Genetically informed analyses of childhood individual differences

Sofieke T. Kevenaar

Genetically informed analyses of childhood individual differences

Sofieke T. Kevenaar

Genetically informed analyses of childhood individual differences

Sofieke T. Kevenaar

Paranymphs
Wonu Akingbuwa
Zenab Tamimy

Cover & lay-out design
Maaike Disco, proefschriftopmaak.nl
Printed by
Ridderprint
ISBN
978-94-6483-386-7
DOI
http://doi.org/10.5463/thesis.358

VRIJE UNIVERSITEIT

Genetically informed analyses of childhood individual differences

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor of Philosophy aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. J.J.G. Geurts,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Gedrags- en Bewegingswetenschappen
op woensdag 18 oktober 2023 om 13.45 uur
in een bijeenkomst van de universiteit,
De Boelelaan 1105

door

Sofieke Thérèse Kevenaar

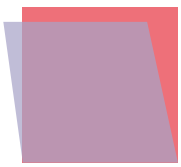
geboren te Vlijmen

promotoren:	prof.dr. D.I. Boomsma prof.dr. C.V. Dolan
copromotoren:	dr. E. van Bergen prof.dr. A.J. Oldehinkel
promotiecommissie:	prof.dr. M.E.J. Raijmakers dr. M. Achterberg prof.dr. P.F. de Jong dr. S.M. van den Berg dr. C.M. van der Laan prof.dr. T. Kretschmer

Table of contents

Chapter 1

General introduction



Introduction

Children differ from each other in many ways. Some children are great at sports, others are very creative, some are tall, and some are great leaders. In this thesis, I tried to explain individual differences in height, cognitive, and non-cognitive variables among school-aged children. During childhood, children become more independent of their home environment, with their school environment becoming an important part of their lives, and we should consider their behavior across multiple environments. For children to thrive, they have to develop physically and psychologically. They need to acquire a variety of skills including new motor skills, they must develop cognitively (e.g., learning to read, write and do arithmetic), non-cognitively (e.g., learning to inhibit disruptive impulses in a classroom), and socially (e.g., making and maintaining friendships).

In my thesis, I studied individual differences in physical characteristics (height), cognitive (school related) traits, and non-cognitive traits (grit, self-control). The goal of this thesis is to increase our insight into the extent to which children differ from each other, and to increase our understanding of why they differ. I applied various methods to achieve this goal: I applied the multivariate classical twin model to estimate heritability and genetic correlations, I combined the classical twin design (CTD) with multilevel modeling, and I used a Bayesian approach to combine evidence for informant differences in reporting on children's behavior across different cohorts.

The phenotypes that I focused on in this dissertation are height, self-control, grit, and school performance. Self-control is the ability to alter dominant responses to adhere to social values and moral norms (Baumeister, Vohs & Tice, 2007), grit is perseverance and passion for long term goals (Duckworth, Peterson, Matthews & Kelly, 2007), and school performance is how well children perform in school according to their grades as reported by their teachers.

Twins as a means to disentangle sources of variation

Phenotypic individual differences in children are influenced by their genotype, which comprises many genetic variants, and their environment, which comprises many different aspects, such as the school environment, parental rearing style, household and neighborhood environments, and other persons' behaviors. During their development, children may learn from interactions with their parents, siblings, their extended family members, their teachers, and their peers. When disentangling sources of variation in children, it is important to note that parents

not only shape the rearing environment of their children, but also provide their children with their genetic material. To obtain a good understanding of why children differ from each other, we need to separate genetic sources of variation from environmental sources of variation.

It is important to do this, because the phenotypic correlation between a given aspect of the home or rearing environment (e.g., "number of children's books in the home") and outcomes in children ("reading ability") is hard to interpret causally. The correlation may be due to the causal effect of the environment on the children's outcomes, or may be due to the genetic correlation between parent and their children (Hart, Little & Van Bergen, 2021). That is, genes that predispose parents to enjoy reading and to attach importance to their offspring's reading ability are transmitted to their offspring. In the offspring the genes may influence the reading ability.

Family designs offer the means to decompose sources of phenotypic individual differences into genetic and environmental components. An important and highly fruitful design is the classical twin design (Boomsma et al., 2002; Polderman, et al. 2015). This design exploits the phenotypic resemblance of monozygotic (MZ) twins, who are genetically (nearly) identical, and dizygotic (DZ) twins, who are genetically as similar as non-twin siblings, i.e., they share on average half of their alleles. DZ and MZ twins differ in genetic resemblance, but otherwise, they share many environmental influences stemming from the home environment (prenatal conditions, the rearing home), and other environments (e.g., classroom and teachers in childhood, geographical region). The classical twin design has been highly productive in advancing our understanding of variation in human phenotypes.

The classical twin design involves a number of univariate and multivariate statistical models, including models for moderation, gene-environment interaction, and phenotypic causality. Typically, these models are fitted to twin data by means of Structural Equation Modeling (SEM), based on path analysis as initially developed in genetics by Wright (1918; 1934). SEM enables us to estimate the parameters (path coefficients, variances, covariances) in a model, test hypotheses about the parameters, and to evaluate overall goodness of fit of the model. In the classical twin model, the parameters of interest are the contributions of the genetic and environmental latent variables to the variance of the measured phenotype. On the genetic side, we distinguish between additive genetic (A) and dominant genetic (D) influences, as sources of variance. Additive genetic variance is due to the additive effect of alleles at the relevant loci. Dominance variance is due to (intra-

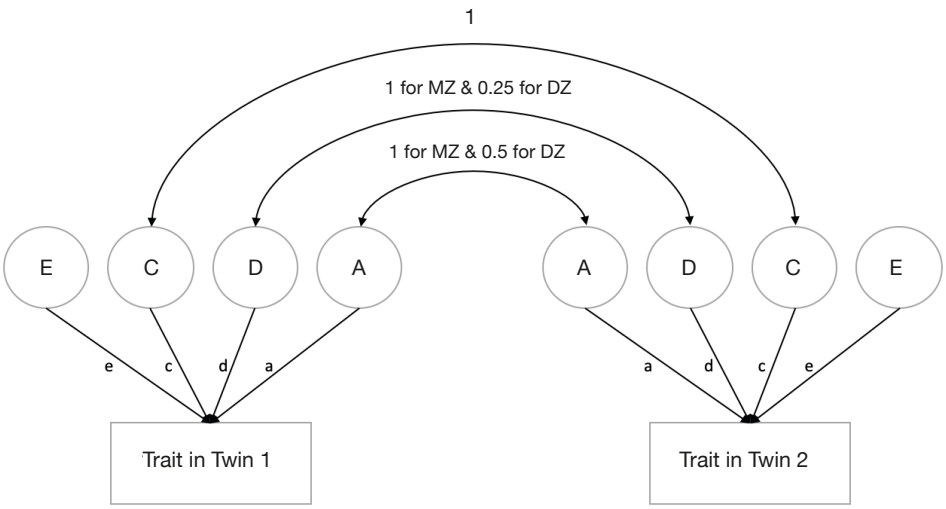


Figure 1. Path diagram of an ACDE classical twin model.

Figure 1. Path diagram of an ACDE classical twin model. The latent variables A, C, D, and E are represented with circles, and the measured phenotypes in rectangles. The model as shown, with four sources of phenotypic variance, is not identified in univariate twin data. The parameters a, d, c and e are path coefficients. These coefficients squared are the variance components, assuming the latent factors are assigned unit variance (which is a standard scaling convention in SEM). Double-headed arrows indicate correlational relations. Single headed arrows represent regression relations.

locus) interactions between the alleles at the relevant loci. On the environmental side, we distinguish common (shared) environment (C) and unique (unshared) environment (E) influences, as sources of variance. Common environmental variance is due to environmental factors that contribute to the phenotypic resemblance of the twins. Unique environmental variance is due to environmental factors that are not shared by the twins. These factors can include influences of friends and school, but also the interpretation and personal experience of the home environment. Usually, the unique environment variance, as estimated in the twin design, includes measurement error. In multivariate twin modeling, it is possible to distinguish between unique environmental variance and measurement error variance.

While we distinguish four sources of phenotypic variance (i.e., A, D, C, E), we can only estimate three using the classical twin model, because this design does not furnish sufficient information to estimate all four (technically, the model with all four is not identified). In practice, a decision between the ACE or ADE model is

made on the basis of the observed MZ and DZ twin correlations (Falconer, 1965). If the MZ twin correlation is greater than twice the DZ correlation, this suggests the presence of dominant genetic effects, and the ADE model is chosen. If the MZ twin correlation is smaller than twice the DZ twin correlation, this suggests the presence of common environment, and the ACE model is chosen.

Figure 1 represents a path diagram of the classical twin model. The observed variables (the measured phenotypic trait in twin 1 and twin 2 of a pair) are depicted in boxes, and the unobserved, latent variables (A, C, D, and E) are depicted in circles. The single-headed arrows represent a unidirectional relationship, which can be interpreted as a regression relationship. So, in Figure 1, we see that the traits are regressed on the latent variables A, C, D, and E. The associated path coefficients, a, c, d, and e, are interpretable as regression coefficients, but are also referred to as loadings. Correlations are represented by double-headed arrows. The correlation between the two latent A factors is fixed to be 1 for MZ twins (because they are genetically identical) and to 0.5 for DZ twins (because on average they share half of their alleles). Dominance effects at a given locus are completely shared by relatives, if they inherited two identical alleles at the locus. This applies to all loci in MZ twins, and to 25% of DZ twins and full siblings. The common environment is shared by the twins (regardless of zygosity), and so this factor contributes 100% to the phenotypic covariance. This is represented by the correlation of 1 between the C factors for both zygosities. Finally, the unique environment is not shared by the twins, so the correlation between the E factors of the twins is zero. As mentioned above, the full ACDE twin model is not identified given univariate data (i.e., a single measured phenotype). Other genetically informative designs, such as the parents and twins design, enable estimation of all four variance components. Sometimes, common environment is indexed by a known or measured exposure that is identical for children from the same family and same age, such as the area in a country where they live, or the socio-economic status of the family. Such exposures create higher-level regional clustering in (childhood) twin data that can be analyzed by combining the classical twin design with multilevel models.

The classical twin design has been highly productive in elucidating the sources of individual differences in many phenotypes (Polderman et al., 2015). Of the phenotypes that feature in this thesis, adult height has been studied most in twin and family design, but also more recently in (ongoing) genome wide studies, which focus on the association between measured genetic variants (single nucleotide polymorphisms) and height. During childhood, the relative contribution of genetic

factors compared to environmental factors explaining variance in height increases with age. Heritability estimates for prenatal length increase from second trimester to birth from 13% to 27%. At 36 months, estimates are 60%, or somewhat higher, with the heritability remaining approximately 60% in childhood (Estourgie-van Burk et al., 2006; Mook et al. 2012; Jelenkovic et al., 2016; Silventoinen et al., 2007).

Measures of reading, writing, and arithmetic ability are subject to relatively large genetic influences in the Netherlands, with heritability estimates as high as .75. (de Zeeuw et al., 2016; Krahpohl et al., 2014). In a large meta-analysis, self-control was estimated to have a heritability of around 60% (Willems et al., 2019). The heritability of grit seems lower, i.e., around 40% (Martinez, et al., 2022; Rimfeld et al., 2016; Tucker-Drob et al., 2016). These results offer a firm basis for follow-up questions about the individual differences in these phenotypes. These questions concern the development of traits across age, from childhood through adolescence and the interrelationship of these phenotypes. For example, with respect to the interrelationship, we can ask whether an association arises because the same genes influence school performance and self-control, because the same environmental influences are important, or because of a direct causal effect of one phenotype (e.g., self-control) on another (e.g., school performance).

Measuring phenotypes in children

Children often do not, or cannot, report on their own behavior, so phenotypic measurement is often based on ratings by a person who knows the child well, e.g., a parent or a teacher. When investigating individual differences in children, it is important to consider how the trait of interest is measured. For some phenotypes, such as height, measurement is relatively straightforward. E.g., we can use a measuring tape as an instrument to determine height. When heights are reported by parents, the reliability of the measure is good (r. = 0.96, Estourgie-van Burk et al., 2006). Outcomes like school performance can be measured by standardized tests, which have the benefit that they are comparable across the Dutch population. In the Netherlands, most children take the same standardized tests, so that the scores, being on the same scale, are commensurate. While standardized tests have many advantages, a disadvantage of a single standardized test can be that it is administered at one specific moment in time, and with the right means, one can be trained to perform well on these standardized tests. Other possible measures of school performance are based on teacher evaluations of the children's school grades. Unlike in standardized tests, these evaluations

may be subject to a rater effect, originating in the teacher's perspective on the child. Like standardized tests, teacher reports of grades can measure school performance across multiple points in time, but teacher reports are likely to better reflect the performance in a real-life classroom setting, and may be less subject to test-specific training than standardized test results.

Behavioral traits like self-control and grit can be measured by survey items, by tasks, or by direct observations in a laboratory or a natural setting. These approaches are considered to reflect different aspects of self-regulation: trait self-control versus state self-control. While direct observations reflect state self-control, i.e., the self-control displayed during a given task, questionnaire items reflect trait self-control. Trait self-control is the tendency to show self-control, and is considered a stable trait (Inzlicht, Werner, Briskin & Roberts, 2021, Malanchini et al., 2019). In this thesis, I focus on this trait aspect of self-control.

With all measures, one has to be aware of the fact that children develop over time. We have to use age-appropriate measures, and compare children of the same age, who are assumed to be in the same developmental phase. Furthermore, with survey data, information about children can come from different informants, such as parents and teachers, but also children themselves. When parents rate their children, the parents of twins typically rate both children. A teacher, however, may rate both twins, if they are in the same class (and so have the same teacher), or single twins, if the twins are in a different class (Webbink et al. 2007; de Zeeuw et al. 2015). These all are important factors to take into consideration when measuring individual differences in children.

Netherlands Twin Register (NTR)
and Consortium of Individual Development (CID) cohorts

I analyzed data from different Dutch population cohorts, which are all included in the Consortium of Individual Development (CID), funded by the NWO (Dutch Research Council) "Gravitation" funds. This dissertation would not have been possible without the Netherlands Twin Register (NTR) and its participants. Since 1986, the NTR has collected survey data on young twins, their parents, and their siblings (Ligthart et al., 2019). The NTR also collects data on adult twins and their relatives. Subgroups of all ages take part in biomaterial data (e.g., DNA) collection, and in dedicated projects, e.g., to collect specific phenotypes, such as IQ and MRI measures. For this dissertation, survey data of mothers, fathers, and teachers, and also of children themselves were collected. The data concern the behavior of 6- to 16-year-old children (Van Beijsterveldt et al., 2013). The DNA data, available in twins who had participated in DNA collection, were a

valuable resource to determine zygosity in same-sex twin pairs. These DNA data were also used to assess the validity of zygosity assessment by questionnaire items. The other Dutch population cohorts from the Consortium of Individual Development (CID) that contributed data to my thesis are TRAILS, Generation R and YOUth. All these cohorts contain large datasets, with measurements on thousands of children and their family members. These resources provided us with the means to make use of a variety of data sources to analyze individual differences in children. TRAILS (Tracking Adolescents' Individual Lives Survey) is based in the north of the Netherlands. TRAILS follows adolescents and young adults from age 11 with two- or three-year intervals, and makes use of surveys, interviews, tests and physical measurements (Oldehinkel et al., 2015). Generation R (Gen R) is a Rotterdam-based study, in which children are assessed prenatally into young adulthood by means of questionnaires, interviews, observations, ultrasound and physical examinations, MRI and biological sampling (Kooijman et al., 2016). YOUth is a Utrecht-based study that follows babies, children from early youth into adolescents. YOUth focusses on neurocognitive development of social competence and behavioral control (Onland-Moret et al., 2020).

