The Journal of Child Psychology and Psychiatry

Journal of Child Psychology and Psychiatry **:* (2021), pp **–** doi:10.1111/jcpp.13519

Using DNA to predict behaviour problems from preschool to adulthood

Agnieszka Gidziela,[1,2] Kaili Rimfeld,[1] Margherita Malanchini,[1,2] Andrea G. Allegrini,[1,3] Andrew McMillan,[1] Saskia Selzam,[1] Angelica Ronald,[4] Essi Viding,[3] Sophie von Stumm,[5] Thalia C. Eley,[1] and Robert Plomin[1]

[1]Social, Genetic & Developmental Psychiatry Centre, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK; [2]School of Biological and Chemical Sciences, Queen Mary University of London, London, UK; [3]Division of Psychology and Language Sciences, University College London, London, UK; [4]Department of Psychological Sciences, Birkbeck University of London, London, UK; [5]Department of Education, University of York, York, UK

Background: One goal of the DNA revolution is to predict problems in order to prevent them. We tested here if the prediction of behaviour problems from genome-wide polygenic scores (GPS) can be improved by creating composites across ages and across raters and by using a multi-GPS approach that includes GPS for adult psychiatric disorders as well as for childhood behaviour problems. **Method:** Our sample included 3,065 genotyped unrelated individuals from the Twins Early Development Study who were assessed longitudinally for hyperactivity, conduct, emotional problems, and peer problems as rated by parents, teachers, and children themselves. GPS created from 15 genome-wide association studies were used separately and jointly to test the prediction of behaviour problems composites (general behaviour problems, externalising, and internalising) across ages (from age 2 to 21) and across raters in penalised regression models. Based on the regression weights, we created multi-trait GPS reflecting the best prediction of behaviour problems. We compared GPS prediction to twin heritability using the same sample and measures. **Results:** Multi-GPS prediction of behaviour problems increased from <2% of the variance for observed traits to up to 6% for cross-age and cross-rater composites. Twin study estimates of heritability, although to a lesser extent, mirrored patterns of multi-GPS prediction as they increased from <40% to 83%. **Conclusions:** The ability of GPS to predict behaviour problems can be improved by using multiple GPS, cross-age composites and cross-rater composites, although the effect sizes remain modest, up to 6%. Our approach can be used in any genotyped sample to create multi-trait GPS predictors of behaviour problems that will be more predictive than polygenic scores based on a single age, rater, or GPS. **Keywords:** Behaviour problems; externalising; internalising; composites; polygenic scores; twin study.

Introduction

Because all behaviour problems in childhood show moderate genetic influence (Cheesman et al., 2017), a next step in genetic research is to find inherited DNA variants responsible for their heritability. The ability to predict behaviour problems from DNA will facilitate research on topics such as how genetic risk unfolds developmentally, gene-environment interaction and correlation, and multivariate issues of genetic heterogeneity and co-morbidity. It will also advance clinical work by identifying problems on the basis of causes rather than symptoms, by moving away from diagnoses towards dimensions, by switching from one-size-fit-all treatments to individually tailored treatments, and by focusing on prevention rather than treatment (Plomin, 2019).

Genome-wide association (GWA) studies identify DNA variants such as single-nucleotide polymorphisms (SNPs) that are associated with complex traits and common disorders (Visscher et al., 2017). Individual SNP associations have small effect sizes, but thousands of SNP associations can be aggregated in genome-wide polygenic scores (GPS) to predict considerably more variance (GPS heritability, aka GPS prediction) for some traits (Martin, Daly, Robinson, Hyman & Neale, 2019).

The most predictive GPS for behavioural traits have been derived from GWA summary statistics for educational attainment (Lee et al., 2018) and general cognitive ability (Savage et al., 2018), with GPS heritabilities up to 16% and 11%, respectively (Allegrini et al., 2019). However, despite substantial twin heritability (a mean of 60%; Cheesman et al., 2017), GPS heritabilities are modest for childhood behaviour problems such as autism spectrum disorder (2.5%; Grove et al., 2019) and ADHD (3.3%; Ronald, de Bode, & Polderman, 2021). In a recent study, GPS prediction of childhood ADHD symptoms, internalising and social problems was reported to be much lower for adult-based GPS of major depression (0.2%), neuroticism (0.1%), insomnia (0.05%), and subjective wellbeing (0.06%) (Akingbuwa et al., 2020). A recent GWA study of childhood and adolescence internalising symptoms predicted 0.4% of the variance in internalising at age 7 and 0.03% at ages 13–18 (Jami et al., 2020). However, the predictive power of GPS is dependent on the size of discovery samples used in GWA studies, which

Conflict of interest statement: No conflicts declared.

needs to be considered when comparing GPS prediction across cognitive and psychiatric traits. For example, for educational attainment (Lee et al., 2018), sample sizes reach up to 1.1 million individuals, whereas some of the GWA studies of psychiatric disorders had sample sizes of less than 20,000 cases (Demontis et al., 2019; Grove et al., 2019). GPS will become more predictive as GWA sample sizes increase and as whole-genome sequencing identifies all DNA variants, rare as well as common, that contribute to heritability (Visscher et al., 2017).

Using existing GPS, we explored ways to increase the prediction of childhood behaviour problems from DNA. Research suggests that using multiple GPS in a multivariate framework can improve prediction (Allegrini, Karhunen, et al., 2020; Allegrini et al., 2019; Grotzinger et al., 2019; Krapohl et al., 2018; Pain et al., 2021). To test the hypothesis that the multi-GPS approach will yield greater GPS heritability than the single-GPS approach, we assessed the joint prediction of 15 GPS in penalised regression models with hold-out evaluation of prediction accuracy (multi-GPS). In addition to GPS for childhood behaviour problems (ADHD; autism spectrum disorder) (Demontis et al., 2019; Grove et al., 2019), we included GPS derived from the much larger GWA studies of adult psychiatric disorders such as schizophrenia (Pardiñas et al., 2018), bipolar disorder (Stahl et al., 2019), and major depressive disorder (Wray et al., 2018) and traits such as neuroticism (Luciano et al., 2018), well-being (Okbay et al., 2016), and risk-taking (Linnér et al., 2019) as they have been shown to predict a variety of childhood phenotypes, including general psychopathology (Allegrini, Cheesman, et al., 2020) and behaviour problems (Akingbuwa et al., 2020).

In both phenotypic and DNA-based analyses of behaviour problems, a general factor of psychopathology has been observed that is known as a 'p-factor' or 'p' (Allegrini, Cheesman, et al., 2020; Caspi et al., 2014), suggesting that diverse behaviour problems share common genetic influences. Accordingly, we created latent composites of general behaviour problems (BPp, externalising and internalising) and used the multi-GPS approach to test two other hypotheses to improve GPS prediction.

First, because age-to-age stability is largely driven genetically (Nivard et al., 2015; Plomin, 2019), we hypothesised that longitudinal composites of behaviour problems composites would yield greater GPS heritability than age-specific observed variables, as suggested by previous genomic research (Cheesman et al., 2018).

Second, building on the assumption that behaviour problems that emerge across situations are more heritable than situation-specific problems, we hypothesised that GPS heritability is greater for behaviour problems composites across raters such as parents, teachers and children themselves who see behaviour problems in different settings than behaviour problems assessed only by one rater (Bartels et al., 2004; Cheesman et al., 2018).

