.530	2.28x10 <sup>-3</sup>	1.000
LRRC25	1.118	0.371	2.62x10 <sup>-3</sup>	1.000
WDR17	0.791	0.266	2.97x10 <sup>-3</sup>	1.000
MRE11A	-0.122	0.041	3.04x10 <sup>-3</sup>	1.000

*Note.* Corrected *p*, Bonferroni corrected *p*-value.

## 5.4 – Discussion

This chapter performed mega-GWASs of four specific PE measures across three European adolescent samples in an attempt to identify associated genetic loci. This is the largest GWAS to date of PEs, and the first to use quantitative measures of specific PEs. The only genetic variant that achieved genome-wide significance was in the mega-GWAS for anhedonia, which was within the gene *IDO2*. Across the four specific PEs, 29 LD independent loci were associated at suggestive significance ( $p=1 \times 10^{-5}$ ). Several genetic variants achieving suggestive significance in the mega-GWASs were within or proximal to genes with prior evidence of association with related outcomes or plausible biological pathways. A full table of suggestive loci with annotation is provided in Table 5.16. This chapter also performed two gene-based analyses (MAGMA and PrediXcan) to identify genes associated with specific adolescent PEs. MAGMA returned no significantly associated genes. PrediXcan identified one gene showing significant predicted differential gene expression in the prefrontal cortex for Cognitive Disorganisation called *HACD2*.

The genome-wide significant SNP for Anhedonia was within the protein-coding gene *IDO2*. *IDO2* is a key enzyme in the regulation of the kynurenine pathway, which upon stimulation by proinflammatory cytokines, converts tryptophan into kynurenine. It has been reported that increased metabolism of tryptophan to kynurenine is associated with increased depressive symptoms via the increased production of cytotoxic kynurenine metabolites (Dantzer, O'Connor, Lawson, & Kelley, 2011; A.-M. Myint et al., 2007; Wichers et al., 2005). In fact, a previous study has reported a significant correlation between kynurenine production and anhedonia in an adolescent sample (Gabbay, Ely, Babb, & Liebes, 2012). These previous studies suggest the association between *IDO2* and anhedonia is plausible. However, this finding should be interpreted

with caution as the SNP was imputed in all three samples, only just achieved genome-wide significance, and failed to replicate in a sufficiently powered independent sample.

Comparison of the mega- and meta- GWAS approaches showed highly similar results when based on the same normalised phenotypic data. When an alternative approach was tried, namely of meta-analysing non-normal phenotypic data, the skew in the phenotypic data led to spurious associations. This was somewhat unexpected, as the GEE method used to account for the presence of related individuals should provide heteroskedasticity-robust standard errors. One explanation for the spurious associations resulting from skewed phenotypic data is that the heteroskedasticity-robustness of GEE only holds under asymptotic conditions, explaining the increased number of spurious associations in the smaller samples. Together these comparisons show that the mega-GWAS approach using normalised traits (randomly splitting ties) provides robust estimates of association.

MAGMA identified no gene region significantly associated with any PE trait after controlling for multiple testing. The most significant association was between *CLYBL* and Cognitive Disorganisation (Bonferroni corrected  $p = 0.11$ ). The *CLYBL* protein is involved in glucose metabolism with both malate synthase and beta-methylmalate synthase activity (Strittmatter et al., 2014). *CLYBL* has been previously implicated in attention deficit hyperactivity disorder (ADHD) (Lasky-Su et al., 2008), supporting its role in cognition-related neurodevelopmental disorders. This gene has also been associated with anticitrullinated peptide antibodies-negative rheumatoid arthritis (Bossini-Castillo et al., 2014).

Analysis of predicted gene-expression in the frontal cortex highlighted one gene with a significant inverse relationship with Cognitive Disorganisation called *HACD2* (3-Hydroxyacyl-CoA Dehydratase 2). It is a protein-coding gene involved in fatty acid metabolism. A previous mouse study reported prenatal stress to cause decreased

*HACD2* expression in the frontal cortex of offspring (Zucchi et al., 2013). Human studies have demonstrated an association between prenatal stress, impaired cognitive development (Laplante et al., 2004; Niederhofer & Reiter, 2004) and schizophrenia (Selten, van der Graaf, van Duursen, Gispén-de Wied, & Kahn, 1999; van Os & Selten, 1998). A possible model unifying these results is that *HACD2* down regulation mediates the association between prenatal stress, cognitive development and schizophrenia. However, this association should be interpreted with caution, as it was not replicated in the independent TEDS replication sample.

The comparison of predicted gene expression in the prefrontal cortex and blood demonstrates the large range in eQTL overlap for a given gene between the blood and brain. The fact that the top associated genes showed similar results in both the blood and brain could indicate that genes associated with adolescent PEs have similar eQTL-mediated regulation in the blood and brain. Alternatively, these genes might just be more reliably imputed in both the brain and blood due to a high heritability of their expression.

This study has tested for both variant- and gene- level associations for a broad range of specific PE traits across three European and adolescent samples. The measures assessing each PE domain within each sample have been carefully harmonised to ensure both content validity and comparability of the scales across the samples (Chapter 4). Alongside careful harmonisation of genotypic data across the three samples, this process has enabled combined analysis across the three samples enabling the largest GWAS of adolescent PEs to date. Although it requires replication, this study has provided the first genome-wide significant association for an adolescent PE. Furthermore, the use of eQTL reference datasets, this study has identified the first gene significantly associated with adolescent PEs.

Although this study has been successful in a number of ways, it demonstrates the ongoing challenge of insufficient power to detect genome-wide significant genetic markers. This lack of power can be interpreted in a number of ways. Two likely interpretations are: 1) individual genetic variants have very small effect sizes requiring larger samples to detect associations at significance, or 2) the phenotypic variance within and across samples explained by tagged genetic variation (SNP-heritability) is very small. One approach to investigate this power issue is by estimating the phenotypic variance of PEs within and across samples that can be explained by common genetic variation. This is explored in the next chapter.

**Table 5.16. Annotation of suggestive loci with prior evidence of association in neuropsychiatric phenotypes.**

***Anhedonia***

CHR	SNP	Nearest gene (distance)	Nearest gene annotation
8	rs149957215	IDO2 (within)	Key enzyme in kynurenine pathway (Fatokun, Hunt, & Ball, 2013). The kynurenine pathway has been implicated in major depression, bipolar disorder, schizophrenia (A. M. Myint, 2012), suicidal behaviour (Bryleva & Brundin, 2017), and antidepressant response (Réus et al., 2015).
11	rs2169485	LRRC4C (within)	Binds netrin G1, which is involved in axon guidance (Lin, Ho, Gurney, & Rosenthal, 2003). LRRC4C is associated with temperament in bipolar disorder (Greenwood, Akiskal, Akiskal, Study, & Kelsoe, 2012), antipsychotic response in schizophrenia (McClay et al., 2011), methylation difference in schizophrenia (Viana et al., 2017), autism, intellectual disability (Sangu et al., 2017).
14	rs12897386	RGS6 (within)	RGS6 has been associated with schizophrenia (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014), and maintenance of dopaminergic neurons (Bifsha, Yang, Fisher, & Drouin, 2014).
9	rs62545506	TRPM3 (within)	Apart of the transient receptor potential channel (TRP) gene family. TRPs are cation-selective channels important for cellular calcium signalling and homeostasis. The TRPM3 region has been repeatedly associated bipolar disorder (Serretti & Mandelli, 2008), and chronic fatigue syndrome (Nguyen, Staines, Nilius, Smith, & Marshall-Gradisnik, 2016). A micro-RNA within TRPM3 has been associated with schizophrenia (Cammaerts et al., 2015).
14	rs34420225	PRIMA1 (35kb)	Transports and anchors acetylcholinesterase to neuronal membranes. Associated with major depression (Sabunciyan et al., 2012), borderline personality disorder (Teschler, Gotthardt, Dammann, & Dammann, 2016), caffeine related sleep disturbance (Byrne et al., 2012), and epilepsy (Hildebrand et al., 2015).
15	rs74519172	UNC13C (90kb)	Associated with attentions deficit hyperactivity disorder (ADHD) (Elia et al., 2010) and post-traumatic stress disorder (PTSD) (Ashley-Koch et al., 2015).

Table 5.16 cont.

***Cognitive Disorganisation***

CHR	SNP	Nearest gene (distance)	Nearest gene annotation
14	rs7147064	MDGA2 (within)	MDGA2 (and its paralog MDGA1) regulate inhibitory synapse development (Pettem, Yokomaku, Takahashi, Ge, & Craig, 2013). MDGA2 has been previously associated with autism (Bucan et al., 2009), neuroticism (van den Oord et al., 2008), and harm avoidance (Heck et al., 2011). MDGA1 has been associated with schizophrenia and bipolar disorder in a Chinese Han population (J. Li et al., 2011).
2	rs7588854	CTNNA2 (within)	This gene acts as a linker between the cadherin adhesion receptors and the cytoskeleton to regulate cell-cell adhesion and differentiation in the nervous system (Abe, Chisaka, Van Roy, & Takeichi, 2004). It regulates morphological plasticity of synapses and cerebellar and hippocampal lamination during development (Park, Falls, Finger, Longo-Guess, & Ackerman, 2002). This gene has been associated with ADHD (Poelmans, Pauls, Buitelaar, & Franke, 2011), alcohol addiction (Song & Zhang, 2014), impulsivity in native Americans (Ehlers et al., 2016), excitement seeking (Terracciano et al., 2011), and smoking in schizophrenia (Mexal et al., 2008).
4	rs1506348	FAT4 (40kb)	FAT4 is a cadherin that maintains planar cell polarity as well as in inhibition of YAP1-mediated neuroprogenitor cell proliferation and differentiation (Cappello et al., 2013). There is evidence of association between FAT4 and cognitive performance (Need et al., 2009), and a combined major depression and bipolar disorder phenotype (Liu et al., 2011).
3	rs185642755	CADM2 (within)	Encodes a member of the synaptic cell adhesion molecule 1 family. Important for synapse organisation, providing regulated trans-synaptic adhesion (Pellissier, Gerber, Bauer, Ballivet, & Ossipow, 2007). CADM2 has been associated with executive function and processing speed (Ibrahim-Verbaas et al., 2016), lifetime cannabis use (Stringer et al., 2016), a range of personality traits (Boutwell et al., 2017), and autism (Casey et al., 2012).

Table 5.16 cont.

***Parent-rated Negative Symptoms***

CHR	SNP	Nearest gene (distance)	Nearest gene annotation
8	rs35428606	SNX31 (within)	Within the nexin gene family, which is are involved in intracellular trafficking. A frameshift mutation in SNX31 was identified in a schizophrenic patient (Balan et al., 2014). Several other SNX genes have been implicated in neuronal or psychiatric phenotypes. A related protein, SNX27, contributes to excitatory synaptic dysfunction via modulation of glutamate receptor recycling in Down syndrome (X. Wang et al., 2013). Reduced expression of SNX7 was associated with positive psychotic symptoms and reduced executive function in bipolar disorder (Sellgren et al., 2016). SNX3 has been associated with schizophrenia (K.-C. Huang, Yang, Lin, Tsao, & Lee, 2014; Mladinov et al., 2016).
7	rs62457829	CHN2 (within)	CHN2 encodes a guanosine triphosphate (GTP)-metabolising protein. This gene has been associated with schizophrenia in men (Hashimoto et al., 2005), atypical psychosis (Kanazawa et al., 2013), and smoking cessation (Uhl et al., 2008).

*Note.* Prior evidence of association in neuropsychiatric phenotypes was determined by scanning functional annotations and related articles on the NCBI Gene and GeneCards search engines.

## 5.5 – Appendix

**Supplementary Table 5.1. Measures of Paranoia and Hallucinations in TEDS, ALSPAC, and CATSS.**

<u><b>TEDS</b></u>	
<b>Leading statement:</b>	How often have you thought...
<b>Response options:</b>	0 = Not at all, 1 = Rarely, 2 = Once a month, 3 = Once a week, 4 = Several times a week, 5 = Daily
<b>Item 1:</b>	I might be being observed or followed?
<b>Item 2:</b>	I can detect coded messages about me in the press/TV/internet?
<b>Item 3:</b>	People might be conspiring against me?
<b>Item 4:</b>	Hear noises or sounds when there is nothing about to explain them?
<b>Item 5:</b>	Hear sounds or music that people near you don't hear?
<b>Item 6:</b>	Hear voices commenting on what you're thinking or doing?
<b>Item 7:</b>	See things that other people cannot?
<b>Item 8:</b>	See shapes, lights, or colours even though there is nothing really there?
<b>Note:</b>	In order to keep the same ratio of items assessing hallucinations to paranoia the same across samples, items 4,5 and 6 were averaged to create a composite auditory hallucinations item, and items 7 and 8 were average to create a composite visual hallucinations item.
<u><b>ALSPAC</b></u>	
<b>Part A response options:</b>	0 = No, never, 1 = Yes, maybe, 2 = Yes, definitely
<b>Part B response options:</b>	0 = Not at all, 1 = Once or twice, 2 = Less than once a month, 3 = More than once a month, 4 = Nearly every day
<b>Item 1a:</b>	Some people believe that other people can read their thoughts. Have other people ever read your thoughts?
<b>Item 1b:</b>	How often have other people read your thoughts since your 15 <sup>th</sup> birthday?
<b>Item 2a:</b>	Have you ever thought you were being followed or spied on?
<b>Item 2b:</b>	How often has this happened since your 15 <sup>th</sup> birthday?
<b>Item 3a:</b>	Have you ever believed that you were being sent special messages though the television or the radio, or that a programme had been arranged just for you alone?
<b>Item 3b:</b>	How often has this happened since your 15 <sup>th</sup> birthday?
<b>Item 4a:</b>	Have you ever seen something or someone that other people could not see?
<b>Item 4b:</b>	How often have you seen something or someone that other people could not see since your 15 <sup>th</sup> birthday?
<b>Item 5a:</b>	Have you ever heard voices that other people couldn't hear
<b>Item 5b:</b>	How often have you heard voices that other people couldn't hear since your 15 <sup>th</sup> birthday?
<b>Note:</b>	The responses to Part A and Part B were summed for each item.
<u><b>CATSS</b></u>	
<b>Leading statement:</b>	Have you ever...
<b>Response options:</b>	0 = Never or rarely, 1 = Sometimes, 2 = Often, 3 = Very Often
<b>Item 1:</b>	Thought you were being followed or spied on?
<b>Item 2:</b>	Thought you were being sent special messages through the television?
<b>Item 3:</b>	Thought other people could read your thoughts?
<b>Item 4:</b>	Seen things other people cannot see?
<b>Item 5:</b>	Heard voices that nobody else can hear?

**Supplementary Table 5.2. Measures of Anhedonia in TEDS and ALSPAC.**

<b><u>TEDS</u></b>	
<b>Response options:</b>	0 = Very false for me, 1 = Moderately false for me, 2 = Slightly false for me, 3 = Slightly true for me, 4 = Moderately true for me, 6 = Very true for me
<b>Item 1:</b>	When something exciting is coming up in my life, I really look forward to it. (R)
<b>Item 2:</b>	When I'm on my way to an amusement park, I can hardly wait to ride the rollercoasters. (R)
<b>Item 3:</b>	When it think about eating my favourite food, I can almost taste how good it is. (R)
<b>Item 4:</b>	I don't look forward to things like eating out at restaurants.
<b>Item 5:</b>	I get so excited the night before a major holiday I can hardly sleep. (R)
<b>Item 6:</b>	When I think of something tasty, like chocolate biscuit, I have to have one. (R)
<b>Item 7:</b>	Looking forward to a pleasurable experience is in itself pleasurable. (R)
<b>Item 8:</b>	I look forward to a lot of things in my life. (R)
<b>Item 9:</b>	When ordering something off a menu, I imagine how good it will taste. (R)
<b>Item 10:</b>	When I hear about a new movie starring my favourite actor, I can't wait to see it. (R)
<b><u>ALSPAC</u></b>	
<b>Response options:</b>	0 = No, never, 1 = Yes, sometimes, 2 = Yes, often, 3 = Yes, nearly always
<b>Item 1:</b>	Have you felt that you experience few or no emotions at important events, such as on your birthday?
<b>Item 2:</b>	Have you felt that you are lacking 'get up and go'?
<b>Item 3:</b>	Have you felt that you have only a few hobbies or interests?
<b>Leading statement:</b>	In the past two weeks...
<b>Response options:</b>	0 = Not true, 1 = Sometimes true, 2 = True
<b>Item 4:</b>	I have been having fun. (R)
<b>Item 5:</b>	I didn't enjoy anything at all.
<b>Item 6:</b>	I felt so tired that I just sat around and did nothing.
<b>Item 7:</b>	I have had a good time. (R)

Note. R, reverse coded.

**Supplementary Table 5.3. Measures of Cognitive Disorganisation in TEDS and CATSS.**

<u><b>TEDS</b></u>	
<b>Response options:</b>	0 = Yes, 1 = No
<b>Item 1:</b>	Are you easily confused if too much happens at the same time?
<b>Item 2:</b>	Do you frequently have difficulty in starting to do things?
<b>Item 3:</b>	Are you a person whose mood goes up and down easily?
<b>Item 4:</b>	Do you dread going into a room by yourself where other people have already gathered and are talking?
<b>Item 5:</b>	Do you find it difficult to keep interested in the same thing for a long time?
<b>Item 6:</b>	Do you find it difficult in controlling your thoughts?
<b>Item 7:</b>	Are you easily distracted from work by daydreams?
<b>Item 8:</b>	Do you ever feel that your speech is difficult to understand because the words are all mixed up and don't make sense?
<b>Item 9:</b>	Are you easily distracted when you read or talk to someone?
<b>Item 10:</b>	Is it hard for you to make decisions?
<b>Item 11:</b>	When in a crowded room, do you often have difficulty in following a conversation?
<u><b>CATSS</b></u>	
<b>Response options:</b>	0 = Never, 1 = Rarely, 2 = Sometimes, 3 = Often, 4 = Very often
<b>Item 1:</b>	How often do you have trouble wrapping up the fine details of a project, once the challenging parts have been done?
<b>Item 2:</b>	When you have a task that requires a lot of thought, how often do you avoid or delay getting started?
<b>Item 3:</b>	How often do you have difficulty keeping your attention when you are doing boring or repetitive work?
<b>Item 4:</b>	How often do you have difficulty concentrating on what people say to you, even when they are speaking to you directly?
<b>Item 5:</b>	How often are you distracted by activity or noise around you?

**Supplementary Table 5.4. Measures of Parent-rated Negative Symptoms in TEDS, ALSPAC, and CATSS.**

<u><b>TEDS</b></u>	
<b>Leading statement:</b>	My child...
<b>Response options:</b>	0 = Not at all true, 1 = Somewhat true, 2 = Mainly true, 3 = Definitely true
<b>Item 1:</b>	Usually gives brief, one word replies to questions, even if encouraged to say more.
<b>Item 2:</b>	Often does not have much to say for him/herself.
<b>Item 3:</b>	Has few or no friends.
<b>Item 4:</b>	Is often inattentive and appears distracted.
<b>Item 5:</b>	Often does not pay attention when being spoken to.
<b>Item 6:</b>	Often sits around for a long time doing nothing.
<b>Item 7:</b>	Has a lack of energy and motivation.
<b>Item 8:</b>	Has very few interests or hobbies.
<b>Item 9:</b>	Often fails to smile or laugh at things others would find funny.
<b>Item 10:</b>	Seems emotionally "flat", for example, rarely changes the emotions he/she shows.
<u><b>ALSPAC</b></u>	
<b>Response options:</b>	0 = Often, 1 = Sometimes, 2 = Hardly ever, 3 = Never
<b>Item 1:</b>	How often does he/she tell you about things that happen at school/college/work?
<b>Item 2:</b>	How often does he/she tell you about things that happen while he's/she's been out?
<b>Response options:</b>	0 = No, 1 = Yes
<b>Item 3:</b>	Thinking back over the last month, has she been feeling tired or felt she had no energy?
<b>Leading statement:</b>	In the past 6 months...
<b>Response options:</b>	0 = Not true, 1 = Somewhat true, 2 = Certainly true, NA = Don't know
<b>Item 4:</b>	He/She did not respond when told to do something.
<b>Item 5:</b>	He/She has at least one good friend. ®
<b>Item 6:</b>	He/She is easily distracted, his/her concentration wanders.
<b>Item 7:</b>	He/She sees tasks through to the end, he/she has good attention span.
<u><b>CATSS</b></u>	
<b>Leading statement:</b>	How accurate are the following statements for your child in the past six months?
<b>Response options:</b>	0 = Not true, 1 = Somewhat true, 2 = Very or often true
<b>Item 1:</b>	Refuses to talk
<b>Item 2:</b>	Secretive, keeps things to self
<b>Item 3:</b>	Has trouble making or keeping friends
<b>Item 4:</b>	Withdrawn, doesn't get involved with others
<b>Item 5:</b>	Fails to finish things he/she should do
<b>Item 6:</b>	Can't concentrate, can't pay attention for long
<b>Item 7:</b>	Underactive, slow moving, or lacks energy
<b>Item 8:</b>	Feels tired without good reason
<b>Item 9:</b>	Stares blankly
<b>Response options:</b>	0 = No, 1 = Yes, to a certain degree, 2 = Yes
<b>Item 10:</b>	Does the twin have difficulties expressing emotions and reactions with facial gestures, prosody, or body language?

# **Chapter 6 - Estimating the SNP-heritability of specific adolescent psychotic experiences using TEDS, ALSPAC and CATSS**

## **6.1 – Introduction**

The amount of variance for a given trait explained by additive and common genetic variation on a microarray (with or without imputation) is called the SNP-heritability (or chip-based heritability) (Yang, Benyamin, McEvoy, Gordon, Henders, Nyholt, et al., 2010). As mentioned in Section 1.4.4, the SNP-heritability is important for determining the amount of power a genome-wide analysis of common variation will have to detect genome-wide significant variation. The SNP-heritability is also informative of the underlying genetic architecture of a trait.

As described in Section 1.5.1, two popular methods for estimating the SNP-heritability of a phenotype are genomic-relatedness-matrix restricted maximum likelihood (GREML) (Yang, Benyamin, McEvoy, Gordon, Henders, & Others, 2010) in Genome-wide Complex Trait Analysis (GCTA) (Yang et al., 2011), and linkage disequilibrium (LD)-score regression (Bulik-Sullivan et al., 2015).

One previous study has estimated the SNP-heritability of specific adolescent PEs (Sieradzka et al., 2015), as described in Chapter 1. This study used GREML to estimate the SNP-heritability for the six specific PE traits captured by the SPEQ (Specific Psychotic Experiences Questionnaire) based on 2,152 unrelated individuals within TEDS (Twins Early Development Sample). This study used three versions of GREML: GREML-SC (single component), GREML-MS (MAF-stratified), and GREML-SC based on LD-pruned genotypic data. The results were fairly consistent across the different GREML-methodologies, concluding that GREML-MS was recommended. GREML-MS returned

evidence of SNP-heritability for Cognitive Disorganisation (23%), Grandiosity (10%) and Anhedonia (32%). GREML-MS estimates were close to zero for Paranoia (6%). GREML-MS estimates for Hallucinations, and Parent-rated Negative Symptoms were 0%. However, due to the limited sample size, these estimates had large standard errors of 25%.

In this chapter, the phenotypic variance of each specific PE was estimated using both GREML in GCTA and LD-score regression. This was performed within and across samples to estimate the phenotypic variance explained by common genetic variation within a homogenous sample, but also across samples. Based on the findings of Sieradzka et al., 2015 (Sieradzka et al., 2015), Paranoia and Hallucinations and Parent-rated Negative Symptom PE domains were predicted to have lower SNP-heritability estimates than Anhedonia and Cognitive Disorganisation PE domains.

## **6.2 – Methods**

### 6.2.1 – Samples

The three samples used in this chapter have been originally described in Section 4.2.1. The exclusion criteria applied within each sample have been detailed in Section 4.2.2.

### 6.2.2 – Measures

The PE measures used in this chapter are based on the results of Chapter 4. The final measures used, and calculation of individual PE scores, are detailed in Sections 5.2.2-5.2.5. As in the previous chapter, the PE scores were normalised using inverse rank-based transformation (randomly splitting tied observations). The following covariates were then regressed out of the normalised PE scores: sex, age, age<sup>2</sup>, sex\*age, sex\*age<sup>2</sup>, study, and the top 8 principal components of ancestry.

### 6.2.3 – DNA collection, genotyping, imputation, and quality control

The details of DNA collection, genotyping, and imputation are in Section 5.2.6 and 5.2.7.

### 6.2.4 – Estimation of SNP-heritability

The SNP-heritability was estimated using the combined sample (including TEDS, ALSPAC and CATSS) to provide estimates of phenotypic variance across the three samples that could be accounted for by common genetic variation (mega-SNP-heritability). If the environmental variance within each sample is equal (although heterogeneous), then we can assume a common value of SNP-heritability. Given the presence of heterogeneity between the three samples, an estimate of SNP-heritability using a combined sample (mega-SNP-heritability) is likely to be downward biased due to an increase in phenotypic variance between individuals that is not attributable to common genetic variation. As such, estimates of SNP-heritability were also calculated within each sample and then inverse variance meta-analysed to provide estimates of SNP-heritability in a homogenous sample (meta-SNP-heritability).