Methodological approaches to explain individual differences

Here, I focus on individual differences as we encounter them in our samples, and, by generalization, in the population. Our aim is to describe and explain individual differences. In this dissertation, I explored several methods to investigate individual differences in children, and to address the challenges and opportunities that researchers encounter to answer questions about individual differences, while working with data obtained from surveys.

When investigating individual differences in a particular population, one factor that one needs to consider is clustering. Individuals are often clustered in different subgroups of the population. For example, children are clustered in families, and these families are often clustered in schools, which are in turn clustered in regions. In **chapter 2** of this thesis, I investigated regional clustering in children's height in the Netherlands using multilevel structural equation modeling. With the approach, which accounts for clustering of children in families, and clustering of families in geographical regions, I quantified how much of the variance in children's height can be explained by geographical region. Also, I investigated the effect of ignoring the clustering has on the genetic and environmental variance components in the model. As regional clustering might in fact reflect ancestry, we investigated the effect of regional clustering after correcting for genetic principal

components. If regional clustering explains no variance additional to variance explained by the genetic principal components, this suggests that the regional differences in height reflects genetic ancestry.

In **chapter 3,** I investigated whether different informants rated children's behavior differently. I did this by combining several datasets using Bayesian evidence synthesis. Different datasets may concern the same phenotypes, but may differ in the specifics, such as the measurement instrument used, the target population (age, geographical location), and the informants (raters), who actually provide the data. Combining information from different datasets enables us to arrive at robust conclusions efficiently, which are less dependent the specifics that may characterize individual datasets. Bayesian evidence synthesis is one approach to combine results, and to aggregate support for competing hypotheses (Kuiper et al., 2012) across different datasets. In chapter 3, I applied Bayesian evidence synthesis to investigate if there are differences between mothers, fathers, teachers, and children in their reports on primary school-aged children's self-control. Parents and teachers see children in different contexts, and have difference reference groups, so they might interpret children's behavior relating to self-control differently. The application of Bayesian evidence synthesis that I applied is unique, because I employed it in a situation, where each of the datasets contributed both common and unique information to test competing hypotheses. Specifically, four different cohorts (NTR, TRAILS, GenR, and YOUth) participated in this study, which each had collected self-control data from multiple informants, but no cohort had information of all informants on all ages. I analyzed data from different informants, at different ages, allowing for different missing data patterns. Bayesian evidence synthesis enabled me to combine evidence from different datasets, which each tested partial hypothesis, and together allowed me to quantify the support for competing hypotheses.

In Chapter 4, I investigated the differential prediction of grit and self-control to explain individual differences in school performance, where I distinguish between genetic and environmental sources of variation making use of genetically informed regression analysis (Boomsma et al., 2021). This method enables the prediction of a dependent trait by multiple correlated predictors by simultaneously fitting a genetic covariance structure model and a regression model to multivariate twin data. Here, I investigated the degree to which variation in school performance was explained by grit and self-control, and I determined the degree to which the explained variance was attributable to genetic and environmental influences. All variables were based on teacher ratings of the children. An important factor

to account for in this chapter was teacher sharing, because some twins were in the same class with the same teacher, while others were in different classes, and consequently were rated by different teachers. To account for the possibility that children rated by the same teacher are judged as being more similar due to rater or other effects, I added a teacher-sharing variance component to the structural equation model to account for teacher sharing. I also considered the effects of censoring in this chapter. Especially the self-control scale shows ceiling effects, stemming from the fact that many children do not experience appreciable problems with self-control.

Chapter 5 follows up on the results in chapter 4 by addressing the question whether grit and self-control are causally related to school performance. Here, I study the direct phenotypic linear relationship between grit and school performance in a bivariate ADE model. This allowed me to study the direct (causal) relationship, while taking into account confounding due to additive genetic effects (pleiotropy: the same genes affecting self-control, grit and school performance), and for confounding due to environmental influences. By confounding, I mean background genetic and environmental influences common to the phenotypes. This chapter addresses the question how much of the association between self-control, grit, and school performance can be attributed to self-control and grit having a direct causal effect on school performance, and what part of the association arises from background genetic and environmental correlations.

Furthermore, I also contributed to the continuation of the longitudinal data collection in the Netherlands Twin Register during my PhD. For all twin analyses, it is essential to know the accurate zygosity of the twin pairs in the data, because the entire twin model is based on the differences between monozygotic and dizygotic twin groups. In the Netherlands Twin Register, the zygosity determination of same sex twin pairs is based on either a resemblance questionnaire or on zygosity determination using DNA or blood tests. **Appendix 1** gives the results of an empirical study in children and adults to assess how well we can determine zygosity on the basis of questionnaire items concerning the similarity of twins. To determine the agreement between zygosity determination by these survey items on twin resemblance and actual zygosity determination based on DNA or blood test, I performed discriminant analyses on the survey items.

This thesis stresses the importance of accounting for the informant, who provides the phenotypic data. To elucidate this aspect of the data collection, **Appendix 2** provides an overview of procedures of the ongoing data collection in teachers

of twins in the Netherlands Twin Register, for which I was responsible during my PhD trajectory. **Appendix 3** contains the emails sent to parents and teachers regarding the data collection in teachers.

Literature

Baumeister, R. F., Vohs, K. D., & Tice, D. M. (2007). The strength model of self-control. *Current directions in psychological science*, *16*(6), 351-355.

Boomsma D, Busjahn A, Peltonen L. Classical twin studies and beyond. Nat Rev Genet. 2002; 3(11):872-82. doi: 10.1038/nrg932

de Zeeuw, E. L., van Beijsterveldt, C. E., Lubke, G. H., Glasner, T. J., & Boomsma, D. I. (2015). Childhood ODD and ADHD behavior: The effect of classroom sharing, gender, teacher gender and their interactions. *Behavior Genetics*, *45*, 394-408.

Boomsma, D. I., Van Beijsterveldt, T. C., Odintsova, V. V., Neale, M. C., & Dolan, C. V. (2021). Genetically informed regression analysis: application to aggression prediction by inattention and hyperactivity in children and adults. *Behavior genetics*, *51*(3), 250-263

de Zeeuw, E. L., van Beijsterveldt, C. E., Glasner, T. J., de Geus, E. J., & Boomsma, D. I. (2016). Arithmetic, reading and writing performance has a strong genetic component: A study in primary school children. *Learning and Individual Differences*, *47*, 156-166.

Derks, E. M., Dolan, C. V., & Boomsma, D. I. (2006). A test of the equal environment assumption (EEA) in multivariate twin studies. *Twin Research and Human Genetics*, *9*(3), 403-411.

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: perseverance and passion for long-term goals. *Journal of personality and social psychology*, *92*(6), 1087.

Estourgie-van Burk, G. F., Bartels, M., Van Beijsterveldt, T. C., Delemarre-van de Waal, H. A., & Boomsma, D. I. (2006). Body size in five-year-old twins: heritability and comparison to singleton standards. *Twin Research and Human Genetics*, *9*(5), 646-655.

Falconer, D. S. (1965). The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annals of human genetics*, *29*(1), 51-76.

Hart, S.A., Little, C. & van Bergen, E. Nurture might be nature: cautionary tales and proposed solutions. *npj Sci. Learn.* **6**, 2 (2021). https://doi.org/10.1038/s41539-020-00079-z

Inzlicht, M., Werner, K. M., Briskin, J. L., & Roberts, B. W. (2021). Integrating models of self-regulation. *Annual review of psychology*, *72*, 319-345.

Jelenkovic, A., Sund, R., Hur, Y. M., Yokoyama, Y., Hjelmborg, J. V. B., Möller, S., ... & Silventoinen, K. (2016). Genetic and environmental influences on height from infancy to early adulthood: An individual-based pooled analysis of 45 twin cohorts. *Scientific reports*, *6*(1), 1-13.

Krapohl, E., Rimfeld, K., Shakeshaft, N. G., Trzaskowski, M., McMillan, A., Pingault, J. B., ... & Plomin, R. (2014). The high heritability of educational achievement reflects many genetically influenced traits, not just intelligence. *Proceedings of the national academy of sciences*, *111*(42), 15273-15278.

Kuiper, R. M., Buskens, V., Raub, W., & Hoijtink, H. (2013). Combining statistical evidence from several studies: A method using Bayesian updating and an example from research on trust problems in social and economic exchange. *Sociological Methods & Research*, *42*(1), 60-81.

Kooijman, M. N., Kruithof, C. J., van Duijn, C. M., Duijts, L., Franco, O. H., van IJzendoorn, M. H., ... & Jaddoe, V. W. (2016). The Generation R Study: design and cohort update 2017. *European journal of epidemiology*, *31*(12), 1243-1264.

Ligthart, L., van Beijsterveldt, C. E., Kevenaar, S. T., de Zeeuw, E., van Bergen, E., Bruins, S., ... & Boomsma, D. I. (2019). The Netherlands Twin Register: longitudinal research based on twin and twin-family designs. *Twin Research and Human Genetics*, *22*(6), 623-636.

Malanchini, M., Engelhardt, L. E., Grotzinger, A. D., Harden, K. P., & Tucker-Drob, E. M. (2019). "Same but different": Associations between multiple aspects of self-regulation, cognition, and academic abilities. *Journal of Personality and Social Psychology*, *117*(6), 1164.

Martinez, K. M., Holden, L. R., Hart, S. A., & Taylor, J. (2022). Examining mindset and grit in concurrent and future reading comprehension: A twin study. *Developmental Psychology.* Advance online publication.

Mook-Kanamori, D. O., Van Beijsterveldt, C. E., Steegers, E. A., Aulchenko, Y. S., Raat, H., Hofman, A., ... & Jaddoe, V. W. (2012). Heritability estimates of body size in fetal life and early childhood. *PLoS One*, *7*(7), e39901.

Oldehinkel, A. J., Rosmalen, J. G., Buitelaar, J. K., Hoek, H. W., Ormel, J., Raven, D., ... & Hartman, C. A. (2015). Cohort profile update: the tracking adolescents' individual lives survey (TRAILS). *International Journal of Epidemiology*, *44*(1), 76-76n.

Plomin R, Willerman L, Loehlin JC. Resemblance in appearance and the equal environments assumption in twin studies of personality traits. Behav Genet. 1976 Jan;6(1):43-52. doi: 10.1007/BF01065677

Polderman, T. J., 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), 702-709.

Onland-Moret, N. C., Buizer-Voskamp, J. E., Albers, M. E., Brouwer, R. M., Buimer, E. E., Hessels, R. S., ... & Kemner, C. (2020). The YOUth study: Rationale, design, and study procedures. *Developmental cognitive neuroscience*, *46*, 100868.

Rimfeld, K., Kovas, Y., Dale, P. S., & Plomin, R. (2016). True grit and genetics: Predicting academic achievement from personality. *Journal of personality and social psychology*, *111*(5), 780.

Silventoinen, K., Bartels, M., Posthuma, D., Estourgie-van Burk, G. F., Willemsen, G., van Beijsterveldt, T. C., & Boomsma, D. I. (2007). Genetic regulation of growth in height and weight from 3 to 12 years of age: a longitudinal study of Dutch twin children. *Twin Research and Human Genetics*, *10*(2), 354-363.

Tucker-Drob, E. M., Briley, D. A., Engelhardt, L. E., Mann, F. D., & Harden, K. P. (2016). Genetically-mediated associations between measures of childhood character and academic achievement. *Journal of Personality and Social Psychology, 111*(5), 790–815.

Van Beijsterveldt, C. E., Groen-Blokhuis, M., Hottenga, J. J., Franić, S., Hudziak, J. J., Lamb, D., ... & Boomsma, D. I. (2013). The Young Netherlands Twin Register (YNTR): longitudinal twin and family studies in over 70,000 children. *Twin Research and Human Genetics*, *16*(1), 252-267.

Webbink, D., Hay, D., & Visscher, P. M. (2007). Does sharing the same class in school improve cognitive abilities of twins?. *Twin Research and Human Genetics*, *10*(4), 573-580.

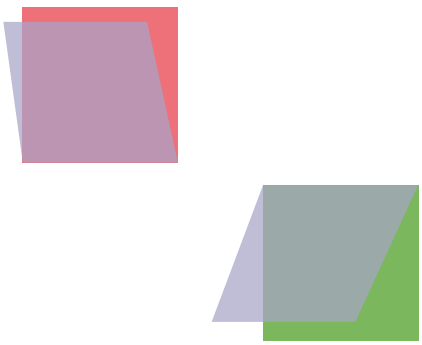
Willems, Y. E., Boesen, N., Li, J., Finkenauer, C., & Bartels, M. (2019). The heritability of self-control: A meta-analysis. *Neuroscience & Biobehavioral Reviews*, *100*, 324-334.

Wright, S. (1918). On the Nature of Size Factors. Genetics, 3, 367-74.

Wright, S. (1934). The Method of Path Coefficients. Annals of Mathematical Statistics, 5, 161-215.

Chapter 2

Multilevel twin models: geographical region as a third level variable



Published as Tamimy, Z., **Kevenaar*, S. T.**, Hottenga, J. J., Hunter, M. D., de Zeeuw, E. L., Neale, M. C., van Beijsterveldt, C. E. M., Dolan, C. V., van Bergen, E., & Boomsma, D. I. (2021). Multilevel twin models: geographical region as a third level variable. Behavior Genetics, 51(3), 319-330.*

Abstract

The classical twin model can be reparametrized as an equivalent multilevel model. The multilevel parameterization has underexplored advantages, such as the possibility to include higher-level clustering variables in which lower levels are nested. When this higher-level clustering is not modeled, its variance is captured by the common environmental variance component. In this paper we illustrate the application of a 3-level multilevel model to twin data by analyzing the regional clustering of 7-year-old children's height in the Netherlands. Our findings show that 1.8%, of the phenotypic variance in children's height is attributable to regional clustering, which is 7% of the variance explained by between-family or common environmental components. Since regional clustering may represent ancestry, we also investigate the effect of region after correcting for genetic principal components, in a subsample of participants with genome-wide SNP data. After correction, region no longer explained variation in height. Our results suggest that the phenotypic variance explained by region might represent ancestry effects on height.