We tested these hypotheses in a sample of 3,065 unrelated individuals from the Twins Early Development Study (TEDS; Rimfeld et al., 2019) for whom we had genotypes and ratings of behaviour problems from early childhood to early adulthood for parents, teachers and the children themselves from age 2 to 21. Because these unrelated individuals were members of twin pairs, we included their co-twins in analyses to estimate heritability using the twin method, testing the hypotheses that cross-age and cross-rater composites increase twin heritability, mirroring the patterns of GPS heritability.

Methods

Our hypotheses and analyses were preregistered in Open Science Framework (OSF) (https://osf.io/27tpj/) prior to accessing the data. Please see Appendix S1 for details. Scripts have been made available on the OSF website.

Participants

Our sample consists of twins born in England and Wales between 1994 and 1996 who were enrolled in the TEDS (for a detailed description of the sample, please refer to the Appendix S2, Table S1 and Rimfeld et al., 2019).

In the current study we investigated heritability of behaviour problems, using data collected when the twins were aged approximately 2, 3, 4, 7, 9, 12, 16 and 21 years old. The sample selected for construction of composites included twins who had at least half of the data on behaviour problems complete across ages and raters. Patterns of missing data were addressed using the full information maximum likelihood. This resulted in a sample of 4,778 twin pairs.

DNA has been genotyped for a subsample of 7,026 unrelated individuals from TEDS (i.e. one twin per pair), out of which 3,065 individuals were included in the present study, which provides a sample size adequate to detect a correlation of 0.10 with more than 99% power (Hulley, Cummings, Browner, Grady, & Newman, 2013). For details on sample sizes per composite, please refer to Table S2.

Genotyping took place on two different genotyping platforms (AffymetrixGeneChip 6.0 and Illumina HumanOmniExpress Exome-8v1.2) in two separate waves. For a detailed genotyping protocol, see Selzam et al. (2018).

Measures

Polygenic scores. Our methods for obtaining DNA, genotyping, quality control and constructing polygenic scores have been described previously (Selzam et al., 2018). In the present analyses, we included 15 GPS of behaviour problems and psychopathology, derived from the most powerful GWA studies, which were used in our previous research (Allegrini, Karhunen, et al., 2020; Allegrini et al., 2019). For the list of polygenic scores, please refer to Appendix S3 and Table S3.

Behaviour problems. We assessed hyperactivity, conduct, emotional and peer problems from early childhood to early adulthood as rated by parents, teachers and the twins themselves. The Preschool Behaviour Questionnaire (PBQ; Behar, 1977) was used to rate hyperactivity, conduct and emotional problems at ages 2 and 3. At ages 4, 7, 9, 12, 16 and 21, the Strengths and Difficulties Questionnaire (SDQ;

Goodman, 1997) assessed peer problems in addition to hyperactivity, conduct and emotional problems. For a description of measure administration and scoring, and an illustration of the four behaviour problems domains across development, please refer to Appendix S4 and Figure S1. We also assessed mental health outcomes reported by the twins at age 21, such as mental health diagnoses and whether they have ever taken a medication for mental health.

Composites

Composites across ages and raters (Figure 1) were constructed using the hierarchical latent factor model, where the two first-order factors (externalising and internalising) loaded on a second-order factor of BPp. The hierarchical modelling was conducted using confirmatory factor analysis, based upon the results of exploratory factor analyses. For details on the exploratory and confirmatory factor analyses and composite construction, please refer to Appendix S5 and S6 and Figures S2–S4.

Using hierarchical confirmatory factor analysis, we constructed cross-age and cross-rater composites of BPp, externalising and internalising. We created cross-age composites from age 2 to 21 separately for each rater, which yielded nine cross-age composites (three rater-specific composites of BPp, three rater-specific composites of externalising, three rater-specific composites of internalising). Cross-rater composites were constructed separately in childhood (ages 2–9), adolescence (ages 12 and 16) and early adulthood (age 21), which yielded nine cross-rater composites (i.e., three age-specific composites each of BPp, externalising, and internalising). The construction of the cross-age and cross-rater composites is summarised in Figure S1A,B, respectively. Phenotypic and genetic correlations between the cross-age and cross-rater composites are presented in Appendix S7 and Figure S5.

To explore whether simultaneously aggregating cross-age and cross-rater effects improves GPS heritability, we constructed cross-age-and-rater composites of BPp, externalising, and internalising, using a three-level hierarchical model. In this model, we analysed behaviour problems at all ages (2–21) rated by parent, teacher and child (cross-age approach) to create the first-order factors of cross-age externalising and internalising, which were then combined across raters to create the second-order factors of cross-age-and-rater externalising and internalising, which subsequently gave rise to the third-order cross-age-and-rater BPp factor (Figure 1C). We validated this approach by combining the behaviour problems scales across raters, but separately in childhood, adolescence and adulthood (cross-rater approach) on the first-order factor level, which yielded similar results.

In addition, we created single-trait composites for the four behaviour problems (hyperactivity, conduct, emotional problems, and peer problems) in order to compare the effects of single-trait composites to BPp, externalising, and internalising composites. Construction and results for the single-trait composites are presented in Appendix S8 and Figures S6 and S7.

Analyses

All variables were regressed on 10 genetic principal components of population structure, genotyping chip, and genotyping batch (Allegrini et al., 2019). The standardised residuals from these regressions were used in all downstream analyses.

GPS heritability

GPS are the estimated effects of thousands of genetic variants on a trait and are calculated as a weighted

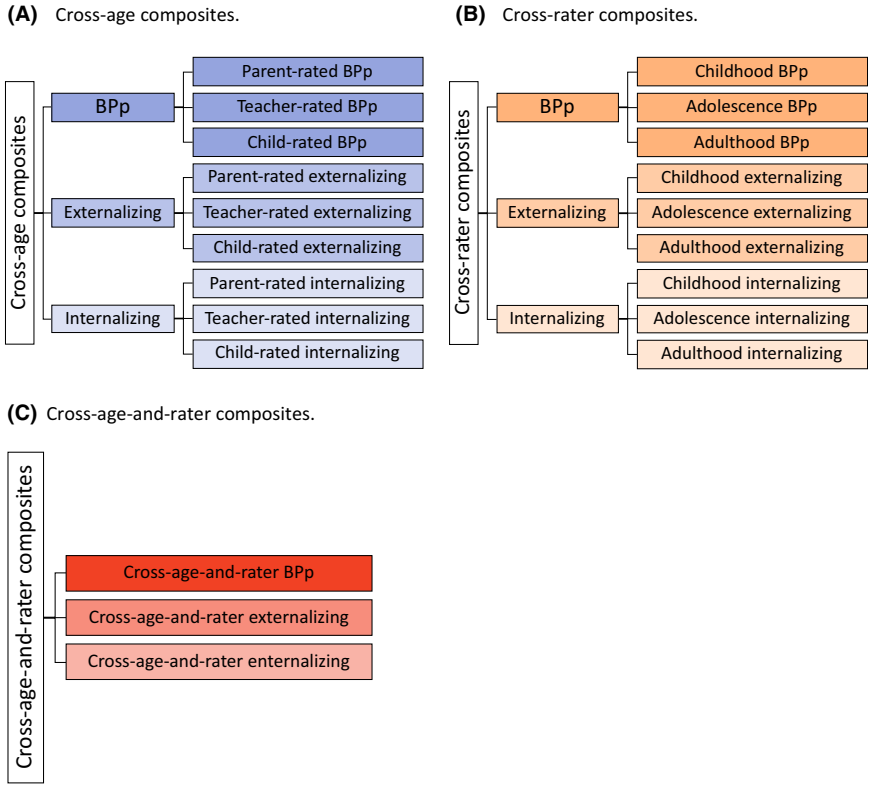

Figure 1 Summary of the construction of the cross-age, cross-rater and cross-age-and-rater composites. This figure illustrates the components of the cross-age and cross-rater composites; it is not the hierarchical model used to create composites

sum of alleles associated with the trait based on summary statistics from GWA studies (Dudbridge, 2013). The GPS were constructed using LD-pred (Vilhjálmsson et al., 2015), with the 1,000 Genomes phase 1 sample as a reference for linkage disequilibrium structure. A detailed description of our LD-pred analytic strategy used to calculate GPS has been published (Allegrini et al., 2019). We report results for GPS created using a fraction of causal markers of 1.0 (i.e., assuming that all SNPs have non-zero effects), although results for GPS fractions 0.3 and 0.01 are presented in Tables S4–S7. In addition, we reported the GPS results separately for males and females (Table S6).