#### 6.2.4.1 - GREML-methodologies

##### 6.2.4.1.1 – GREML-SC

GREML was performed using a single component (GREML-SC), meaning that one GRM, based on all tagged genetic variation was used. However, GREML-SC estimates of SNP-heritability may be downward biased if the causal variants have a different minor allele frequency (MAF) spectrum than the variants used in the analysis (Speed, Hemani, Johnson, & Balding, 2012; Wray et al., 2013).

##### 6.2.4.1.2 – GREML-MS

GREML-MS (MAF-stratified)(S Hong Lee et al., 2013), using multiple GRMs based on different MAF bins, is an alternative approach that is robust to a difference in the MAF

spectrum between causal variants and those used in the analysis. However, GREML-MS requires larger samples than GREML-SC and therefore GREML-MS could only be performed when combining the three samples (mega-GREML-MS).

#### 6.2.4.1.3 – GREML-LDMS

GREML-LDMS (LD-score and MAF stratified) (Yang et al., 2015) is an alternative method that controls for both difference in MAF spectrum between causal and tagged variants, and region specific heterogeneity in LD. GREML-LDMS has larger sample size requirements than GREML-MS and so it was also performed in the combined sample only (mega-GREML-LDMS).

#### 6.2.4.1.4 – Incorporating related individuals in GREML analysis

All GREML-analyses included both unrelated and related individuals using a method proposed by Zaitlen and colleagues that enables the simultaneous estimation of SNP-heritability and family-based narrow sense heritability (Zaitlen et al., 2013). GRMs were adjusted for prediction error as recommended by the software developers (Yang, Benyamin, McEvoy, Gordon, Henders, & Others, 2010).

#### 6.2.4.1.5 – Comparison of constrained and unconstrained GREML estimates

The default in GCTA is to constrain estimates of variance explained to between zero and one. If the estimates are unconstrained, in scenarios where the true SNP-heritability is close to zero or above one, estimates of SNP-heritability can be negative or above one respectively. In addition to the default constrained estimates of variance explained unconstrained estimates were also calculated using the --reml-no-constrain option. This was to assess in which scenarios constrained and unconstrained estimates differ.

When analysing the ALSPAC sample alone for meta-GREML analyses, individuals that were cryptically related were removed using a threshold of 0.05 as recommended by the developers of GCTA (Yang, 2016).

To investigate the effect of normalising the dependent variable (PEs) on the SNP-heritability, within sample GREML-SC SNP-heritability estimates were also calculated using untransformed PE residuals.

#### 6.2.4.2 – LD-score regression

LD-score regression was based on summary statistics from mega- and meta- genome-wide analyses in Chapter 5. There was no evidence of confounding in any analysis (Figure 5.6) so the intercept was constrained to 1. As advised by the developers of LD-score regression, the effective sample size was used in LD-score regression analyses, thus matching the sample in the GWAS. The effective sample size was calculated as follows:  $2 * \text{sample size} / 1 + \text{correlation between siblings}$  (Minica et al., 2014).

In summary SNP-heritability was estimated using five approaches: meta-GREML-SC, mega-GREML-SC, mega-GREML-MS, meta-LD-score regression, and mega-LD-score regression.

### **6.3 – Results**

#### 6.3.1 – Comparison of within (meta-) and across (mega-) sample estimates of SNP-heritability

SNP-heritability estimates from constrained meta-GREML-SC and meta-LDSC were between 2.8-8.8% and 6.6-21.5% respectively suggesting some variance in PEs is explained by common additive genetic effects (Table 6.1). SNP-heritability estimates from mega-GREML-SC and mega-LDSC were between 0.0-5.0% and -0.2-13.6% respectively (Table 6.1). No meta- or mega-GREML-SC SNP-heritability estimates were

significantly non-zero (Table 6.1). However, the meta- and mega-LDSC SNP-heritability estimates for Anhedonia and Cognitive Disorganisation, and the meta-LDSC SNP-heritability estimate for Parent-rated Negative Symptoms, were significantly non-zero (Table 6.1).

Estimates of SNP-heritability from GREML-SC and LD-score regression for each sample individually are in Tables 6.2 and 6.3 respectively.

### 6.3.2 – Comparison of different GREML methodologies

#### 6.3.2.1 – Comparison of GREML-SC, -MS and -LDMS estimates

This comparison was only possible for mega-SNP-heritability estimates due to sample size restrictions for GREML analyses using stratified GRMs. Mega-GREML-MS estimates were 0.1% - 6.5%, on average 58% greater than mega-GREML-SC estimates, which were 0.0% - 5.0% (Table 6.4). For traits with available estimates, mega-GREML-LDMS estimates were 10% - 11.1%, on average 54% larger than mega-GREML-MS estimates, which were 4.9% - 6.5% (Table 6.4). Mega-GREML-LDMS estimates were only possible for Anhedonia and Cognitive Disorganisation. Over stratification of the low SNP-heritability for Paranoia and Hallucinations and Parent-rated Negative Symptoms led to over half the components being constrained to  $1 \times 10^{-6}$  and the analysis terminating.

#### 6.3.2.2 – Comparison of constrained and unconstrained GREML estimates

Unconstrained SNP-heritability estimates were consistently lower than constrained estimates (Table 6.5). Estimates from unconstrained mega-GREML-SC and -MS for Paranoia and Hallucinations, and Parent-rated Negative Symptoms were negative. Comparing constrained and unconstrained estimates from mega-GREML-SC, -MS and -LDMS for Anhedonia and Cognitive Disorganisation showed that as the analysis became more stratified (i.e. a larger number of components) the difference between constrained and unconstrained estimates increased. For example, the difference between

constrained and unconstrained SNP-heritability estimates for Anhedonia was 0.3% when using GREML-MS, and 3% when using GREML-LDMS. Comparison of constrained and unconstrained meta-GREML estimates showed that when the SNP-heritability was close to zero, meta-analysis of constrained estimates were larger than those from unconstrained (Table 6.6).

#### 6.3.2.3 – Effect of phenotypic normalisation on GREML estimates

Comparison of within sample GREML-SC SNP-heritability estimates when using normalised and untransformed PEs showed that normalisation typically had no effect on SNP-heritability estimates (Table 6.7). The largest discrepancy between normalised and non-normalised estimates was for Paranoia and Hallucinations in the CATSS sample, which was a difference of 7.1%.

#### 6.3.3 – Distribution of genetic effects across MAF bins

A breakdown of mega-GREML-MS SNP-heritability estimates into MAF bins is in Table 6.8. For both Anhedonia and Cognitive Disorganisation over half the variance was explained by genetic variants with a MAF less than 5% with the remainder of variance explained coming from more common genetic variants. The low SNP-heritability in Paranoia and Hallucinations was accounted for by genetic variants within MAF bins 0.1-0.2 and 0.4-0.5. Due to the very low SNP-heritability estimate for Parent-rated Negative Symptoms, it is difficult to identify different contributions across the MAF spectrum.

#### 6.3.4 – Comparison of LD-score regression and GREML estimates

SNP-heritability estimates from mega- and meta-LD-score regression (Table 6.1) were typically larger than those from GREML-SC, -MS and -LDMS (Table 6.1 and 6.5). For example, unconstrained estimates of SNP-heritability for Anhedonia were 3.3% from mega-GREML-SC, 4.6% from mega-GREML MS, and 7% from mega-GREML LDMS,

whereas mega-LD-score regression estimated the SNP-heritability for Anhedonia as 9.6%.

**Table 6.1. Mega- and meta- SNP-heritability estimates for specific PEs from GREML and LD-score regression.**

PE	<i>Mega-GREML-SC</i>			<i>Meta-GREML-SC</i>			<i>Mega-LDSC</i>			<i>Meta-LDSC</i>		
	<i>SNP-h<sup>2</sup></i>	<i>SE</i>	<i>p</i>									
Paranoia and Hallucinations	1.00x10 <sup>-6</sup>	0.03	0.5	0.028	0.05	0.293	-8.20x10 <sup>-3</sup>	0.039	0.417	0.066	0.068	0.168
Anhedonia	0.033	0.039	0.198	0.088	0.055	0.057	0.096	0.053	0.036	0.204	0.078	4.51x10 <sup>-3</sup>
Cognitive Disorganisation	0.05	0.046	0.138	0.059	0.063	0.176	0.136	0.068	0.022	0.215	0.085	5.79x10 <sup>-3</sup>
Parent-Rated Negative Symptoms	1.00x10 <sup>-6</sup>	0.026	0.5	0.059	0.045	0.096	-0.028	0.035	0.216	0.119	0.061	0.026

Note: *SNP-h<sup>2</sup>* = *SNP-heritability*.

**Table 6.2. Within sample GREML-SC estimates of SNP-heritability for specific PEs.**

PE	TEDS		ALSPAC		CATSS	
	<i>SNP-heritability</i>	<i>SE</i>	<i>SNP-heritability</i>	<i>SE</i>	<i>SNP-heritability</i>	<i>SE</i>
Paranoia and Hallucinations	1.00x10 <sup>-6</sup>	0.098	0.048	0.066	1.00x10 <sup>-6</sup>	0.127
Anhedonia	0.144	0.095	0.059	0.068	NA	NA
Cognitive Disorganisation	0.125	0.098	NA	NA	0.013	0.083
Parent-Rated Negative Symptoms	1.00x10 <sup>-6</sup>	0.106	0.107	0.062	7.74x10 <sup>-3</sup>	0.084

**Table 6.3. Within sample LD-score regression estimates of SNP-heritability for specific PEs.**

PE	TEDS		ALSPAC		CATSS	
	<i>SNP-heritability</i>	<i>SE</i>	<i>SNP-heritability</i>	<i>SE</i>	<i>SNP-heritability</i>	<i>SE</i>
Paranoia and Hallucinations	0.111	0.130	0.023	0.092	0.129	0.163
Anhedonia	0.440	0.135	0.085	0.096	NA	NA
Cognitive Disorganisation	0.189	0.131	NA	NA	0.234	0.112
Parent-rated Negative Symptoms	0.063	0.136	0.133	0.083	0.133	0.121

**Table 6.4. SNP-heritability estimates for specific psychotic experiences from mega-GREML-SC, -MS and -LDMS.**

PE	<u><i>mega-GREML-SC</i></u>		<u><i>mega-GREML-MS</i></u>		<u><i>mega-GREML-LDMS</i></u>	
	<i>SNP-heritability</i>	<i>SE</i>	<i>SNP-heritability</i>	<i>SE</i>	<i>SNP-heritability</i>	<i>SE</i>
Paranoia and Hallucinations	1.00x10 <sup>-6</sup>	0.030	0.021	0.032	NA	NA
Anhedonia	0.033	0.039	0.049	0.040	0.100	0.053
Cognitive Disorganisation	0.050	0.046	0.065	0.047	0.111	0.061
Parent-Rated Negative Symptoms	1.00x10 <sup>-6</sup>	0.026	9.15x10 <sup>-3</sup>	0.028	NA	NA

**Table 6.5. Constrained and unconstrained estimates of SNP-heritability from GREML analyses.**

PE	<u>Mega-GREML-SC</u>		<u>Mega-GREML-MS</u>		<u>Mega-GREML-LDMS</u>	
	Constrained	Unconstrained	Constrained	Unconstrained	Constrained	Unconstrained
Paranoia and Hallucinations	1.00x10 <sup>-6</sup>	-3.14x10 <sup>-2</sup>	0.021	-0.015	NA	NA
Anhedonia	0.033	0.033	0.049	0.046	0.100	0.070
Cognitive Disorganisation	0.050	0.050	0.065	0.044	0.111	0.063
Parent-Rated Negative Symptoms	1.00x10 <sup>-6</sup>	-2.43x10 <sup>-2</sup>	9.15x10 <sup>-3</sup>	-4.20x10 <sup>-2</sup>	NA	NA

**Table 6.6. Constrained and unconstrained meta-GREML-SC estimates of SNP-heritability for specific psychotic experiences.**

PE	<u>Constrained</u>		<u>Unconstrained</u>	
	<i>SNP-heritability</i>	SE	<i>SNP-heritability</i>	SE
Paranoia and Hallucinations	0.028	0.050	2.95x10 <sup>-4</sup>	0.050
Anhedonia	0.088	0.055	0.088	0.055
Cognitive Disorganisation	0.059	0.063	0.059	0.063
Parent-Rated Negative Symptoms	0.059	0.045	0.029	0.044

**Table 6.7. Within sample GREML-SC SNP-heritability estimates for normalised and untransformed specific psychotic experiences.**

PE	<u>TEDS</u>		<u>ALSPAC</u>		<u>CATSS</u>	
	<i>Normalised</i>	<i>Untransformed</i>	<i>Normalised</i>	<i>Untransformed</i>	<i>Normalised</i>	<i>Untransformed</i>
Paranoia and Hallucinations	1.00x10 <sup>-6</sup>	1.00x10 <sup>-6</sup>	0.048	0.038	1.00x10 <sup>-6</sup>	0.071
Anhedonia	0.144	0.143	0.059	0.081	NA	NA
Cognitive Disorganisation	0.125	0.116	NA	NA	0.013	0.014
Parent-Rated Negative Symptoms	1.00x10 <sup>-6</sup>	1.00x10 <sup>-6</sup>	0.107	0.086	0.008	1.00x10 <sup>-6</sup>

**Table 6.8. MAF-breakdown of mega-GREML-MS SNP-heritability estimates.**

<b>Paranoia and Hallucinations</b>	<b>MAF bin</b>	<b>V(G)/Vp</b>	<b>SE</b>
	<0.05	1.00x10 <sup>-6</sup>	0.020
	0.05 - 0.10	1.00x10 <sup>-6</sup>	0.015
	0.10 - 0.20	8.10x10 <sup>-3</sup>	0.021
	0.20 - 0.30	1.00x10 <sup>-6</sup>	0.017
	0.30 - 0.40	1.00x10 <sup>-6</sup>	0.017
	0.40 - 0.50	0.013	0.015
	All	0.021	0.032

<b>Anhedonia</b>	<b>MAF bin</b>	<b>V(G)/Vp</b>	<b>SE</b>
	<0.05	0.027	0.028
	0.05 - 0.10	4.99x10 <sup>-3</sup>	0.022
	0.10 - 0.20	2.89x10 <sup>-3</sup>	0.026
	0.20 - 0.30	1.00x10 <sup>-6</sup>	0.018
	0.30 - 0.40	0.014	0.023
	0.40 - 0.50	1.00x10 <sup>-6</sup>	0.019
	All	0.049	0.040

<b>Cognitive Disorganisation</b>	<b>MAF bin</b>	<b>V(G)/Vp</b>	<b>SE</b>
	<0.05	0.040	0.031
	0.05 - 0.10	1.00x10 <sup>-6</sup>	0.025
	0.10 - 0.20	1.98x10 <sup>-3</sup>	0.030
	0.20 - 0.30	0.010	0.027
	0.30 - 0.40	0.013	0.025
	0.40 - 0.50	1.00x10 <sup>-6</sup>	0.023
	All	0.065	0.047

<b>Parent-rated Negative Symptoms</b>	<b>MAF bin</b>	<b>V(G)/Vp</b>	<b>SE</b>
	<0.05	1.00x10 <sup>-6</sup>	0.019
	0.05 - 0.10	1.88x10 <sup>-3</sup>	0.013
	0.10 - 0.20	1.00x10 <sup>-6</sup>	0.018
	0.20 - 0.30	5.14x10 <sup>-3</sup>	0.017
	0.30 - 0.40	2.13x10 <sup>-3</sup>	0.015
	0.40 - 0.50	1.00x10 <sup>-6</sup>	0.013
	All	9.15x10 <sup>-3</sup>	0.028

*Note.* MAF bin, minor allele frequency range of genetic variation in variance component; V(G)/Vp, phenotypic variance explained by variance component.

## 6.4 – Discussion

This chapter has estimated the SNP-heritability of specific PEs in a number of ways to assess the variance in PEs explained by common genetic variation, and to compare different methodologies.

Although there was some discrepancy between the methods used to estimate the mega-SNP-heritability estimates for specific PEs, there was consistent evidence of non-zero variance explained by common genetic factors for both Anhedonia and Cognitive Disorganisation. Conversely, all mega-SNP-heritability estimates for Paranoia and Hallucinations, and Parent-rated Negative Symptoms suggest little or no phenotypic variance in common across the three samples can be explained by common genetic variation. This pattern of results is congruent with the findings of Sieradzka et al., 2015 (Sieradzka et al., 2015).

For the following reasons, this study deems the most accurate estimates of SNP-heritability for adolescent PEs are those of constrained meta-GREML-SC, providing estimates of 3% - 9%. LD-score regression is less appropriate when based on GWAS summary statistics with a mean chi-square of less than 1.02, and therefore GREML estimates were thought to be more reliable. Meta-SNP-heritability estimates are proposed to be more accurate as they are the average SNP-heritability estimates for a homogenous sample. It is unclear whether constrained estimates are more accurate than unconstrained, but given that a negative SNP-heritability is not realistic, constrained estimates were thought to be accurate.

In comparison to the twin heritability estimates for adolescent PEs of ~30%-50% (Zavos et al., 2014), the SNP-heritability estimates of 3% - 9% means that common genetic variation accounts for ~10% - 19% of the twin-based heritability. The discrepancy between the twin- and SNP-based heritability is well known. One reason for this discrepancy is that SNP-based heritability estimates only consider common additive

genetic effects, whereas twin-based heritability includes common additive genetic effects as well as the effects of rare genetic variation. However, the twin- and SNP-based heritability discrepancy for physical and cognitive phenotypes are in general smaller than the missing heritability for behavioural phenotypes (Trzaskowski, Dale, & Plomin, 2013). The increased discrepancy between twin heritability estimates and SNP-based heritability estimates is particularly prevalent for behavioural traits, as opposed to diagnostic categories (Cross-Disorder Group of the Psychiatric Genomics Consortium, 2013; Thapar & Harold, 2014). A number of explanations have been put forward including increased non-additive genetic effects that are hidden in twin studies by inflated twin similarity, increased gene-environment interaction, or assortative mating leading to larger additive genetic effects (Thapar & Harold, 2014; Trzaskowski, Dale, et al., 2013). Although each of these are reasonable explanations, they are yet to be formally tested. The discrepancy between twin- and SNP-based heritability for adolescent PE is larger than that of schizophrenia, which has a twin heritability is ~80% (Cardno & Gottesman, 2000; P. F. Sullivan, Kendler, & Neale, 2003) and SNP-heritability from GREML-MS is 30% (S Hong Lee et al., 2013).

One feature of Paranoia and Hallucinations, and Parent-rated Negative Symptoms that may contribute to their low SNP-heritability is that they were originally highly skewed. The process of randomly separating ties and subsequent rank-based normalisation of the dependent variable will decrease the correlation between phenotypic similarity and genotypic similarity, which could therefore lead to downward biased estimates of SNP-heritability. This may explain why traits that were originally close to normality (Anhedonia and Cognitive Disorganisation) have larger estimates of SNP-heritability. Comparison of SNP-heritability estimates when using transformed or untransformed PEs demonstrated that the normalisation process rarely had an effect on SNP-heritability estimates. However, it has been reported that GREML underestimates the

SNP-heritability of skewed traits (Nivard et al., 2016). Further simulation studies are required to investigate the effect of skew on estimates of SNP-heritability.

The phenotypic variance that cannot be explained by genetic or shared environmental factors is attributed to non-shared environmental factors which include measurement error. Therefore, the stability of the phenotype being measured will influence its heritability estimate. The degree of measurement error will decrease heritability estimates from both SNP- and twin based approaches and is therefore unlikely to explain the difference between these estimates. Nonetheless, strategies to decrease measurement error and increase the heritability of adolescent PEs should be employed to improve statistical power in future genetic studies. One strategy would be to use measures across multiple time points as genetic factors play an important role in the stability of behaviours across time (Hannigan, Walaker, Waszczuk, McAdams, & Eley, 2017). Adolescent PEs have been reported to have a test-retest reliability of 0.65-0.74 across a 9-month period (Ronald et al., 2014) suggesting that use of PE measures across multiple time points could increase heritability estimates. Other strategies for decreasing measurement error and increasing heritability have been discussed elsewhere (Cheesman et al., 2017).

Mega-SNP-heritability estimates from GREML varied in range of 0% - 10% depending on whether estimates were constrained to be between 0% - 100%, or whether the genetic data was stratified to account for known sources of confounding. Although constraining estimates is the default of the software and limits results to only realistic outcomes, if the true heritability is low, constraining estimates may inflate the total or average heritability estimates when performing stratified or meta- GREML analyses. However, constrained estimates from GREML-LDMS were most similar to estimates from LD-score regression. This agreement between methods could suggest that constrained estimates are more accurate.

The standard errors of meta-SNP-heritability estimates were consistently larger than those of mega-SNP-heritability estimates for both GREML and LD-score regression methodology. For GREML methodologies this difference occurs due to meta-analysis not utilising the relationships between individuals across samples. For LD-score regression this difference may be a result of non-linear relationship between statistical power and sample size.

The comparison of meta- and mega- SNP-heritability estimates provide insight into the extent of heterogeneity between samples. Constrained SNP-heritability estimates across the three samples (mega-SNP heritability) were approximately a third smaller than estimates from meta-analysis of within sample estimates (meta-SNP heritability). This indicates some degree of heterogeneity between the three samples that will lead to reduced statistical power when testing for associations across samples. However, the phenotypic variance in common between the three samples may capture a core aspect of each PE trait. The extent of heterogeneity between samples did vary for different PEs. Cognitive Disorganisation showed the least heterogeneity, as the mega- and meta-SNP heritability estimates were very similar (5% and 6% respectively). In contrast, Parent-rated Negative Symptoms showed the largest extent of heterogeneity with meta-SNP-heritability of 6% but a mega-SNP-heritability of 0%. However it is difficult to compare the differences in SNP-heritability estimates as the standard errors were often overlapping. Due to sample size restrictions, it was not possible to assess genetic heterogeneity between samples by estimating the genetic correlation between measures within each sample.

In summary, this chapter has provided evidence that within a homogenous sample, adolescent PEs can in part be explained by common genetic variation. However, across the different samples used in this chapter, the presence of heterogeneity across samples leads to lower estimates of SNP-heritability. PE traits that were originally more skewed

showed smaller SNP-heritability estimates, possibly reflecting the reduced statistical power when normalising skewed traits. Further work should investigate the effect on SNP-heritability of using models that do not require normality of quantitative traits.

# **Chapter 7 - Assessing the genetic relationship between specific adolescent psychotic experiences in TEDS, ALSPAC and CATSS samples, and major psychiatric disorders**

## **7.1 – Introduction**

Investigating the aetiology of adolescent psychotic experiences (PEs) is important for two key reasons: understanding the complex interplay of genetic and environmental factors across development underlying adolescent behavioural traits, and providing insight into the factors affecting mental health. As previously described in Section 1.2.2, adolescent PEs have been identified as risk factors for a number of health related factors and outcomes including suicide attempts and ideation, and substance abuse in adolescence (Cederlöf et al., 2016; Kelleher et al., 2013, 2014), and psychotic and non-psychotic psychiatric disorders in adulthood (McGrath et al., 2016; Poulton et al., 2000; S. A. Sullivan et al., 2014; Van Os et al., 2009; Welham et al., 2009). The relationships between adolescent PEs and schizophrenia, bipolar disorder, and major depression are the focus of this chapter due to past epidemiological links, their ostensible connection in terms of similarity of phenotype (e.g. paranoia, anhedonia), and the availability of well powered genome-wide association summary statistics for schizophrenia, bipolar disorder and major depression. Although there is a robust phenotypic relationship between adolescent PEs and these psychiatric disorders, evidence of shared genetic factors is limited (described in Section 1.5.3). This chapter will investigate the common genetic association between adolescent PEs and psychiatric disorders using common genotyped variation.

Before considering the evidence of a genetic association between two phenotypes, we must first demonstrate that the two phenotypes of interest can be partly explained by genetic variation. As described in Section 1.5.1, twin studies can be used to estimate the phenotypic variance explained by all genetic factors (additive, dominant, epistatic), and gene-environment correlation and interaction effects (Plomin et al., 2013). Although twin-based heritability estimates are informative, SNP-based heritability estimates are of greater relevance to our ability to test for a common genetic association between two traits. Schizophrenia, bipolar disorder and major depression have all been reported as showing a strong common genetic basis based on analysis in GCTA (Genome-wide Complex Trait Analysis) (Cross-Disorder Group of the Psychiatric Genomics Consortium, 2013) and LD (linkage disequilibrium)-score regression (Bulik-Sullivan et al., 2015). In the previous chapter, the SNP-heritability of specific adolescent PEs was estimated within and across three samples: TEDS (Twins Early Development Study), ALSPAC (Avon Longitudinal Study of Parents and Children), and CATSS (Child and Adolescent Twin study in Sweden).

As previously described in Section 1.6, two samples have been used to investigate the common genetic association between adolescent PEs and schizophrenia using a schizophrenia polygenic risk score (PRS) derived from the schizophrenia PGC2 (Psychiatric Genomics Consortium 2) GWAS. The first study (Sieradzka et al., 2014), using the TEDS sample, reported no positive association between the schizophrenia PRS and any of the six specific PEs. In fact there was evidence of a negative association for several of the specific PE domains. This study reported similar findings using the PGC bipolar disorder PRS. The second study (H. J. Jones et al., 2016), using the ALSPAC sample, reported a significant positive association between the schizophrenia PRS and negative symptoms, but identified no association with positive symptoms (including paranoia, hallucinations and delusions). The mixed findings for negative symptoms could be explained by a number of differences between the studies. One reason may be

that the SNP-heritability of the negative symptom measures used in these two studies differed. Although the measures used in these previous PRS studies vary from those used in the last chapter, evidence from the previous chapter indicates the SNP-heritability for negative symptoms is higher in ALSPAC than it is in the TEDS sample. There has been no published analysis assessing the relationship between adolescent PEs in the CATSS sample and the schizophrenia PRS.