Key words: Multilevel Model, Classical Twin Design, OpenMx, Region, Ancestry, Height

Introduction

The classical twin model (CTM) is often approached from a structural equation modeling (SEM) framework (Bentler and Stein, 1992; Boomsma and Molenaar, 1986; Heath et al., 1989; Neale & Cardon, 1992; Rijsdijk and Sham, 2002). In this framework, it is a one-level model with family as level one sampling unit. The analysis of twin data can, however, also be approached from a multilevel model (MLM) perspective. MLMs were developed specifically for the analysis of clustered data (Goldstein, 2011; Laird and Ware, 1982; Longford, 1993; Paterson and Goldstein, 1991). Classical examples are children (level 1 units), who are clustered in classes (level 2) within schools (level 3; Sellström and Bremberg, 2006). Other examples are fMRI measures (level 1) that are clustered in individuals (level 2), who are clustered in scanner type (level 3; Chen et al., 2012), or biomarker data (level 1) that are clustered in measurement batches (level 2; Scharpf et al., 2011). The classical twin design is based on data that also have natural clustering, namely, twins are clustered within pairs. For this reason, the MLM framework can accommodate the CTM (Guo and Wang, 2002; McArdle and Prescott, 2005; Rabe-Hesketh et al., 2008; Van den Oord, 2001). Hunter (2021) provides a detailed account of the CTM in the MLM framework with example code and several extensions. While the MLM specification of the CTM is equivalent to the SEM approach, it also has some interesting, yet underexplored, advantages. In this paper we aim to elaborate on these advantages, and to provide an empirical illustration of a multilevel twin model, where we study the clustering of children's height in geographical regions in the Netherlands, and consider the role therein of genetic ancestry.

In the SEM approach to the CTM, the covariance structure of twin-pairs is modelled to decompose phenotypic variance into multiple components that represent genetic and non-genetic influences. Given the biometrical underpinning of the twin model (Eaves et al. 1978; Falconer and MacKay, 1996; Fisher, 1918), the phenotypic variance can be decomposed into additive genetic variance (A), non-additive or dominance genetic variance (D), common environmental variance (C), and unique environmental variance (E) components. Variance decomposition is based on the premise that monozygotic (MZ) twins share 100% of their DNA and dizygotic (DZ) twins share on average 50% of their segregating genes. Hence, additive and non-additive genetic variance is fully shared by MZ twins, whereas additive and non-additive variance components are shared for 50% and 25% by DZ twins. In the CTM, all influences that are not captured by segregating genetic variants are labeled as "environment". These influences can be categorized as

common environment (i.e., shared by twins from the same family) or unique or unshared environment (i.e., creating variation among members from the same family). These are also referred to as between and within family environmental influences. The full ACDE model is not identified when analyzing one phenotype per twin, and only three of the four components can be simultaneously estimated. In this SEM approach to modeling twin data, the variance decomposition is based on the bivariate data observed in twin pairs (i.e., one phenotype for twin 1, and one for twin 2, which are both level 1 units).

In the MLM framework the phenotypic variance can be decomposed into a within-pair (level 1) and a between-pair (or family; level 2) components. This requires reparameterization of the model into level 1 and level 2 variance components. Because the E component captures variance that is not shared by twins, this component is an individual level 1 variance component. The C component is by definition shared by twins, regardless of zygosity, and is a family level 2 variance component. The A component, however, is more complicated, as it is a level 2 component in MZ twin pairs, but both a level 1 and a level 2 component in DZ twin pairs. To account for this, the A-component is divided into two orthogonal components, unique additive (A_U) and common additive (A_C). Here, A_U is a first-level component representing the A variance at the individual level (within pairs or within families), while A_C is a second-level component (between pairs or between families), representing the A variance at the twin-pair level. These definitions are consistent with the classical notations in which A_C refers to within family genetic variance known as A_1 (Boomsma and Molenaar, 1986; Martin and Eaves 1977), or the average breeding value variance (Barton et al., 2017), while A_U refers to the between family genetic variance known as A_2 (Boomsma and Molenaar, 1986; Martin & Eaves 1977), or the segregating genetic variance (Barton et al., 2017). In MZs, the A_U variance component is 0, since all the variance explained by A is shared by both twins from a pair. For DZ twins, the variance of both A_C and A_U are constrained to equal 0.5, since on average 50% of the A variance is shared by the individuals and 50% of the A variance is unique for the individual.

An important, yet underexplored, advantage of the MLM approach, is the possibility to include higher-level variables in which lower-levels are nested. By including these higher-level variables, we can identify variance components which are attributable to higher-level clustering. Such clusters may be a consequence of data acquisition or design, e.g., clustering of biomarker data that are measured in batches, or clustering of brain imaging data by fMRI scanner type. They may also occur naturally, for example, families in regions, neighborhoods or schools. If the higher-level variable is not included in the variance decomposition models,

the variance that it explains will be captured as part of the C-component, since both twins, regardless of zygosity, share the higher-level variable (i.e., the twin pair is nested in the higher-level variable).

Within the SEM framework, higher-level variables can be included in the model as a fixed effect on the individual level (i.e., covariate) by means of (linear) regression. For nominal covariates (i.e., factors in the ANOVA sense), this approach requires the variable to be dummy coded, which may be impractical, for example when the number of assays for a biomarker or the number of schools that twins are enrolled in is large. In the MLM framework, however, the higher-level variable is treated as a random rather than a fixed effect, and this reduces the number of parameters to one single variance component. That is, given a factor with L categories, the fixed effects approach requires L-1 additional parameters, whereas the random effects approach requires one additional parameter (a variance component). In addition, the MLM approach is more suitable than the SEM framework in dealing with unequal group sizes (Gelman, 2005). Finally, an MLM approach allows us to evaluate the contribution of the higher-level component to the C-component, as estimated in the standard twin model. This can be achieved by comparing the C-component estimate of the two-level model (i.e., the standard twin model) to the estimate of the three-level model.

In this paper, we illustrate the use of multilevel twin models by investigating the regional clustering of children's height with twin data from the Netherlands Twin Register (Boomsma et al., 1992; Ligthart et al., 2019). Height serves as an indicator of the general development of a country, and is known to decrease in times of scarcity and increase in times of prosperity (Baten & Blum, 2014; Baten & Komlos, 1998). Also, children's height is an indicator of overall development, where height is associated with cognitive development and school achievement (Karp et al., 1992; Spears, 2012). In 7-year-old children, resemblance between family members for height is explained by additive genetic (approximately 60%) and common environmental (approximately 20%) factors (Jelenkovic et al., 2016; Silventoinen et al., 2004; Silventoinen et al., 2007).

In the Netherlands, the association between height and geographical region is well established (Abdellaoui et al. 2013), which makes this a clustering variable of interest. Inhabitants of different geographical region may display genetic and environmental differences. Location is associated with genetic differences (e.g. Abdellaoui et al., 2019) and differences in social and cultural traditions, diet, socio-economic status, and living circumstances (e.g., rural vs urban, e.g. Colodro-Conde et al. 2018). By analyzing height and geographical region data

in a three-level MLM, we can determine whether variance in children's height is associated with geographic region, and estimate the proportion of the common environmental or between-family variance that can be explained by these regional effects.

In a subsample of 7-year-old participants, we investigated the extent to which regional clustering may be due to genetic ancestry by including the first three genetic principal components (PCs; Hotelling, 1933). The genetic PCs are obtained through principal component analysis of the covariance matrix of the genotype Single Nucleotide Polymorphism (SNP) data (Reich et al., 2008). In the Netherlands, the first genetic PC is associated with a north-south height gradient (Abdellaoui et al., 2013; Boomsma et al., 2014). This gradient is likely a result of social, geographical and historical divisions between the north and the south. Southern regions were conquered by the Roman empire, adopted Catholicism, and were geographically separated from the northern regions by five large rivers in the Netherlands (Schalekamp, 2009). This first Dutch PC also shows a strong correlation with the European PC that differentiates northern from southern European populations (1000 Genomes PC4; Abdellaoui et al., 2013). The second PC is associated with the east-west division of the Netherlands. This PC may reflect differences between rural and urban environments, since the east of the Netherlands is characterized by less populous and rural areas, while the west includes the largest concentration of urban areas in the Netherlands. Alternatively, it could also be a result of geographical separation by the IJssel river or the Veluwe hillridge. The third PC is associated with the more central regions of the country (Abdellaoui et al. 2013). By adding the PCs to our models, we assessed the role of genetic ancestry of individuals between regions.

In this paper, we first considered regional clustering of children's height in a large data set of MZ and DZ twins (N = 7,436). Secondly, we considered the model within a subgroup of children who were genotyped on genome-wide SNP arrays (N = 1,375). Subsequently, we determined whether the region effects represent genetic ancestry. And finally, we analyzed the relationship between the three PCs and height in 7-year-old children, and included the genetic PCs that show an association as an individual level (level 1) covariate in the model.

Methods

Participants and procedure

The data were obtained from the Netherlands Twin Register (NTR), which has collected data on multiple-births and their family members since 1987 (Ligthart et al., 2019). In the longitudinal NTR surveys of phenotypes in children, parents were asked to complete questionnaires on their children's health, growth, and behavior with intervals of approximately two years.

For the present study, we included data on 6- and 7-year-old twin children (range 6 years and 0 months to 7 years and 11 months). The sample included 7,346 twin children (50.3% girls) in 3,724 families. The twins were 7.4 (SD 0.3) years old on average, when their mothers reported their height. Of these children, 1,375 (18.7% of total) were genotyped. Genotyping largely took place independent of phenotype criteria. The 1,375 genotyped individuals were from 714 families, 52.4% of this subsample were girls and the average age was 7.4 (SD 0.3).

We included data from 2002 onwards, as that was when active collection of postal code data began. In approximately 1% of the questionnaires that were sent out after 2002, postal code was missing and approximately 20% of the parents did not report their children's height at age 7. We only included participants with both height and postal code information at age 7 in our initial selection. Next, children with severe handicaps were excluded, as were multiple twin pairs per family, twins born before 34 weeks of gestation, and twins outside the 6-8 age range. A flowchart outlining the sample size after every step of exclusion is displayed in Figure 1. Zygosity was determined by DNA polymorphisms or by a parent-reported zygosity questionnaire on twin similarity. The zygosity determination by questionnaire has an accuracy of over 95% (Ligthart et al., 2019). Table I displays the descriptive statistics of the phenotypic data by zygosity for the total and for the genotyped sample.

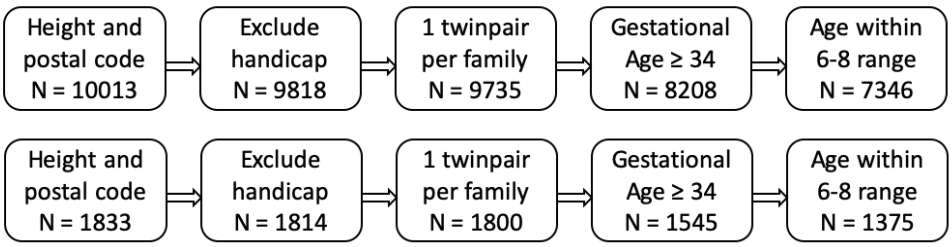


Figure 1. Flowchart containing sample size for the total sample (upper row) and the genotyped subsample (lower row) after every step of exclusion.

Table I. The number of twins, the mean, standard deviation and the twin correlation per zygosity group for the total sample and the genotyped subsample.

	MZm	DZm	MZf	DZf	DZmf	DZfm
Total sample						
N (individuals)	1283	1228	1338	1208	1163	1126
Mean height	128.3	128.4	127.3	127.8	128.6	128.4
SD height	6.1	5.8	5.8	5.8	5.7	6.2
Twin correlation	0.95	0.61	0.94	0.63	0.58	0.68
Genotyped subsample						
N (individuals)	350	167	251	221	136	150
Mean height	128.5	129.1	127.7	127.4	128.2	128.0
SD height	6.0	5.8	5.7	6.1	5.6	5.7
Twin-correlation	0.97	0.71	0.95	0.68	0.57	0.69

Measures

Height

Mothers reported child height in centimeters and the date of measurement. Estourgie-van Burk et al. (2006) demonstrated that the correlation between maternal report and height measured in the laboratory was .96 in 5-year-old children in NTR. Mothers reported the age of their children at the moment of completing the survey and the date of the height measurement. In 5% of the children, the date at the time of height measurement was not available. Therefore, in this 5%, we took the age at the time of questionnaire completion. The correlation between age at questionnaire completion and age at height measurement is 0.95, and the mean difference in age is 0.01 years.

Region

At the time of reporting height, parents also reported the four digits of the postal code of their current address. In the Netherlands, postal codes map to geographical locations. The postal code consists of four digits and two letters, where the first two digits map to region and the second two digits and letters map to city, neighborhood within the city, and street. In our analyses, region is specified by the first two digits of the postal code, resulting in 90 regions which are displayed in Figure 2. They cover on average 462 km^2 and have a mean population of around 192,000 (total area of the country is 41,543 km^2, including

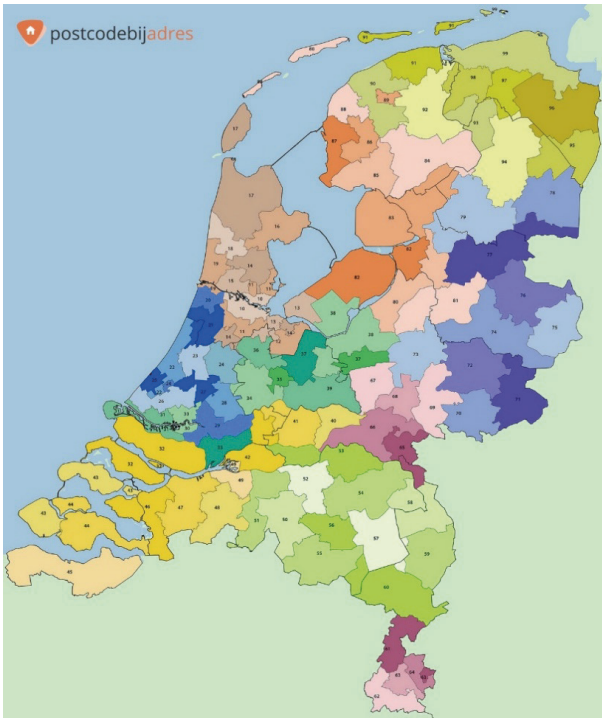


Figure 2. Map of the 90 regions in the Netherlands based on first two digits of the postal code.
Note. This figure is reprinted from 'Postcodekaart van Nederland' by postcodebijadres, retrieved July 29, 2020, from https://postcodebijadres.nl/postcodes-nederland

~19% water bodies). Most regions encompass several municipalities. In the total sample, the number of children per postal code unit ranged from 10 to 194 (M= 81.6, SD = 38.4). In the genotyped sample, the number of children per postal code unit ranged from 1 to 43 (M= 15.6, SD = 8.6).