We estimated the joint prediction of the 15 GPS (multi-GPS heritability) in a penalised regression elastic net model (Zou & Hastie, 2005) with hold-out evaluation of prediction accuracy. For details on the elastic net regularisation analytic procedure, please refer to Appendix S9 and Allegrini, Karhunen, et al. (2020).

Multi-GPS effects

To investigate whether a multi-GPS approach improved prediction as compared to a single-GPS approach, we compared the joint prediction of behaviour problems by the 15 GPS (multi-GPS heritability) to individual predictions yielded by each of the 15 GPS alone (single-GPS heritability). The multi-GPS heritability was estimated in elastic net regularisation models and multiple regression models (using adjusted R^2). Single-GPS heritability was estimated using squared correlations (r^2) between each of the 15 GPS and composites.

Compositing effects

We compared the multi-GPS heritability for the composites to the mean multi-GPS heritability for the individual constituent behaviour problem traits that comprise these composites (that is, the age-specific and rater-specific traits, which we will refer to as observed traits) (Tables S7 and S8). For example, the multi-GPS heritability of the cross-age parent-rated externalising composite was compared to the mean of multi-GPS heritabilities of parent-rated hyperactivity and conduct scales across ages 2–21. Although the focus of this article is to present a broad picture of the effect sizes, rather than formally testing for significant differences, in order to present the 95% confidence intervals of the estimates that index significance of differences, we also used a meta-analytic approach (Appendix S10 and Figures S8 and S9).

Analysis of extremes

In addition to continuous analyses, we investigated the ability of GPS to predict differences in behaviour problems at the decile extremes of the multi-trait GPS, using the cross-age-and-rater composites of BPp, externalising and internalising as an example. We created multi-trait GPS scores based on the individual predictor GPS coefficients from the elastic net regularisation models (Table S9), using the following formula:

$$\mathrm{GPSmulti\_trait}_i = \sum_{j=1}^{k} \mathrm{GPS}_{ij}\beta_j$$

where GPSmulti_trait is the multi-trait GPS for individual i in the full sample, j {1, 2, …, 15} and denotes the GPS value for the k GPS for individual i and β indicates the elastic net coefficient of the association between the jth predictor GPS and the composite that was learnt in the training set (see Appendix S9 for details).

After assigning multi-trait GPS scores to each individual for BPp, externalising and internalising, we divided the sample into deciles and compared their mean phenotypic scores for BPp, externalising, and internalising as well as for other mental health outcomes.

Twin heritability

We compared the multi-GPS heritability results to heritability results from twin analyses (Tables S8 and S10). The classical univariate twin design was employed to estimate broad heritability (additive and non-additive genetic variance) for individual behaviour problems as compared to composites. We performed twin analyses using OpenMx 2.0 for *R* (Neale et al., 2016; R Core Team, 2021). Additionally, we report the univariate twin model estimates separately for males and females (Table S11).

In order to investigate the impact of compositing on twin heritability, we contrasted twin heritability estimates for composites to the mean twin heritabilities for the observed traits. Significance of these differences was assessed using a meta-analytic approach (Appendix S10).

Results

Multi-GPS heritability: cross-age and cross-rater composites

Results of the multi-GPS prediction with elastic net regularisation are shown in Figure 2 for cross-age composites (Figure 2A) and cross-rater composites (Figure 2B). As shown in Appendix S7, cross-age and cross-rater composites were substantially correlated phenotypically and genetically (Table S12).

Compositing across ages increased multi-GPS heritability as compared to the mean multi-GPS heritability of observed traits for BPp, externalising and internalising (Figure 2A). The greatest cross-age effect was found for parent-rated BPp, with the

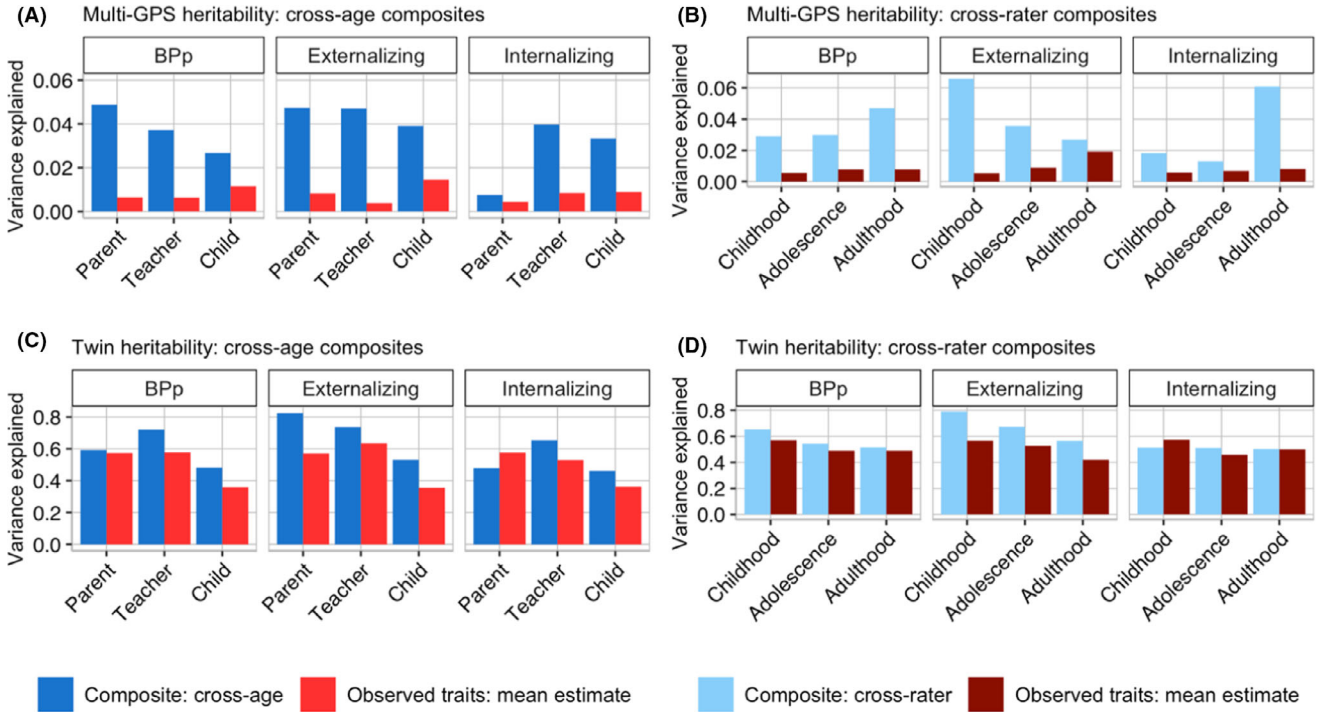

Figure 2 Multi-GPS and twin heritability of cross-age composites and cross-rater composites, compared to the mean multi-GPS and twin heritability of observed traits

multi-GPS predicting 4.9% of the variance, as compared to the mean estimate of 0.6% when considering observed behaviour problems. Parent-rated multi-GPS heritabilities were 4.7% versus 0.8% for externalising, but only 0.7% versus 0.4% for internalising. For teacher ratings, the multi-GPS heritabilities were 3.7% versus 0.6% for BPp, 4.7% versus 0.4% for externalising problems and 4% versus 0.8% for internalising problems. Finally, for child ratings, the multi-GPS heritabilities were 2.7% versus 1.2% for BPp, 3.9% versus 1.4% for externalising problems and 3.3% versus 0.9% for internalising problems.