There has been no previous study assessing the common genetic association between adolescent PEs and major depression.

This chapter aims to improve upon these previous studies by assessing the common genetic overlap between schizophrenia, bipolar disorder, and major depression, and a broad range of quantitative and specific adolescent PE domains across multiple samples. This was achieved using both PRS-based analysis and LD-score regression. Given the previously reported positive association between adolescent PEs and these psychiatric disorders on a phenotypic level, a positive genetic association was predicted.

## **7.2 – Methods**

### 7.2.1 – Samples

The three samples used in this chapter have been described in Section 4.2.1. The exclusion criteria applied within each sample have been detailed in Section 4.2.2.

### 7.2.2 – Measures

The measures used in this chapter are based on the results of Chapter 4. The final measures used and calculation of individuals PE scores are detailed in Sections 5.2.2 - 5.2.5. Therefore, the PE scores used in analyses were normally distributed and uncorrelated with the following covariates: sex, age, age<sup>2</sup>, sex\*age, sex\*age<sup>2</sup>, study, and the top 8 principal components of ancestry.

### 7.2.3 – Genotypic data

The details of DNA collection, genotyping, and imputation are in Section 5.2.6 and 5.2.7.

### 7.2.4 – Polygenic risk score analysis

The polygenic risk score for an individual is typically calculated as the sum of risk alleles that individuals carries, weighted by the effect size. Using PRSice (Euesden, Lewis, & O'Reilly, 2015), polygenic risk scores (PRSs) of schizophrenia, bipolar disorder, and major depression were calculated in the adolescent sample using the log of the odds ratios from the latest Psychiatric Genomics Consortium GWAS of schizophrenia (PGC2) (Ripke et al., 2014), bipolar disorder (PGC Bipolar Disorder Working Group, 2011), and major depression (Ripke, Wray, et al., 2013). LD was controlled for using LD-based clumping using the typical  $r^2$ -cutoff of 0.1 within a 250-kb window (Palla & Dudbridge, 2015; Purcell et al., 2007). For each individual, scores were generated using SNPs with the following  $p$ -value thresholds ( $p$ Ts) to define alleles included in the polygenic risk scores: 0.001, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, and 0.5. Linear regression was performed in R, using generalized estimating equations (GEE) with an exchangeable correlation matrix to account for related individuals (Minică et al., 2015).

Logistic regressions comparing PRSs in low and high psychotic experience domain groups (defined as bottom and top 25% of raw psychotic experience sum scores) were performed to confirm the results of linear analyses using normalised PE traits.

The linear associations between the PRSs for schizophrenia, bipolar disorder and major depression, and specific PEs were also calculated within each sample to highlight the contribution of each sample to the across sample results.

### 7.2.5 – Analysis of non-linear polygenic risk score effects

Quantile plots and local polynomial regression were used to examine the linearity of associations. If there was evidence of a non-linear relationship, then the non-linear relationship was formalised by performing linear regression in a subset of individuals.

### 7.2.6 – Analysis of within sample polygenic risk score effects

To demonstrate consistency with previous studies using TEDS and ALSPAC samples, and to investigate differences between all three samples, polygenic risk score associations were also performed within each sample separately.

### 7.2.7 – Estimation of genetic covariance

To estimate the genetic covariance between psychotic experience domains and schizophrenia, bipolar disorder, and major depression, both LD-score regression and AVENGEME (Additive Variance Explained and Number of Genetic Effects Method of Estimation) were used (Bulik-Sullivan et al., 2015; Palla & Dudbridge, 2015).

AVENGEME uses the results of polygenic risk score analyses across multiple significance thresholds to estimate the model parameters including the genetic covariance.

AVENGEME estimates 95% confidence intervals using the profile likelihood method. To improve the accuracy of the estimates of genetic covariance derived from the AVENGEME analysis, the SNP-heritability of liability for schizophrenia, bipolar disorder, and major depression were constrained to the LD-score regression estimates of SNP-heritability (see Table 7.1). The prevalence of schizophrenia, bipolar disorder and major depression were set to 0.01, 0.01, and 0.15 respectively (Cross-Disorder Group of the Psychiatric Genomics Consortium, 2013). When specifying the sample size of the PE sample, the effective sample size was used to account for the presence of related individuals. The effective sample size was calculated as follows:  $2 * \text{sample size} / (1 + \text{correlation between siblings})$  (Minica et al., 2014).

There was no evidence of confounding or sample overlap in the mega-GWAS summary statistics, as such the heritability-intercept was constrained to 1 and the genetic covariance intercept was set to 0 in LD-score regression. These parameters were constrained to reduce the standard error of estimates of genetic covariance and correlation.

Power calculations using the 'polygenescore' function in AVENGEME showed the power to detect moderate levels of genetic covariance varied from 0.09 to 0.85, mainly dependent on the estimated SNP-heritability of each PE (Supplementary Tables 7.1-7.3). LD-score regression's power is dependent on the sample sizes used to calculate summary statistics for both phenotypes, whereas AVENGEME can provide unidirectional estimates that are only dependent on the size of the discovery sample. As a result, LD-score regression analysis will have less power due to the relatively small sample size used to calculate summary statistics for adolescent PEs. Therefore, in these analyses, LD-score regression is thought to have less power.

### 7.2.8 – Estimation of genetic correlation

LD-Score regression automatically estimates the genetic correlation when calculating the genetic covariance. As described above, the heritability-intercept was constrained to 1 and the genetic covariance intercept was set to 0 in LD-score regression.

Genetic covariance estimates from AVENGEME were standardised into genetic correlations using the following formula:

$$r_{G_{12}} = \text{cov}(G_{12})/\text{sqrt}(V_{g_1} * V_{g_2}),$$

where  $\text{cov}(G_{12})$  equals the genetic covariance estimate of either AVENGEME, and  $V_{g_1}$  and  $V_{g_2}$  equal the SNP-heritability estimate of phenotype 1 and 2 respectively. SNP-heritability estimates for schizophrenia, bipolar disorder and major depression were estimated by LD-score regression on a liability scale using the same PGC summary

statistics. Mega-LD-score regression estimates of SNP-heritability for specific PEs were used (Chapter 6). It is not currently possible to calculate the error of genetic correlation estimates from AVENGEME (F. Dudbridge, personal communication, 16 May, 2017).

**Table 7.1. Parameters used in AVENGEME analysis.**

Psych	$N_1$	Samp	Prev	SNP- $h^2_1$	$n$ SNP	Psychotic Experience subscale	$N_2$
SCZ	77096	0.447	0.01	0.2618	102323	Paranoia and Hallucinations	7970
SCZ	77096	0.447	0.01	0.2618	102323	Paranoia and Hallucinations Excl. zero-scorers	3845
SCZ	77096	0.447	0.01	0.2618	102323	Anhedonia	6068
SCZ	77096	0.447	0.01	0.2618	102323	Cognitive Disorganization	5083
SCZ	77096	0.447	0.01	0.2618	102323	Parent-rated Negative Symptoms	8763
BD	16731	0.443	0.01	0.2548	67299	Paranoia and Hallucinations	7970
BD	16731	0.443	0.01	0.2548	67299	Anhedonia	6068
BD	16731	0.443	0.01	0.2548	67299	Cognitive Disorganization	5083
BD	16731	0.443	0.01	0.2548	67299	Parent-rated Negative Symptoms	8763
MDD	18759	0.493	0.15	0.1919	63107	Paranoia and Hallucinations	7970
MDD	18759	0.493	0.15	0.1919	63107	Anhedonia	6068
MDD	18759	0.493	0.15	0.1919	63107	Cognitive Disorganization	5083
MDD	18759	0.493	0.15	0.1919	63107	Parent-rated Negative Symptoms	8763

*Note.* Psych, psychiatric disorder; SCZ, schizophrenia; BD, bipolar disorder; MDD, major depression;  $N_1$ , Number of individuals in discovery sample; Samp, proportion of cases in training sample; Prev, prevalence of disorder in general population; SNP- $h^2_1$ , LDSC estimate of SNP-heritability on a liability scale in training sample;  $n$ SNP, number of LD-independent genetic variants overlapping between discovery and target samples;  $N_2$ , effective sample size of target sample.

## 7.3 – Results

### 7.3.1 - Polygenic risk score association

The schizophrenia PRS significantly and positively predicted Anhedonia ( $p = 0.030$  at  $pT = 0.10$ ), Cognitive Disorganization ( $p = 0.035$  at  $pT = 0.01$ ) and Parent-rated Negative Symptoms ( $p = 5.41 \times 10^{-3}$  at  $pT = 0.05$ ) (Table 7.2; Figure 7.1). The bipolar disorder PRS significantly and *negatively* predicted Paranoia and Hallucinations only ( $p = 2.47 \times 10^{-3}$  at  $pT = 0.010$ ) (note opposite direction to expected) (Table 7.2; Figure 7.1). The major depression PRS significantly and positively predicted Anhedonia ( $p = 0.010$  at  $pT = 0.5$ ) and Parent-rated Negative Symptoms ( $p = 8.29 \times 10^{-3}$  at  $pT = 0.001$ ) (Table 7.2; Figure 7.1). Figures 7.2-7.4 and Supplementary Tables 7.4-7.6 show the full results of these analyses.

Logistic regression comparing PRSs in low and high psychotic experience domain groups (defined as bottom and top 25% of raw psychotic experience sum scores) were congruent with linear analyses (Supplementary Table 7.7, Figure 7.5).

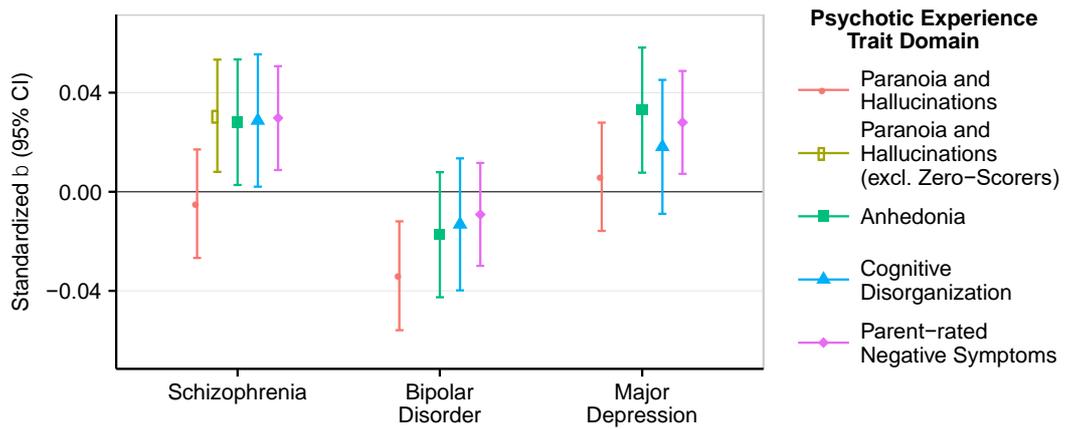
### 7.3.2 – Non-linear effects in Paranoia and Hallucinations scale

Quantile plots showing the mean PRS within subsets of the PE distributions highlighted one non-linear relationship between the schizophrenia PRS and Paranoia and Hallucinations (Figures 7.6-7.7). No other PE-psychiatric disorder PRS quantile plot showed a non-linear relationship. The non-linear relationship between the schizophrenia PRS and Paranoia and Hallucinations was U-shaped with the point of inflection at the median. The majority of individuals (81%) below the median had a raw score of zero. Post-hoc removal of individuals with a raw Paranoia and Hallucinations score of zero led to the schizophrenia PRS positively predicting Paranoia and Hallucinations ( $p = 7.90 \times 10^{-3}$  at  $pT = 0.001$ ) (Table 7.2; Figure 7.1; Supplementary Table

7.4). Logistic regression comparing low and high groups of non-zero scoring individuals supported these findings (Supplementary Table 7.7, Figure 7.5).

### 7.3.3 - Within sample results

Within sample PRS associations with specific PEs are shown in Supplementary Figures 7.1-7.3. Within sample analyses show the positive association between the schizophrenia PRS and Anhedonia, and Parent-rated Negative Symptoms is mainly driven by the ALSPAC sample (Supplementary Figure 7.1). The positive association between schizophrenia and Cognitive Disorganisation is mainly driven by CATSS (Supplementary Figure 7.1). The positive association between schizophrenia PRS, and Paranoia and Hallucinations after excluding zero scorers is driven by both TEDS and ALSPAC (Supplementary Figure 7.1). For bipolar disorder and major depression, the within sample analyses show consistent effects across samples (Supplementary Figures 7.2-7.3).



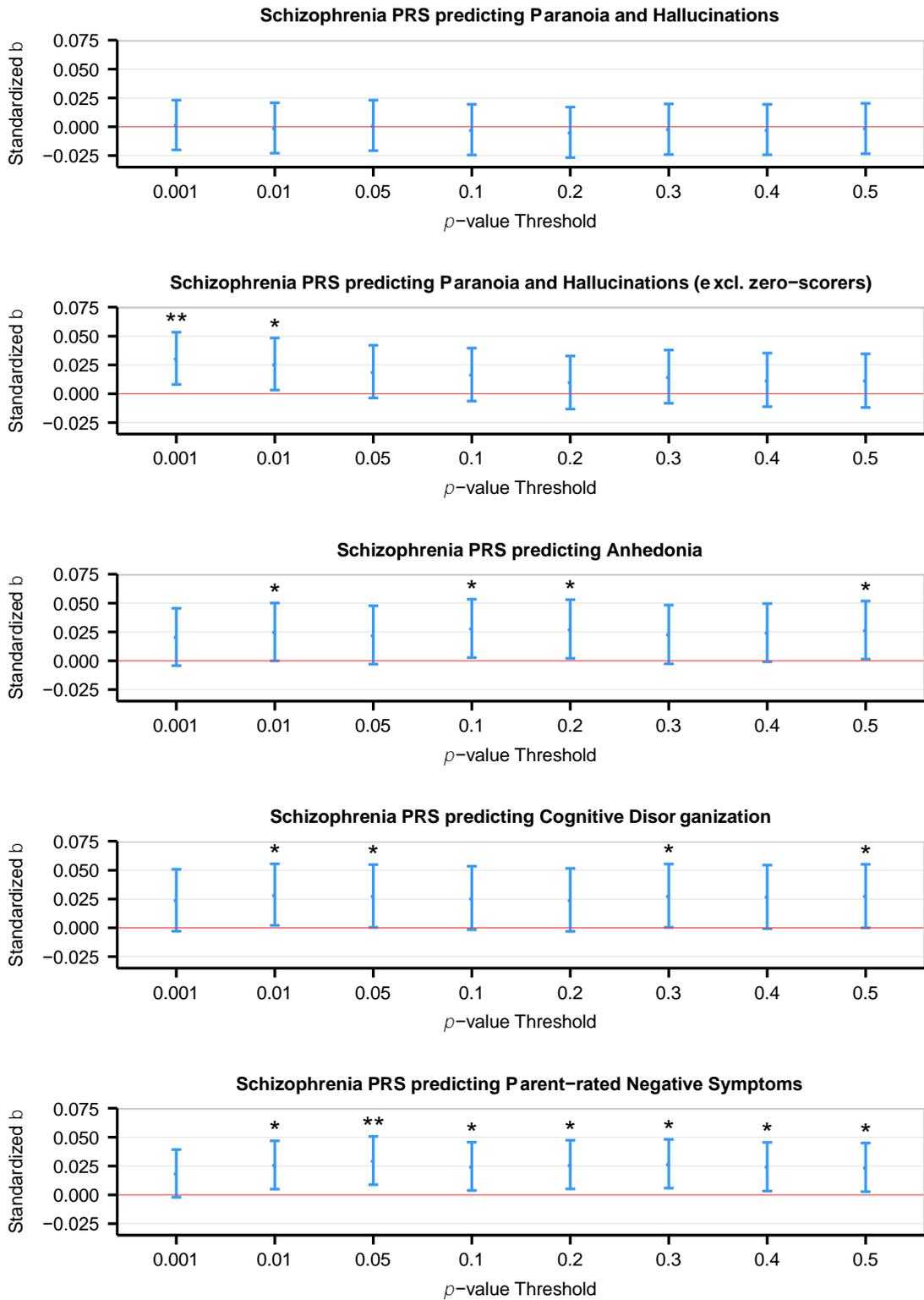
**Figure 7.1. Polygenic risk scores for schizophrenia, bipolar disorder, and major depression predict adolescent psychotic experience domains.**

Note. This figure shows results for polygenic risk scores at the most predictive  $p$ -value threshold for each trait. Error bars are 95% confidence intervals (95% CI).

**Table 7.2. Schizophrenia, bipolar disorder, and major depression polygenic risk scores predicting psychotic experience domains in adolescents.**

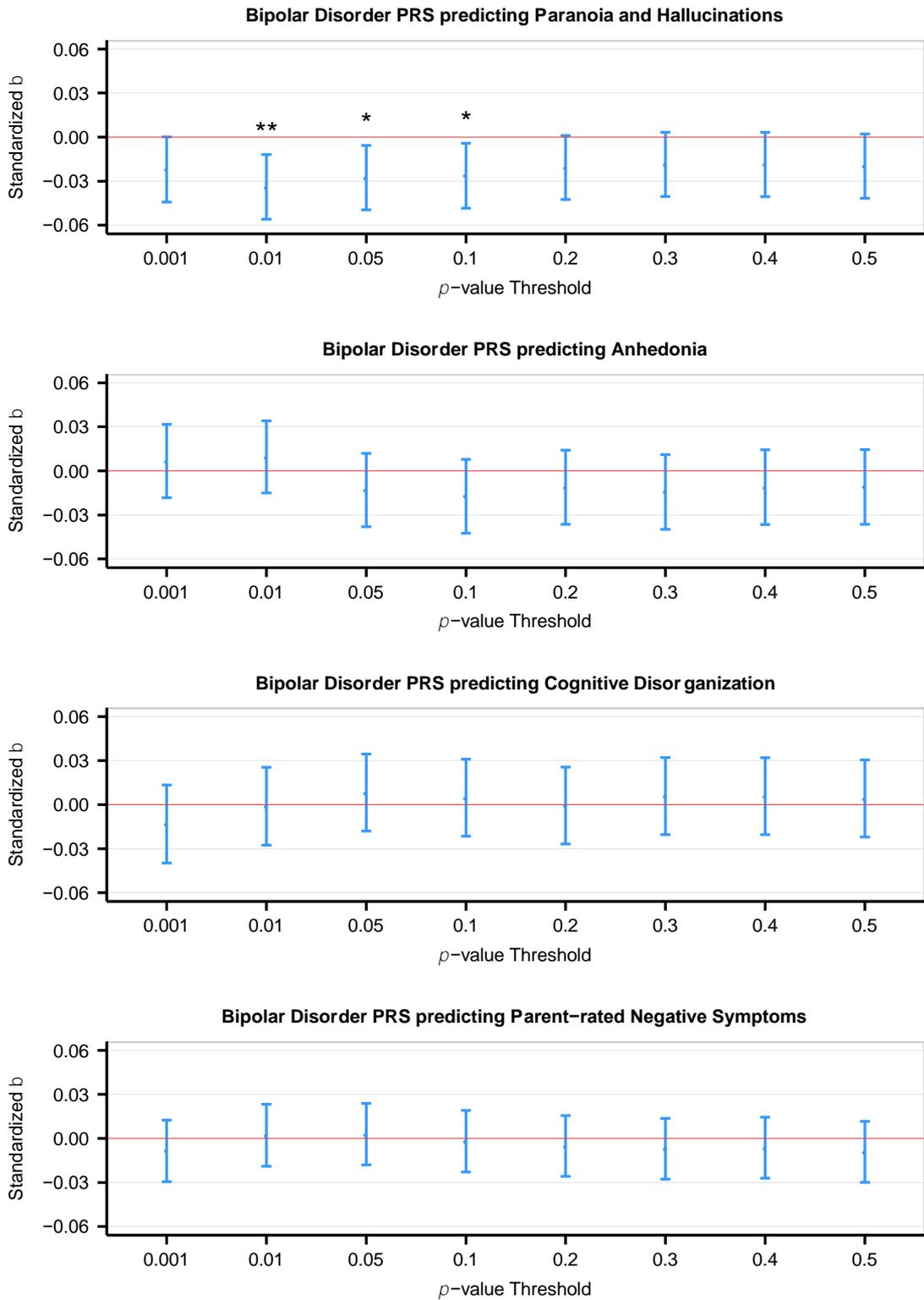
<i>Schizophrenia</i>					
<b>Specific PE</b>	<b><math>\beta</math></b>	<b>SE</b>	<b><math>p</math></b>	<b><math>r^2</math></b>	<b><math>pT</math></b>
Paranoia and Hallucinations	-0.005	0.011	0.664	0.002%	0.2
Paranoia and Hallucinations excl. zero-scorers	0.031	0.012	$7.90 \times 10^{-3}$	0.094%	0.001
Anhedonia	0.028	0.013	0.030	0.079%	0.10
Cognitive Disorganisation	0.029	0.014	0.035	0.083%	0.01
Parent-rated Negative Symptoms	0.030	0.011	$5.41 \times 10^{-3}$	0.088%	0.05
<i>Bipolar Disorder</i>					
<b>Specific PE</b>	<b><math>\beta</math></b>	<b>SE</b>	<b><math>p</math></b>	<b><math>r^2</math></b>	<b><math>pT</math></b>
Paranoia and Hallucinations	-0.034	0.011	$2.47 \times 10^{-3}$	0.115%	0.01
Anhedonia	-0.017	0.013	0.178	0.030%	0.100
Cognitive Disorganisation	-0.013	0.014	0.333	0.017%	0.001
Parent-rated Negative Symptoms	-0.009	0.011	0.388	0.008%	0.5
<i>Major Depression</i>					
<b>Specific PE</b>	<b><math>\beta</math></b>	<b>SE</b>	<b><math>p</math></b>	<b><math>r^2</math></b>	<b><math>pT</math></b>
Paranoia and Hallucinations	0.006	0.011	0.589	0.004%	0.1
Anhedonia	0.033	0.013	0.010	0.109%	0.5
Cognitive Disorganisation	0.018	0.014	0.189	0.033%	0.05
Parent-rated Negative Symptoms	0.028	0.011	$8.29 \times 10^{-3}$	0.078%	0.001

*Note.* This table shows results for polygenic risk scores at the most predictive  $p$ -value threshold for each trait. Specific PE, specific psychotic experience;  $\beta$ , standardised beta value;  $pT$ ,  $p$ -value threshold for selecting genetic variants included in the polygenic risk score.



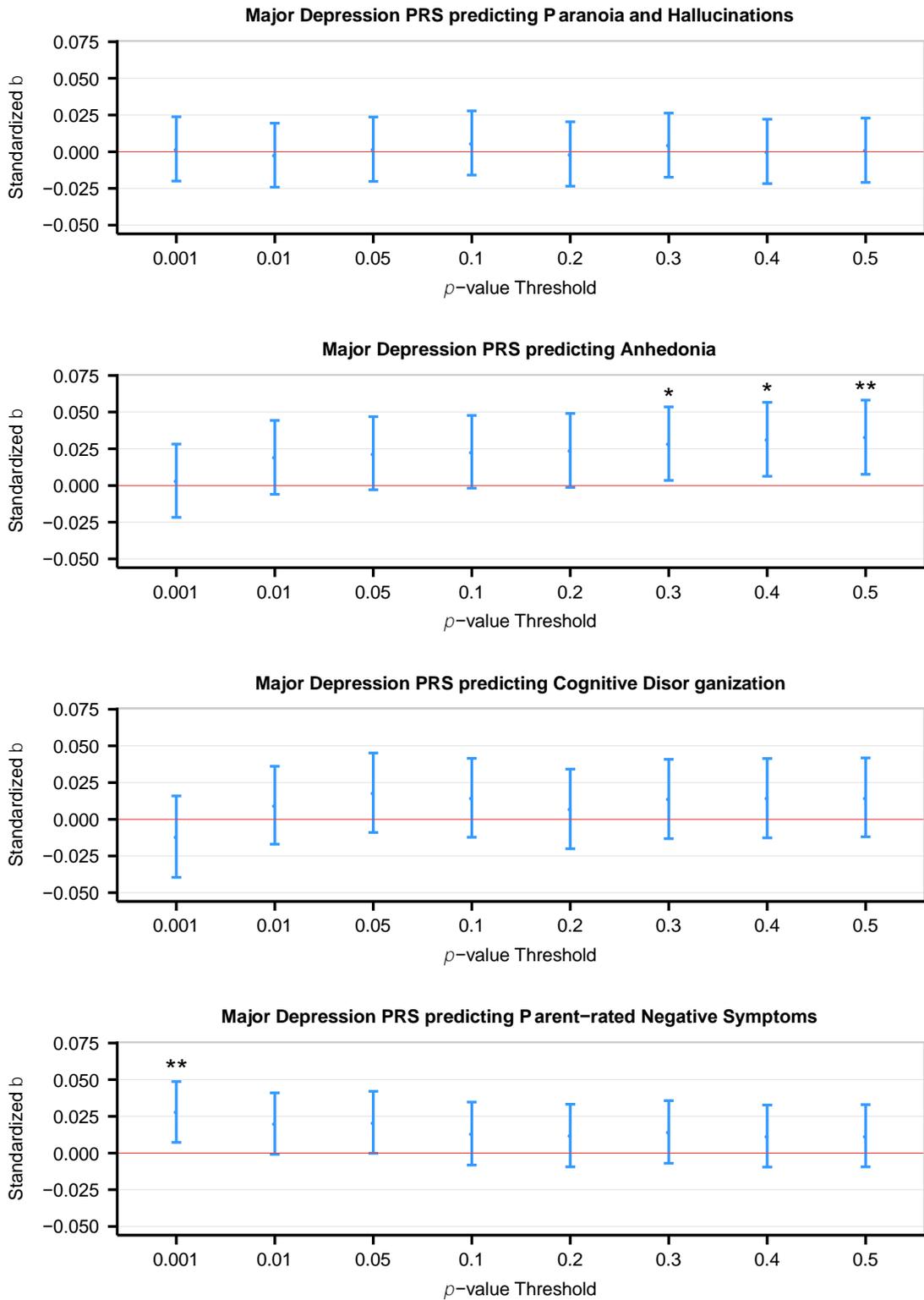
**Figure 7.2. Schizophrenia polygenic risk score predicting psychotic experience domains in adolescence.**

*Note.* \*,  $p \leq 0.05$ ; \*\*,  $p \leq 0.01$ . Linear regression results are shown for polygenic risk scores at all  $p$ -value thresholds. Figure corresponds to results shown in Supplementary Table 7.4. Error bars indicate 95% confidence intervals.