Principal components

Genotype data in 1375 individuals were collected by the following genotype platforms: Affymetrix 6, Axiom and Perlegen, Illumina 1M, 660 and GSA-NTR. The SNP data obtained on the 6 platforms were pruned in Plink to be independent, with additional filters to ensure Minor Allele Frequency (MAF) > 0.01, Hardy-Weinberg Equilibrium (HWE) p > 0.0001 and call rate over 95%. Subsequently, long range Linkage Disequilibrium (LD) regions were excluded as described in Abdellaoui et al. (2013), because elevated levels of LD result in overrepresentation of these loci in the PCs, disguising genome-wide patterns that reflect ancestry. For each platform, the NTR data were merged with the

data of the individuals from the 1000 Genomes reference panel for the same SNPs, and Principal Components were calculated using SMARTPCA (Prince et al., 2006), where the 1000 genomes populations were projected onto the NTR participants (Privé et al., 2020). Population outliers were identified using pairwise PC plots. People who were identified as outliers from the central population on the basis of visual inspection of these pairwise PC plots, were excluded, rendering the final clustering homogeneous. The NTR platform genotype data of this cluster were aligned to the GoNL reference panel V4 (The Genome of the Netherlands Consortium, 2014), merged into a single dataset, and then imputed in MaCH-Admix (Liu et al., 2013). From the imputed data, SNPs were selected that satisfied $R^2 \geq 0.90$, and that were genotyped on at least one platform. These SNPs were subsequently filtered on MAF < 0.025, HWE $p < 0.0001$, call rate ≥ 98%, and the absence of Mendelian errors. Again, the long-range LD regions were removed from these SNP data. With this selection of SNPs, 20 new PCs were calculated with SMARTPCA (Prince et al., 2006), to indicate the residual Dutch genetic stratification.

Models

The Classical Twin Model

In the classical twin model, the phenotypic variance can be decomposed into three components: Additive genetic (A), Common environmental (C) and unique Environment (E) component, which includes measurement error. As in most earlier publications, we will not consider genetic dominance variance for height (but see Joshi et al., 2015).

Assuming A, C, and E are mutually independent, we have the following decomposition of phenotypic variance:

$$var(y) = \sigma_A^2 + \sigma_C^2 + \sigma_E^2 .$$

The variance component model can be written as a path model in which A, C and E are standardized to have unit variance (see Figure 3):

$$y_1 = \mu + a * A_1 + c * C_1 + e * E_1,$$
$$y_2 = \mu + a * A_2 + c * C_2 + e * E_2,$$

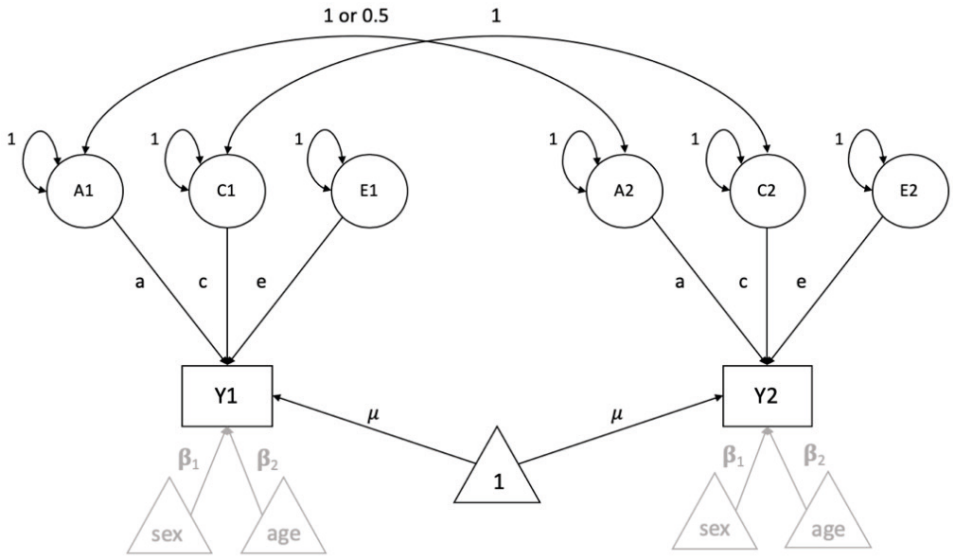


Figure 3. The Classical Twin Model including three latent factors per person, representing Additive genetic, Common and unique Environmental influences. Two additional covariates, age and sex, are presented in a schematic way in grey.

where y_1 represents the phenotype of first twin and y_2 of the second twin in a twin pair. A, C and E represent individual factor scores for twin 1 and twin 2, and a, c, e represent population specific factor loadings or path coefficients.

If A, C and E have unit variance, the variance decomposition is:

$$var(y) = a^2 + c^2 + e^2.$$

In terms of the path coefficient model, the covariance between the twins equals $\sigma_{mz} = a^2 + c^2$ in MZ twins, and $\sigma_{dz} = \frac{1}{2}a^2 + c^2$ in DZ twins.

Multilevel Twin Model

When specifying a CTM as an MLM, the variance components of the CTM are parametrized as within and between family components. The additive genetic variance is separated into two parts: a part that is shared by the members of a twin pair on the second level, A_C, and a part that is unique to each individual on the first level, A_U. The path coefficients associated with the A_C and A_U are equal.

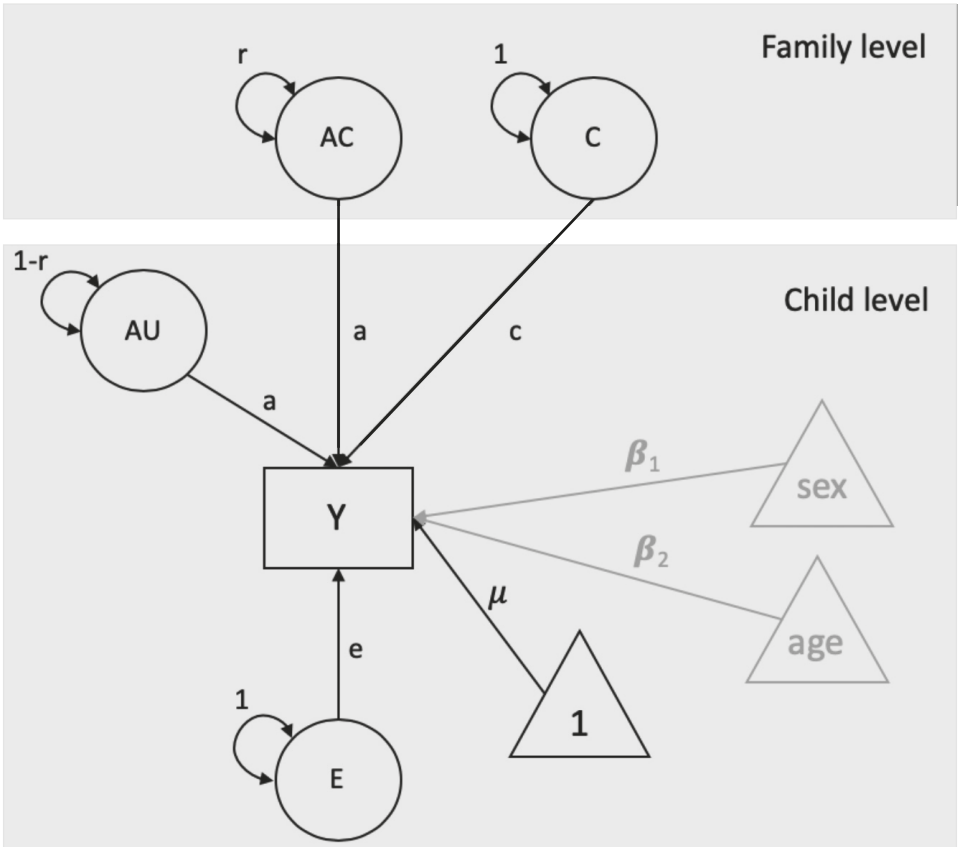


Figure 4. Multilevel parametrization of the ACE model, where Y represents the phenotype, latent variables A_U represent the unique additive genetic influences and E unique environment. μ is the intercept, sex and age are covariates, presented in a schematic way. On the family level, C is common environment and A_C common genetic influences. The path coefficients, a, c, e, β_1 and β_2 represent regression coefficients. The r parameter represents variance (1 for MZ twins and 0.5 for DZ twins).

The variance of the common genetic factor (r) and the unique genetic factor (1-r) depend on the zygosity of the twin pair: for MZ r = 1.00, while for DZ r = 0.50. The common environmental factor, representing between family influences, is a level two component. Unique environmental factors E represent within family, level one, influences. The means (intercepts) m are specified on the first level and are assumed to be equal for first- and second-born twins and zygosity. The ACE model in multilevel parametrization is illustrated in Figure 4. Here, we included age at the individual level, because it represents the age at reported height measure and thus could differ between twins.

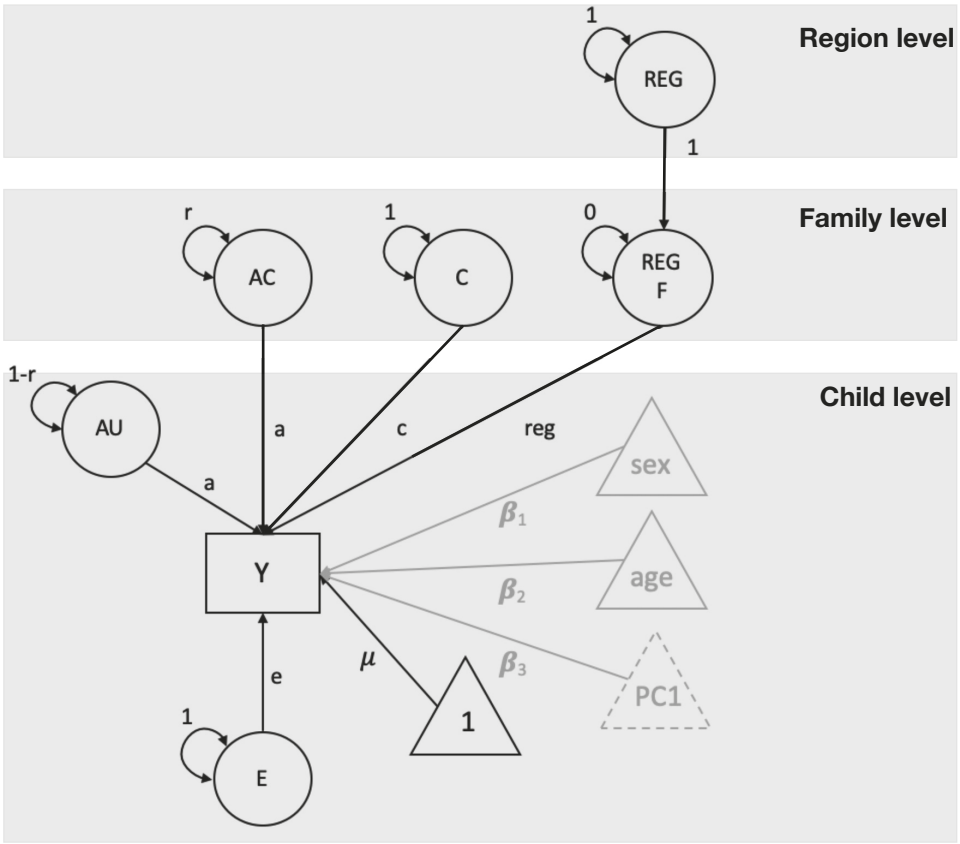


Figure 5. Multilevel parametrization of the ACE model with Region as a third level, which loads on the region variable REG F on the family level, on which the observed variable Y is regressed with its coefficient estimating the effect of region. Sex and age are covariates. We compared the model where we included PC1 as a covariate with the model where we did not include PC1 as a covariate (we tested PC1, PC2, and PC3, but depict only PC1 to avoid clutter, and because our final model included only PC1). Note that the dummy REG F latent variable serves as a placeholder to stress the nesting of families in regions, but technically, it is not needed.

Multilevel Twin Model with third level clustering variable and individual level covariates

Other clustering variables can be added to this model, as displayed in Figure 5. A higher-order clustering variable can be added to the third level of this model in two steps. On the third level, the higher-order clustering variable is added with a variance of 1 and a path loading of 1 to a latent variable on the second level, which has a variance of 0 and a freely estimated path loading from Region (reg) to the observed phenotype. Although the Region latent variable could directly affect the child-level phenotype and does not need to pass through the family level,

we draw it here to indicate the nesting that region-level effects pass through the family level before impacting the child level. The same 3-level model which also includes PC1 as a fixed covariate is displayed in Figure 5.

Analyses

All analyses were performed in R (R Core Team, 2020) with the package OpenMx (Boker et al., 2011; Neale at al., 2016; Pritkin et al., 2017). Age at measurement was converted to z-scores. Due to scaling, the variance of PC scores is extremely low compared to the variance of the other variables in the model. Therefore, we multiplied these scores by 1000 to avoid ill-conditioning in the parameter covariance matrix, since ill-conditioning can cause optimization problems. First, in the full sample, a variance decomposition of the variance in height was obtained in the regular genetic covariance structure modeling. We included the z-scores of age at measurement and sex as covariates. Then, we repeated the analysis in the multilevel model to illustrate the equivalence of the two approaches. Following this, we added region as a third level in the multilevel parametrization. We repeated these steps in the genotyped group to investigate the representativeness of this subsample. Finally, in the genotyped subset, we added the PC scores as individual level covariates in the 3- level model.

We tested the contribution of region to the variance of height by comparing the difference of fit in the 3-level model and the 2-level model without region with the log-likelihood ratio test. Under certain regularity conditions (Steiger et al., 1985), the difference in fit between these models is distributed as chi-squared with one degree of freedom. For all analysis we employed an alpha level of 0.01

Results

The plot of the average height by region revealed a north-south trend, with the children in the northern regions being taller than those in the southern regions of the Netherlands (of the 12 provinces in the Netherlands, the northern province Drenthe had the highest mean height (M = 129.40) and the southern province Noord-Brabant had the lowest mean height (M = 127.01)). Figure 6 displays the mean height of 7-year-olds per region. In the genotyped group, height correlated with PC1 (i.e., the PC showing a north-south gradient) (r = 0.16), but not with other PCs (r = 0.01 for PC2, r = -0.01 for PC3). Therefore, we incorporated PC1 into subsequent analyses and omitted PC2 and PC3.The 2-level model fitted significantly worse than the 3-level model with region as level three clustering

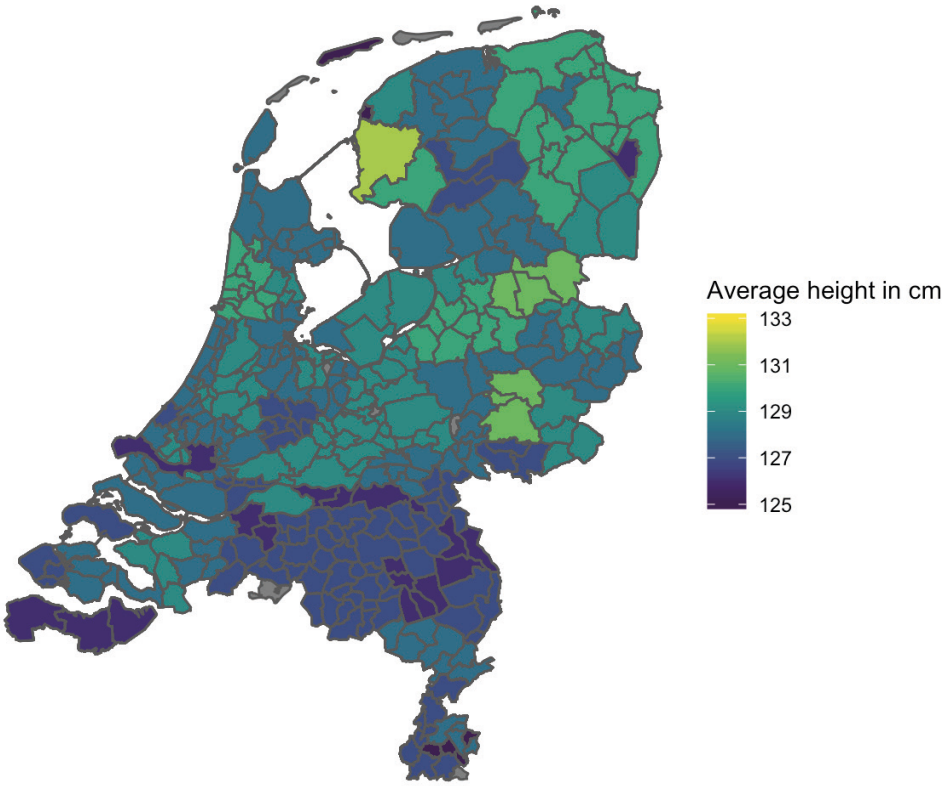


Figure 6. Mean height (in centimeters) of 7-year-old children by region in the Netherlands.

variable (Δ-2LL = 22.93, Δdf = 1, p <.001). So, region in the Netherlands accounts for a statistically significant proportion (1.8%) of the variance in height in 7-year-olds. Table II displays the parameter estimates and the standardized variance components of the models. Comparing the parameter estimates of the models shows that the variance attributable to region in the 3-level model was captured by the C-component in the 2-level model.