Compositing across raters also increased multi-GPS heritability in childhood, adolescence, and adulthood, as shown in Figure 2B. For BPp, multi-GPS heritability for cross-rater composites was 2.9% as compared to the mean of 0.5% for the observed traits in childhood, 3.0% versus 0.8% in adolescence and 4.7% versus 0.8% in adulthood. For externalising problems, multi-GPS heritabilities were 6.6% versus 0.5% in childhood, 3.6% versus 0.9% in adolescence and 2.7% versus 1.9% in adulthood. For internalising problems, multi-GPS heritabilities were 1.8% versus 0.6% in childhood, 1.3% versus 0.7% in adolescence and 6.0% versus 0.8% in adulthood. The greatest cross-rater effect was found for externalising problems in childhood, with the multi-GPS prediction of 6.6% as compared to 0.8% for observed traits. The analogous twin heritabilities (Figure 2C, D) are discussed later.

Figure 3 compares the multi-GPS approach to the single-GPS approach in prediction of cross-age and cross-rater composites. The first row of each of the six panels in Figure 3 repeats the results in Figure 2 showing the multi-GPS prediction using elastic net regularisation for cross-age composites (Figure 3A) and cross-rater composites (Figure 3B). The second row shows that in most cases the elastic net regularisation performed better than adjusted R^2 from simple multiple regressions. The rest of each panel shows the variance explained (correlation squared) by each of the 15 GPS alone.

For BPp and externalising problems, the ADHD GPS was the most predictive GPS for cross-age and cross-rater composites, predicting up to 2.6% of the variance in the cross-rater composite of adulthood BPp and 2.5% of the variance in the cross-age composite of teacher-rated externalising. Other than the ADHD GPS, none of the individual GPS predicted more than 1.5% of the variance. For internalising problems, the most predictive GPS was the neuroticism GPS, which predicted up to 1.4% of the variance in cross-age composites of child-rated internalising and cross-rater childhood internalising and 1.3% in cross-rater composites of childhood and adolescence internalising.

Twin heritability: cross-age and cross-rater composites

Figure 2C,D summarises twin heritability estimates for cross-age composites (Figure 2C) and cross-rater composites (Figure 2D) as compared to the mean estimates of twin heritability of the observed traits. In general, cross-age and cross-rater composites

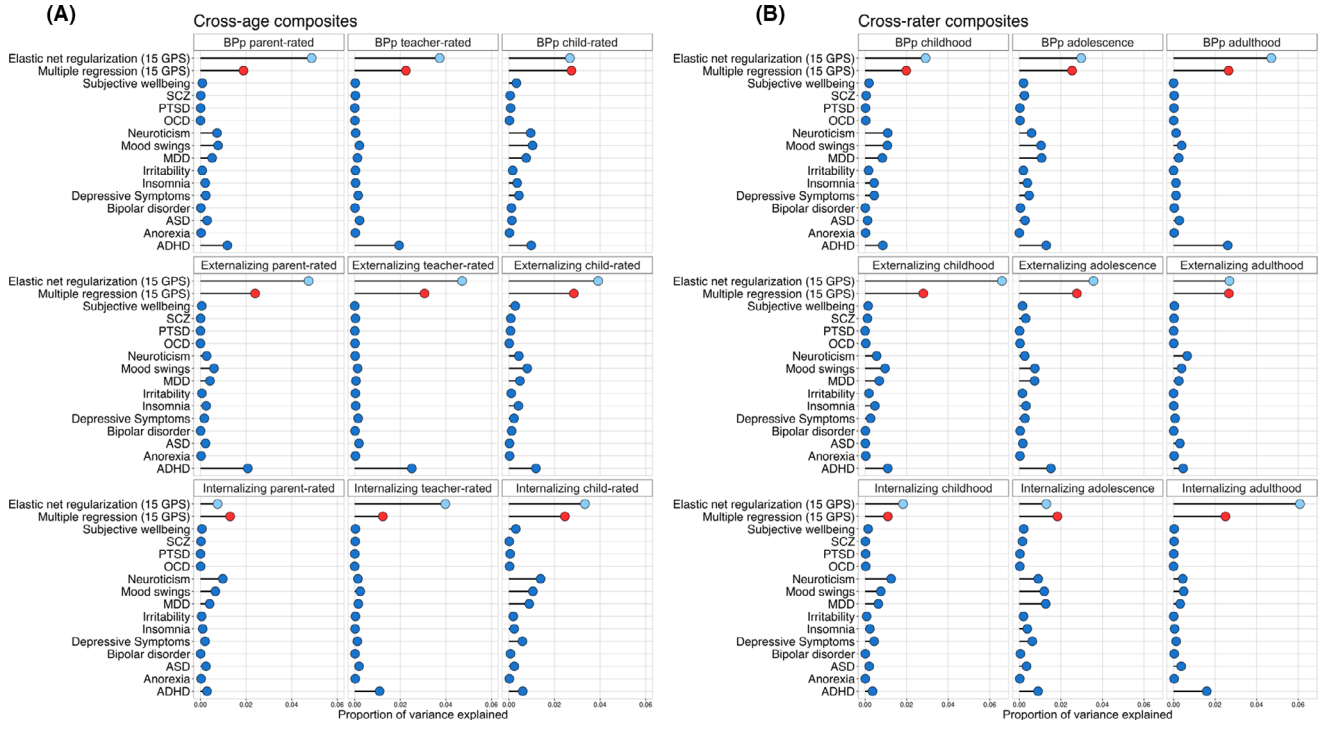

Figure 3 Multi-GPS prediction as compared to single-GPS prediction of cross-age composites and cross-rater composites

yielded greater twin heritability estimates than the observed traits.

The average heritability for cross-age composites was 61% as compared to 50% for the observed traits (Figure 2C). The largest difference was found for parent-rated externalising problems (82% vs. 57%). The pattern of cross-age effects for twin heritability largely mirrored the multi-GPS heritability results, with the notable exception that twin heritability showed no increase for parent ratings of BPp, whereas this was one of the largest cross-age effects for multi-GPS heritability.

For cross-rater composites, the average heritability was 58% as compared to 51% for the observed traits (Figure 2D). The average cross-rater effect across the three ages was strongest for externalising problems (68% vs. 55%), weaker for BPp (57% vs. 54%) and absent for internalising problems (51% vs. 53%). The strongest cross-rater effect was observed for externalising problems in childhood (79% vs. 57%), which is consistent with the multi-GPS results. Similar to multi-GPS heritability, twin heritability for cross-rater externalising problems decreased from childhood (79%) to adolescence (67%) to adulthood (57%).