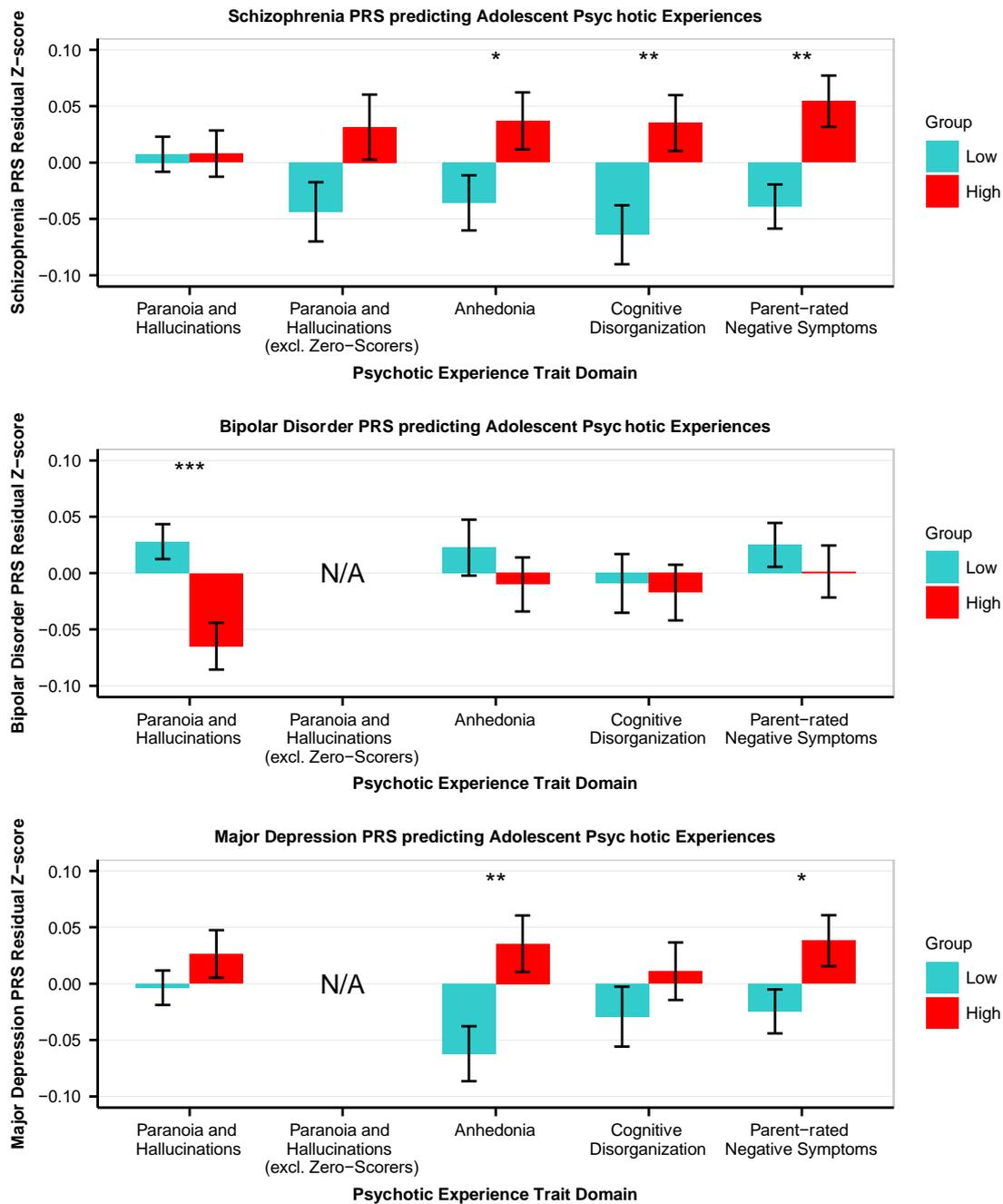
**Figure 7.3. Bipolar disorder polygenic risk score predicting psychotic experience domains in adolescence.**

*Note.* \*,  $p \leq 0.05$ ; \*\*,  $p \leq 0.01$ . Linear regression results are shown for polygenic risk scores at all p-value thresholds. Figure corresponds to results shown in Supplementary Table 7.5. Error bars indicate 95% confidence intervals.



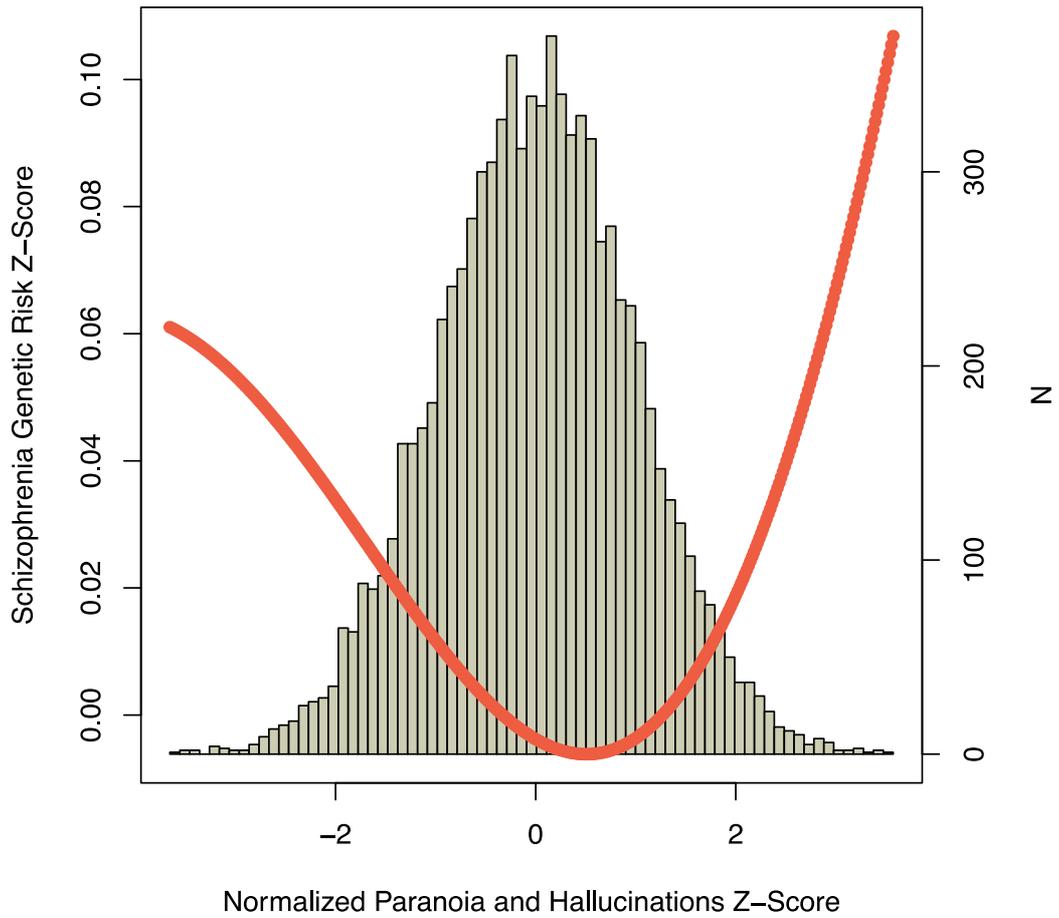
**Figure 7.4. Major depression polygenic risk score predicting psychotic experience domains in adolescence.**

*Note.* \*,  $p \leq 0.05$ ; \*\*,  $p \leq 0.01$ . Linear regression results are shown for polygenic risk scores at all p-value thresholds. Figure corresponds to results shown in Supplementary Table 7.6. Error bars indicate 95% confidence intervals.



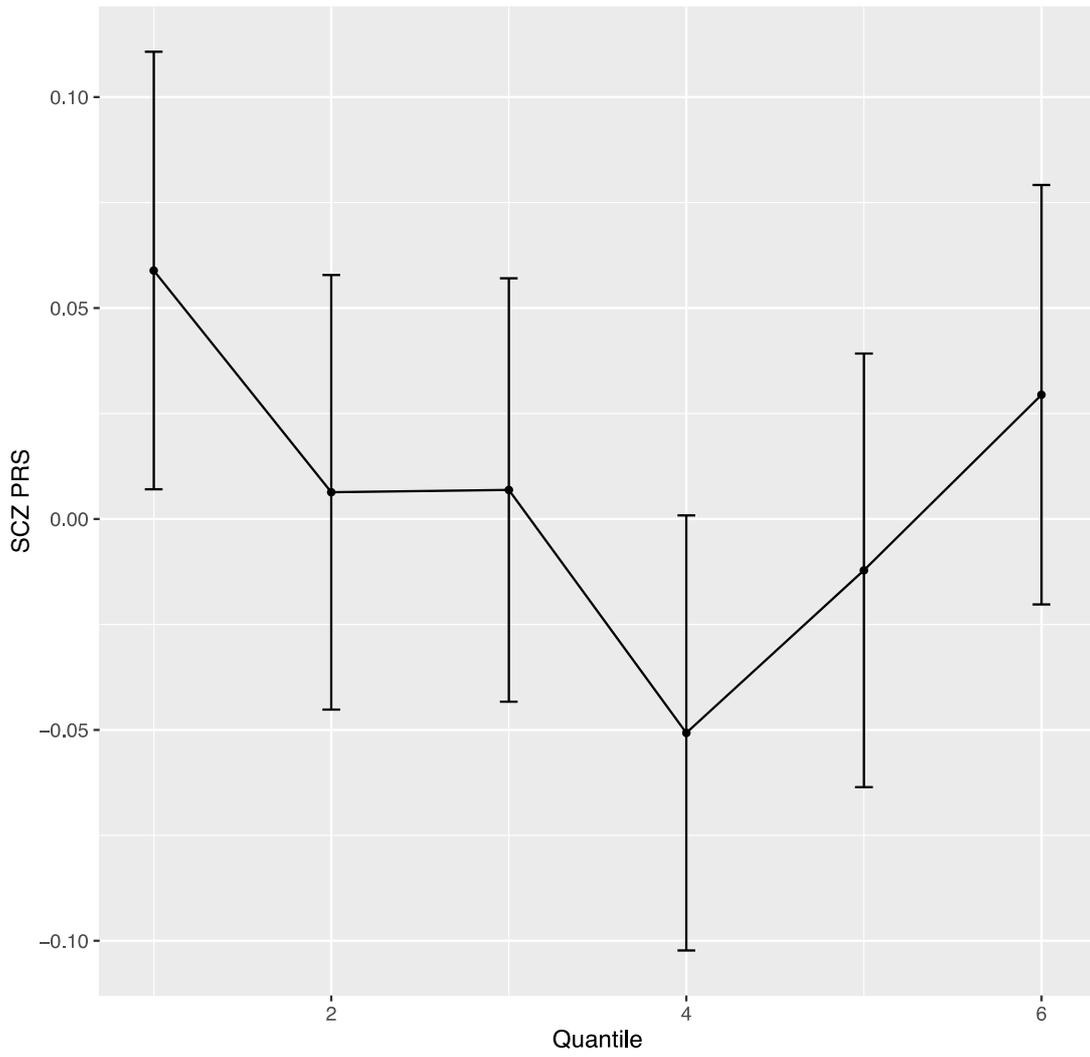
**Figure 7.5. Schizophrenia, bipolar disorder, and major depression polygenic risk score mean differences between low- and high-scoring psychotic experience domain groups.**

Note. \*,  $p \leq 0.05$ ; \*\*,  $p \leq 0.01$ ; \*\*\*,  $p \leq 0.001$ ; PRS, polygenic risk score. Polygenic risk scores were adjusted to control for covariate effects. Low- and high-scoring groups determined as the bottom and top 25% of raw psychotic experience domain sum scores. This plot shows mean differences for polygenic risk scores at the most predictive  $p$ -value threshold for each trait. Significance of mean difference was determined using logistic regression (results shown in Supplementary Table 7.7). Error bars indicate the standard error of the mean.



**Figure 7.6. Local polynomial regression of schizophrenia polygenic risk score ( $p$ -value threshold of  $p < 0.3$ ) and Paranoia and Hallucinations.**

*Note.* The red line indicates the schizophrenia polygenic risk score (left y-axis) of individuals across the Paranoia and Hallucinations distribution. Histogram in background shows number of individuals (N, right y-axis) across the Paranoia and Hallucinations distribution.



**Figure 7.7. Mean schizophrenia polygenic risk score (SCZ PRS) in six quantiles of Paranoia and Hallucinations scores.**

#### 7.3.4 – Estimates of genetic covariance

Table 7.3 presents the AVENGEME estimates of genetic covariance, which were highly congruent with the PRS analysis results. AVENGEME pools evidence across  $p$ -value thresholds tested from PRS analysis. As such, even when there is consistent but non-significant evidence of association at individual  $p$ -value thresholds, AVENGEME genetic covariance estimates can be significant. Consequently, there were two significant results that were not shown by the PRS analyses; between the Anhedonia PE domain and bipolar disorder (genetic covariance = -0.022, 95%CI = -0.041 – -0.002), and the Cognitive Disorganization PE domain and major depression (genetic covariance = 0.033, 95%CI = 0.005 – 0.062).

Table 7.3 presents estimates of genetic covariance from LD-score regression. LD-score regression mirrored over half of the significant associations shown between the equivalent polygenic risk scores and psychotic experience domains in Table 7.2. The genetic covariance between the schizophrenia PRS and non-zero scorers on Paranoia and Hallucinations could not be estimated because the genome-wide association analysis included zero scorers. Unlike for the equivalent results in Table 7.2, the genetic covariance between schizophrenia and anhedonia psychotic experience domain, and between major depression and parent-rated negative symptoms domain, were not significant.

#### 7.3.5 – Estimates of genetic correlation

Genetic correlations between Anhedonia and Cognitive Disorganisation with schizophrenia, bipolar disorder, and major depression are in Table 7.4 (Genetic correlation estimates were not possible for Paranoia and Hallucinations, and Parent-rated Negative Symptoms as the mega-LDSC SNP-heritability estimates were negative).

It is not currently possible to estimate the errors of genetic correlations derived from AVENGEME estimates of genetic covariance. Consistent with the genetic covariance results from LD-score regression, genetic correlation estimates were significant between Cognitive Disorganisation and schizophrenia (LD-score regression:  $r_G = 0.205$ ,  $SE = 0.090$ ,  $p = 0.023$ ; AVENGEME:  $r_G = 0.134$ ), and Anhedonia and major depression (LD-score regression:  $r_G = 0.432$ ,  $SE = 0.193$ ,  $p = 0.025$ ; AVENGEME:  $r_G = 0.477$ ).

**Table 7.3. Genetic covariance between each psychotic experience domain and schizophrenia, bipolar disorder, and major depression.**

**Schizophrenia**

Specific PE	AVENGEME			LDSC		
	<i>cov(G)</i>	Low 95% CI	High 95% CI	<i>cov(G)</i>	SE	<i>p</i>
Paranoia and Hallucinations	-0.002	-0.013	0.008	-0.003	0.014	0.844
P and H excl. zero-scorers	<b>0.019</b>	<b>0.011</b>	<b>0.029</b>	NA	NA	NA
Anhedonia	<b>0.024</b>	<b>0.013</b>	<b>0.038</b>	0.021	0.016	0.184
Cognitive Disorganisation	<b>0.025</b>	<b>0.013</b>	<b>0.040</b>	<b>0.054</b>	<b>0.018</b>	<b>3.52x10<sup>-3</sup></b>
Parent-rated Negative Symptoms	<b>0.025</b>	<b>0.015</b>	<b>0.036</b>	<b>0.047</b>	<b>0.014</b>	<b>5.57x10<sup>-4</sup></b>

**Bipolar Disorder**

Specific PE	AVENGEME			LDSC		
	<i>cov(G)</i>	Low 95% CI	High 95% CI	<i>cov(G)</i>	SE	<i>p</i>
Paranoia and Hallucinations	<b>-0.032</b>	<b>-0.048</b>	<b>-0.018</b>	<b>-0.045</b>	<b>0.022</b>	<b>0.039</b>
Anhedonia	<b>-0.022</b>	<b>-0.041</b>	<b>-0.002</b>	0.008	0.028	0.767
Cognitive Disorganisation	0.008	-0.013	0.029	0.021	0.028	0.451
Parent-rated Negative Symptoms	-0.010	-0.026	0.006	0.009	0.021	0.669

**Major Depressive Disorder**

Specific PE	AVENGEME			LDSC		
	<i>cov(G)</i>	Low 95% CI	High 95% CI	<i>cov(G)</i>	SE	<i>p</i>
Paranoia and Hallucinations	0.004	-0.018	0.026	0.013	0.020	0.523
Anhedonia	<b>0.065</b>	<b>0.037</b>	<b>0.094</b>	<b>0.061</b>	<b>0.023</b>	<b>7.50x10<sup>-3</sup></b>
Cognitive Disorganisation	<b>0.033</b>	<b>0.005</b>	<b>0.062</b>	0.015	0.028	0.594
Parent-rated Negative Symptoms	<b>0.023</b>	<b>0.010</b>	<b>0.042</b>	0.031	0.020	0.119

Note. PE, psychotic experience; *cov(G)*, genetic covariance; CI, confidence interval; AVENGEME, additive variance explained and number of genetic effects method of estimation; LDSC, linkage-disequilibrium score regression.

**Table 7.4. Estimates of genetic correlation between specific adolescent psychotic experiences, and schizophrenia, bipolar disorder, and major depression.**

Psych	PE	<i>AVENGEME</i>	<i>LDSC</i>		
		rG	rG	SE	<i>p</i>
SCZ	Paranoia and Hallucinations	NA	NA	NA	NA
SCZ	P and H excl. zero-scorers	NA	NA	NA	NA
SCZ	Anhedonia	<b>0.154</b>	0.097	0.074	0.188
SCZ	Cognitive Disorganisation	<b>0.134</b>	<b>0.205</b>	<b>0.090</b>	<b>0.023</b>
SCZ	Parent-rated Negative Symptoms	NA	NA	NA	NA
BD	Paranoia and Hallucinations	NA	NA	NA	NA
BD	Anhedonia	-0.138	0.038	0.128	0.768
BD	Cognitive Disorganisation	0.041	0.087	0.116	0.452
BD	Parent-rated Negative Symptoms	NA	NA	NA	NA
MDD	Paranoia and Hallucinations	NA	NA	NA	NA
MDD	Anhedonia	<b>0.477</b>	<b>0.432</b>	<b>0.193</b>	<b>0.025</b>
MDD	Cognitive Disorganisation	0.204	0.096	0.182	0.600
MDD	Parent-rated Negative Symptoms	NA	NA	NA	NA

Note. Psych, psychiatric disorder; SCZ, schizophrenia; BD, bipolar disorder; MDD, major depression; PE, psychotic experience; P and H excl. zero-scorers, Paranoia and Hallucinations excluding zero-scorers; rG, genetic correlation; AVENGEME, additive variance explained and number of genetic effects method of estimation; LDSC, linkage-disequilibrium score regression.

## 7.4 – Discussion

This study has tested for a common genetic overlap between a broad range of adolescent PEs and schizophrenia, bipolar disorder and major depression in the largest sample to date. This is the first study to find significant genetic overlap between schizophrenia and such a wide range of PEs.

Results showed significant and positive genetic covariance between schizophrenia and Paranoia and Hallucinations (in non-zero scorers only), Anhedonia, Cognitive Disorganisation, and Parent-rated Negative Symptoms. LD-score regression results were congruent with those of AVENGEME. The best fitting schizophrenia PRSs predicted between 0.08 – 0.09% of the variance in Paranoia and Hallucinations (in non-zero scorers only), Anhedonia, Cognitive Disorganisation, and Parent-rated Negative Symptoms. Although only a small amount variance in PEs was explained by the schizophrenia PRS, genetic correlation estimates, where possible, were approximately 0.15.

The use of quantitative measures of PEs allowed investigation of non-linear effects. Our study finds evidence that Paranoia and Hallucinations during adolescence are only associated with schizophrenia genetic risk if the individual reports at least some degree of paranoia or hallucinations. Individuals reporting no Paranoia and Hallucinations can exist anywhere on the schizophrenia genetic liability spectrum. This finding requires further investigation with longitudinal data. An explanation may lie in the fact that age of onset of paranoia and hallucinations varies widely among individuals: our study was focused on PEs in mid to late adolescence.

The combined analysis of three samples has provided a larger sample size than previous studies and as a result more power to detect associations at significance. Furthermore, the use of multiple samples means the overall effects observed are more generalizable and likely to hold true for other samples. The within sample results relating to

schizophrenia are consistent with those of previous studies in TEDS and ALSPAC (H. J. Jones et al., 2016; Sieradzka et al., 2014), but the combined analysis of a broad range of specific and quantitative measures has provided some solidity to previously mixed findings, and has offered several new insights into the relationship between adolescent PEs and schizophrenia. These new insights include the genetic association between the schizophrenia and non-zero Paranoia and Hallucinations, Anhedonia, and Cognitive Disorganisation.

In addition to testing for a common genetic association between adolescent PEs and schizophrenia, this study also examined the common genetic relationship between adolescent PEs and major depression and bipolar disorder.

Significant and positive genetic covariance between major depression and Anhedonia, Cognitive Disorganisation, and Parent rated Negative Symptoms was found, and LD-score regression results were consistent in direction of effect, although not all estimates were significant. The genetic correlation between major depression and Anhedonia was significant with estimates  $\sim 0.45$ . This finding is in accordance with a previous study reporting that subclinical depressive symptoms (including anhedonia) are a strong predictor of major depressive episodes in adulthood (Pine, Cohen, Cohen, & Brook, 1999). Anhedonia is present as a symptom of both schizophrenia and depression in psychiatric diagnoses, and our research shows that as a trait dimension in adolescence it shares common genetic underpinnings with both schizophrenia and depression.

Contrary to our hypothesis, significant negative genetic covariance was found between bipolar disorder and Paranoia and Hallucinations, and Anhedonia. The significant negative genetic covariance between bipolar disorder and Paranoia and Hallucinations conflicts with previous reports of increased paranoia, hallucinations, and delusions prior to the onset of bipolar disorder (McGrath et al., 2016). Although the negative association is somewhat surprising, the absence of a positive association is less so given

that the bipolar disorder PRS can only explain 2.83% of the variance in bipolar disorder (PGC Bipolar Disorder Working Group, 2011).

The previous chapter estimated the SNP-heritability of adolescent PEs, with zero SNP-heritability estimates when looking across the three samples for Paranoia and Hallucinations, and Parent-rated Negative Symptoms. Given these zero SNP-heritability estimates, it is somewhat counterintuitive that these traits can have a significant genetic covariance with other phenotypes. However, when estimating the SNP-heritability within each sample separately, thereby removing the between sample heterogeneity, the SNP-heritability estimates for Paranoia and Hallucinations, and Parent-rated Negative Symptoms were 3% and 5 % respectively. Therefore, the effects within each of the samples could drive the genetic covariance observed in this study. Another explanation is that the across sample SNP-heritability of Paranoia and Hallucinations, and Parent-rated Negative Symptoms are not zero, but are very close to zero.

It should be noted that the significant genetic covariance estimates reported here could be partly explained by the presence of adolescents with diagnosed relatives. If an adolescent has a relative with schizophrenia for example, the adolescent may have increased PEs due to their shared environment with the relative. Future studies should investigate the effect of parental diagnosis on the genetic covariance between adolescent PEs and psychiatric disorders.

Collectively these findings provide evidence of a common genetic overlap between a broad range of adolescent PEs, schizophrenia and major depression, suggesting that investigating the common genetic basis of adolescent PEs could provide insight into the development of these major psychiatric disorders. However, given the heterogeneity between within-sample PRS analysis results, the relationship between specific adolescent PE and adult psychiatric disorders should be further explored. The clinical, theoretical and genetic implications of these findings will be considered in the following

chapter. Furthermore, schizophrenia polygenic risk appears to be only predictive of Paranoia and Hallucinations among individuals reporting at least some experiences of paranoia or hallucinations. This relationship was observed in all three samples and should be further investigated to understand the factors underlying it.