Results of analyses for genotyped sample

In the genotyped group, region explained 1.6% of the variance, which almost equals the percentage 1.8% reported above. The likelihood ratio test of this component was not significant: Δ-2LL = 0.85, Δdf = 1, p = .36. However, we ascribed this to a lack of power given the appreciably smaller sample size (in terms of individuals, N = 7,346, vs. N = 1,375). The parameter estimates and standardized variance components are displayed in Table III.

Table II. Results of CTM and 2-level and 3-level MLM analyses for the full sample: path coefficient estimates with standard errors (SE) and standardized variance components of the 2-level and the 3-level models (with age and sex as covariates). N = 7346 twins in 3724 families.

	parameter	CTM	2-level model	3-level model
Intercept	Intercept (SE)	128.4 (0.11)	128.4 (0.11)	128.5 (0.14)
Covariates	β_{sex} (SE)	-0.62 (0.12)	-0.62 (0.12)	-0.62 (0.12)
	β_{age} (SE)	1.43 (0.08)	1.42 (0.08)	1.42 (0.08)
a, c, e, region path loadings	a (SE)	4.70 (0.08)	4.70 (0.08)	4.70 (0.08)
	c (SE)	**2.90 (0.16)**	**2.90 (0.16)**	**2.80 (0.16)**
	e (SE)	1.39 (0.03)	1.39 (0.03)	1.39 (0.03))
	region (SE)			92.33 (2.13)
	Total variance ($a^2 + c^2 + e^2$ (+ region^2))	32.46	32.46	32.47
	A (standardized)	0.681	0.681	0.681
	C (standardized)	**0.259**	**0.259**	**0.241**
	E (standardized)	0.060	0.060	0.060
	REGION (standardized)			0.018

Table III. Results of 2-level and 3-level MLM analyses in the genotyped sample (N = 1375 twins in 714 families). Path coefficient estimates with standard errors (SE) and standardized variance components of the 2- and 3-level model (with age and sex as covariates).

	parameter	2-level model	3-level model
Intercept	Intercept (SE)	128.4 (0.26)	128.5 (0.27)
Covariates	β_{sex} (SE)	-0.62 (0.31)	-0.63 (0.31)
	β_{age} (SE)	1.16 (0.18)	1.15 (0.18)
a, c, e, region path loadings	a (SE)	4.53 (0.20)	4.53 (0.20)
	c (SE)	**3.21 (0.34)**	**3.13 (0.16)**
	e (SE)	1.13 (0.04)	1.13 (0.04))
	Region (SE)		0.71 (2.13)
	Total variance ($a^2 + c^2 + e^2$ (+ region^2))	32.11	32.11
	A (standardized)	0.640	0.640
	C (standardized)	**0.320**	**0.305**
	E (standardized)	0.040	0.040
	REGION (standardized)		0.016

Table IV. Results of 2-level and 3-level MLM analyses for the genotyped sample with PC covariate (N = 1375 twins in 714 families). Path coefficient estimates with standard errors (SE) and standardized variance components of the 2- and 3-level ACER model, including PC1.

	parameter	2-level model	3-level model
Intercept	Intercept (SE)	128.5 (0.26)	128.5 (0.26)
Covariates	β_{sex} (SE)	-0.74 (0.31)	-0.74 (0.31)
	β_{age} (SE)	1.16 (0.18)	1.17 (0.18)
	β_{PC1} (SE)	0.11 (0.02)	0.11 (0.02)
a, c, e, region path loadings	a (SE)	4.55 (0.20)	4.55 (0.20)
	c (SE)	3.03 (0.36)	3.03 (0.36)
	e (SE)	1.13 (0.04)	1.13 (0.04))
	region (SE)		$8.97 * 10^{-6}$ (0.76)
	Total variance	31.19	31.19
	($a^2 + c^2 + e^2$ (+ region^2))		
	A (standardized)	0.664	0.664
	C (standardized)	0.295	0.295
	E (standardized)	0.041	0.041
	REGION (standardized)		$2.58 * 10^{-12}$

Results of analyses for genotyped sample with PC1 as covariate

Table IV displays the parameter estimates and standardized variance components of the 2- and 3-level model with PC1 included as a fixed covariate. When we included PC1 in the 3-level model, the variance explained by region went from 1.6% (before inclusion of PC1; see previous section) to <0.001%. This indicates that when PC1 is included as a covariate, region no longer explains any phenotypic variance in height. This was confirmed by the likelihood ratio test comparing the 2-level and 3-level model. As expected, with PC1 as a covariate the 2-level model fitted equally well as the 3-level model (Δ-2LL < 0.001, Δdf = 1), suggesting no effects of region after inclusion of PC1 in the model.

Discussion

In this paper we specified a multilevel twin model in OpenMx and fitted it to data on children's height. We added a higher-level variable, region in the Netherlands, in which the twin pairs were nested. Adding a third level variable enabled us to determine whether part of the variance in children's height can be explained by differences in geographical region.

We found that 68% of the variance in 7-year-old children's height is attributable to additive genetic factors. Common environmental factors accounted for 26%, and unshared environmental factors (including measurement error) for 6% of the variance. We found that regional differences accounted for a significant 1.8% of the phenotypic variance in the complete sample (1.6% in the genotype subsample). In a standard multilevel ACE-twin model, ignoring regional clustering, this variance was captured by the C-component. This is expected, because the common environmental component captures between-family variance, regardless of its source. At age 7, cohabiting MZ and DZ twins necessarily share region, so that the effect of region will contribute to C variance.

In a subsample of children with genetic PC scores, i.e., the genotyped subsample, we found a statistically significant correlation (r = 0.16) between height and the first genetic PC, representing the geographical north-south gradient in the Netherlands. This correlation is similar to previous results for height in a Dutch sample of adults and in line with the findings in European samples, where northern populations are on average taller than southern populations (Abdellaoui et al., 2013). The correlations between the second and third PC and children's height were negligible. After the inclusion of the first PC in the multilevel model, region no longer explained any variance.

This last result indicates that the variance in children's height that is explained by region is attributable to differences in genetic ancestry. That is, although unmodeled regional clustering manifests as C, it does not mean that the inflation of the common environmental variance is due to genuine shared environmental factors like region. When we included the first PC, which reflects differences in allele frequencies between regions, no variance was explained by geographical region above and beyond what was already explained by the PC. Because offspring from the same family share their ancestry, a proportion of the variance that is captured in the C-component of the CTM is actually of a genetic nature. This does not, however, entirely exclude the presence of environmental effects that are explanatory of regional clustering in height. The PC representing the

north-south gradient could be correlated with environmental factors that might contribute to the relationship between PC1, height and regional clustering.

We note the following limitations of our study. First, we did not explicitly model qualitative differences in genetic architecture between boys and girls. There is some evidence that the additive genetic correlation in opposite-sex twins is lower than 0.50, suggesting that partly different genes operate in 7-year-old boys and girls (Silventoinen et al., 2007). However, the twin correlations in our sample did not suggest the presence of qualitative sex differences (we observed correlations of .61 and .63 in the DZ male and DZ female, versus .58 and .68 in the DZ opposite sex male-female and DZ opposite sex female-male twins, respectively).

Secondly, we surmise that the power to detect the region effect in the genotyped sample was low, given the sample size (*N*=1,375 in the genotyped sample). However, the effect sizes in both samples were very similar (1.8% vs 1.6%), and in the full sample (*N*=7,346) effect was statistically significance. Therefore, we trust that the regional effect is real.

A final limitation to note is that the current approach assumes that lower levels are fully nested in the higher-level. That is, members of a twin pair cannot differ on the clustering variable. It is therefore not possible to define a third-level clustering variable, when the variable of interest differs within a twin pair (e.g., adult twins who do not live in the same region). It is possible, however, to include variables in which both twins are not nested as a lower-level variance component. When the clustering variable is not specified as a higher-level (i.e., nesting) variable, the effect of clustering can also be manifested as any of the other variance components (i.e., A/C/D/E) when unmodeled. Furthermore, missing data for higher-level clustering variable (here: region) is not allowed. The higher-level variable needs to have a sufficient number of units for the model to have enough power to detect the effect of the higher-level variable (e.g., postal codes in our region example; Goldstein, 2011).

The current study showed that when data are nested in a higher-level variable, adding this higher-level variable to a multilevel model for twin data provides opportunities to further decompose the phenotypic variance. Clustering can be due to unwanted confounding, for example, batch effects. Applying a multilevel model to identify the nuisance variance that is explained by higher-level clustering would in this case serve as a correction. However, as is shown within this paper, the MLM can also be used to empirically study clustering.

Literature

Abdellaoui, A., Hottenga, J. J., Knijff, P. D., Nivard, M. G., Xiao, X., Scheet, P., ... & Boomsma, D. I. (2013). Population structure, migration, and diversifying selection in the Netherlands. European journal of human genetics, 21(11), 1277-1285. https://doi.org/10.1038/ejhg.2013.48

Abdellaoui, A., Hugh-Jones, D., Yengo, L., Kemper, K. E., Nivard, M. G., Veul, L., ... & Visscher, P. M. (2019). Genetic correlates of social stratification in Great Britain. Nature human behaviour, 3(12), 1332-1342.https://doi.org/10.1038/s41562-019-0757-5

Barton, N. H., Etheridge, A. M., & Véber, A. (2017). The infinitesimal model: Definition, derivation, and implications. Theoretical population biology, 118, 50-73. https://doi.org/10.1016/j.tpb.2017.06.001

Baten, J., & Blum, M. (2014). In: van Zanden JL (Ed.), How was life?: Global well-being since 1820, OECD Publishing, Paris, pp 117-137. https://doi.org/10.1787/9789264214262-11-en

Baten, J., & Komlos, J. (1998). Height and the Standard of Living. The Journal of Economic History, 58(3), 866-870. https://doi.org/10.1017/S0022050700021239

Bentler, P. M., & Stein, J. A. (1992). Structural equation models in medical research. Statistical methods in medical research, 1(2), 159-181. https://doi.org/10.1177/096228029200100203

Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., ... & Fox, J. (2011). OpenMx: an open source extended structural equation modeling framework. Psychometrika, 76, 306-317. https://doi.org/10.1007/s11336-010-9200-6

Boomsma, D. I., Orlebeke, J. F., & Van Baal, G. C. M. (1992). The Dutch Twin Register: Growth data on weight and height. Behavior Genetics, 22, 247-251. https://doi.org/10.1007/BF01067004

Boomsma, D. I., & Molenaar, P. C. (1986). Using LISREL to analyze genetic and environmental covariance structure. Behavior Genetics, 16(2), 237-250. https://doi.org/10.1007/BF01070799

Boomsma, D. I., Wijmenga, C., Slagboom, E. P., Swertz, M. A., Karssen, L. C., Abdellaoui, A., ... & van Duijn, C. M. (2014). The Genome of the Netherlands: design, and project goals. European Journal of Human Genetics, 22(2), 221-227. https://doi.org/10.1038/ejhg.2013.118

Chen, G., Saad, Z. S., Nath, A. R., Beauchamp, M. S., & Cox, R. W. (2012). FMRI group analysis combining effect estimates and their variances. Neuroimage, 60(1), 747-765. https://doi.org/10.1016/j.neuroimage.2011.12.060

Colodro-Conde, L., Couvy-Duchesne, B., Whitfield, J. B., Streit, F., Gordon, S., Kemper, K. E., ... & Martin, N. G. (2018). Association between population density and genetic risk for schizophrenia. JAMA psychiatry, 75(9), 901-910. https://doi.org/10.1001/jamapsychiatry.2018.1581

Eaves, L. J., Last, K. A., Young, P. A., & Martin, N. G. (1978). Model-fitting approaches to the analysis of human behaviour. Heredity, 41(3), 249-320. https://doi.org/10.1038/hdy.1978.101

Estourgie-van Burk, G. F., Bartels, M., Van Beijsterveldt, T. C., Delemarre-van de Waal, H. A., & Boomsma, D. I. (2006). Body size in five-year-old twins: heritability and comparison to singleton standards. Twin Research and Human Genetics, 9(5), 646-655. https://doi.org/10.1375/twin.9.5.646

Falconer, D.S. & McKay, T.F.C. (1996). Introduction to Quantitative Genetics, Burnt Mill, England. https://doi.org/10.2307/2529912

Fisher, R. A. (1919). The correlation between relatives on the supposition of Mendelian inheritance. Earth and Environmental Science Transactions of the Royal Society of Edinburgh, 52(2), 399-433. https://doi.org/10.1017/S0080456800012163

Gelman, A. (2005). Analysis of variance—why it is more important than ever. The annals of statistics 33(1), 1-53. https://doi.org/10.1214%2F009053604000001048

Goldstein, H. (2011). Multilevel statistical models (Vol. 922). John Wiley & Sons, Chichester, UK. https://doi.org/10.1002/9780470973394

Guo, G., & Wang, J. (2002). The mixed or multilevel model for behavior genetic analysis. Behavior genetics, 32, 37-49. https://doi.org/10.1023/A:1014455812027

Heath, A. C., Neale, M. C., Hewitt, J. K., Eaves, L. J., & Fulker, D. W. (1989). Testing structural equation models for twin data using LISREL. Behavior genetics, 19(1), 9-35. https://doi.org/10.1007/BF01065881

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. Journal of educational psychology, 24(6), 417. https://doi.org/10.1037/h0071325

Hunter, M. D. (2021). Multilevel modeling in classical twin and modern molecular behavior genetics. Behavior Genetics, 51(3), 301-318.