Aggregated cross-age-and-rater effects

Figure 4 compares the multi-GPS heritability and twin heritability obtained for the combined cross-age-and-rater composites. Multi-GPS heritabilities of the cross-age-and-rater composites were similar to multi-GPS heritability of the cross-age composites (Figure 4A) and cross-rater composites (Figure 4B). Combining traits across ages and raters did not significantly improve GPS heritability. The variance explained by the GPS in the combined cross-age-and-rater composites (3.3%) was similar to the average prediction yielded by cross-age and cross-rater composites (3.6%).

The twin analyses also showed that the benefits of cross-age and cross-rater compositing are not additive (Figure 4C,D, respectively). The twin heritability for cross-age-and-rater composites (63%) was similar to the mean twin heritability yielded by cross-age and cross-rater composites (60%).

Analysis of multi-trait GPS decile extremes

Multi-trait GPS scores for BPp, externalising and internalising were created for each individual as explained earlier. We used these multi-trait GPS scores to divide the sample into deciles. Figure 5 shows box plots, presenting the z-standardised scores for cross-age-and-rater BPp, externalising, and internalising as a function of the multi-trait GPS deciles. Mean behaviour problems increase linearly from the lowest to the highest GPS deciles, with a scatterplot of scores as expected from the modest correlations between the multi-trait GPS and BPp (r = .19), externalising (r = .20), and internalising (r = .16). At the lowest and highest decile extremes, the differences are substantial: the mean standard score difference between the lowest and highest GPS deciles is 0.61 for BPp, 0.67 for externalising and 0.51 for internalising.

Differences between the lowest and highest deciles were reflected in mental health outcomes. For

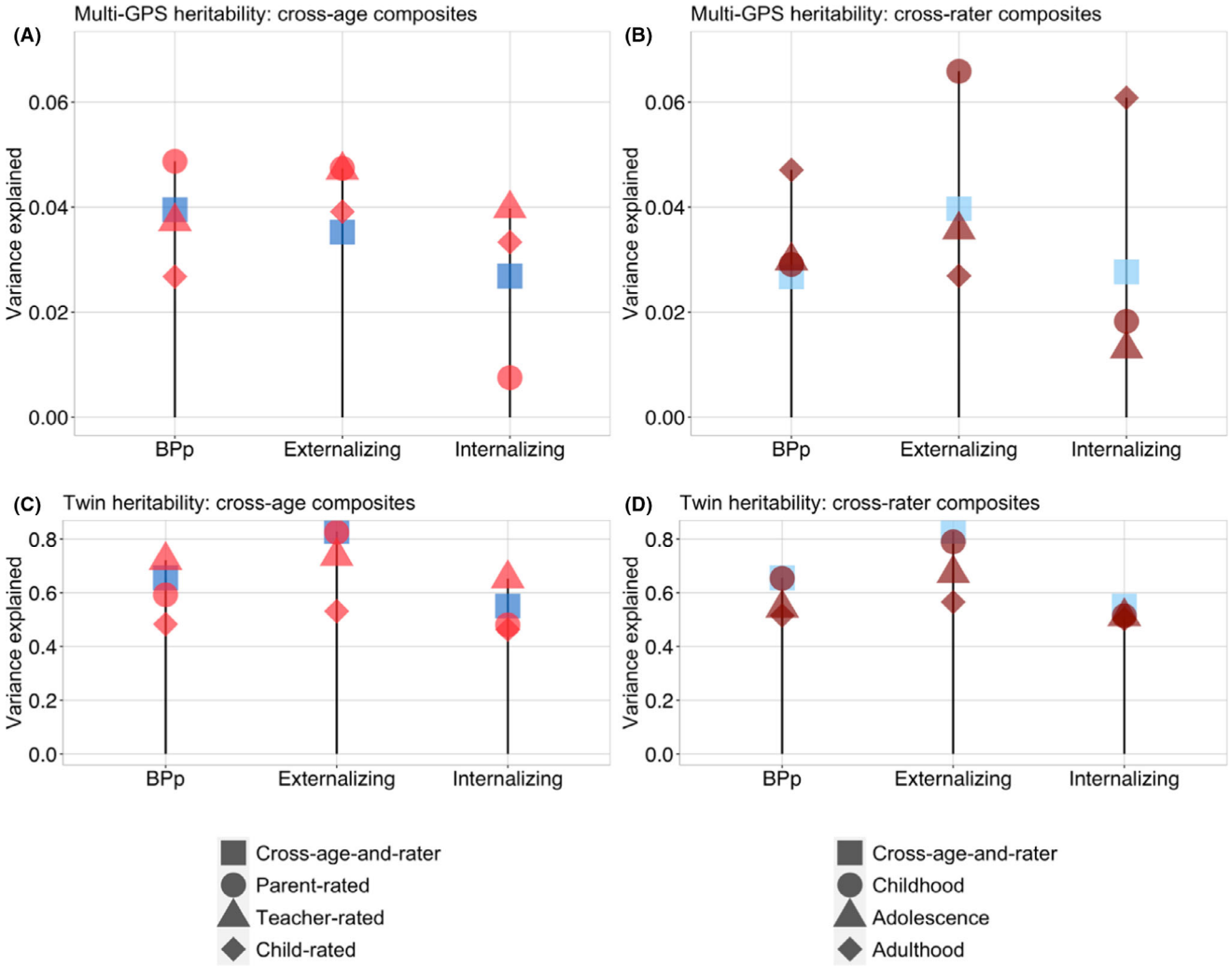

Figure 4 Multi-GPS heritability and twin heritability of cross-age-and-rater composites as compared to cross-age composites and to cross-rater composites. Dark blue squares signifies the cross-age-and-rater composites constructed using the cross-age approach; lights blue squares signifies the cross-age-and-rater composites constructed using the cross-rater approach (see Methods)

example, 15% of individuals in the lowest multi-trait GPS decile for BPp and 15% in the lowest multi-trait GPS decile for externalising have taken medication for mental health, compared to 20% in the highest BPp and 21% in the highest externalising decile, although these differences are not statistically significant [odds ratio and 95% confidence intervals: 1.47 (0.84, 2.57) for BPp and 1.01 (0.58, 1.78) for externalising]. For the multi-trait internalising GPS, 12% of individuals in the lowest decile have been diagnosed with depression, compared to 19% in the highest decile [odds ratio and 95% confidence intervals: 1.82 (1.03, 3.24)], while 7% of individuals in the lowest decile have been diagnosed with anxiety disorder, compared to 19% in the highest decile [odds ratio and 95% confidence intervals: 2.90 (1.53, 5.75)].

Discussion

Our findings indicate that a multi-GPS approach using cross-age and cross-rater composites doubles the prediction estimates for general behaviour problems. These results are bolstered by twin analyses showing, although to a lesser extent, increased heritability for cross-age and cross-rater composites. The twin heritability estimates can be viewed as the prediction ceiling for GPS because the twin design assesses the effect of all inherited DNA differences, not just SNPs shown to be associated with behavioural problems.