## 7.5 – Appendix

### Supplementary Tables:

**Supplementary Table 7.1. Power calculations for genetic covariance analysis between PEs and schizophrenia.**

PE	$N_1$	$h^2_1$	$N_2$	$h^2_2$	rG	Alpha	Power
<b>Paranoia and Hallucinations</b>	77096	0.2618	7970	0.028	0.2	0.05	0.26
<b>Anhedonia</b>	77096	0.2618	6068	0.088	0.2	0.05	0.52
<b>Cognitive Disorganization</b>	77096	0.2618	5083	0.059	0.2	0.05	0.33
<b>Parent-rated Negative Symptoms</b>	77096	0.2618	8763	0.059	0.2	0.05	0.52
<b>Paranoia and Hallucinations</b>	77096	0.2618	7970	0.028	0.3	0.05	0.51
<b>Anhedonia</b>	77096	0.2618	6068	0.088	0.3	0.05	0.85
<b>Cognitive Disorganization</b>	77096	0.2618	5083	0.059	0.3	0.05	0.63
<b>Parent-rated Negative Symptoms</b>	77096	0.2618	8763	0.059	0.3	0.05	0.85

*Note.* PE, psychotic experience;  $N_1$ , sample size of discovery sample;  $h^2_1$ , SNP-heritability of disorder in training sample on a liability scale;  $N_2$ , sample size of target sample;  $h^2_2$ , SNP-heritability of psychotic experience in target sample. Set as meta-GREML-SC estimate from Chapter 6; rG, genetic correlation. Estimates calculated using AVENGEME, assuming 100,000 LD independent genetic markers overlap between samples, 70% of genetic markers have no effect on the training trait, and an alpha of 0.05.

**Supplementary Table 7.2. Power calculations for genetic covariance analysis between PEs and bipolar disorder.**

PE	$N_1$	$h^2_1$	$N_2$	$h^2_2$	rG	Alpha	Power
<b>Paranoia and Hallucinations</b>	16731	0.2548	7970	0.028	0.2	0.05	0.09
<b>Anhedonia</b>	16731	0.2548	6068	0.088	0.2	0.05	0.16
<b>Cognitive Disorganization</b>	16731	0.2548	5083	0.059	0.2	0.05	0.11
<b>Parent-rated Negative Symptoms</b>	16731	0.2548	8763	0.059	0.2	0.05	0.15
<b>Paranoia and Hallucinations</b>	16731	0.2548	7970	0.028	0.3	0.05	0.15
<b>Anhedonia</b>	16731	0.2548	6068	0.088	0.3	0.05	0.30
<b>Cognitive Disorganization</b>	16731	0.2548	5083	0.059	0.3	0.05	0.19
<b>Parent-rated Negative Symptoms</b>	16731	0.2548	8763	0.059	0.3	0.05	0.29

Note. PE, psychotic experience;  $N_1$ , sample size of discovery sample;  $h^2_1$ , SNP-heritability of disorder in training sample on a liability scale;  $N_2$ , sample size of target sample;  $h^2_2$ , SNP-heritability of psychotic experience in target sample. Set as meta-GREML-SC estimate from Chapter 6; rG, genetic correlation. Estimates calculated using AVENGEME, assuming 100,000 LD independent genetic markers overlap between samples, 70% of genetic markers have no effect on the training trait, and an alpha of 0.05.

**Supplementary Table 7.3. Power calculations for genetic covariance analysis between PEs and major depression.**

PE	$N_1$	$h^2_1$	$N_2$	$h^2_2$	rG	Alpha	Power
<b>Paranoia and Hallucinations</b>	18759	0.1919	7970	0.028	0.2	0.05	0.09
<b>Anhedonia</b>	18759	0.1919	6068	0.088	0.2	0.05	0.14
<b>Cognitive Disorganization</b>	18759	0.1919	5083	0.059	0.2	0.05	0.10
<b>Parent-rated Negative Symptoms</b>	18759	0.1919	8763	0.059	0.2	0.05	0.14
<b>Paranoia and Hallucinations</b>	18759	0.1919	7970	0.028	0.3	0.05	0.14
<b>Anhedonia</b>	18759	0.1919	6068	0.088	0.3	0.05	0.26
<b>Cognitive Disorganization</b>	18759	0.1919	5083	0.059	0.3	0.05	0.17
<b>Parent-rated Negative Symptoms</b>	18759	0.1919	8763	0.059	0.3	0.05	0.25

*Note.* PE, psychotic experience;  $N_1$ , sample size of discovery sample;  $h^2_1$ , SNP-heritability of disorder in training sample on a liability scale;  $N_2$ , sample size of target sample;  $h^2_2$ , SNP-heritability of psychotic experience in target sample. Set as meta-GREML-SC estimate from Chapter 6; rG, genetic correlation. Estimates calculated using AVENGEME, assuming 100,000 LD independent genetic markers overlap between samples, 70% of genetic markers have no effect on the training trait, and an alpha of 0.05.

**Supplementary Table 7.4. Schizophrenia polygenic risk score predicting psychotic experience domains at 8  $p$ -value thresholds.**

<b>Paranoia and Hallucinations</b>				
$pT$	$\beta$	SE	$p$	$r^2$
0.001	0.002	0.011	0.889	0.000%
0.01	-0.001	0.011	0.926	0.000%
0.05	0.001	0.011	0.915	0.000%
0.1	-0.002	0.011	0.824	0.001%
0.2	-0.005	0.011	0.664	0.002%
0.3	-0.002	0.011	0.848	0.000%
0.4	-0.002	0.011	0.832	0.001%
0.5	-0.002	0.011	0.892	0.000%
<b>Paranoia and Hallucinations (excl. zero-scorers)</b>				
$pT$	$\beta$	SE	$p$	$r^2$
0.001	0.031	0.012	0.008	0.094%
0.01	0.026	0.012	0.025	0.067%
0.05	0.019	0.012	0.098	0.037%
0.1	0.017	0.012	0.156	0.028%
0.3	0.015	0.012	0.208	0.022%
0.4	0.012	0.012	0.311	0.014%
0.5	0.011	0.012	0.336	0.013%
0.2	0.010	0.012	0.404	0.010%
<b>Anhedonia</b>				
$pT$	$\beta$	SE	$p$	$r^2$
0.001	0.021	0.013	0.104	0.043%
0.01	0.025	0.013	0.050	0.063%
0.05	0.022	0.013	0.083	0.050%
0.1	0.028	0.013	0.030	0.079%
0.2	0.028	0.013	0.034	0.076%
0.3	0.023	0.013	0.077	0.052%
0.4	0.024	0.013	0.059	0.059%
0.5	0.027	0.013	0.039	0.071%
<b>Cognitive Disorganisation</b>				
$pT$	$\beta$	SE	$p$	$r^2$
0.001	0.024	0.014	0.081	0.057%
0.01	0.029	0.014	0.035	0.083%
0.05	0.028	0.014	0.048	0.076%
0.1	0.026	0.014	0.065	0.067%
0.2	0.024	0.014	0.082	0.059%
0.3	0.028	0.014	0.048	0.077%
0.4	0.027	0.014	0.056	0.072%
0.5	0.028	0.014	0.050	0.076%

**Supplementary Table 7.4 cont.**

<b>Negative Symptoms</b>				
<b><math>pT</math></b>	<b><math>\beta</math></b>	<b>SE</b>	<b><math>p</math></b>	<b><math>r^2</math></b>
0.001	0.019	0.011	0.078	0.034%
0.01	0.026	0.011	0.015	0.067%
0.05	0.030	0.011	0.005	0.088%
0.1	0.025	0.011	0.021	0.061%
0.2	0.026	0.011	0.014	0.069%
0.3	0.027	0.011	0.012	0.073%
0.4	0.024	0.011	0.023	0.059%
0.5	0.024	0.011	0.027	0.057%

*Note.*  $pT$ , p-value threshold used to select genetic variation included in risk score;  $\beta$ , standardised beta value.

**Supplementary Table 7.5. Bipolar disorder polygenic risk score predicting psychotic experience domains at 8  $p$ -value thresholds.**

<b>Paranoia and Hallucinations</b>				
$pT$	$\beta$	SE	$p$	$r^2$
0.001	-0.022	0.011	0.051	0.049%
0.01	-0.034	0.011	0.002	0.115%
0.05	-0.028	0.011	0.014	0.076%
0.1	-0.026	0.011	0.019	0.069%
0.2	-0.021	0.011	0.062	0.043%
0.3	-0.019	0.011	0.095	0.035%
0.4	-0.019	0.011	0.094	0.035%
0.5	-0.020	0.011	0.077	0.039%
<b>Anhedonia</b>				
$pT$	$\beta$	SE	$p$	$r^2$
0.001	0.007	0.013	0.600	0.004%
0.01	0.010	0.013	0.448	0.009%
0.05	-0.013	0.013	0.304	0.017%
0.1	-0.017	0.013	0.178	0.030%
0.2	-0.011	0.013	0.385	0.013%
0.3	-0.014	0.013	0.269	0.021%
0.4	-0.011	0.013	0.391	0.012%
0.5	-0.011	0.013	0.399	0.012%
<b>Cognitive Disorganisation</b>				
$pT$	$\beta$	SE	$p$	$r^2$
0.001	-0.013	0.014	0.333	0.017%
0.01	-0.001	0.014	0.937	0.000%
0.05	0.008	0.013	0.541	0.007%
0.1	0.005	0.013	0.725	0.002%
0.2	-0.001	0.013	0.964	0.000%
0.3	0.006	0.013	0.665	0.003%
0.4	0.006	0.013	0.670	0.003%
0.5	0.004	0.013	0.754	0.002%
<b>Parent-rated Negative Symptoms</b>				
$pT$	$\beta$	SE	$p$	$r^2$
0.001	-0.009	0.011	0.425	0.007%
0.01	0.002	0.011	0.837	0.000%
0.05	0.003	0.011	0.786	0.001%
0.1	-0.002	0.011	0.858	0.000%
0.2	-0.005	0.011	0.626	0.003%
0.3	-0.007	0.011	0.506	0.005%
0.4	-0.006	0.011	0.551	0.004%
0.5	-0.009	0.011	0.388	0.008%

Note.  $pT$ ,  $p$ -value threshold used to select genetic variation included in risk score;  $\beta$ , standardised beta value.

**Supplementary Table 7.6. Major depression polygenic risk score predicting psychotic experience domains at 8  $p$ -value thresholds.**

<b>Paranoia and Hallucinations</b>				
$pT$	$\beta$	SE	$p$	$r^2$
0.001	0.002	0.011	0.861	0.000%
0.01	-0.002	0.011	0.835	0.001%
0.05	0.002	0.011	0.877	0.000%
0.1	0.006	0.011	0.589	0.004%
0.2	-0.002	0.011	0.892	0.000%
0.3	0.004	0.011	0.687	0.002%
0.4	0.000	0.011	0.980	0.000%
0.5	0.001	0.011	0.926	0.000%
<b>Anhedonia</b>				
$pT$	$\beta$	SE	$p$	$r^2$
0.001	0.003	0.013	0.794	0.001%
0.01	0.019	0.013	0.135	0.037%
0.05	0.022	0.013	0.083	0.048%
0.1	0.023	0.013	0.069	0.053%
0.2	0.024	0.013	0.062	0.057%
0.3	0.029	0.013	0.025	0.082%
0.4	0.032	0.013	0.014	0.100%
0.5	0.033	0.013	0.010	0.109%
<b>Cognitive Disorganisation</b>				
$pT$	$\beta$	SE	$p$	$r^2$
0.001	-0.012	0.014	0.403	0.014%
0.01	0.010	0.014	0.479	0.009%
0.05	0.018	0.014	0.189	0.033%
0.1	0.015	0.014	0.285	0.021%
0.2	0.007	0.014	0.610	0.005%
0.3	0.014	0.014	0.316	0.019%
0.4	0.014	0.014	0.296	0.021%
0.5	0.015	0.014	0.276	0.022%
<b>Parent-rated Negative Symptoms</b>				
$pT$	$\beta$	SE	$p$	$r^2$
0.001	0.028	0.011	0.008	0.078%
0.01	0.020	0.011	0.061	0.040%
0.05	0.021	0.011	0.051	0.044%
0.1	0.013	0.011	0.226	0.018%
0.2	0.012	0.011	0.271	0.014%
0.3	0.014	0.011	0.187	0.021%
0.4	0.012	0.011	0.283	0.013%
0.5	0.012	0.011	0.273	0.014%

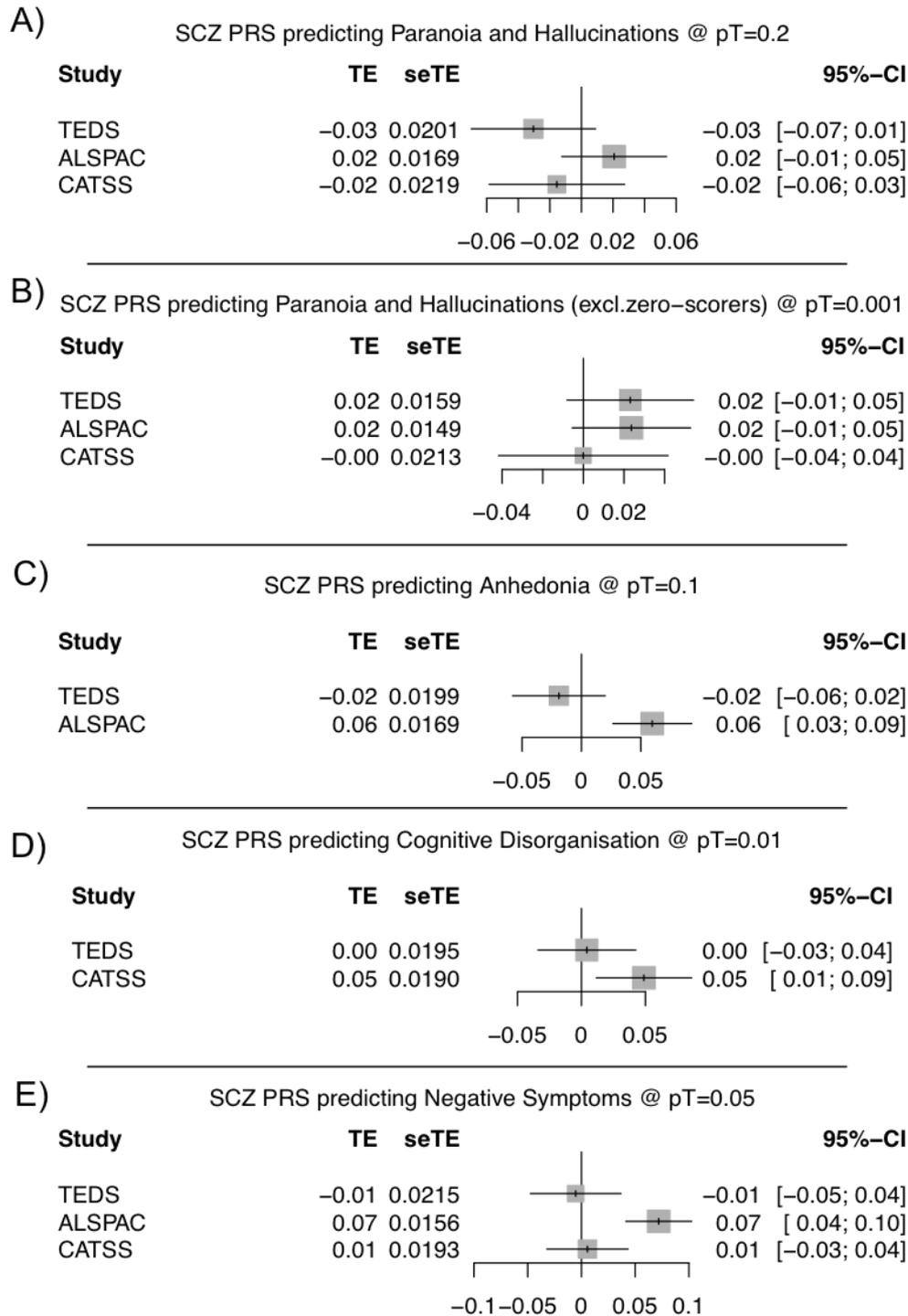
Note.  $pT$ ,  $p$ -value threshold used to select genetic variation included in risk score;  $\beta$ , standardised beta value.

**Supplementary Table 7.7. Comparison of schizophrenia, bipolar disorder, and major depression polygenic risk scores in low and high psychotic experience domain groups.**

<b><i>Schizophrenia</i></b>			
<b>PE</b>	<b>OR</b>	<b>CI 95%</b>	<b><i>p</i></b>
Paranoia and Hallucinations	0.997	0.049	0.894
P and H excl. zero-scorers	1.077	0.077	0.059
Anhedonia	1.073	0.067	0.039
Cognitive Disorganization	1.110	0.070	3.82x10 <sup>-3</sup>
Parent-rated Negative Symptoms	1.084	0.059	7.42x10 <sup>-3</sup>
<b><i>Bipolar Disorder</i></b>			
<b>PE</b>	<b>OR</b>	<b>CI 95%</b>	<b><i>p</i></b>
Paranoia and Hallucinations	0.903	0.050	5.90x10 <sup>-5</sup>
Anhedonia	0.967	0.068	0.338
Cognitive Disorganization	0.999	0.071	0.968
Parent-rated Negative Symptoms	0.974	0.058	0.384
<b><i>Major Depression</i></b>			
<b>PE</b>	<b>OR</b>	<b>CI 95%</b>	<b><i>p</i></b>
Paranoia and Hallucinations	1.032	0.050	0.218
Anhedonia	1.102	0.067	4.59x10 <sup>-3</sup>
Cognitive Disorganization	1.061	0.071	0.099
Parent-rated Negative Symptoms	1.062	0.058	0.042

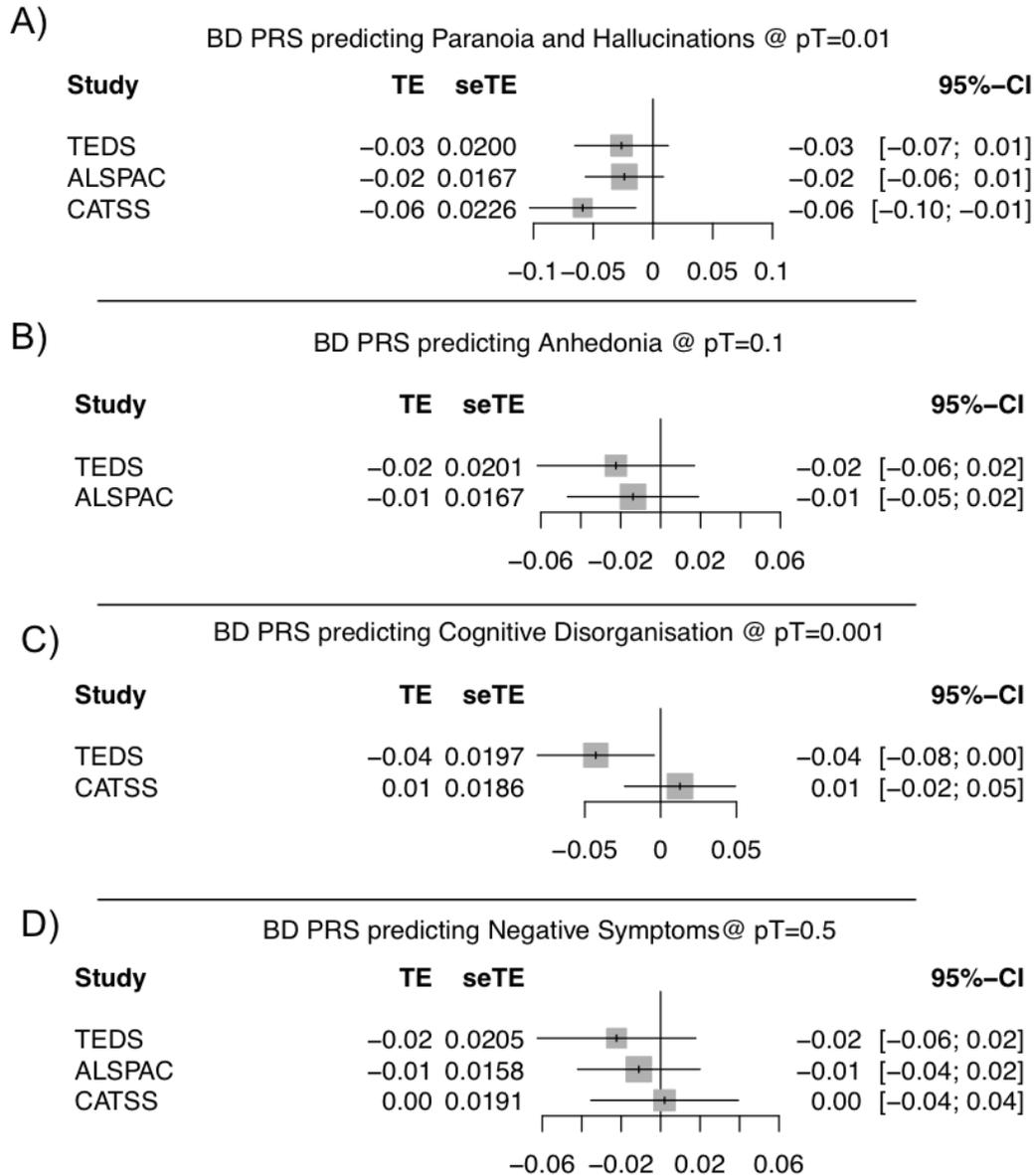
*Note.* PE, psychotic experience; P and H excl. zero-scorers, Paranoia and Hallucinations excluding zero scorers; OR, odds ratio; CI 95%, 95% confidence interval of odds ratio. Low and high groups were defined as the bottom and top 25% of raw psychotic experience domain sum scores. This table shows results when using polygenic risk scores at the most predictive p-value threshold for each trait. Linear regression results for the same p-value thresholds are shown in Table 7.2 of the main text.

**Supplementary Figures:**



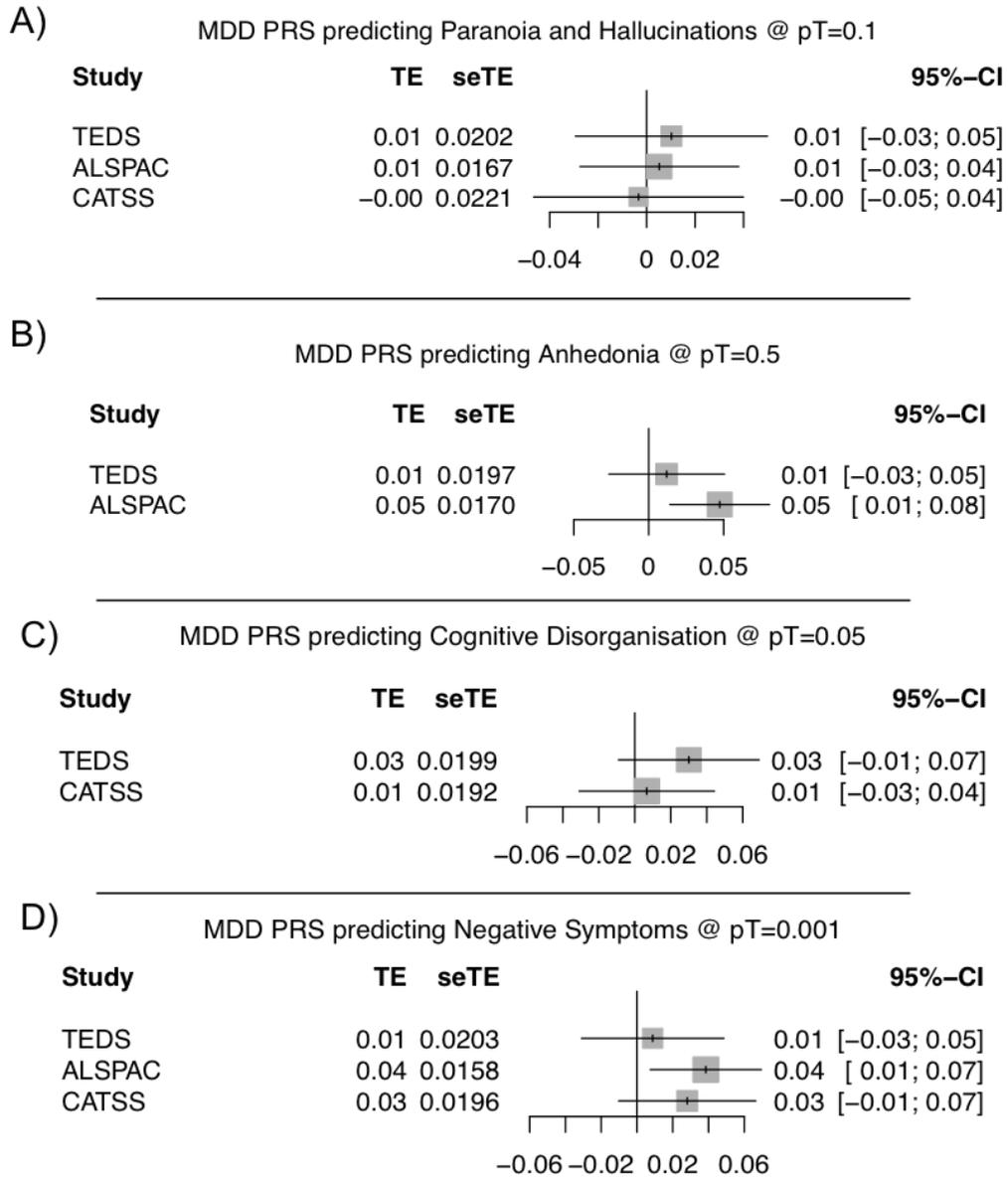
**Supplementary Figure 7.1. Schizophrenia (SCZ) PRS predicting specific adolescent PEs within TEDS, ALSPAC and CATSS samples.**

*Note.* This figure shows results for polygenic risk scores at the most predictive across sample  $p$ -value threshold for each trait. TE, effect size; seTE, Effect size standard error; 95%-CI, 95% confidence intervals. Error bars represent 95% confidence intervals.