Jelenkovic, A., Sund, R., Hur, Y. M., Yokoyama, Y., Hjelmborg, J. V. B., Möller, S., ... & Silventoinen, K. (2016). Genetic and environmental influences on height from infancy to early adulthood: An individual-based pooled analysis of 45 twin cohorts. Scientific reports, 6(1), 28496. https://doi.org/10.1038/srep28496

Joshi, P. K., Esko, T., Mattsson, H., Eklund, N., Gandin, I., Nutile, T., ... & Kuusisto, J. (2015). Directional dominance on stature and cognition in diverse human populations. Nature, 523(7561), 459-462. https://doi.org/10.1038/nature14618

Karp, R., Martin, R., Sewell, T., Manni, J., & Heller, A. (1992). Growth and academic achievement in inner-city kindergarten children: the relationship of height, weight, cognitive ability, and neurodevelopmental level. Clinical Pediatrics, 31(6), 336-340. https://doi.org/10.1177/000992289203100604

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. Biometrics, 963-974. https://doi.org/10.2307/2529876

Ligthart, L., van Beijsterveldt, C. E., Kevenaar, S. T., de Zeeuw, E., van Bergen, E., Bruins, S., ... & Boomsma, D. I. (2019). The Netherlands Twin Register: longitudinal research based on twin and twin-family designs. Twin Research and Human Genetics, 22(6), 623-636. https://doi.org/10.1017/thg.2019.93

Liu, E. Y., Li, M., Wang, W., & Li, Y. (2013). MaCH-Admix: genotype imputation for admixed populations. Genetic epidemiology, 37(1), 25-37. https://doi.org/10.1002/gepi.21690

Longford, N. T. (1993). Regression analysis of multilevel data with measurement error. British Journal of Mathematical and Statistical Psychology, 46(2), 301-311. https://doi.org/10.1111/j.2044-8317.1993.tb01018.x

Martin, N. G., & Eaves, L. J. (1977). The genetical analysis of covariance structure. Heredity, 38(1), 79-95. https://doi.org/10.1038/hdy.1977.9

McArdle, J. J., & Prescott, C. A. (2005). Mixed-effects variance components models for biometric family analyses. Behavior genetics, 35, 631-652. https://doi.org/10.1007/s10519-005-2868-1

Neale, M. C., & Cardon, L. R. (1992). NATO ASI series D: Behavioural and social sciences, Vol. 67. Methodology for genetic studies of twins and families. New York, NY, US. https://doi.org/10.1007/978-94-015-8018-2

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. https://doi.org/10.1007/s11336-014-9435-

Paterson, L., & Goldstein, H. (1991). New statistical methods for analysing social structures: an introduction to multilevel models. British educational research journal, 17(4), 387-393. https://doi.org/10.1080/0141192910170408

Postcodebijadres (2020). Postcodekaart van Nederland. Retrieved from https://postcodebijadres.nl/postcodes-nederland. Accessed July 29, 2020.

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. https://doi.org/10.1038/ng1847

Pritikin, J. N., Hunter, M. D., von Oertzen, T., Brick, T. R., & Boker, S. M. (2017). Many-level multilevel structural equation modeling: An efficient evaluation strategy. Structural equation modeling: a multidisciplinary journal, 24(5), 684-698. https://doi.org/10.1080/10705511.2017.1293542

Privé, F., Luu, K., Blum, M. G., McGrath, J. J., & Vilhjálmsson, B. J. (2020). Efficient toolkit implementing best practices for principal component analysis of population genetic data. Bioinformatics, 36(16), 4449-4457. https://doi.org/10.1093/bioinformatics/btaa520

R Development Core Team (2020). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org (ISBN 3-900051-07-0)

Schalekamp, J. C. (2009). Bataven en buitenlanders: 20 eeuwen immigratie in Nederland. Wind Publishers.: Huizen, the Netherlands, pp 15–40.

Scharpf, R. B., Ruczinski, I., Carvalho, B., Doan, B., Chakravarti, A., & Irizarry, R. A. (2011). A multilevel model to address batch effects in copy number estimation using SNP arrays. Biostatistics, 12(1), 33-50. https://doi.org/10.1093/biostatistics/kxq043

Sellström, E., & Bremberg, S. (2006). Is there a “school effect” on pupil outcomes? A review of multilevel studies. Journal of Epidemiology & Community Health, 60(2), 149-155. https://doi.org/10.1136/jech.2005.036707

Silventoinen, K., Bartels, M., Posthuma, D., Estourgie-van Burk, G. F., Willemsen, G., van Beijsterveldt, T. C., & Boomsma, D. I. (2007). Genetic regulation of growth in height and weight from 3 to 12 years of age: a longitudinal study of Dutch twin children. Twin Research and Human Genetics, 10(2), 354-363. https://doi.org/10.1375/twin.10.2.354

Silventoinen, K., Krueger, R. F., Bouchard, T. J., Kaprio, J., & McGue, M. (2004). Heritability of body height and educational attainment in an international context: Comparison of adult twins in Minnesota and Finland. American Journal of Human Biology: The Official Journal of the Human Biology Association, 16(5), 544-555. https://doi.org/10.1002/ajhb.20060

Steiger, J. H., Shapiro, A., & Browne, M. W. (1985). On the multivariate asymptotic distribution of sequential chi-square statistics. Psychometrika, 50, 253-263. https://doi.org/10.1007/BF02294104.

Rabe-Hesketh, S., Skrondal, A., & Gjessing, H. K. (2008). Biometrical modeling of twin and family data using standard mixed model software. Biometrics, 64(1), 280-288. https://doi.org/10.1111/j.1541-0420.2007.00803.x

Reich, D., Price, A. L., & Patterson, N. (2008). Principal component analysis of genetic data. Nature genetics, 40(5), 491-492.. https://doi.org/10.1038/ng0508-491

Rijsdijk, F. V., & Sham, P. C. (2002). Analytic approaches to twin data using structural equation models. Briefings in bioinformatics, 3(2), 119-133. https://doi.org/10.1093/bib/3.2.119

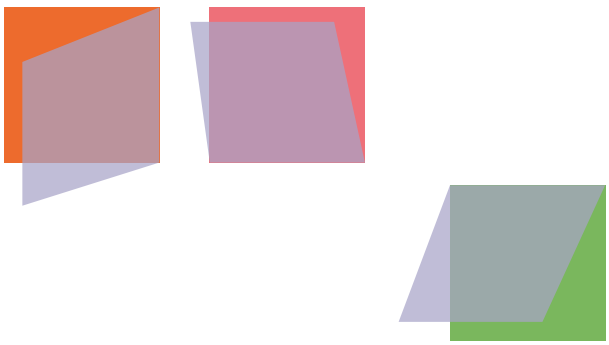
Spears, D. (2012). Height and cognitive achievement among Indian children. Economics & Human Biology, 10(2), 210-219.. https://doi.org/10.1016/j.ehb.2011.08.005

The Genome of the Netherlands Consortium (2014). Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nature genetics, 2014, 46.8: 818-825. https://doi.org/10.1038/ng.3021

van den Oord, E. J. (2001). Estimating effects of latent and measured genotypes in multilevel models. Statistical Methods in Medical Research, 10(6), 393-407. https://doi.org/10.1177%2F096228020101000603

Chapter 3

Bayesian evidence synthesis in case of multi-cohort datasets: An illustration by multi-informant differences in self-control



Published as: **Kevenaar**, S. T., Zondervan-Zwijnenburg, M. A., Blok, E., Schmengler, H., Fakkel, M. T., De Zeeuw, E. L., van Bergen, E., Onland-Moret, N. C., Peters, M., Hillegers, M. H. J., Boomsma, D. I., & Oldehinkel, A. J. (2021). Bayesian evidence synthesis in case of multi-cohort datasets: An illustration by multi-informant differences in self-control. *Developmental Cognitive Neuroscience, 47*, 100904.

Abstract

The trend toward large-scale collaborative studies gives rise to the challenge of combining data from different sources efficiently while facilitating hypothesis testing. Here, we demonstrate how Bayesian evidence synthesis can be used to quantify and compare support for competing hypotheses and to aggregate this support over studies. We applied this method to study the ordering of multi-informant scores on the ASEBA Self Control Scale (ASCS), employing a multi-cohort design with data from four Dutch cohorts. Self-control reports were collected from mothers, fathers, teachers and children themselves. The available set of reporters differed between cohorts, so in each cohort varying components of the overarching hypotheses were evaluated. We found consistent support for the partial hypothesis that parents reported more self-control problems in children than teachers. Furthermore, the aggregated results indicate most support for the combined hypothesis that children report most problem behaviors, followed by their mothers and fathers, and that teachers report the fewest problems. However, there was considerable inconsistency across cohorts regarding the rank order of children's reports. This article illustrates Bayesian evidence synthesis as a method when some of the cohorts only have data to evaluate a partial hypothesis. With Bayesian evidence synthesis, these cohorts can still contribute to the aggregated result.

Keywords: Multiple cohorts, Multiple Informants, Self-Control, Bayesian Evidence Synthesis, Multiple Imputation by chained equations (MICE).

Introduction

There is a growing awareness of the limited reliability of single-study findings, in Developmental Cognitive Neuroscience and other fields of empirical research (Open Science Collaboration, 2015). This awareness has contributed to the call for replication and the need to synthesize findings across studies. Consortia, such as the Consortium on Individual Development (CID), have been established to combine research efforts of different groups to study a particular subject. This raises the challenge to do so in a way that includes and does justice to each study's unique qualities, and still allows conclusions based on accumulated evidence.

A common way to synthesize research findings is meta-analysis, where the results of several previously conducted studies concerning a particular research question, topic, or theory are combined (Rosenthal & DiMatteo, 2002). Meta-analysis has notable advantages, such as the possibility to base the analysis on summary statistics, but has also limitations. Three limitations are (1) that meta-analysis does not allow additional inference on the level of the individual studies, (2) that meta-analysis is prone to the effects of searching strategies and publication bias, (3) and that meta-analysis can only include studies employing comparable models and parameters.

In this article, we apply the alternative strategy of Bayesian evidence synthesis to reach robust conclusions by combining results derived from different sources. Here, the different data sources are four Dutch population cohort studies. Bayesian evidence synthesis can be used to combine results by aggregating their evidence for competing hypotheses (Kuiper, Buskens, Raub & Hoijtink, 2012; Zondervan-Zwijnenburg et al., 2019). In this manner, studies covering various contexts and measurement instruments can be combined (Zondervan-Zwijnenburg et al., 2019, 2020). This approach is also suitable to combine the results of structural equation modeling (Zondervan-Zwijnenburg et al., 2019, 2020). The main assumptions of Bayesian evidence synthesis are that all sources of information provided by individual studies contribute to the overarching research question, and that all samples are representative of the population of interest (Veldkamp et al., 2020).

In the current study, we demonstrate that Bayesian research synthesis can be used even if not all parameters relevant to the hypotheses are estimated in all cohorts. More specifically, our overarching hypothesis concerns the ordering of mean raters obtained from four raters of child self-control: teachers, fathers, mothers and children. However, some cohorts only have data of three or fewer

raters, and provide partial information concerning the ordering of the mean ratings. So while the comprehensive hypothses may concern the ordering of several means, the information provided by some cohort may be limited to a subset of the means. For example, consider the assessment of differences among multiple neuropsychological tasks that are assumed to assess the same process, brain areas that are activated by a task, or, as in our case, informants that rate a specific trait or state. In these cases, the Bayesian synthesis approach offers the advantage that it enables statements about the support for specific hypotheses concerning the ordering of parameters, and the possibility to aggregate results, given incomplete information (results) in one or more of the studies. To our best knowledge, this application of Bayesian evidence synthesis is new.

We demonstrate the opportunities and challenges of Bayesian evidence synthesis for a comparison of multiple groups using multi-informant scores of self-control. Self-control is a key topic within the Dutch Consortium on Individual Development (CID). Self-control is the ability to enforce appropriate subdominant responses and inhibit inappropriate dominant impulses (Friedman & Mayake, 2004; Nigg, 2017). Self-control is viewed as an effortful, top-down process in behavioral control. It has been related to, inter alia, the dorsal anterior cingulate cortex, dorsolateral prefrontal cortex, and cortical structures (Bridgett et al., 2015). We assessed self-control in 8- to 12-year-old children using the self-control scale (ASCS) in the Achenbach System of Empirically Based Assessment (ASEBA), which was filled in by four different informants: mothers, fathers, teachers and the children themselves. The ASCS was constructed by Willems et al. (2018) based on items of the ASEBA checklists, which are available in parent-, teacher- and self-report versions (Achenbach, Ivanova & Rescorla, 2017; Willems et al., 2018). It is well-established that in completeing questionnaires like the ASEBA scales, different raters have different perspectives, and consequently provide different information (see for example Van der Ende, Verhulst & Tiemeier, 2012). Here, we make use of Bayesian research synthesis to assess hypotheses regarding differences between the raters with respect to the ASCS. We assessed the support for competing hypotheses regarding the ordering of the informants in four CID cohorts: the Netherlands Twin Register (NTR), Tracking Adolescents' Individual Lives Survey (TRAILS), Generation R (GenR), and YOUth, in primary school-aged children aged 8 to 12 years. The competing informative hypotheses and the literature supporting these hypotheses are discussed in the section "Formulation of competing informative hypotheses" below.

Methods

Participants

The participants came from four of the cohort studies that are part of the Consortium on Individual Development: The Netherlands Twin Register (NTR; Bartels et al., 2007; Ligthart et al., 2019), Generation R (GenR; Kooijman et al., 2016), Tracking Adolescents' Individual Lives Survey (TRAILS; Huisman et al., 2008; Oldehinkel et al., 2015), and YOUth (Onland-Moret et al., 2020). The NTR is a national register based in Amsterdam in which twins, other multiples and their families participate. It was established in 1987 and includes children and adults. Children are registered by their parents at birth or any time after birth. About every two years, parents, and, once the children are old enough, teachers and the children themselves, are invited to fill out questionnaires about the children´s health and behavior (Bartels et al., 2007; Ligthart et al., 2019). The NTR sample used in the present study largely overlaps with the sample used by Willems et al. (2018) to develop the ASCS. GenR is a cohort study that follows individuals born in Rotterdam from fetal life to adulthood. Mothers with a delivery date between April 2002 and July 2006 were enrolled in the study. During the primary school years, questionnaires were administered twice (Kooijman et al., 2016). TRAILS concerns a population cohort, established in 2000/2001, which has followed children from the Northern parts of the Netherlands from the age of 11 onwards (Oldehinkel et al., 2015). Finally, YOUth is a prospective cohort study established in 2015. In the primary school years, questionnaires were administered at ages 6, 9 and 12 (Onland-Moret et al., 2020).

During development, children display different levels of behavioral problems (Verhulst & Van der Ende, 1995). The developmental trends may be informant-specific, that is, trends may be characterized by parameters, such as intercept and slope(s), that vary over informants (Van der Ende & Verhulst, 2005). We do not formally test the development of informant differences here, but explore the presence of such differences by defining two age groups: a younger group consisting of 8.5 – 10.5-year-olds and an older age group of 10.5 – 12.5-year-olds. Table 1 breaks down, by cohort, and age group, the number of individuals, number of ASCS observations (total and per informant) mean age, and percentage of boys. As this table shows, in some cohorts, some raters are missing, i.e., there is systematic missingness in the ratings. Self-reports were especially scarce in the younger age group, because pre-adolescents often are not asked to report on their own behavior. Within each age group, the same participant was only

included once. In all cohorts except the TRAILS cohort, the participants could be present in both the younger and the older age group (i.e., given longitudinal designs, children participated repeated at different ages). This does not pose a problem, because the data are analyzed and results are aggregated within age groups only. In case of multiple participants in the same nuclear family (e.g. siblings), we randomly selected one to be included in the analyses

Table 1. Number of ASCS observations, means and standard deviations (SD) of age, and percentage boys per informant, cohort and age.