The multi-GPS weights for our cross-age-and-rater composites that simultaneously composite across age and across raters provide the best polygenic prediction currently available for children's BPp, externalising and internalising problems (Table S4). These multi-GPS beta weights may be useful as genetic predictors of behaviour problems for other samples with DNA regardless of whether behaviour problems data are available. Just as GPS can be created from DNA for any sample, our sets of multi-GPS weights can be used to create the strongest genetic estimates of BPp, externalising and

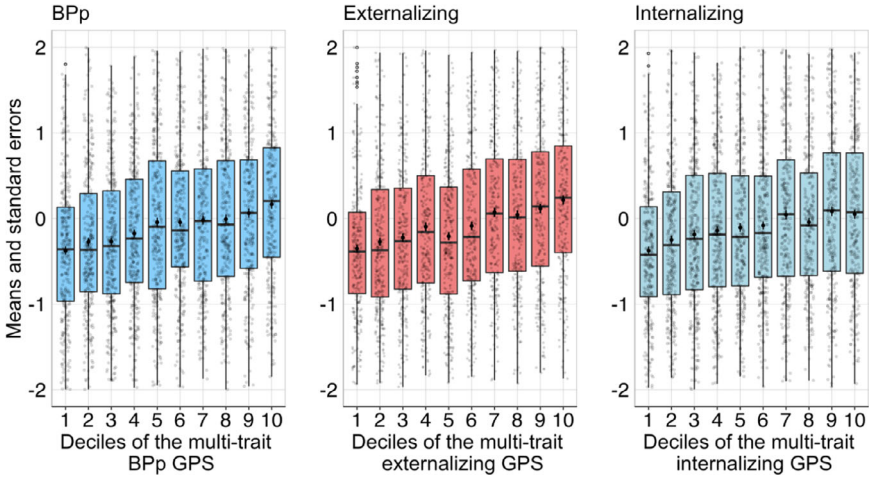

Figure 5 Box plots showing z-standardised means and distributions of cross-age-and-rater multi-trait GPS scores for BPp, externalising, and internalising. The boxes enclose 50% of the distribution of each GPS decile. Horizontal lines in boxes indicate the median values. Dots and error bars (vertical bars going through dots) in boxes indicate means and standard errors. Vertical lines outside the boxes indicate the normal distribution of GPS deciles. Point contours indicate outliers

internalising based on cross-age, cross-rater, and cross-age-and-rater composites. These multi-trait GPS can facilitate developmental, multivariate, and gene-environment interplay research because they are more predictive of behaviour problems than polygenic scores based on a single age, rater, or trait. However, external validation in other samples is necessary to determine the degree to which these weights can be considered optimal.

Our goal of increasing GPS heritability led us to focus on compositing across ages, raters, and traits, which should not be seen to denigrate the continued search for specific genetics effects for each age, rater, or trait. Although we present results for the cross-age-and-rater multi-trait GPS for conceptual consistency, we report weights for all the composites, which will allow researchers to construct developmental stage-specific and rater-specific multi-trait GPS. However, it should be noted that the TEDS sample is largely of European ancestry, so are the samples involved in GWA studies from which the GPS were derived, and the reported GPS results are likely to be less predictive in other ancestral populations (Peterson et al., 2019).

In order to condense the results, we focused on the second-order factors of externalising and internalising and a third-order factor representing BPp. However, we also present multi-GPS weights for the single-trait cross-age and cross-rater composites of hyperactivity, conduct, emotional and peer problems (Table S4). Although these traits generally showed increased GPS and twin heritability for cross-age and cross-rater composites, results for these trait-specific factors are subject to more measurement error; hence, the results are less consistent than for the general factors representing BPp, externalising and internalising problems.

More research is needed to identify the mechanisms by which compositing increases GPS prediction. We had assumed that compositing across ages captures new genetic effects that come on board at later ages and that compositing across raters captures trans-situational genetic effects in the home for parent ratings and in school for teacher ratings. However, if different mechanisms are responsible for increasing GPS prediction for cross-age and cross-rater composites, we would expect that the effects of compositing across ages and across raters would be additive. Instead, we found that the combined cross-age-and-rater composites do not show increased GPS heritability nor increased twin heritability as compared to the cross-age and cross-rater composites. Notably, we found that cross-age effects differ depending on rater, and, similarly, cross-rater effects depend on developmental stage. Furthermore, age and rater effects may correlate within, but not between developmental stages. In childhood, ratings were made mostly by parents, with teacher ratings appearing from age 7 and self-ratings appearing only at age 9. In adolescence, behaviour problems were rated equally by parents, teacher, and self-report, while in adulthood the teacher-ratings were no longer available. These interactions might explain in part why cross-age and cross-rater effects do not add up. However, going against this interaction hypothesis is the strong phenotypic overlap (~0.60) and genetic overlap (~0.65) between cross-age and cross-rater effects (Appendix S7), which suggests that to a large extent the same mechanisms are responsible for increasing heritability for cross-age and cross-rater composites. A likely candidate is increased reliability, which could increase heritability for both cross-age and cross-rater composites. However, this reliability hypothesis requires the added assumption that compositing either across ages or across raters reaches a ceiling of reliability, so that there is no additional increase in heritability for cross-age-and-rater composites.

Our results are limited to existing GWA studies and will need to be updated as new GWA studies are reported. A more specific limitation is that we focused on the 15 most powerful GWA of psychopathology regardless of whether the GWA analysis targeted childhood disorders (autism spectrum disorder and ADHD) or disorders in adulthood (e.g., schizophrenia and depression). It is reasonable to expect that GWA studies targeted on childhood disorders will add disproportionately to the multi-GPS prediction of childhood behaviour problems. Supporting this expectation is our finding that the ADHD GPS was by far the strongest single GPS predictor of behaviour problems, especially for parent and teacher ratings of the BPp factor and externalising problems (Figure 3A). Nonetheless, multi-GPS predicted twice as much variance, with adult-based GPS for neuroticism, mood swings, and major depressive disorder contributing to the prediction from the ADHD GPS, which could imply sequential comorbidity, where childhood ADHD can be predictive of both externalising and internalising problems in adolescence and emerging adulthood. The predictive power of ADHD GPS across developmental stages, raters, and behaviour problems may also point towards overlapping longitudinal processes underlying both early risk and later externalising and internalising problems. Our approach is atheoretical and empirical in the sense that we would include any GPS, child-based or adult-based, that adds to the multi-GPS prediction of behaviour problems.

There is special value in focusing on GPS derived from adult-based GWA studies because they predict adult psychiatric disorders from childhood regardless of their associations with childhood behaviour problems. We chose not to do this at this time because our aim was to increase the DNA prediction of childhood behaviour problems, and we show that multi-GPS limited to extant adult-based GWA studies are weak predictors of childhood behaviour problems.

Although compositing doubles the predictive power of GPS, the effect sizes remained modest (<6%), suggesting that it is still a long way before we will reach levels of prediction that can be useful clinically in diagnosis, treatment, or prevention. Nonetheless, even with their current effect sizes, GPS can be useful in clinical research. For example, we show (Figure 5) that sizeable (Cohen's $d = .5$) mean differences in behaviour problems are observable at the multi-trait GPS decile extremes, such as the twofold greater risk of a depression diagnosis for individuals in the highest versus lowest decile of the internalising GPS, although it should be noted that these results might not apply to other samples.

Increasing the power of GPS to predict behaviour problems is the first step to exploring the biological and environmental mechanisms that mediate this prediction so that the predictive power of GPS can be brought into more actionable space and, eventually, to prevention. This ultimately depends on bigger GWA studies that can scoop up SNP associations of miniscule effect sizes, and whole-genome sequencing that can detect all differences in inherited DNA sequence, not just common SNPs (Wainschtein, Jain, Yengo, Zheng, & Visscher, 2019). Our results indicate that GWA studies can also increase their power to detect effects by conducting GWA analyses using cross-age or cross-rater composites instead of age- and rater-specific measures, to capture longitudinal and trans-situational effects, minimising the measurement error.