**Supplementary Figure 7.2. Bipolar disorder (BD) PRS predicting specific adolescent PEs within TEDS, ALSPAC and CATSS samples.**

*Note.* This figure shows results for polygenic risk scores at the most predictive across sample  $p$ -value threshold for each trait. TE, effect size; seTE, Effect size standard error; 95%-CI, 95% confidence intervals. Error bars represent 95% confidence intervals.



**Supplementary Figure 7.3. Major depression (MDD) PRS predicting specific adolescent PEs within TEDS, ALSPAC and CATSS samples.**

*Note.* This figure shows results for polygenic risk scores at the most predictive across sample  $p$ -value threshold for each trait. TE, effect size; seTE, Effect size standard error; 95%-CI, 95% confidence intervals. Error bars represent 95% confidence intervals.

## Chapter 8 - Discussion

This chapter will first provide a short reintroduction of the pre-existing literature surrounding adolescent psychotic experiences (PEs) that motivated the overarching aims of this thesis. Second, a brief overview of the six empirical chapters will be provided. Finally, key points relating to this thesis will be discussed in a wider context, including the genetic architecture of adolescent PEs, the relationship between adolescent PEs and schizophrenia, bipolar disorder, and major depression, and the pooling of information across multiple samples.

### *8.1 – Motivation for this thesis*

The following section will summarise information discussed in Sections 1.2.1, 1.2.2, 1.5.2, 1.5.3, and 1.6.1.

Adolescent PEs have been phenotypically associated with psychotic and non-psychotic psychiatric disorders in adulthood, as well as several aspects of adolescent mental health. Adolescent PEs have been shown to exist as quantitative and specific traits, typically separating into positive, cognitive and negative symptoms domains. Twin studies estimates that a third to half of the variance in adolescent PEs are accounted for by genetic factors. One previous study has provided evidence that for some PEs, approximately half of this heritability is explained by common genetic variation. There has been one genome-wide association study (GWAS) of adolescent PEs, which returned no genome-wide significant variation ( $p < 5 \times 10^{-8}$ ). Family-based studies have reported that adolescent PEs may share genetic factors with schizophrenia and affective disorders. The common genetic relationship between adolescent PEs and psychiatric disorders has been explored using two samples for schizophrenia, and one sample for bipolar disorder. For schizophrenia, studies suggested no common genetic relationship between positive or cognitive PEs and schizophrenia, but there were mixed findings

relating to negative symptoms. For bipolar disorder, there was no common genetic relationship with adolescent PEs.

This previous research relating to common genetic variation has provided useful insight into the common genetic architecture of adolescent PEs and their relationship with some psychiatric disorders. However, these previous studies had several limitations. One limitation is the use of binary and non-specific PE measures. This approach reduces statistical power due to phenotypic heterogeneity within groups, and doesn't allow the investigation of non-linear effects. Another limitation of these previous studies is the use of one moderately small sample ( $N = 3,483$ ) thus limiting statistical power and the findings are less generalizable.

Given the importance of characterising the common genetic basis of adolescent PEs and their relationship with psychiatric disorders, and the limitations of previous studies, this thesis had the following specific aims: 1) Harmonise genetic data from samples with adolescent PE data to enable combined analysis, 2) Harmonise measures of specific PEs between samples, 3) Estimate SNP-heritability of specific PEs. 4) Perform GWAS of specific PEs to identify associated genetic variation and biological pathways. 5) Estimate the genetic correlation between specific PEs and typically adult-onset psychiatric disorders including schizophrenia, bipolar disorder and major depressive disorder.

## 8.2 – Summary of methods and results

### 8.2.1 – GWAS of specific adolescent PEs in TEDS

Chapter 2 performed the second ever GWAS of adolescent PEs within the TEDS (Twins Early Development Study) sample alone. This study used quantitative and specific measures of adolescent PEs to improve statistical power. Another step taken to improve statistical power was the inclusion of ungenotyped siblings of monozygotic individuals using a generalised estimating equation. Across the six specific PE genome-wide

association analyses three independent loci were associated at genome-wide significance, two for Cognitive Disorganisation, and one for Parent-rated Negative Symptoms. One of the associations for Cognitive Disorganisation was within *CSMD1*, a gene previously associated with schizophrenia. Limitations of this study were sample size and lack of a replication sample. Subsequent chapters worked towards using multiple samples to investigate the genetic basis of adolescent PEs.

### 8.2.2 – Effect of normalising residuals

In preparation for harmonising data between samples to improve statistical power for subsequent analyses, the effect of normalisation before or after correcting for covariates was investigated in Chapter 3. Correcting for covariates before normalisation has the practical advantage of separating tied observations, improving the efficacy of rank-based normalisation. However, this chapter showed that normalisation of residuals can re-introduce a correlation with covariates, resulting in confounding effects. Rank-based normalisation (randomly splitting ties) before correcting for covariates was shown to be an alternative approach with many of the practical advantages of the previous approach, but without causing confounding. This approach was therefore employed when preparing phenotypic data for analysis in subsequent chapters.

### 8.2.3 – Harmonisation of PE measures of samples

In Chapter 4, the items assessing aspects of PEs within each sample were identified with the help of clinicians. Using principal components analysis the correlation structure between PE items was investigated within each sample to identify specific domains. Four comparable domains were identified in each sample confirming the existence of distinct and replicable PE domains: Paranoia and Hallucinations, Anhedonia, Cognitive Disorganisation, and Parent-rated Negative Symptoms. Items within these domains were then used to create scales.

#### 8.2.4 – GWAS of specific adolescent PEs in TEDS, ALSPAC, and CATSS

Chapter 5 aimed to perform a genome-wide association study of these harmonised PE measures across the three samples. This involved genotypic harmonisation of the samples through careful quality control, imputation to a common reference, and estimation of population structure covariates. The four mega-GWASs returned one variant achieving genome-wide significance. This genome-wide significant variant for Anhedonia was within a biologically plausible gene but was not replicated in the replication sample, possibly suggesting that it was a false positive. Gene based association analysis was also used to identify associated genes. One gene achieved significance but did not replicate in the replication sample. The absence of many genome-wide significant associations demonstrated a lack of power possibly due to overall sample size, between sample heterogeneity, or a low amount of PE variance that can be explained by common genetic variation.

#### 8.2.5 – Estimating SNP-heritability of specific adolescent PEs in TEDS, ALSPAC, and CATSS

Chapter 6 aimed to characterise the genetic architecture of adolescent PEs within a homogenous sample, but also to estimate the SNP-heritability of PEs across the three samples. This chapter provided evidence that ~3-9% of the heritability of adolescent PE could be accounted for by common additive genetic variation, with estimates of SNP-heritability for Anhedonia and Cognitive Disorganisation being significantly non-zero. The SNP-heritability was higher for more normally distributed traits (Anhedonia and Cognitive Disorganisation) than skewed traits (Paranoia and Hallucinations and Parent-rated Negative Symptoms). This could reflect differences in underlying genetic architecture, or the appropriateness of current methods for estimating SNP-heritability of skewed traits. Across-sample (mega) estimates of SNP-heritability were lower than within-sample (meta) estimates, suggesting heterogeneity between the samples.

### 8.2.6 – Estimating genetic association between psychiatric disorders and specific adolescent PEs in TEDS, ALSPAC, and CATSS

Chapter 7 aimed to test for evidence of a SNP-based genetic covariance between adolescent PEs and schizophrenia, bipolar disorder and major depression. This chapter demonstrated significant positive genetic covariance between a range of adolescent PE and schizophrenia and major depression, but a negative genetic covariance with bipolar disorder. These results were broadly supported by LD-score regression analysis.

### 8.3 – Wider implications of research

#### 8.3.1 – The genetic architecture of adolescent PEs

This thesis reports that 3%-9% of the variance in adolescent PEs can be explained by common genetic variation. Estimates of SNP-heritability varied across specific PE domains, with higher estimates for Anhedonia and Cognitive Disorganisation than for Paranoia and Hallucinations and Parent-rated Negative Symptoms. These estimates of SNP heritability accounted for 10%-19% of the twin-based heritability. This discrepancy is a result of several factors as described in Section 6.4.

The items within Paranoia and Hallucinations and Parent-rated Negative measures were more rarely endorsed in comparison to the items within Anhedonia and Cognitive Disorganisation measures. This thesis potentially supports a difference in genetic architecture between these two groups of PEs as well, with Anhedonia and Cognitive Disorganisation showing larger amounts of variance explained by common genetic effects than Paranoia and Hallucinations and Parent-rated Negative Symptoms. There are a number of possible reasons for this difference.

The first reason could be a methodological one. Although SNP-heritability estimates were calculated using two different methodologies, both were based on normalised PE scores. The process of normalisation can introduce artificial variance (i.e. noise) and

thereby reduce the proportion of variance that can be explained by common genetic variation. Normalisation will have a larger effect on more highly skewed variables, which could therefore explain why the more skewed traits show a lower SNP-heritability. Although GREML SNP-heritability estimates were consistent when using untransformed PE scores, it has been reported that GREML underestimates the SNP-heritability of skewed traits (Nivard et al., 2016). Further simulation studies are required to investigate the effect of skew on estimates of SNP-heritability.

A second reason that SNP-heritability estimates vary between adolescent PEs, in spite of similar twin heritability estimates, is that different parts of the minor allele frequency spectrum are contributing to their variance. Investigation of rare genetic variation associated with adolescent PEs was not within the scope of this thesis. However, a future study investigating the contribution of genetic variation with a frequency below 0.01 would be of interest.

A third reason for a difference in SNP-heritability between adolescent PEs could be due to contributions of non-SNP variation (i.e. CNVs and InDels) varying across PEs. Although some CNVs and InDels were available in the genetic data used when estimating SNP-heritability, the coverage of non-SNP variation was low. Therefore, it is possible that an increased contribution of non-SNP variation for a given PE could lead to smaller SNP-heritability estimates.

### *8.3.2 – The genetic relationship between adolescent PEs and schizophrenia, bipolar disorder and major depression*

This thesis reports a significant genetic covariance between a range of specific adolescent PEs and schizophrenia, bipolar disorder, and major depression. This supports the notion that adolescent PEs share biological pathways with adult psychiatric disorders, and are not merely epiphenomena. As suggested by previous epidemiological studies, the genetic association between adolescent PEs, schizophrenia,

and major depression was positive. Prior to this thesis, only the positive common genetic association between adolescent Negative Symptoms and schizophrenia had been demonstrated. The common genetic association between schizophrenia and adolescent Paranoia and Hallucinations when excluding zero scorers had not been identified in previous studies that used two of the same samples, highlighting the importance of investigating non-linear effects. Contrary to previous research (McGrath et al., 2016), the genetic association between adolescent PEs and bipolar disorder was negative, indicating that genetic variation increasing the presence of adolescent PEs actually decrease the risk of developing bipolar disorder. Although the genetic association between adolescent PEs and bipolar disorder is not positive, the strength of the association implies again that PEs are meaningful to clinical outcomes.

Although non-zero estimates of genetic covariance are informative, the magnitude of the genetic overlap is important as it indicates the amount of variance in one trait that can be explained by the genetic variants associated with another. For example, this thesis reports that the SNP-based genetic correlation between schizophrenia and Cognitive Disorganisation is  $\sim 0.15$ , this means that if all of the SNP-heritability in Cognitive Disorganisation could be explained, then  $\sim 2.2\%$  of the SNP-heritability of schizophrenia could also be explained. When considering the value of studying Anhedonia to gain insight into major depression, the genetic correlation of  $\sim 0.45$  suggests that if the SNP-heritability of Anhedonia could be fully explained, then 20% of the SNP-heritability of major depression could also be explained. This thesis was unable to calculate the magnitude of genetic correlation for Paranoia and Hallucinations and Parent-rated Negative Symptoms due to the apparent zero heritability (across samples). These findings indicate that the common genetic basis of specific adolescent PEs are to some degree informative of the common genetic basis of psychiatric disorders.

Given the evidence of shared aetiology between certain adolescent PEs and psychiatric disorders, in theory adolescent PEs could be used as a predictor for the onset of these psychiatric disorders. However, the amount of the total phenotypic variance in psychiatric disorders that could be explained by only the common genetic variation underlying adolescent PEs is likely to be low in most cases as the SNP-heritability of these psychiatric disorders is 19-26%. For example, if all of the SNP-heritability in Cognitive Disorganisation could be explained, then ~2.2% of the SNP-heritability of schizophrenia could also be explained. Given that common genetic variation only accounts for 26% of the variance in schizophrenia, the common genetic variation underlying adolescent Anhedonia would explain 0.6% of the total phenotypic variance in schizophrenia. For another example, if all of the SNP-heritability of Anhedonia could be explained, 20% of the SNP-heritability of major depression could also be explained. Given that common genetic variation only accounts for 19% of the variance in major depression, the common genetic variation underlying adolescent Anhedonia would explain 3.8% of the total phenotypic variance in major depression. Given the sample sizes required to fully account for the SNP-heritability of adolescent PEs, using the common genetic basis of adolescent PEs alone to predict later psychiatric disorders will likely be inefficient. However, given the complex aetiology of common psychiatric disorders it is unlikely to find many predictors (apart from family history) with large effect size, and therefore, even if the variance explained by the common genetic basis of adolescent PEs is small, it could make a useful contribution to the prediction of psychiatric disorders in combination with other predictors. Although the non-zero genetic covariance estimates presented here indicate that genetic variation associated with adolescent PEs may be useful for the prediction of certain adult psychiatric disorders, we view strong statements about causality as impossible (Pickrell et al., 2016). For example, a correlation between two variables might be mediated by a third variable that is correlated with both outcomes. The same concept applies to the

interpretation of genetic correlation estimates. Therefore, the non-zero genetic covariance estimates between PEs and psychiatric disorders indicate an overlap in associated heritable factors.

### *8.3.3 – The pooling of information across multiple samples*

Pooling information across samples has two key advantages. Firstly, increasing sample size improves statistical power to identify associations at significance. This is very important when studying the common genetic basis of complex traits, such as adolescent PEs, as the effect size of individual genetic variants is very small. Secondly, by including multiple samples in an analysis, it reduces the likelihood of identifying sample specific effects. Therefore the results from combined sample analyses are likely to be more generalisable to other populations.

This thesis has indicated that increasing sample size does not always increase statistical power to detect genetic associations. When studying adolescent PEs in the TEDS sample alone, the genome-wide association study (GWAS) identified several genome-wide associations. Whereas in the combined sample GWAS, only one genetic association was identified at genome-wide significance. In theory, as sample size increases, our ability to identify small effect sizes at significance also increases. One reason why fewer genome-wide significant associations were identified when using larger samples could be that the likelihood of detecting false positives increases in smaller samples. Another reason may be that combining multiple samples leads to an increase in heterogeneity (thereby decreasing effect sizes) that outweighs the increase in statistical power. The comparison of within and across sample SNP-heritability estimates in Chapter 6 demonstrates the presence of heterogeneity between samples, reducing the variance in PEs that can be explained by common genetic variation. The inverse relationship between magnitude of effect size and sample size has been seen in other areas of genetics as well. For example, a recent RNA sequencing study of schizophrenia that used a sample 10-fold larger than

any previous RNA-sequencing study reported that differential gene expression was far more subtle (i.e. smaller effect sizes) than previous smaller studies had reported (Fromer et al., 2016). Unlike previous smaller RNA-sequencing studies of schizophrenia, this large RNA-sequencing identified no gene as significantly differentially expressed (Fromer et al., 2016). It is likely that there is a trade off between the increase in sample size and the increase in heterogeneity when adding samples to an analysis.

In terms of improving the generalizability of results, pooling information from multiple samples in this thesis has been successful. The harmonisation of the specific PE measures within each sample has enabled direct comparison of the results from each sample, and the combined analysis across samples has provided more accurate estimates of effects. Prior to this thesis, the SNP-heritability of specific adolescent PEs had only been estimated in one sample. This thesis estimated the SNP-heritability of specific PEs in three samples, highlighting differences in SNP-heritability estimates within each sample, but also providing more accurate estimates of SNP-heritability by averaging estimates across samples (meta-SNP-heritability). Furthermore, prior to this thesis, the relationship between adolescent PEs and schizophrenia had been explored using two samples. These samples reported contrasting results that could not be directly compared due to differences in the measures used, meaning that an overall effect could not be estimated. This thesis enabled the direct comparison of within sample effects and provided overall effect size estimates. These are just two examples demonstrating the utility of analysing multiple phenotypically harmonised samples.

#### 8.4 – Considerations for future research

##### 8.4.1 – Measurement

A key strength of this thesis was the derivation of psychometrically-sound quantitative individual PE domains: these were derived using principal component analysis and had content validity. Greater power was achieved compared to past research through

combining independent samples. However, as mentioned in the previous section, differences between the measures across samples impact the amount of phenotypic variance across the three samples that can be explained by common genetic variation. One approach to overcome this issue would be to alter the measures within each sample to optimise (increase) the across sample (mega-) SNP-heritability.

#### 8.4.2 – Models for non-normal data

For the more skewed PE domains, with larger numbers of tied individuals, the process of randomly ranking tied individuals during normalisation will have introduced noise and thus downward biased SNP-heritability estimates and other parameter estimates. Given that normalisation is essential for combined analysis of multiple samples (mega-analysis) and the standard linear regression, a more powerful approach may be to use a model that does not assume normality (or is heteroskedasticity robust) to estimate effects within each sample, and then perform meta-analysis of the results. Although using more complex models present their own practical limitations, such as often not being available in genome-wide analysis software, application of models that do not require the normality assumption should be explored further.

#### 8.4.3 – Genetic relationship between adolescent PEs and other traits/disorders

This thesis has only investigated the common genetic relationship between adolescent PEs and three psychiatric disorders. It would be useful to estimate the genetic relationship between adolescent PEs and other phenotypes, to understand the position of adolescent PEs in the aetiological landscape of health and disease. Schizophrenia, bipolar disorder and major depression were chosen due to past epidemiological links, their ostensible connection in terms of similarity of phenotype (e.g. paranoia, anhedonia), and the availability of well-powered genome-wide association summary statistics. There are other phenotypes that have past epidemiological links with adolescent PEs that should also be investigated such as autistic traits (R. B. Jones,

Thapar, Lewis, & Zammit, 2012; Taylor, Robinson, et al., 2015) and sleep disturbances (Taylor, Gregory, Freeman, & Ronald, 2015; Thompson et al., 2015). Phenotypes that are correlated with adolescent PEs on a phenotypic level are expected to also correlate at a common genetic level.

The genetic correlation between two phenotypes can also be used to infer a phenotypic association. Given that the genetic correlation between two phenotypes can be estimated even when the phenotype has been measured in separate samples, this provides an opportunity to identify novel phenotypic associations between adolescent PEs and outcomes assessed in separate samples.

#### 8.4.4 – Developmental stages of PEs

This thesis has focused on PEs in adolescence for reasons described in Section 1.1 and 1.2.2. However, investigation of the relationship between PEs in adulthood and psychiatric disorders would be of interest as it could shine light on the factors that change specific PEs in healthy individuals into symptoms of a pathology. Furthermore, the relationship between adolescent PEs and adult PEs would also be of interest.

#### 8.4.5 – Rare genetic variation

As previously mentioned in Section 8.3.1, this thesis has focused on common genetic variation only. Given that common genetic variation only accounts for a part of the variance in adolescent PEs, other sources of variance should also be explored, including rarer genetic variation, environmental factors, and the interplay between genetic and environmental factors.

Ideally, investigation of rare genetic variation underlying adolescent PEs would be achieved using DNA sequence data. However, a cheaper and more practical way of investigating the contribution of rarer genetic variation would be to impute genotype array data using a larger genomic reference panel, such as the Haplotype Reference

Consortium (McCarthy et al., 2016). Genomic reference panels with large samples can more accurately impute rarer genetic variation (J. Huang et al., 2015).

There have already been several studies investigating the environmental factors underlying adolescent PEs, for example cannabis use and stressful life events (Shakoor et al., 2015, 2016). To improve power to detect environmental effects and uncover interplay between genetic and environmental factors, future studies investigating environmental factors could incorporate genetic data in their analyses by stratifying by genotype or incorporating interaction effects in the model.

### *8.5 – Conclusion and future directions*

In conclusion, this thesis has provided a robust characterisation of the common genetic basis of specific adolescent PEs through the harmonisation and combined analysis of data from three European samples. It has provided evidence that common genetic variation accounts for 3-9% of the variance in adolescent PEs, with some suggestion that the rarer PEs may also have a rarer genetic architecture. This thesis has performed the largest GWAS of adolescent PEs to date, identifying one genome-wide significant variant for Anhedonia. Additionally, predicted gene expression association analysis identified one gene significantly associated with Cognitive Disorganisation. This thesis has also provided robust evidence of a genetic overlap between adolescent PEs and schizophrenia and major depression.

This thesis has highlighted many research avenues for future studies. First, investigation into the different contributions of common genetic variation across specific adolescent PEs requires further investigation. Does the genetic architecture of specific adolescent PEs vary in terms of the frequency or type of underlying genetic variation? Or is the difference in SNP-heritability between specific PEs a reflection of methodological issues relating to the skew of traits. Second, although effects were often consistent between the different samples used in this thesis, investigation of the heterogeneity between these

three samples could help understand instances where effects differed. This process could uncover factors underlying observed effects, and could also aid in the planning and interpretation of future studies. Third, investigation of the genetic association between adolescent PEs and other phenotypes should occur. This thesis has supported the notion that adolescent PEs are clinically meaningful, and it would therefore be interesting to estimate their genetic (and phenotypic) relationship with other outcomes.