Age group	Cohort	Mother	Father	Teacher	Self	Mean (SD) age	% boys	Total obs.	Number of individuals (N)*
Younger	NTR	9,904	6,821	6,971	-	9.79 (0.43)	49.7	23,696	12,514
(8.5 –	GenR	4,516	3,269	713	-	9.50 (0.27)	51.6	8,498	4,972
10.5)	TRAILS	232	-	-	252	10.32 (0.13)	49.0	484	259
	YOUth	504	-	201	-	9.47 (0.58)	42.9	705	513
Older	NTR	6,403	4,633	5,355	562	12.08 (0.23)	50.4	16,953	9,095
(10.5 –	GenR	102	90	-	-	11.11 (0.53)	54.0	192	154
12.5)	TRAILS	1,713	-	-	1,935	11.24 (0.52)	49.0	3,648	1,953
	YOUth	139	-	73	-	10.82 (0.20)	47.1	212	140

*Note that there are missing data because but not all participants have data from all available informants. See Table 5 for the samples sizes used in the analyses.

Measures

Self-control was measured using the ASEBA self-control scale (ASCS; Willems, 2018). The ASEBA system includes questionnaires for different informants: the Child Behavior Checklist 6-18 (CBCL) for parents, the Teacher's Report Form (TRF) for teachers, and the Youth Self-Report (YSR) for the children. In these questionnaires, problem behaviors are rated on a three-point scale with the response options *not true (0)*, *somewhat or sometimes true (1),* and *very true or often true (2)*. In all cohorts, the ASCS was administered as part of the entire ASEBA. The content of the eight items in the ASCS are displayed in Table 2. Four items come from the attention problem scale (item 4, 8, 41, and 78), three from the aggressive behavior scale (item 86, 87, and 95), and one from the rule breaking behavior scale (item 28). The sum scores of the ASCS range from 0 to 16. The psychometric properties of the scale are reported in Willems et al.,

2018. The inter-rater reliability for each of the participating cohorts is displayed in Supplementary Table S1. Inter-rater reliability was highest between mother and father ratings, and lowest between self- and mother-ratings. Table 3 contains the ASCS means and standard deviations for each age group, informant and cohort.

Table 2. Items of the ASEBA self-control scale (ASCS).

Item number	Item
4	Fails to finish things he/she starts
8	Can't concentrate, can't pay attention for long
28	Breaks rules at home, school or elsewhere
41	Impulsive or acts without thinking
78	Inattentive or easily distracted
86	Stubborn, sullen or irritable
87	Sudden changes in mood or feelings
95	Temper tantrums or hot temper

Table 3. Means (SD) of the ASEBA self-control scale (ASCS) per informant, cohort, and age group.

		Rater / informant			
		Mother	Father	Teacher	Self
Younger	NTR	3.36 (3.17)	2.88 (2.97)	2.26 (2.93)	-
(8.5 - 10.5)	GenR	2.89 (2.87)	2.94 (2.79)	3.11 (3.66)	-
	TRAILS	4.62 (3.25)	-	-	3.81 (2.85)
	YOUth	4.08 (3.25)	-	2.08 (2.52)	-
Older	NTR	3.01 (3.00)	2.66 (2.86)	2.02 (2.75)	4.21 (3.06)
(10.5 - 12.5)	GenR	3.09 (3.05)	3.64 (2.86)	-	-
	TRAILS	4.65 (3.33)	-	-	3.95 (2.65)
	YOUth	3.92 (3.38)	-	2.41 (3.16)	-

Bayesian evidence synthesis

Bayesian evidence synthesis consists of four steps, which are be explained in detail below. In the first step, informative hypotheses are formulated, based on available literature. The second step is to fit the model of interest in all datasets

separately. In the third step, Bayesian informative hypothesis testing is employed. The fourth and final step involves the actual Bayesian evidence synthesis, in which the support for each hypothesis is aggregated across all cohorts.

Formulation of competing informative hypotheses

Bayesian evidence synthesis starts with a specification of a set of informative hypotheses about the model parameters (Hoijtink, 2012). When formulating informative hypotheses, the inclusion of all plausible hypotheses supported by literature, expert knowledge, or other sources is recommended. Whereas the classical frequentist null hypothesis testing if one or more model parameters deviate significantly from a given value (usually zero), informative hypotheses may also stipulate an ordering of parameters or range constraints.

We formulated competing informative hypotheses based on literature on informant differences in the measurement of self-control. Informants see the children in different contexts (e.g., at school or at home) and may have different relationships with the child. These differences may give rise to differences in perspective on the child's behaviour, and to differences in reference (i.e., a teacher may rate a child relative to other children in the class, whereas a father may rate a child relative to its siblings). Thus, informants have different perspectives on the child's behavior, and may display varying levels of agreement concerning the child's behavior. Several studies have focused on informant differences in problem behaviors, with diverging results. For self-control assessed with the ASCS, Willems et al. (2018) reported the highest average scores for self-reports, followed by, respectively, mother-, father-, and teacher-reports in data from 7- to 16-year-olds in the Netherlands Twin Register (i.e., $\mu_{self} > \mu_{mother} > \mu_{father} > \mu_{teacher}$). Note that their data partly overlap with the NTR data used in the present study. Comparable results were found for the ASEBA total problems scale (Grigorenko, Geiser, Slobodskaya & Francis, 2010; Rescorla et al., 2013; Van der Ende & Verhulst, 2005) attention problems (Bartels et al., 2018), and rule-breaking behaviors, (Bartels et al., 2018; Noordhof, Oldehinkel, Verhulst & Ormel, 2008). With regard to self- and mother-ratings of aggressive problems, Noordhof et al. (2008) reported the opposite pattern (i.e., $\mu_{self} < \mu_{mother}$). Noordhof's sample overlapped with the TRAILS data used in the present study. An alternative hypothesis is that the means of all raters are equal (i.e., $\mu_{self} = \mu_{mother} = \mu_{father} = \mu_{teacher}$). This cannot be ruled out as in most studies the mean differences between the raters were not tested. Thus, based on literature discussed above we formulated the following competing hypotheses, which were evaluated across cohorts:

H1: $\mu_{self} = \mu_{mother} = \mu_{father} = \mu_{teacher}$;

H2: $\mu_{self} > \mu_{mother} > \mu_{father} > \mu_{teacher}$;

H3: $\mu_{self} < \mu_{mother} < \mu_{father} < \mu_{teacher}$.

Hc: complement of H1 – H3; any ordering not specified by the three hypotheses above. This hypothesis is included to test if there is any support for possible configurations of differences in means not included in the set H1 to H3.

Model fitting in each cohort separately

The second step is to fit the model of interest in all datasets separately. That is, we fitted a within-subjects linear model, in which we estimated the mean ASCS sum scores of the informants separately in each cohort and age group.

Bayesian informative hypothesis testing

After specification of the competing informative hypotheses and fitting of the model, the relative support for each of the hypotheses is evaluated for each cohort separately, by means of Bayesian informative hypothesis testing (Hoijtink, 2012). Contrary to the frequentist approach - where only support against the null hypothesis is obtained - the Bayesian approach quantifies support for each of the competing hypotheses, including the null-hypothesis, in terms of posterior model probabilities.

We note that the available data in each cohort determines which components of the hypotheses can be tested. Table 4 contains an overview of which components of each hypothesis are tested in each cohort and age group. For example, the support for H1 in NTR younger age group represents the support for $\mu_{mother} = \mu_{father} = \mu_{teacher}$ only, i.e., does not include childrens' self-reports. Hc, the fail-safe hypothesis capturing orderings not specified by the other hypotheses, can only be tested in cohorts with three or four informants (i.e. GenR in the young age group and NTR in both age groups in the complete case analyses and only in NTR in the analyses based on imputed data), because in cohorts with fewer informants all combinations were covered by the specified hypotheses.

The R package bain (version 0.2.2) was used to compute Bayes Factors to assess the support of two competing hypotheses (Gu, Hoijtink, Mulder & Van Lissa, 2019). For example, a Bayes Factor of $BF_{12} = 10$ means that the support in the data for hypothesis 1 is 10 times greater than the support for hypothesis 2 (Lavine & Schervish, 1999). A priori, all hypotheses were considered equally likely in our

Table 4. Partial hypotheses tested by each cohort, each age group and analysis method.
Complete case analyses

		H1: $\mu_{self} = \mu_{mother} = \mu_{father} = \mu_{teacher}$	H2: $\mu_{self} > \mu_{mother} > \mu_{father} > \mu_{teacher}$	H3: $\mu_{self} < \mu_{mother} < \mu_{father} < \mu_{teacher}$	Hc
Younger	NTR	$\mu_{mother} = \mu_{father} = \mu_{teacher}$	$\mu_{mother} > \mu_{father} > \mu_{teacher}$	$\mu_{mother} < \mu_{father} < \mu_{teacher}$	Yes
	GenR	$\mu_{mother} = \mu_{father} = \mu_{teacher}$	$\mu_{mother} > \mu_{father} > \mu_{teacher}$	$\mu_{mother} < \mu_{father} < \mu_{teacher}$	Yes
	TRAILS	$\mu_{self} = \mu_{mother}$	$\mu_{self} > \mu_{mother}$	$\mu_{self} < \mu_{mother}$	No
	YOUth	$\mu_{mother} = \mu_{teacher}$	$\mu_{mother} > \mu_{teacher}$	$\mu_{mother} < \mu_{teacher}$	No
Older	NTR	$\mu_{self} = \mu_{mother} = \mu_{father} = \mu_{teacher}$	$\mu_{self} > \mu_{mother} > \mu_{father} > \mu_{teacher}$	$\mu_{self} < \mu_{mother} < \mu_{father} < \mu_{teacher}$	Yes
	GenR	$\mu_{mother} = \mu_{father}$	$\mu_{mother} > \mu_{father}$	$\mu_{mother} < \mu_{father}$	No
	TRAILS	$\mu_{self} = \mu_{mother}$	$\mu_{self} > \mu_{mother}$	$\mu_{self} < \mu_{mother}$	No
	YOUth	$\mu_{mother} = \mu_{teacher}$	$\mu_{mother} > \mu_{teacher}$	$\mu_{mother} < \mu_{teacher}$	No
Analyses based on imputed data					
All	NTR	$\mu_{mother} = \mu_{father} = \mu_{teacher}$	$\mu_{mother} > \mu_{father} > \mu_{teacher}$	$\mu_{mother} < \mu_{father} < \mu_{teacher}$	Yes
	GenR	$\mu_{mother} = \mu_{father}$	$\mu_{mother} > \mu_{father}$	$\mu_{mother} < \mu_{father}$	No
	TRAILS	$\mu_{self} = \mu_{mother}$	$\mu_{self} > \mu_{mother}$	$\mu_{self} < \mu_{mother}$	No

study, so were assigned the same prior model probability. Given equal priors, Bayes Factors can be easily translated to posterior model probabilities (PMPs), which express the relative support for each of the tested hypotheses (Kuiper et al., 2012). The closer to zero the PMP of a specific hypothesis is, the less likely it is that the hypothesis is true. The PMPs add up to 1.0 over all hypotheses (Lavine & Chervish, 1999). PMPs were calculated for each cohort individually, so the PMPs express support for the partial hypothesis in each cohort. For example, in the younger age group the PMP of Hypothesis 1 reflects support for $\mu_{mother} = \mu_{father} = \mu_{teacher}$ in NTR, $\mu_{mother} = \mu_{father} = \mu_{teacher}$ in GenR, $\mu_{self} = \mu_{mother}$ in TRAILS, and $\mu_{mother} = \mu_{teacher}$ in YOUth. The hypothesis that received most support was considered to describe the data the best in that cohort and age group. If the PMPs of two hypotheses differed less than 0.1, we judged the hypotheses to be equally likely.

Bayesian evidence synthesis

In the final step, the cohort-specific PMPs are aggregated across cohorts to obtain the posterior model probabilities that represent the relative probability of a hypothesis being supported by all cohorts simultaneously (Kuiper et al., 2012; Zondervan-Zwijnenburg et al., 2019). Hence, the approach adopted makes it possible to compare the quantified support for each hypothesis both within studies, and accumulated over studies. By combining the cohort-specifc PMPs that each represent relative support for different components of a specific hypothesis, the aggregated PMP covers the full hypothesis, because every informant is available in the combined partial hypotheses at least once, there is enough overlap in informants across cohorts, and the cohorts are representative of the same population. For example, in the younger age group the synthesized support for Hypothesis 1 ($\mu_{self} = \mu_{mother} = \mu_{father} = \mu_{teacher}$) represents support for $\mu_{mother} = \mu_{father} = \mu_{teacher}$ in NTR and GenR and for $\mu_{self} = \mu_{mother}$ in TRAILS and for $\mu_{mother} = \mu_{teacher}$ in YOUth. While this is justified statistically, it is important to realize that the overall support represents a combination of different components tested in different cohorts, and that some components (e.g. the comparison between mother- and teacher-reports) are tested in more cohorts than other components. We used equal prior model probabilities for all hypotheses as a starting point for the first cohort. For the subsequent cohorts, the PMP of the previous cohort was used as a prior model probability, until all cohorts were added. The order of updating is irrelevant for the final results. The details of this procedure can be found in Kuiper et al. (2012).

Because larger sample sizes lead to more precision, Bayes Factors based on

larger samples show clearer evidence for or against the hypotheses of interest. This is reflected in greater differences in the PMPs of hypotheses in cohorts with larger sample sizes. This stronger evidence will have a larger impact on the final PMP. The impact of a cohort on the result is thus determined by the strength of the BF, which can be affected by sample size.

In addition to sample size, PMPs of a given hypothesis close to zero also affect the the aggregated results over all cohorts. A hypothesis with a near-zero PMP (i.e., close to zero support) in one or more of the cohorts is likely near zero support in the results aggregated results, even if this hypothesis is well supported by other cohorts (i.e., PMP appreciably greater than zero). This is because the support is used as a multiplier in the updating process. In theory, this is a desirable quality of the method because the goal is to reach robust, broadly supported conclusions. However, the updated results over cohorts may provide a picture that appears to be at variance with the the results of the individual cohorts.

Missing data

In the current study, we had to deal with missing data within and across cohorts and with missing data on the item level and on the sumscore level. There are several ways to deal with missing data. Here we provide an account of what we considered to be the best stragegy to handle the missing data in the present study.

On the item level, we allowed for missingness in three or fewer items. That is, within each cohort, we computed sum scores of the ASCS only if three or fewer items were missing. We used person-mean imputation in calculating the sum scores of a particular person at a particular age per rater (as suggested in Willems et al., 2018).

To handle the missing data at the sumscore level, we used two missing data handling methods, complete case analysis and multiple imputaton, and analyzed the data given both methods. Both methods have their own advantages and disadvantages. We used both methods to establish that our conclusions did not depend on the method used. It is important to distinguish between sumscores that are not available at all in the a certain cohort (for example, self-reports in YOUth), and actual missing data on sumscores that were available in that cohort (for example, a participant for whom mother-report was missing in YOUth). We call the former *systematic* missingness and the latter *incidental* missingness. Here, we applied two methods to handle *incidental* missingness Systematic

missingness does not call for imputation. *Given systematic* missingness (e.g., self-reports in YOUth), we tested the partial hypotheses based on the available data.