It seems likely that GPS will eventually be sufficiently powerful predictors that they will affect not only clinical work but also society more generally (von Stumm & Plomin, 2021). DNA testing has already been incorporated in the national health services of Finland and Estonia and is being trialled in the United Kingdom. The next step will be DNA testing at birth. Francis Collins, the head of the US National Institutes of Health and leader of the Human Genome Project, predicted: 'I am almost certain that complete genome sequencing will become part of newborn screening in the next few years.... It is likely that within a few decades people will look back on our current circumstance with a sense of disbelief that we screened for so few conditions' (Collins, 2010, p. 50). The current five-year plan of the Chinese government is to sequence the DNA of at least 50% of the 15 million babies born each year in China (Metzl, 2019).

Medical uptake of DNA testing is driven by its potential to predict and prevent rare single-gene disorders as well as preventable common medical disorders such as cardiovascular disease. However, the same genomic results from DNA testing can also be used to create GPS for many other traits, including behaviour problems. Now is the time to discuss how to maximise clinical benefits and minimise risks.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article:

Appendix S1. Hypotheses from our OSF (Open Science Framework) statement (https://osf.io/27tpj/).
Appendix S2. Sample description.
Appendix S3. Polygenic scores.
Appendix S4. Additional information about the behaviour problems measures: SDQ and PBQ.
Appendix S5. Exploratory factor analysis (EFA).
Appendix S6. Confirmatory factor analysis (CFA).
Appendix S7. Phenotypic and genetic correlations between cross-age and cross-rater composites.
Appendix S8. Construction and results for the single-trait composites.
Appendix S9. Elastic net regularization.

Appendix S10. Meta-analytic approach to comparing multi-GPS heritability between composites and observed traits.
Table S1. Representativeness of the selected sample.
Table S2. Behaviour problems composites: sample characteristics.
Table S3. Polygenic scores and sample sizes.
Table S4. Behaviour problems composites: model fit indices and predictions from elastic net regularization.
Table S5. Behaviour problems composites: model fit indices and predictions from multiple regression.
Table S6. Behaviour problems composites: model fit indices and predictions from multiple regression, for males and females separately.
Table S7. SDQ scales (observed traits): model fit indices and predictions from elastic net regularization.
Table S8. SDQ scales (observed traits): univariate twin model fitting results.
Table S9. Weights for the individual polygenic scores from elastic net regularization.
Table S10. Behaviour problems composites: univariate twin model fitting results.
Table S11. Behaviour problems composites: univariate twin model fitting results, for males and females separately.
Table S12. Cross-age and cross-rater behaviour problems composites: bivariate twin model fitting results.
Figure S1. The sunburst plot showing the observed variables at each age.
Figure S2. Exploratory factor analyses for parent, teacher and child-rated data.
Figure S3. Hierarchical and bifactor cross-age model of BPp, externalizing and internalizing.
Figure S4. Correlations between hierarchical and bifactor composites of BPp, externalizing and internalizing for parent, teacher and child ratings.
Figure S5. Phenotypic and genetic correlations between cross-age and cross-rater composites of BPp, externalizing and internalizing.
Figure S6. Summary of the construction of the single-trait cross-age and single-trait cross-rater composites.
Figure S7. Multi-GPS and twin heritability results for single-trait cross-age composites and single-trait cross-rater composites as compared to the mean multi-GPS and twin heritability of observed traits.
Figure S8. Multi-GPS correlation and twin heritability results for cross-age composites and cross-rater composites as compared to the grand mean multi-GPS correlation and twin heritability of observed traits.
Figure S9. Multi-GPS correlation and twin heritability results for single-trait cross-age composites and single-trait cross-rater composites as compared to the grand mean multi-GPS correlation and twin heritability of observed traits.

Acknowledgements

The authors gratefully acknowledge the on-going contribution of the participants in the TEDS and their families. TEDS is supported by a programme grant to R.P. from the UK Medical Research Council (MR/M021475/1 and previously G0901245), with additional support from the US National Institutes of Health (AG046938) and the European Commission (602768; 295366). The authors have declared that they have no competing or potential conflicts of interest.

Correspondence

Agnieszka Gidziela, School of Biological and Chemical Sciences, Queen Mary University of London, London, UK; Email: a.bubel@qmul.ac.uk

Key points

- Genome-wide polygenic scores (GPS) can be used to predict behaviour problems in childhood, but the effect sizes are generally less than 3.5%.
- DNA-based prediction models achieve greater accuracy if aggregation approaches are employed, that is cross-trait, longitudinal and trans-situational approaches.
- The prediction of childhood behaviour problems can be improved by using multiple GPS to predict composites that aggregate behaviour problems across ages and across raters.
- Our results yield weights that can be applied to GPS in any study to create
- Multi-trait GPS predictors of behaviour problems based on cross-age and cross-rater composites.
- As compared to individuals in the lowest multi-trait GPS decile, nearly three times as many individuals in the highest internalising multi-trait GPS decile were diagnosed with anxiety disorder, and 25% more individuals in the highest general behaviour problems and externalising multi-trait GPS deciles have taken medication for mental health.

References

Akingbuwa, W.A., Hammerschlag, A.R., Jami, E.S., Allegrini, A.G., Karhunen, V., Sallis, H., … & Hagenbeek, F.A. (2020). Genetic associations between childhood psychopathology and adult depression and associated traits in 42 998 individuals: A meta-analysis. *JAMA Psychiatry*, *77*, 715–728.

Allegrini, A.G., Cheesman, R., Rimfeld, K., Selzam, S., Pingault, J.B., Eley, T.C., & Plomin, R. (2020). The p factor: genetic analyses support a general dimension of psychopathology in childhood and adolescence. *Journal of Child Psychology and Psychiatry*, *61*, 30–39.

Allegrini, A.G., Karhunen, V., Coleman, J.R.I., Selzam, S., Rimfeld, K., von Stumm, S., … & Plomin, R. (2020). Multivariable GE interplay in the prediction of educational achievement. *PLoS Genetics*, *16*, e1009153.

Allegrini A. G., Selzam S., Rimfeld K., von Stumm S., Pingault J. B., & Plomin R. (2019). Genomic prediction of cognitive traits in childhood and adolescence. *Molecular Psychiatry*, *24*, 819–827.

Bartels, M., Boomsma, D.I., Hudziak, J.J., Rietveld, M.J., van Beijsterveldt, T.C., & van den Oord, E.J. (2004). Disentangling genetic, environmental, and rater effects on internalizing and externalizing problem behavior in 10-year-old twins. *Twin Research and Human Genetics*, *7*, 162–175.

Behar, L.B. (1977). The preschool behavior Questionnaire. *Journal of Abnormal Child Psychology*, *5*, 265–275.

Caspi, A., Houts, R.M., Belsky, D.W., Goldman-Mellor, S.J., Harrington, H.L., Israel, S., … & Moffitt, T.E. (2014). The p factor: One general psychopathology factor in the structure of psychiatric disorders? *Clinical Psychological Science*, *2*, 119–137.