## Bibliography

- Abe, K., Chisaka, O., Van Roy, F., & Takeichi, M. (2004). Stability of dendritic spines and synaptic contacts is controlled by  $\alpha$ N-catenin. *Nature Neuroscience*, 7(4), 357–363.
- Abecasis, G. R., Cherny, S. S., Cookson, W. O., & Cardon, L. R. (2002). Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics*, 30(1), 97–101.
- American Psychiatric Association. (2013a). *Diagnostic and Statistical Manual of Mental Disorders*. (5th ed.). Washington, DC: American Psychiatric Association.
- American Psychiatric Association. (2013b). Schizophrenia spectrum and other psychotic disorders. In *Diagnostic and Statistical Manual of Mental Disorders*. (5th ed., pp. 89–122). Washington, DC: American Psychiatric Association.
- Anckarsäter, H., Lundström, S., Kollberg, L., Kerekes, N., Palm, C., Carlström, E., ... Bölte, S. (2011). The child and adolescent twin study in Sweden (CATSS). *Twin Research and Human Genetics*, 14(6), 495–508.
- Andreassen, O. A., Harbo, H. F., Wang, Y., Thompson, W. K., Schork, A. J., Mattingsdal, M., ... Kelsoe, J. R. (2015). Genetic pleiotropy between multiple sclerosis and schizophrenia but not bipolar disorder: differential involvement of immune-related gene loci. *Molecular Psychiatry*, 20(2), 207–214.
- Angold, A., Costello, E. J., Messer, S. C., & Pickles, A. (1995). Development of a short questionnaire for use in epidemiological studies of depression in children and adolescents. *International Journal of Methods in Psychiatric Research*.
- Ardlie, K. G., Kruglyak, L., & Seielstad, M. (2002). Patterns of linkage disequilibrium in the human genome. *Nature Reviews Genetics*, 3(4), 299–309.
- Ashley-Koch, A. E., Garrett, M. E., Gibson, J., Liu, Y., Dennis, M. F., Kimbrel, N. A., ... Hauser, M. A. (2015). Genome-wide association study of posttraumatic stress disorder in a cohort of Iraq–Afghanistan era veterans. *Journal of Affective Disorders*, 184, 225–234.
- Aulchenko, Y. S., Ripke, S., Isaacs, A., & Van Duijn, C. M. (2007). GenABEL: an R library for genome-wide association analysis. *Bioinformatics*, 23(10), 1294–1296.
- Australian Institute of Health and Welfare. (2011). *Comorbidity of mental disorders and physical conditions, 2007*. [https://doi.org/Cat. no. PHE 155](https://doi.org/Cat.no.PHE155)
- Bacanu, S.-A., Devlin, B., & Roeder, K. (2000). The power of genomic control. *The American Journal of Human Genetics*, 66(6), 1933–1944.
- Balan, S., Iwayama, Y., Toyota, T., Toyoshima, M., Maekawa, M., & Yoshikawa, T. (2014). 22q11. 2 deletion carriers and schizophrenia-associated novel variants. *The British Journal of Psychiatry*, bjp-bp.
- Baranzini, S. E., Wang, J., Gibson, R. A., Galwey, N., Naegelin, Y., Barkhof, F., ... Johnson, M. R. (2009). Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis. *Human Molecular Genetics*, 18(4), 767–778.
- Beasley, T. M., Erickson, S., & Allison, D. B. (2009). Rank-based inverse normal transformations are increasingly used, but are they merited? *Behavior Genetics*, 39(5), 580.
- Berry, W. D. (1993). *Understanding regression assumptions: Series quantitative applications in the social sciences*. Newbury Park, CA: Sage.
- Bifsha, P., Yang, J., Fisher, R. A., & Drouin, J. (2014). Rgs6 is required for adult

- maintenance of dopaminergic neurons in the ventral substantia nigra. *PLoS Genet*, *10*(12), e1004863.
- Binbay, T., Drukker, M., Elbi, H., Tanık, F. A., Özkınay, F., Onay, H., ... Alptekin, K. (2012). Testing the psychosis continuum: differential impact of genetic and nongenetic risk factors and comorbid psychopathology across the entire spectrum of psychosis. *Schizophrenia Bulletin*, *38*(5), 992–1002.
- Bossini-Castillo, L., de Kovel, C., Kallberg, H., van't Slot, R., Italiaander, A., Coenen, M., ... Huizinga, T. (2014). A genome-wide association study of rheumatoid arthritis without antibodies against citrullinated peptides. *Annals of the Rheumatic Diseases*, annrheumdis-2013.
- Boutwell, B., Hinds, D., Tielbeek, J., Ong, K., Day, F., Perry, J., & Team, 23andMe Research. (2017). Replication and characterization of CADM2 and MSRA genes on human behavior. *bioRxiv*, 110395.
- Boyd, A., Golding, J., Macleod, J., Lawlor, D. A., Fraser, A., Henderson, J., ... Smith, G. D. (2012). Cohort profile: the “children of the 90s”—the index offspring of the Avon Longitudinal Study of Parents and Children. *International Journal of Epidemiology*, dys064.
- Bryleva, E. Y., & Brundin, L. (2017). Kynurenine pathway metabolites and suicidality. *Neuropharmacology*, *112*, 324–330.
- Bucan, M., Abrahams, B. S., Wang, K., Glessner, J. T., Herman, E. I., Sonnenblick, L. I., ... Bradfield, J. P. (2009). Genome-wide analyses of exonic copy number variants in a family-based study point to novel autism susceptibility genes. *PLoS Genet*, *5*(6), e1000536.
- Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., ... Consortium, S. W. G. of the P. G. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, *47*(3), 291–295.
- Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3).
- Byrne, E. M., Johnson, J., McRae, A. F., Nyholt, D. R., Medland, S. E., Gehrman, P. R., ... Chenevix-Trench, G. (2012). A genome-wide association study of caffeine-related sleep disturbance: confirmation of a role for a common variant in the adenosine receptor. *Sleep*, *35*(7), 967–975.
- Cammaerts, S., Strazisar, M., Smets, B., Weckhuysen, S., Nordin, A., De Jonghe, P., ... Del Favero, J. (2015). Schizophrenia-associated MIR204 regulates noncoding RNAs and affects neurotransmitter and ion channel gene sets. *PloS One*, *10*(12), e0144428.
- Cappello, S., Gray, M. J., Badouel, C., Lange, S., Einsiedler, M., Srour, M., ... Morgan, T. (2013). Mutations in genes encoding the cadherin receptor-ligand pair DCHS1 and FAT4 disrupt cerebral cortical development. *Nature Genetics*, *45*(11), 1300–1308.
- Cardno, A. G., & Gottesman, I. I. (2000). Twin studies of schizophrenia: from bow-and-arrow concordances to star wars Mx and functional genomics. *American Journal of Medical Genetics Part A*, *97*(1), 12–17.
- Casey, J. P., Magalhaes, T., Conroy, J. M., Regan, R., Shah, N., Anney, R., ... Bacchelli, E. (2012). A novel approach of homozygous haplotype sharing identifies candidate genes in autism spectrum disorder. *Human Genetics*, *131*(4), 565–579.
- Cederlöf, M., Kuja-Halkola, R., Larsson, H., Sjölander, A., Östberg, P., Lundström, S., ... Lichtenstein, P. (2016). A longitudinal study of adolescent psychotic experiences and later development of substance use disorder and suicidal behavior.

*Schizophrenia Research.*

- Chang, X., Yamada, R., & Yamamoto, K. (2005). Inhibition of antithrombin by hyaluronic acid may be involved in the pathogenesis of rheumatoid arthritis. *Arthritis Res Ther*, 7(2), R268.
- Cheesman, R., Selzam, S., Ronald, A., Dale, P. S., McAdams, T. A., Eley, T. C., & Plomin, R. (2017). Childhood behaviour problems show the greatest gap between DNA-based and twin heritability. *Translational Psychiatry*, 7(12), 1284.
- Chen, W. J., Hsiao, C. K., & Lin, C. C. H. (1997). Schizotypy in community samples: The three-factor structure and correlation with sustained attention. *Journal of Abnormal Psychology*, 106(4), 649.
- Cochrane, M., Petch, I., & Pickering, A. D. (2010). Do measures of schizotypal personality provide non-clinical analogues of schizophrenic symptomatology? *Psychiatry Research*, 176(2), 150–154.
- Cole, S. R., Platt, R. W., Schisterman, E. F., Chu, H., Westreich, D., Richardson, D., & Poole, C. (2009). Illustrating bias due to conditioning on a collider. *International Journal of Epidemiology*, 39(2), 417–420.
- Collier, D. A., Stöber, G., Li, T., Heils, A., Catalano, M., Di Bella, D., ... Bengel, D. (1996). A novel functional polymorphism within the promoter of the serotonin transporter gene: possible role in susceptibility to affective disorders. *Molecular Psychiatry*, 1(6), 453–460.
- Cross-Disorder Group of the Psychiatric Genomics Consortium. (2013). Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nature Genetics*, 45(9), 984–994.
- Dahl, R. E. (2004). Adolescent brain development: a period of vulnerabilities and opportunities. Keynote address. *Annals of the New York Academy of Sciences*, 1021(1), 1–22.
- Dantzer, R., O'Connor, J. C., Lawson, M. A., & Kelley, K. W. (2011). Inflammation-associated depression: from serotonin to kynurenine. *Psychoneuroendocrinology*, 36(3), 426–436.
- de Bock, L., Somers, K., Fraussen, J., Hendriks, J. J. A., van Horssen, J., Rouwette, M., ... Espino, M. (2014). Sperm-associated antigen 16 is a novel target of the humoral autoimmune response in multiple sclerosis. *The Journal of Immunology*, 193(5), 2147–2156.
- de Leeuw, C. A., Mooij, J. M., Heskes, T., & Posthuma, D. (2015). MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput Biol*, 11(4), e1004219.
- Ding, C., & Jin, S. (2009). High-throughput methods for SNP genotyping. *Single Nucleotide Polymorphisms: Methods and Protocols*, 245–254.
- Docherty, S. J., Davis, O. S. P., Kovas, Y., Meaburn, E. L., Dale, P. S., Petrill, S. A., ... Plomin, R. (2010). A genome-wide association study identifies multiple loci associated with mathematics ability and disability. *Genes, Brain and Behavior*, 9(2), 234–247.
- Dudbridge, F. (2016). Polygenic epidemiology. *Genetic Epidemiology*, 40(4), 268–272.
- Dudbridge, F., & Gusnanto, A. (2008). Estimation of significance thresholds for genomewide association scans. *Genetic Epidemiology*, 32(3), 227–234.
- Ehlers, C. L., Gizer, I. R., Bizon, C., Slutske, W., Peng, Q., Schork, N. J., & Wilhelmsen, K. C. (2016). Single nucleotide polymorphisms in the REG-CTNNA2 region of chromosome 2 and NEIL3 associated with impulsivity in a Native American sample. *Genes, Brain and Behavior*, 15(6), 568–577.

- Elia, J., Gai, X., Xie, H. M., Perin, J. C., Geiger, E., Glessner, J. T., ... Lantieri, F. (2010). Rare structural variants found in attention-deficit hyperactivity disorder are preferentially associated with neurodevelopmental genes. *Molecular Psychiatry*, *15*(6), 637–646.
- Ericson, M., Tuvblad, C., Raine, A., Young-Wolff, K., & Baker, L. A. (2011). Heritability and longitudinal stability of schizotypal traits during adolescence. *Behavior Genetics*, *41*(4), 499–511.
- Euesden, J., Lewis, C. M., & O'Reilly, P. F. (2015). PRSice: polygenic risk score software. *Bioinformatics*, *31*(9), 1466–1468.
- Fatokun, A. A., Hunt, N. H., & Ball, H. J. (2013). Indoleamine 2, 3-dioxygenase 2 (IDO2) and the kynurenine pathway: characteristics and potential roles in health and disease. *Amino Acids*, *45*(6), 1319–1329.
- Feingold, E. (2002). Regression-Based Quantitative-Trait-Locus Mapping in the 21 st Century. *The American Journal of Human Genetics*, *71*(2), 217–222.
- Fonseca-Pedrero, E., Paño-Piñeiro, M., Lemos-Giráldez, S., Villazón-García, Ú., & Muñiz, J. (2009). Validation of the Schizotypal Personality Questionnaire—Brief Form in adolescents. *Schizophrenia Research*, *111*(1), 53–60.
- Fossati, A., Raine, A., Carretta, I., Leonardi, B., & Maffei, C. (2003). The three-factor model of schizotypal personality: invariance across age and gender. *Personality and Individual Differences*, *35*(5), 1007–1019.
- Fromer, M., Roussos, P., Sieberts, S. K., Johnson, J. S., Kavanagh, D. H., Perumal, T. M., ... Shah, H. R. (2016). Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nature Neuroscience*, *19*(11), 1442.
- Gabbay, V., Ely, B. A., Babb, J., & Liebes, L. (2012). The possible role of the kynurenine pathway in anhedonia in adolescents. *Journal of Neural Transmission*, *119*(2), 253–260.
- Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V, Aquino-Michaels, K., Carroll, R. J., ... Cox, N. J. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, *47*(9), 1091–1098.
- Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., ... Shen, Y. (2003). The international HapMap project. *Nature*, *426*(6968), 789–796.
- Goodbourn, P. T., Bosten, J. M., Bargary, G., Hogg, R. E., Lawrance-Owen, A. J., & Mollon, J. D. (2014). Variants in the 1q21 risk region are associated with a visual endophenotype of autism and schizophrenia. *Genes, Brain and Behavior*, *13*(2), 144–151.
- Green, H., McGinnity, A., Meltzer, H., Ford, T., & Goodman, R. (2005). *Mental health of children and young people in Great Britain, 2004*.
- Greenwood, T. A., Akiskal, H. S., Akiskal, K. K., Study, B. G., & Kelsoe, J. R. (2012). Genome-wide association study of temperament in bipolar disorder reveals significant associations with three novel Loci. *Biological Psychiatry*, *72*(4), 303–310.
- Hallett, V., Ronald, A., Rijdsdijk, F., & Eley, T. C. (2009). Phenotypic and genetic differentiation of anxiety-related behaviors in middle childhood. *Depression and Anxiety*, *26*(4), 316–324.
- Hamera, E. K., Schneider, J. K., Potocky, M., & Casebeer, M. A. (1996). Validity of self-administered symptom scales in clients with schizophrenia and schizoaffective disorders. *Schizophrenia Research*, *19*(2), 213–219.
- Hannigan, L. J., Walaker, N., Waszczuk, M. A., McAdams, T. A., & Eley, T. C. (2017).

- Aetiological influences on stability and change in emotional and behavioural problems across development: a systematic review. *Psychopathology Review*, 4(1), 52.
- Hashimoto, R., Yoshida, M., Ozaki, N., Yamanouchi, Y., Iwata, N., Suzuki, T., ... Kunugi, H. (2005). A missense polymorphism (H204R) of a Rho GTPase-activating protein, the chimerin 2 gene, is associated with schizophrenia in men. *Schizophrenia Research*, 73(2), 383–385.
- Haworth, C. M. A., Davis, O. S. P., & Plomin, R. (2013). Twins Early Development Study (TEDS): a genetically sensitive investigation of cognitive and behavioral development from childhood to young adulthood. *Twin Research and Human Genetics*, 16(1), 117–125.
- Heck, A., Pfister, H., Czamara, D., Müller-Myhsok, B., Pütz, B., Lucae, S., ... Ising, M. (2011). Evidence for associations between MDGA2 polymorphisms and harm avoidance—replication and extension of a genome-wide association finding. *Psychiatric Genetics*, 21(5), 257–260.
- Hemphälä, M., & Hodgins, S. (2014). Do psychopathic traits assessed in mid-adolescence predict mental health, psychosocial, and antisocial, including criminal outcomes, over the subsequent 5 years? *Canadian Journal of Psychiatry*, 59(1), 40–49.
- Hildebrand, M. S., Tankard, R., Gazina, E. V., Damiano, J. A., Lawrence, K. M., Dahl, H. M., ... Marini, C. (2015). PRIMA1 mutation: a new cause of nocturnal frontal lobe epilepsy. *Annals of Clinical and Translational Neurology*, 2(8), 821–830.
- Hill, W. G., Goddard, M. E., & Visscher, P. M. (2008). Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet*, 4(2), e1000008.
- Hoggart, C. J., Clark, T. G., De Iorio, M., Whittaker, J. C., & Balding, D. J. (2008). Genome-wide significance for dense SNP and resequencing data. *Genetic Epidemiology*, 32(2), 179–185.
- Howie, B., Marchini, J., & Stephens, M. (2011). Genotype imputation with thousands of genomes. *G3: Genes, Genomes, Genetics*, 1(6), 457–470.
- Howie, B. N., Donnelly, P., & Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*, 5(6), e1000529.
- Huang, J., Howie, B., McCarthy, S., Memari, Y., Walter, K., Min, J. L., ... Zheng, H.-F. (2015). Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nature Communications*, 6.
- Huang, K.-C., Yang, K.-C., Lin, H., Tsao, T. T.-H., & Lee, S.-A. (2014). Transcriptome alterations of mitochondrial and coagulation function in schizophrenia by cortical sequencing analysis. *BMC Genomics*, 15(9), S6.
- Hubbard, A. E., Ahern, J., Fleischer, N. L., Van der Laan, M., Lippman, S. A., Jewell, N., ... Satariano, W. A. (2010). To GEE or not to GEE: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiology*, 21(4), 467–474.
- Hur, Y.-M., Cherny, S. S., & Sham, P. C. (2012). Heritability of hallucinations in adolescent twins. *Psychiatry Research*, 199(2), 98–101.
- Ibrahim-Verbaas, C. A., Bressler, J., Debette, S., Schuur, M., Smith, A. V., Bis, J. C., ... Wolf, C. (2016). GWAS for executive function and processing speed suggests involvement of the CADM2 gene. *Molecular Psychiatry*, 21(2), 189–197.
- Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer Science & Business Media. Retrieved from

[https://books.google.co.uk/books/about/Principal\\_Component\\_Analysis.html?id=\\_olByCrhjwIC&pgis=1](https://books.google.co.uk/books/about/Principal_Component_Analysis.html?id=_olByCrhjwIC&pgis=1)

- Jones, H. J., Stergiakouli, E., Tansey, K. E., Hubbard, L., Heron, J., Cannon, M., ... Jones, P. B. (2016). Phenotypic manifestation of genetic risk for schizophrenia during adolescence in the general population. *JAMA Psychiatry*, *73*(3), 221–228.
- Jones, P. B. (2013). Adult mental health disorders and their age at onset. *The British Journal of Psychiatry. Supplement*, *54*, s5-10.  
<https://doi.org/10.1192/bjp.bp.112.119164>
- Jones, R. B., Mars, B., Collishaw, S., Potter, R., Thapar, A., Craddock, N., ... Zammit, S. (2016). Prevalence and correlates of psychotic experiences amongst children of depressed parents. *Psychiatry Research*, *243*, 81–86.
- Jones, R. B., Thapar, A., Lewis, G., & Zammit, S. (2012). The association between early autistic traits and psychotic experiences in adolescence. *Schizophrenia Research*, *135*(1), 164–169.
- Jones, S. E., Tyrrell, J., Wood, A. R., Beaumont, R. N., Ruth, K. S., Tuke, M. A., ... Hayward, C. (2016). Genome-wide association analyses in 128,266 individuals identifies new morningness and sleep duration loci. *PLoS Genet*, *12*(8), e1006125.
- Kanazawa, T., Ikeda, M., Glatt, S. J., Tsutsumi, A., Kikuyama, H., Kawamura, Y., ... Takeda, M. (2013). Genome-wide association study of atypical psychosis. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *162*(7), 679–686.
- Kelleher, I., Cederlöf, M., & Lichtenstein, P. (2014). Psychotic experiences as a predictor of the natural course of suicidal ideation: a Swedish cohort study. *World Psychiatry*, *13*(2), 184–188.
- Kelleher, I., Corcoran, P., Keeley, H., Wigman, J. T. W., Devlin, N., Ramsay, H., ... Cannon, M. (2013). Psychotic symptoms and population risk for suicide attempt: a prospective cohort study. *JAMA Psychiatry*, *70*(9), 940–8.  
<https://doi.org/10.1001/jamapsychiatry.2013.140>
- Kelleher, I., Harley, M., Murtagh, A., & Cannon, M. (2009). Are screening instruments valid for psychotic-like experiences? A validation study of screening questions for psychotic-like experiences using in-depth clinical interview. *Schizophrenia Bulletin*, *37*(2), 362–369.
- Kelleher, I., Keeley, H., Corcoran, P., Lynch, F., Fitzpatrick, C., Devlin, N., ... Harley, M. (2012). Clinicopathological significance of psychotic experiences in non-psychotic young people: evidence from four population-based studies. *The British Journal of Psychiatry*, *201*(1), 26–32.
- Knevel, R., Klein, K., Somers, K., Ospelt, C., Houwing-Duistermaat, J. J., van Nies, J. A. B., ... Schonkeren, J. (2013). Identification of a genetic variant for joint damage progression in autoantibody-positive rheumatoid arthritis. *Annals of the Rheumatic Diseases*, annrheumdis-2013.
- Koiliari, E., Roussos, P., Pasparakis, E., Lencz, T., Malhotra, A., Siever, L. J., ... Bitsios, P. (2014). The CSMD1 genome-wide associated schizophrenia risk variant rs10503253 affects general cognitive ability and executive function in healthy males. *Schizophrenia Research*, *154*(1), 42–47.
- Kraemer, H. C. (2007). DSM categories and dimensions in clinical and research contexts. *International Journal of Methods in Psychiatric Research*, *16*(S1).
- Krapohl, E., Euesden, J., Zabaneh, D., Pingault, J. B., Rimfeld, K., Von Stumm, S., ... Plomin, R. (2016). Phenome-wide analysis of genome-wide polygenic scores. *Molecular Psychiatry*, *21*(9), 1188–1193.

- Kraus, D. M., Elliott, G. S., Chute, H., Horan, T., Pfenninger, K. H., Sanford, S. D., ... Holers, V. M. (2006). CSMD1 is a novel multiple domain complement-regulatory protein highly expressed in the central nervous system and epithelial tissues. *The Journal of Immunology*, *176*(7), 4419–4430.
- Laplante, D. P., Barr, R. G., Brunet, A., Du Fort, G. G., Meaney, M. L., Saucier, J.-F., ... King, S. (2004). Stress during pregnancy affects general intellectual and language functioning in human toddlers. *Pediatric Research*, *56*(3), 400–410.
- Lasky-Su, J., Neale, B. M., Franke, B., Anney, R. J. L., Zhou, K., Maller, J. B., ... Buitelaar, J. (2008). Genome-wide association scan of quantitative traits for attention deficit hyperactivity disorder identifies novel associations and confirms candidate gene associations. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *147*(8), 1345–1354.
- Laurens, K. R., Hodgins, S., Maughan, B., Murray, R. M., Rutter, M. L., & Taylor, E. A. (2007). Community screening for psychotic-like experiences and other putative antecedents of schizophrenia in children aged 9–12 years. *Schizophrenia Research*, *90*(1), 130–146.
- Lee, S. H., Yang, J., Chen, G.-B., Ripke, S., Stahl, E. A., Hultman, C. M., ... Goddard, M. E. (2013). Estimation of SNP heritability from dense genotype data. *American Journal of Human Genetics*, *93*(6), 1151.
- Lee, S. H., Yang, J., Goddard, M. E., Visscher, P. M., & Wray, N. R. (2012). Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics*, *28*(19), 2540–2542. <https://doi.org/10.1093/bioinformatics/bts474>
- Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., ... Cummings, B. B. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), 285–291.
- Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., ... Denisov, G. (2007). The diploid genome sequence of an individual human. *PLoS Biol*, *5*(10), e254.
- Li, J., Liu, J., Feng, G., Li, T., Zhao, Q., Li, Y., ... He, L. (2011). The MDGA1 gene confers risk to schizophrenia and bipolar disorder. *Schizophrenia Research*, *125*(2), 194–200.
- Li, P., Xiang, T., Li, H., Li, Q., Yang, B., Huang, J., ... Ren, G. (2015). Hyaluronan synthase 2 overexpression is correlated with the tumorigenesis and metastasis of human breast cancer. *International Journal of Clinical and Experimental Pathology*, *8*(10), 12101.
- Lin, J. C., Ho, W.-H., Gurney, A., & Rosenthal, A. (2003). The netrin-G1 ligand NGL-1 promotes the outgrowth of thalamocortical axons. *Nature Neuroscience*, *6*(12), 1270–1276.
- Linscott, R. J., & Van Os, J. (2013). An updated and conservative systematic review and meta-analysis of epidemiological evidence on psychotic experiences in children and adults: on the pathway from proneness to persistence to dimensional expression across mental disorders. *Psychological Medicine*, *43*(6), 1133–1149.
- Liu, Y., Blackwood, D. H., Caesar, S., de Geus, E. J. C., Farmer, A., Ferreira, M. A. R., ... Green, E. K. (2011). Meta-analysis of genome-wide association data of bipolar disorder and major depressive disorder. *Molecular Psychiatry*, *16*(1), 2–4.
- Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., ... Yang, J. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature*, *518*(7538), 197–206.
- Manchia, M., Cullis, J., Turecki, G., Rouleau, G. A., Uher, R., & Alda, M. (2013). The impact

- of phenotypic and genetic heterogeneity on results of genome wide association studies of complex diseases. *PLoS One*, 8(10), e76295.
- Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorf, L. A., Hunter, D. J., ... Chakravarti, A. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265), 747–753.
- Martin, N., Boomsma, D., & Machin, G. (1997). A twin-pronged attack on complex traits. *Nature Genetics*, 17(4), 387–392.
- Mason, O., Claridge, G., & Jackson, M. (1995). New scales for the assessment of schizotypy. *Personality and Individual Differences*, 18(1), 7–13.
- McCarthy, S., Das, S., Kretschmar, W., Delaneau, O., Wood, A. R., Teumer, A., ... Sharp, K. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*.
- McClay, J. L., Adkins, D. E., Åberg, K., Bukszár, J., Khachane, A. N., Keefe, R. S. E., ... Vann, R. E. (2011). Genome-wide pharmacogenomic study of neurocognition as an indicator of antipsychotic treatment response in schizophrenia. *Neuropsychopharmacology*, 36(3), 616–626.
- McGrath, J. J., Saha, S., Al-Hamzawi, A., Alonso, J., Bromet, E. J., Bruffaerts, R., ... Fayyad, J. (2015). Psychotic experiences in the general population: a cross-national analysis based on 31 261 respondents from 18 countries. *JAMA Psychiatry*, 72(7), 697–705.
- McGrath, J. J., Saha, S., Al-Hamzawi, A., Andrade, L., Benjet, C., Bromet, E. J., ... Demyttenaere, K. (2016). The bidirectional associations between psychotic experiences and DSM-IV mental disorders. *American Journal of Psychiatry*.
- McKenzie, M., Henders, A. K., Caracella, A., Wray, N. R., & Powell, J. E. (2014). Overlap of expression quantitative trait loci (eQTL) in human brain and blood. *BMC Medical Genomics*, 7(1), 31.
- Mexal, S., Berger, R., Pearce, L., Barton, A., Logel, J., Adams, C. E., ... Leonard, S. (2008). Regulation of a novel  $\alpha$ N-catenin splice variant in schizophrenic smokers. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 147(6), 759–768.
- Miller, A. H., & Raison, C. L. (2016). The role of inflammation in depression: from evolutionary imperative to modern treatment target. *Nature Reviews Immunology*, 16(1), 22–34.
- Minica, C. C., Boomsma, D. I., Vink, J. M., & Dolan, C. V. (2014). MZ twin pairs or MZ singletons in population family-based GWAS[quest] More power in pairs. *Mol Psychiatry*, 19(11), 1154–1155. Retrieved from <http://dx.doi.org/10.1038/mp.2014.121>
- Minică, C. C., Dolan, C. V., Kampert, M. M. D., Boomsma, D. I., & Vink, J. M. (2015). Sandwich corrected standard errors in family-based genome-wide association studies. *European Journal of Human Genetics*, 23(3), 388–394.
- Mladinov, M., Sedmak, G., Fuller, H. R., Babić Leko, M., Mayer, D., Kirincich, J., ... Šimić, G. (2016). Gene expression profiling of the dorsolateral and medial orbitofrontal cortex in schizophrenia. *Translational Neuroscience*, 7(1), 139–150.
- Muller, N., & J Schwarz, M. (2010). The role of immune system in schizophrenia. *Current Immunology Reviews*, 6(3), 213–220.
- Myint, A.-M., Kim, Y. K., Verkerk, R., Scharpé, S., Steinbusch, H., & Leonard, B. (2007). Kynurenine pathway in major depression: evidence of impaired neuroprotection. *Journal of Affective Disorders*, 98(1), 143–151.
- Myint, A. M. (2012). Kynurenines: from the perspective of major psychiatric disorders.