Cheesman, R., Purves, K.L., Pingault, J.-B., Breen, G., Rijsdij k, F., Plomin, R., & Eley, T.C. (2018). Extracting stability increases the SNP heritability of emotional problems in young people. *Translational Psychiatry*, *8*, 1–9.

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*, 1–9.

Collins, F.S. (2010). *The language of life: DNA and the revolution in personalized medicine*. New York: Harper Collins.

Demontis, D., Walters, R.K., Martin, J., Mattheisen, M., Als, T.D., Agerbo, E., … & Neale, B.M. (2019). Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nature Genetics*, *51*, 63–75.

Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. *PLoS Genetics*, *9*, e1003348.

Goodman, R. (1997). The strengths and difficulties Questionnaire: A research note. *Journal of Child Psychology and Psychiatry*, *38*, 581–586.

Grotzinger, A.D., Rhemtulla, M., de Vlaming, R., Ritchie, S.J., Mallard, T.T., Hill, W.D., … & Tucker-Drob, E.M. (2019). Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nature Human Behaviour*, *3*, 513–525.

Grove, J., Ripke, S., Als, T.D., Mattheisen, M., Walters, R.K., Won, H., … & Børglum, A.D. (2019). Identification of common genetic risk variants for autism spectrum disorder. *Nature Genetics*, *51*, 431–444.

Hulley, S., Cummings, S., Browner, W., Grady, D., & Newman, T. (2013). *Designing clinical research* (vol. *4*). Philadelphia, PA: Lippincott Williams & Wilkins, Wolters Kluwer.

Jami, E.S., Hammerschlag, A.R., Ip, H.F., Allegrini, A.G., Benyamin, B., Border, R., … & Middeldorp, C.M. (2020). Genome-wide association meta-analysis of childhood and adolescent internalising symptoms. *medRxiv*, 2020.2009.2011.20175026.

Krapohl, E., Patel, H., Newhouse, S., Curtis, C.J., von Stumm, S., Dale, P.S., … & Plomin, R. (2018). Multi-polygenic score approach to trait prediction. *Molecular Psychiatry*, *23*, 1368–1374.

Lee, J.J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., … & Fontana, M.A. (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature Genetics*, *50*, 1112–1121.

Linnér, R.K., Biroli, P., Kong, E., Meddens, S.F.W., Wedow, R., Fontana, M.A., … & Nivard, M.G. (2019). Genome-wide association analyses of risk tolerance and risky behaviours in over 1 million individuals identify hundreds of loci and shared genetic influences. *Nature Genetics*, *51*, 245–257.

Luciano, M., Hagenaars, S.P., Davies, G., Hill, W.D., Clarke, T.-K., Shirali, M., … & Deary, I.J. (2018). Association analysis in over 329,000 individuals identifies 116 independent variants influencing neuroticism. *Nature Genetics*, *50*, 6–11.

Martin, A.R., Daly, M.J., Robinson, E.B., Hyman, S.E., & Neale, B.M. (2019). Predicting polygenic risk of psychiatric disorders. *Biological Psychiatry*, *86*, 97–109.

Metzl, J. (2019). *Hacking Darwin: Genetic engineering and the future of humanity*. Sourcebooks.

Neale, M.C., Hunter, M.D., Pritikin, J.N., Zahery, M., Brick, T.R., Kirkpatrick, R.M., … & Boker, S.M. (2016). OpenMx 2.0: Extended structural equation and statistical modeling. *Psychometrika*, *81*, 535–549.

Nivard, M.G., Dolan, C.V., Kendler, K.S., Kan, K.-J., Willemsen, G., van Beijsterveldt, C.E.M., … & Boomsma, D.I. (2015). Stability in symptoms of anxiety and depression as a function of genotype and environment: a longitudinal twin study from ages 3 to 63 years. *Psychological Medicine*, *45*, 1039–1049.

Okbay, A., Baselmans, B.M.L., De Neve, J.-E., Turley, P., Nivard, M.G., Fontana, M.A., … & Cesarini, D. (2016). Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nature Genetics*, *48*, 624–633.

Pain, O., Glanville, K.P., Hagenaars, S.P., Selzam, S., Fürtjes, A.E., Gaspar, H.A., … & Lewis, C.M. (2021). Evaluation of polygenic prediction methodology within a reference-standardized framework. *PLoS Genetics*, *17*, e1009021.

Pardiñas, A.F., Holmans, P., Pocklington, A.J., Escott-Price, V., Ripke, S., Carrera, N., … & Walters, J.T.R. (2018). Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nature Genetics*, *50*, 381–389.

Peterson, R.E., Kuchenbaecker, K., Walters, R.K., Chen, C.-Y., Popejoy, A.B., Periyasamy, S., … & Duncan, L.E. (2019). Genome-wide association studies in ancestrally diverse populations: Opportunities, methods, pitfalls, and recommendations. *Cell*, *179*, 589–603.

Plomin, R. (2019). *Blueprint: How DNA makes us who we are*. London, UK: Allen Lane (Penguin Press).

R Core Team (2021). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. https://www.R-project.org/

Rimfeld, K., Malanchini, M., Spargo, T., Spickernell, G., Selzam, S., McMillan, A., … & Plomin, R. (2019). Twins Early Development Study: A genetically sensitive investigation into behavioural and cognitive development from infancy to emerging adulthood. *Twin Research and Human Genetics*, *22*, 508–513.

Ronald, A., de Bode, N., & Polderman, T.J.C. (2021). Systematic review: How the attention-deficit/hyperactivity disorder polygenic risk score adds to our understanding of ADHD and associated traits. *Journal of the American Academy of Child & Adolescent Psychiatry*. S0890-8567(21)00070-8. https://doi.org/10.1016/j.jaac.2021.01.019. Epub ahead of print.

Savage, J.E., Jansen, P.R., Stringer, S., Watanabe, K., Bryois, J., de Leeuw, C.A., … & Posthuma, D. (2018). Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nature Genetics*, *50*, 912–919.

Selzam, S., McAdams, T.A., Coleman, J.R., Carnell, S., O'Reilly, P.F., Plomin, R., & Llewellyn, C.H. (2018). Evidence for gene-environment correlation in child feeding: Links between common genetic variation for BMI in children and parental feeding practices. *PLoS Genetics*, *14*, e1007757.

Stahl, E.A., Breen, G., Forstner, A.J., McQuillin, A., Ripke, S., Trubetskoy, V., … & Sklar, P. (2019). Genome-wide association study identifies 30 loci associated with bipolar disorder. *Nature Genetics*, *51*, 793–803.

Vilhjálmsson, B., Yang, J., Finucane, H., Gusev, A., Lindström, S., Ripke, S., … & Zheng, W. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The American Journal of Human Genetics*, *97*, 576–592.

Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., & Yang, J. (2017). 10 years of GWAS discovery: biology, function, and translation. *The American Journal of Human Genetics*, *101*, 5–22.

Von Stumm, S., & Plomin, R. (2021). Using DNA to predict intelligence. *Intelligence*, *86*, 101530.

Wainschtein, P., Jain, D.P., Yengo, L., Zheng, Z., TOPMed Anthropometry Working Group, Trans-Omics for Precision Medicine Consortium, … & Visscher, P.M. (2019). Recovery of trait heritability from whole genome sequence data [preprint]. *Biorxiv*.

Wray, N.R., Ripke, S., Mattheisen, M., Trzaskowski, M., Byrne, E.M., Abdellaoui, A., … & Sullivan, P.F. (2018). Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nature Genetics*, *50*, 668–681.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*, 301–320.

Accepted for publication: 13 July 2021