*FEBS Journal*, 279(8), 1375–1385.

- Need, A. C., Attix, D. K., McEvoy, J. M., Cirulli, E. T., Linney, K. L., Hunt, P., ... Shianna, K. V. (2009). A genome-wide study of common SNPs and CNVs in cognitive performance in the CANTAB. *Human Molecular Genetics*, 18(23), 4650–4661.
- Nguyen, T., Staines, D., Nilius, B., Smith, P., & Marshall-Gradisnik, S. (2016). Novel identification and characterisation of Transient receptor potential melastatin 3 ion channels on Natural Killer cells and B lymphocytes: effects on cell signalling in Chronic fatigue syndrome/Myalgic encephalomyelitis patients. *Biological Research*, 49(1), 27.
- Niederhofer, H., & Reiter, A. (2004). Prenatal maternal stress, prenatal fetal movements and perinatal temperament factors influence behavior and school marks at the age of 6 years. *Fetal Diagnosis and Therapy*, 19(2), 160–162.
- Nivard, M. G., Middeldorp, C. M., Lubke, G., Hottenga, J.-J., Abdellaoui, A., Boomsma, D. I., & Dolan, C. V. (2016). Detection of gene–environment interaction in pedigree data using genome-wide genotypes. *European Journal of Human Genetics*, 24(12), 1803–1809.
- Okuda, H., Kobayashi, A., Xia, B., Watabe, M., Pai, S. K., Hirota, S., ... Fukuda, K. (2012). Hyaluronan Synthase HAS2 Promotes Tumor Progression in Bone by Stimulating the Interaction of Breast Cancer Stem–Like Cells with Macrophages and Stromal Cells. *Cancer Research*, 72(2), 537–547.
- Palla, L., & Dudbridge, F. (2015). A fast method that uses polygenic scores to estimate the variance explained by genome-wide marker panels and the proportion of variants affecting a trait. *The American Journal of Human Genetics*, 97(2), 250–259.
- Park, C., Falls, W., Finger, J. H., Longo-Guess, C. M., & Ackerman, S. L. (2002). Deletion in *Catn2*, encoding  $\alpha$ N-catenin, causes cerebellar and hippocampal lamination defects and impaired startle modulation. *Nature Genetics*, 31(3), 279–284.
- Patel, V., Flisher, A. J., Hetrick, S., & McGorry, P. (2007). Mental health of young people: a global public-health challenge. *Lancet*, 369(9569), 1302–1313. [https://doi.org/10.1016/S0140-6736\(07\)60368-7](https://doi.org/10.1016/S0140-6736(07)60368-7)
- Patton, G. C., Viner, R. M., Linh, L. C., Ameratunga, S., Fatusi, A. O., Ferguson, B. J., & Patel, V. (2010). Mapping a global agenda for adolescent health. *Journal of Adolescent Health*, 47(5), 427–432.
- Pe'er, I., Yelensky, R., Altshuler, D., & Daly, M. J. (2008). Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genetic Epidemiology*, 32(4), 381–385.
- Pedrero, E. F., & Debbané, M. (2017). Schizotypal traits and psychotic-like experiences during adolescence: An update. *Psicothema*, 29(1), 5–17.
- Pellissier, F., Gerber, A., Bauer, C., Ballivet, M., & Ossipow, V. (2007). The adhesion molecule Necl-3/SynCAM-2 localizes to myelinated axons, binds to oligodendrocytes and promotes cell adhesion. *BMC Neuroscience*, 8(1), 90.
- Peng, B., Robert, K. Y., DeHoff, K. L., & Amos, C. I. (2007). Normalizing a large number of quantitative traits using empirical normal quantile transformation. In *BMC proceedings* (Vol. 1, p. S156). BioMed Central.
- Pettem, K. L., Yokomaku, D., Takahashi, H., Ge, Y., & Craig, A. M. (2013). Interaction between autism-linked MDGAs and neuroligins suppresses inhibitory synapse development. *J Cell Biol*, jcb-201206028.
- PGC Bipolar Disorder Working Group. (2011). Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nature*

- Genetics*, 43(10), 977–983.
- Pickrell, J. K., Berisa, T., Liu, J. Z., Séguérel, L., Tung, J. Y., & Hinds, D. A. (2016). Detection and interpretation of shared genetic influences on 42 human traits. *Nature Genetics*, 48(7), 709.
- Pine, D. S., Cohen, E., Cohen, P., & Brook, J. (1999). Adolescent depressive symptoms as predictors of adult depression: moodiness or mood disorder? *American Journal of Psychiatry*.
- Plomin, R., DeFries, J. C., Knopik, V. S., & Neiderhiser, J. (2013). *Behavioral genetics*. Palgrave Macmillan.
- Poelmans, G., Pauls, D. L., Buitelaar, J. K., & Franke, B. (2011). Integrated genome-wide association study findings: identification of a neurodevelopmental network for attention deficit hyperactivity disorder. *American Journal of Psychiatry*, 168(4), 365–377.
- Polderman, T. J. C., Benyamin, B., de Leeuw, C. A., Sullivan, P. F., van Bochoven, A., Visscher, P. M., & Posthuma, D. (2015). Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Genetics*, 47(7), 702–9. <https://doi.org/10.1038/ng.3285>
- Poulton, R., Caspi, a, Moffitt, T. E., Cannon, M., Murray, R., & Harrington, H. (2000). Children’s self-reported psychotic symptoms and adult schizophreniform disorder: a 15-year longitudinal study. *Archives of General Psychiatry*, 57(11), 1053–1058. <https://doi.org/10.1001/archpsyc.57.11.1053>
- Power, R. A., Tansey, K. E., Buttenschøn, H. N., Cohen-Woods, S., Bigdeli, T., Hall, L. S., ... Steinberg, S. (2017). Genome-wide association for major depression through age at onset stratification: Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium. *Biological Psychiatry*, 81(4), 325–335.
- Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8), 904–909.
- Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2), 945–959.
- Purcell, S., Cherny, S. S., & Sham, P. C. (2003). Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics*, 19(1), 149–150.
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., ... Daly, M. J. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3), 559–575.
- R Core Team. (2015). R: A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved from <http://www.r-project.org>
- Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, 1(3), 385–401.
- Raven, J. C., Raven, J. C., & Court, J. H. (1996). *Standard Progressive Matrices: Sets A, B, C, D & E*. Oxford Psychologists Press Oxford, England.
- Raven, J. C., Raven, J. E., & Court, J. H. (1989). *Mill Hill vocabulary scale*. Psychological Corporation.
- Réus, G. Z., Jansen, K., Titus, S., Carvalho, A. F., Gabbay, V., & Quevedo, J. (2015). Kynurenine pathway dysfunction in the pathophysiology and treatment of depression: Evidences from animal and human studies. *Journal of Psychiatric*

- Research*, 68, 316–328.
- Revelle, W. (2015). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Evanston, Illinois. Retrieved from <http://cran.r-project.org/package=psych>
- Ripke, S., Neale, B. M., Corvin, A., Walters, J. T. R., Farh, K.-H., Holmans, P. a., ... O'Donovan, M. C. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511, 421–427. <https://doi.org/10.1038/nature13595>
- Ripke, S., O'Dushlaine, C., Chambert, K., Moran, J. L., Kähler, A. K., Akterin, S., ... Fromer, M. (2013). Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nature Genetics*, 45(10), 1150–1159.
- Ripke, S., Wray, N. R., Lewis, C. M., Hamilton, S. P., Weissman, M. M., Breen, G., ... Cichon, S. (2013). A mega-analysis of genome-wide association studies for major depressive disorder. *Molecular Psychiatry*, 18(4), 497–511.
- Risch, N. J. (2000). Searching for genetic determinants in the new millennium. *Nature*, 405(6788), 847–856.
- Risch, N., & Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science*, 273(5281), 1516–1517.
- Ronald, A. (2015). Recent quantitative genetic research on psychotic experiences: new approaches to old questions. *Current Opinion in Behavioral Sciences*, 2, 81–88. <https://doi.org/10.1016/j.cobeha.2014.10.001>
- Ronald, A., Sieradzka, D., Cardno, A. G., Haworth, C. M. a, McGuire, P., & Freeman, D. (2014). Characterization of psychotic experiences in adolescence using the specific psychotic experiences questionnaire: Findings from a study of 5000 16-Year-Old Twins. *Schizophrenia Bulletin*, 40(4), 868–877. <https://doi.org/10.1093/schbul/sbt106>
- Sabunciyani, S., Aryee, M. J., Irizarry, R. A., Rongione, M., Webster, M. J., Kaufman, W. E., ... Feinberg, A. P. (2012). Genome-wide DNA methylation scan in major depressive disorder. *PLoS One*, 7(4), e34451.
- Sangu, N., Shimojima, K., Takahashi, Y., Ohashi, T., Tohyama, J., & Yamamoto, T. (2017). A 7q31.33q32.1 microdeletion including LRRC4 and GRM8 is associated with severe intellectual disability and characteristics of autism. *Human Genome Variation*, 4, 17001.
- Sawyer, S. M., Afifi, R. A., Bearinger, L. H., Blakemore, S.-J., Dick, B., Ezech, A. C., & Patton, G. C. (2012). Adolescence: a foundation for future health. *The Lancet*, 379(9826), 1630–1640.
- Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium. (2011). Genome-wide association study identifies five new schizophrenia loci. *Nature Genetics*, 43(10), 969–976.
- Schizophrenia Working Group of the Psychiatric Genomics Consortium. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511(7510), 421–427.
- Sellgren, C. M., Kegel, M. E., Bergen, S. E., Ekman, C. J., Olsson, S., Larsson, M., ... Sklar, P. (2016). A genome-wide association study of kynurenic acid in cerebrospinal fluid: implications for psychosis and cognitive impairment in bipolar disorder. *Molecular Psychiatry*, 21(10), 1342–1350.
- Selten, J.-P., van der Graaf, Y., van Duursen, R., Gispen-de Wied, C. C., & Kahn, R. S. (1999). Psychotic illness after prenatal exposure to the 1953 Dutch Flood Disaster. *Schizophrenia Research*, 35(3), 243–245.

- Selten, J.-P., Wiersma, D., & van den Bosch, R. J. (2000). Clinical predictors of discrepancy between self-ratings and examiner ratings for negative symptoms. *Comprehensive Psychiatry*, *41*(3), 191–196.
- Serretti, A., & Mandelli, L. (2008). The genetics of bipolar disorder: genome “hot regions,” genes, new potential candidates and future directions. Nature Publishing Group.
- Servin, B., & Stephens, M. (2007). Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet*, *3*(7), e114.
- Shakoor, S., Zavos, H. M. S., Haworth, C. M. A., McGuire, P., Cardno, A. G., Freeman, D., & Ronald, A. (2016). Association between stressful life events and psychotic experiences in adolescence: evidence for gene–environment correlations. *The British Journal of Psychiatry*, bjp-bp.
- Shakoor, S., Zavos, H. M. S., McGuire, P., Cardno, A. G., Freeman, D., & Ronald, A. (2015). Psychotic experiences are linked to cannabis use in adolescents in the community because of common underlying environmental risk factors. *Psychiatry Research*, *227*(2), 144–151.
- Sieradzka, D., Power, R. A., Freeman, D., Cardno, A. G., Dudbridge, F., & Ronald, A. (2015). Heritability of Individual Psychotic Experiences Captured by Common Genetic Variants in a Community Sample of Adolescents. *Behavior Genetics*. <https://doi.org/10.1007/s10519-015-9727-5>
- Sieradzka, D., Power, R. A., Freeman, D., Cardno, A. G., McGuire, P., Plomin, R., ... Ronald, A. (2014). Are genetic risk factors for psychosis also associated with dimension-specific psychotic experiences in adolescence? *PLoS ONE*, *9*(4). <https://doi.org/10.1371/journal.pone.0094398>
- Siever, L. J., Kalus, O. F., & Keefe, R. S. (1993). The boundaries of schizophrenia. *Psychiatric Clinics of North America*.
- Silverman, W. K., Fleisig, W., Rabian, B., & Peterson, R. A. (1991). Childhood anxiety sensitivity index. *Journal of Clinical Child and Adolescent Psychology*, *20*(2), 162–168.
- Sitskoorn, M. M., Aleman, A., Ebisch, S. J. H., Appels, M. C. M., & Kahn, R. S. (2004). Cognitive deficits in relatives of patients with schizophrenia: a meta-analysis. *Schizophrenia Research*, *71*(2), 285–295.
- Song, C., & Zhang, H. (2014). TARV: Tree-based Analysis of Rare Variants Identifying Risk Modifying Variants in CTNNA2 and CNTNAP2 for Alcohol Addiction. *Genetic Epidemiology*, *38*(6), 552–559.
- Speed, D., Hemani, G., Johnson, M. R., & Balding, D. J. (2012). Improved heritability estimation from genome-wide SNPs. *The American Journal of Human Genetics*, *91*(6), 1011–1021.
- Steinberg, L. (2005). Cognitive and affective development in adolescence. *Trends in Cognitive Sciences*, *9*(2), 69–74.
- Stringer, S., Minică, C. C., Verweij, K. J. H., Mbarek, H., Bernard, M., Derringer, J., ... Maciejewski, D. F. (2016). Genome-wide association study of lifetime cannabis use based on a large meta-analytic sample of 32 330 subjects from the International Cannabis Consortium. *Translational Psychiatry*, *6*(3), e769.
- Strittmatter, L., Li, Y., Nakatsuka, N. J., Calvo, S. E., Grabarek, Z., & Mootha, V. K. (2014). CLYBL is a polymorphic human enzyme with malate synthase and  $\beta$ -methylmalate synthase activity. *Human Molecular Genetics*, *23*(9), 2313–2323.
- Subramanian, S. V., & O'Malley, A. J. (2010). Modeling neighborhood effects: the futility of

- comparing mixed and marginal approaches. *Epidemiology (Cambridge, Mass.)*, 21(4), 475.
- Sullivan, P. F., Kendler, K. S., & Neale, M. C. (2003). Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Archives of General Psychiatry*, 60(12), 1187–1192.
- Sullivan, S. A., Wiles, N., Kounali, D., Lewis, G., Heron, J., Cannon, M., ... Zammit, S. (2014). Longitudinal Associations between Adolescent Psychotic Experiences and Depressive Symptoms. *PLoS ONE*, 9(8), e105758. <https://doi.org/10.1371/journal.pone.0105758>
- Taylor, M. J., Gregory, A. M., Freeman, D., & Ronald, A. (2015). Do sleep disturbances and psychotic-like experiences in adolescence share genetic and environmental influences? *Journal of Abnormal Psychology*, 124(3), 674.
- Taylor, M. J., Robinson, E. B., Happé, F., Bolton, P., Freeman, D., & Ronald, A. (2015). A longitudinal twin study of the association between childhood autistic traits and psychotic experiences in adolescence. *Molecular Autism*, 6(1), 44.
- Terracciano, A., Esko, T., Sutin, A. R., De Moor, M. H. M., Meirelles, O., Zhu, G., ... Realo, A. (2011). Meta-analysis of genome-wide association studies identifies common variants in CTNNA2 associated with excitement-seeking. *Translational Psychiatry*, 1(10), e49.
- Teschler, S., Gotthardt, J., Dammann, G., & Dammann, R. H. (2016). Aberrant DNA Methylation of rDNA and PRIMA1 in Borderline Personality Disorder. *International Journal of Molecular Sciences*, 17(1), 67.
- Thapar, A., & Harold, G. (2014). Editorial Perspective: Why is there such a mismatch between traditional heritability estimates and molecular genetic findings for behavioural traits? *Journal of Child Psychology and Psychiatry*, 55(10), 1088–1091.
- The Pathway Analysis Subgroup of the Psychiatric Genomics Network Consortium. (2015). Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways. *Nature Neuroscience*, 18(2), 199–209.
- Thompson, A., Lereya, S. T., Lewis, G., Zammit, S., Fisher, H. L., & Wolke, D. (2015). Childhood sleep disturbance and risk of psychotic experiences at 18: UK birth cohort. *The British Journal of Psychiatry*, bjp-bp.
- Traylor, M., Bevan, S., Rothwell, P. M., Sudlow, C., Dichgans, M., Markus, H. S., & Lewis, C. M. (2013). Using Phenotypic Heterogeneity to Increase the Power of Genome-Wide Association Studies: Application to Age at Onset of Ischaemic Stroke Subphenotypes. *Genetic Epidemiology*, 37(5), 495–503.
- Trotman, H. D., Holtzman, C. W., Ryan, A. T., Shapiro, D. I., MacDonald, A. N., Goulding, S. M., ... Walker, E. F. (2013). The development of psychotic disorders in adolescence: a potential role for hormones. *Hormones and Behavior*, 64(2), 411–419.
- Trzaskowski, M., Dale, P. S., & Plomin, R. (2013). No genetic influence for childhood behavior problems from DNA analysis. *Journal of the American Academy of Child & Adolescent Psychiatry*, 52(10), 1048–1056.
- Trzaskowski, M., Eley, T. C., Davis, O. S. P., Doherty, S. J., Hanscombe, K. B., Meaburn, E. L., ... Plomin, R. (2013). First genome-wide association study on anxiety-related behaviours in childhood. *PloS One*, 8(4), e58676.
- Uhl, G. R., Liu, Q.-R., Drgon, T., Johnson, C., Walther, D., Rose, J. E., ... Lerman, C. (2008). Molecular genetics of successful smoking cessation: convergent genome-wide association study results. *Archives of General Psychiatry*, 65(6), 683–693.
- Van den Berg, S. M., De Moor, M. H. M., McGue, M., Pettersson, E., Terracciano, A., Verweij,

- K. J. H., ... Van Grootheest, G. (2014). Harmonization of Neuroticism and Extraversion phenotypes across inventories and cohorts in the Genetics of Personality Consortium: an application of Item Response Theory. *Behavior Genetics*, 44(4), 295–313.
- van den Oord, E. J. C. G., Kuo, P.-H., Hartmann, A. M., Webb, B. T., Möller, H.-J., Hettema, J. M., ... Rujescu, D. (2008). Genomewide association analysis followed by a replication study implicates a novel candidate gene for neuroticism. *Archives of General Psychiatry*, 65(9), 1062–1071.
- Van Os, J., Linscott, R. J., Myin-Germeys, I., Delespaul, P., & Krabbendam, L. (2009). A systematic review and meta-analysis of the psychosis continuum: evidence for a psychosis proneness–persistence–impairment model of psychotic disorder. *Psychological Medicine*, 39(2), 179–195.
- van Os, J., & Selten, J.-P. (1998). Prenatal exposure to maternal stress and subsequent schizophrenia. The May 1940 invasion of The Netherlands. *The British Journal of Psychiatry*, 172(4), 324–326.
- Verdoux, H., & van Os, J. (2002). Psychotic symptoms in non-clinical populations and the continuum of psychosis. *Schizophrenia Research*, 54(1), 59–65.
- Viana, J., Hannon, E., Dempster, E., Pidsley, R., Macdonald, R., Knox, O., ... Turecki, G. (2017). Schizophrenia-associated methylomic variation: molecular signatures of disease and polygenic risk burden across multiple brain regions. *Human Molecular Genetics*, 26(1), 210–225.
- Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery. *The American Journal of Human Genetics*, 90(1), 7–24.
- Visscher, P. M., Hemani, G., Vinkhuyzen, A. A. E., Chen, G.-B., Lee, S. H., Wray, N. R., ... Yang, J. (2014). Statistical power to detect genetic (co) variance of complex traits using SNP data in unrelated samples. *PLoS Genet*, 10(4), e1004269.
- Vollema, M. G., & Hoijtink, H. (2000). The multidimensionality of self-report schizotypy in a psychiatric population: an analysis using multidimensional Rasch models. *Schizophrenia Bulletin*, 26(3), 565–575.
- Wain, L. V., Shrine, N., Miller, S., Jackson, V. E., Ntalla, I., Artigas, M. S., ... Cook, J. P. (2015). Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. *The Lancet Respiratory Medicine*, 3(10), 769–781.
- Wang, K., & Huang, J. (2002). A score-statistic approach for the mapping of quantitative-trait loci with sibships of arbitrary size. *The American Journal of Human Genetics*, 70(2), 412–424.
- Wang, X., Zhao, Y., Zhang, X., Badie, H., Zhou, Y., Mu, Y., ... Yang, B. (2013). Loss of sorting nexin 27 contributes to excitatory synaptic dysfunction by modulating glutamate receptor recycling in Down's syndrome. *Nature Medicine*, 19(4), 473–480.
- Welham, J., Scott, J., Williams, G., Najman, J., Bor, W., O'Callaghan, M., & McGrath, J. (2009). Emotional and behavioural antecedents of young adults who screen positive for non-affective psychosis: a 21-year birth cohort study. *Psychological Medicine*, 39(4), 625–634. <https://doi.org/10.1017/S0033291708003760>
- Wichers, M. C., Koek, G. H., Robaey, G., Verkerk, R., Scharpe, S., & Maes, M. (2005). IDO and interferon- $\alpha$ -induced depressive symptoms: a shift in hypothesis from tryptophan depletion to neurotoxicity. *Molecular Psychiatry*, 10(6), 538–544.
- Wigman, J. T. W., Vollebergh, W. A. M., Jacobs, N., Wichers, M., Derom, C., Thiery, E., ... van Os, J. (2012). Replication of the five-dimensional structure of positive psychotic

- experiences in young adulthood. *Psychiatry Research*, 197(3), 353–355.
- Willer, C. J., Li, Y., & Abecasis, G. R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, 26(17), 2190–2191.
- World Health Organisation. (2012). No Title. Retrieved February 1, 2017, from [http://www.who.int/maternal\\_child\\_adolescent/epidemiology/adolescence/en/](http://www.who.int/maternal_child_adolescent/epidemiology/adolescence/en/)
- World Health Organization. (2009a). *Global health risks: mortality and burden of disease attributable to selected major risks*. World Health Organization.
- World Health Organization. (2009b). *Women and health: today's evidence tomorrow's agenda*. World Health Organization.
- World Health Organization. (2012). Public health action for the prevention of suicide: a framework.
- Wray, N. R., Yang, J., Hayes, B. J., Price, A. L., Goddard, M. E., & Visscher, P. M. (2013). Pitfalls of predicting complex traits from SNPs. *Nature Reviews Genetics*, 14(7), 507–515.
- Xu, W., Cohen-Woods, S., Chen, Q., Noor, A., Knight, J., Hosang, G., ... Muglia, P. (2014). Genome-wide association study of bipolar disorder in Canadian and UK populations corroborates disease loci including SYNE1 and CSMD1. *BMC Medical Genetics*, 15(1), 2.
- Yang, J. (2016). GCTA Forum. Retrieved from <http://gcta.freeforums.net/thread/260/grm-cut-off-move-most-samples>
- Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A. A. E., Lee, S. H., ... van Vliet-Ostapchouk, J. V. (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nature Genetics*, 47(10), 1114–1120.
- Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., ... Montgomery, G. W. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 42(7), 565–569.
- Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., & Others. (2010). Common {SNPs} explain a large proportion of the heritability for human height. *Nat Gen*, 42(7), 565–569. <https://doi.org/10.1038/ng.608>. Common
- Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1), 76–82.
- Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M., & Price, A. L. (2014). Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics*, 46(2), 100–106.
- Yoshida, M., Sai, S., Marumo, K., Tanaka, T., Itano, N., Kimata, K., & Fujii, K. (2004). Expression analysis of three isoforms of hyaluronan synthase and hyaluronidase in the synovium of knees in osteoarthritis and rheumatoid arthritis by quantitative real-time reverse transcriptase polymerase chain reaction. *Arthritis Res Ther*, 6(6), R514.
- Zaitlen, N., Kraft, P., Patterson, N., Pasaniuc, B., Bhatia, G., Pollack, S., & Price, A. L. (2013). Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS Genet*, 9(5), e1003520.
- Zammit, S., Hamshere, M., Dwyer, S., Georgiva, L., Timpson, N., Moskvina, V., ... O'Donovan, M. C. (2014). A Population-Based Study of Genetic Variation and Psychotic Experiences in Adolescents. *Schizophrenia Bulletin*, 1–9. <https://doi.org/10.1093/schbul/sbt146>

- Zavos, H. M. S., Freeman, D., Haworth, C. M. a., McGuire, P., Plomin, R., Cardno, A. G., & Ronald, A. (2014). Consistent Etiology of Severe, Frequent Psychotic Experiences and Milder, Less Frequent Manifestations. *JAMA Psychiatry*, 1–10.  
<https://doi.org/10.1001/jamapsychiatry.2014.994>
- Zhang, Z., Zariwala, M. A., Mahadevan, M. M., Caballero-Campo, P., Shen, X., Escudier, E., ... Gerton, G. L. (2007). A Heterozygous Mutation Disrupting the SPAG16 Gene Results in Biochemical Instability of Central Apparatus Components of the Human Sperm Axoneme 1. *Biology of Reproduction*, 77(5), 864–871.
- Zucchi, F. C. R., Yao, Y., Ward, I. D., Ilnytskyy, Y., Olson, D. M., Benzie, K., ... Metz, G. A. S. (2013). Maternal stress induces epigenetic signatures of psychiatric and neurological diseases in the offspring. *PloS One*, 8(2), e56967.