In the complete case analysis, also known as listwise deletion, a participant with any missing data was excluded. Depending on the cohort and the age group, this resulted in a reduction of the sample sizes ranging from 12% to 95% and may result in in bias (depending on the exact cause of the *incidental* missingness). On the other hand, this complete case approach enabled us to test our hypotheses in the younger and older age groups separately, thus providing an indication of stability of the results over the two age groups. Furthermore, there was no loss of informants in the complete case analysis, as only participants that had data of all availabe informants for that cohort and age group were included in this method.

The second method was multiple imputation. We adopted this strategy, because we believe that imputation quality cannot be guaranteed when more than half of the data is missing. In case of a percentage of missing data greater than 50%, the ratings of the informant were discarded from further analyses (see the sample sizes per informant relative to the total number of individuals in Table 1). Consequently, the multiple imputation approach included substantially more participants, but fewer informants than the complete case analysis approach (see Table 5). In the YOUth cohort, following this procedure, the remaining data was limited to only one informant, so that the informative hypotheses could not be evaluated in this cohort. In sum, multiple imputation maximized the sample size and reduced the number of partical hypotheses that could be tested. We pooled the data of the two age groups in carrying out multiple imputation to optimize the total number of participants. If we would have decided to impute and analyze the data for the age groups seperately, some of the cohorts would have again included a very small amount of participants. In case a participant had participated repeatedly, we randomly selected one assessment. Multiple imputation was performed using the R-package mice (multiple imputation by chained equations, version 3.7.0; Van Buuren & Groothuis-Oudshoorn, 2011) in R (version 3.6.1; R core team, 2019). Sumscores were imputed for each cohort separately by means of predictive mean matching (Van Buuren, 2018). The predicted value of the target variable was calculated by the specified imputation model. For each missing value, the method identifies a set of donors from the complete cases, who have predicted values closest to the predicted value for the missing value. One of these donors is randomly selected, and the observed value of the donor is used to replace the missing value (van Buren, 2018). Imputations were based on the gender

of the child and the other informants' ASCS scores. An initial predictor matrix for imputation was created based on minimum correlations of 0.20 between all combinations of variables. For each imputation, 15 iterations were performed and missing data points were imputed 50 times (Azur, Stuart, Frangakis & Leaf, 2011). The within-subject lineair regressions were performed on each imputed dataset, and the results pooled by the R-package semTools (version 0.5.2; Jorgensen et al., 2019). The final sample sizes given the two mothods, the complete case analyses and the analyses based on imputed data, are given in Table 5.

Results

The means and sample sizes for the complete case analyses and for the analyses based on imputed data can be found in Table 5.

The top part of Table 6 shows the posterior model probabilities (PMPs) of each hypothesis, within each cohort and age group given the first msising data approach, i.e., the complete case analysis. Note that in all the analyses, each cohort tests a component of the hypotheses of interest, i.e., partial hypotheses. First, we evaluated support for the hypotheses H1, H2 and H3. At age 8.5 - 10.5, the support for the components of hypothesis 2 was the greatest in NTR ($\mu_{mother} > \mu_{father} > \mu_{teacher}$), GenR ($\mu_{mother} > \mu_{father} > \mu_{teacher}$) and YOUth, ($\mu_{mother} > \mu_{teacher}$). In TRAILS, partial hypothesis 3 ($\mu_{self} < \mu_{mother}$) received most support. The aggregated support was greatest for hypothesis 2 ($\mu_{self} > \mu_{mother} > \mu_{father} > \mu_{teacher}$). At age 10.5 - 12.5, the aggregated support was again strongest for hypothesis 2, but there was more variation in support across cohorts. In NTR, which included all four informants at this age, the support for hypothesis 2 was greatest. In GenR, hypothesis 1 ($\mu_{mother} = \mu_{father}$) received most support and in TRAILS, hypothesis 3 ($\mu_{self} < \mu_{mother}$) received most support.

Subsequently, to evaluate any patterns not captured by our informative hypotheses, we evaluated support for any hypothesis other than our hypotheses H1 to H3, we evaluated the support for hypothesis Hc in the cohorts and age groups with at least three informants, i.e., GenR at age 8.5 - 10.5 and NTR at both age groups in the complete case analyses and only in NTR in the analyses based on multiple inputation. In these cohorts, there was little support for the Hc hypothesis (PMP of Hc ≤ 0.001), but for age 8.5-10.5, Hc received most support, with a PMP of 0.738 (Table 7). A post hoc inspection of the mean values in Table 5 suggests that the Hc hypothesis represents the hypothesis $\mu_{mother} = \mu_{father} > \mu_{teacher}$ here.

Table 5. Means (with 95% confidence intervals (CI)) and sample size for the complete case analyses (age groups 8.5-10.5 and 10.5-12.5 years) and for the analyses based on imputed data (ages 8.5-12.5 years).

Complete case analyses

Age	Informant	Mother Mean (95% CI)	Father Mean (95% CI)	Teacher Mean (95% CI)	Self Mean (95% CI)	Sample size
8.5 - 10.5	NTR	3.21 (3.11 – 3.32)	2.85 (2.74 – 2.95)	2.09 (1.99 – 2.18)	-	3,229
	GenR	2.77 (2.41 – 3.13)	2.96 (2.61 – 3.32)	1.89 (1.53 – 2.25)	-	230
	TRAILS	4.63 (4.20 – 5.06)	-	-	3.90 (3.52 – 4.28)	225
	YOUth	3.86 (3.40 – 4.31)	-	2.16 (1.80 – 2.52)	-	192
10.5 - 12.5	NTR	3.27 (2.83 – 3.70)	2.91 (2.50 – 3.31)	1.86 (1.45 – 2.27)	3.96 (3.52 – 4.40)	186
	GenR	3.31 (2.04 – 4.24)	3.33 (2.28 – 4.38)	-	-	38
	TRAILS	4.64 (4.49 – 4.80)	-	-	3.96 (3.83 – 4.08)	1,695
	YOUth	4.28 (3.50 – 5.05)	-	2.43 (1.70 – 3.16)	-	72

Analyses based on imputed data

Age	Informant	Mother Mean (95% CI)	Father Mean (95% CI)	Teacher Mean (95% CI)	Self Mean (95% CI)	Sample size
8.5 - 10.5	NTR	3.21 (3.11 – 3.32)	3.04 (1.96 – 3.11)	2.17 (2.10– 2.45)	-	15,884
	GenR	2.89 (2.80 – 2.99)	3.00 (2.90 – 3.10	-	-	4,778
	TRAILS	4.27 (4.14 – 4.40)	-	-	4.02 (3.91 – 4.13)	2,205

Table 6. Posterior model probabilities (PMPs) of the hypotheses concerning the rank ordering of mean ASCS scores from different informants for the complete case analyses (age groups 8.5-10.5 and 10.5-12.5 years) and for the analyses based on imputed data (ages 8.5-12.5 years).

Complete case analyses

Age 8.5 - 10.5	Informants	H1	H2	H3
NTR	m, f, t	< 0.001	1.000	< 0.001
GenR	m, f, t	< 0.001	1.000	< 0.001
TRAILS	s, m	0.089	< 0.001	0.910
YOUth	m, t	< 0.001	1.000	< 0.001
Aggregated		< 0.001	1.000	< 0.001

Age 10.5 - 12.5	Informants	H1	H2	H3
NTR	s, m, f, t	< 0.001	1.000	< 0.001
GenR	m, f	0.736	0.086	0.178
TRAILS	s, m	< 0.001	< 0.001	1.000
YOUth	m, t	< 0.001	1.000	< 0.001
Aggregated		< 0.001	1.000	< 0.001

Analyses based on imputed data

Age 8.5 - 12.5	Informants	H1	H2	H3
NTR	m, f, t	< 0.001	1.000	< 0.001
GenR	m, f	0.033	< 0.001	0.967
TRAILS	s, m	0.078	< 0.001	0.922
Aggregated		< 0.001	1.000	< 0.001

Note: H1: $\mu_{self} = \mu_{mother} = \mu_{father} = \mu_{teacher}$; H2: $\mu_{self} > \mu_{mother} > \mu_{father} > \mu_{teacher}$; H3: $\mu_{self} < \mu_{mother} < \mu_{father} < \mu_{teacher}$. The aggregated support reflects the support for the combined partial hypotheses.

The bottom part of Table 6 shows the posterior model probabilities for each hypothesis based on the imputed datasets. The general pattern is similar to that of the complete case analyses. Overall, hypothesis 2 again received most support. In NTR, hypothesis 2 ($\mu_{mother} > \mu_{father} > \mu_{teacher}$) received most support. In GenR, hypothesis 3 ($\mu_{mother} < \mu_{father}$) was judged to be the best hypothesis and as

Table 7. Posterior model probabilities for the hypotheses concerning the rank ordering of the mean ASCS scores from different raters, including the catch-all hypothesis (Hc).

Age 8.5 - 10.5	H1	H2	H3	Hc
NTR	< 0.001	1.000	< 0.001	< 0.001
GenR	< 0.001	0.262	< 0.001	0.738
Aggregated	< 0.001	1.000	< 0.001	< 0.001

Age 10.5 - 12.5	H1	H2	H3	Hc
NTR	< 0.001	1.000	< 0.001	< 0.001

Analyses based on imputed data

Age 8.5 - 12.5	H1	H2	H3	Hc
NTR	< 0.001	1.000	< 0.001	< 0.001

was the case for TRAILS ($\mu_{self} < \mu_{mother}$).

Summarizing, we found the strongest evidence for the hypothesis that children themselves report most self-control problems, followed by mothers, fathers and teachers (i.e., H2 $\mu_{self} > \mu_{mother} > \mu_{father} > \mu_{teacher}$). However, we found some inconsistent results across cohorts. The most consistent difference between informants was that parents reported less self-control problems than teachers did. Although this hypothesis (i.e. $\mu_{self} > \mu_{mother} > \mu_{father} > \mu_{teacher}$) received the strongest overall support, it was not the preferred ordering when considering each study separately. Again, it is important to realize that the synthesized result demonstrates which hypothesis is best supported by all cohorts simultaneously, and that this can be different from the hypothesis that is most often preferred within cohorts.

Discussion

The trend towards large-scale collaborative studies involving consortia, such as CID, gives rise to the challenge of combining data from different sources efficiently in a manner that facilitates comprehensive hypothesis testing. Here,

we presented Bayesian evidence synthesis as a method to combine data from different sources and to quantify support for competing informative hypotheses, both within and across cohorts. We illustrated the use of Bayesian evidence synthesis in the situation that different components of the hypotheses were tested in different cohorts.

Overall, our results show most support for the hypothesis that children on average report most problem behaviors, followed by their mothers and fathers, and that on average, teachers report the fewest problems (H2: $\mu_{self} > \mu_{mother} > \mu_{father} > \mu_{teacher}$). The most consistent evidence was found for the conclusion that parents report more self-control problems than teachers. The aggregated findings should be interpreted in relation to the findings within each cohort. Observing different findings across cohorts may call for (post hoc) inspection of the exact differences between the cohorts that gave rise to the inconsistent results. In Bayesian evidence synthesis, we make the assumption that the samples are representative of the same target population, in our case, the population of 8- to 12-year-old Dutch children. In our illustration, the cohorts are all assumed to be selected from the general Dutch population, but differ, for example, in the regions of the Netherlands covered and the periods of data collection. Furthermore, one of the cohorts included twins. It is important to take into account differences between the samples and how these might relate to the concept under investigation when interpreting differences in results. Differences in cohort samples should be evaluated in the light of their relevance with regards to the phenomenon of interest, so the implications of sample differences vary from study to study.

Results from the analyses on the complete cases and on the imputed data favored the same hypothesis. The approaches we used to handle missing data have advantages and disadvantages, but the aggregated results supported the same ordering pattern of means. This indicates that the conclusions about the ordering of the means do not depend on the missing data approach.

The ordering of the sumscores of the different informants was the same in 8.5-10.5-year-olds and 10.5-12.5-years-olds, indicating a constant rank ordering in the two age groups. On the cohort level, the only difference in best supported hypothesis between the younger and older age group concerned GenR. This difference likely is due to the fact that teacher data was available only in the 8.5-10.5 group. A post hoc inspection of the mean differences suggests that H2 (partial hypothesis $\mu_{mother} > \mu_{father} > \mu_{teacher}$) in GenR was likely to be preferred in the younger age group, in view of the big difference in teacher ratings and parent ratings. In the 10.5-12.5 group, only ratings of mothers and fathers were

available, and these differed much less than the differences between parents and teachers. Post hoc inspection of the means suggest that the differences in means between parents are much smaller, hence, H1 (partial hypothesis $\mu_{mother} = \mu_{father}$) receives most support here. Hence, which components of the hypotheses are tested in a specific sample can have an impact on which hypothesis received the most support.

A novel aspect of Bayesian research synthesis is that it can accommodate partial hypotheses given the available data in the cohorts. We illustrated that this method can be used if the information in cohiorts is limited to partial hypotheses, while the synthesized information for all cohorts did address the (complete) hypotheses of interest. In previous studies that used Bayesian research synthesis to combine results over cohorts, all aspects of the hypotheses were tested in all cohorts, even though the measurement instrument might differ (Veldkamp et al., 2020; Zondervan-Zwijnenburg et al., 2019, 2020). Statistically, Bayesian research synthesis is suitable to to assess and combine the support for partial hypothesis. As mentioned above, it is important to interpret the support for each hypothesis in a particular cohort as the support for the particular component of the hypothesis that was actually tested in that cohort. In the present application, combining the support for partial hypotheses with Bayesian evidence synthesis was feasible because there was sufficient overlap between the partial hypotheses that were tested in each cohort. While the different cohorts each addressed only a part of the hypothesized orderings, together the data contained information with regard to all comparisons between informants. Put simply, the present overlap between the partial hypotheses was sufficient to arrive at a comprehensive interpretation of the aggregated PMPs.

Bayesian evidence synthesis has several advantages. One advantage is that this approach, in contrast to meta-analysis, is not influenced by publication bias as it is not dependent on published results (Sutton et al., 2000). Although publication bias may affect the formulation of competing informative hypotheses, the determination of prior model probabilities, and the inclusion of particular datasets, it does not play a role in the actual updating process. If the hypotheses cover all orderings, all hypotheses are considered equally likely a priori, and no datasets are excluded based on published findings, Bayesian evidence synthesis is not affected by publication bias. Furthermore, Bayesian evidence synthesis does not require previous investigations to form hypotheses, as it is equally suitable to address new research questions. Here, we included data of all Dutch cohorts that track children's self-control with the ASCS. As we included a complement

hypothesis (Hc), assigned equal prior model probability to all hypotheses and, to our best knowledge, included all ASCS data collected in the Netherlands, publication bias plays no role in the current study. A disadvantage of Bayesian evidence synthesis is that, contrary to classical meta-analysis, it requires access to the raw data. However, we note that the analysis of individual participant data is more reliable than aggregate data in meta-analysis (Riley, Lambert & Abo-Zaid, 2010).

A major advantage of Bayesian evidence synthesis is that it provides the degree of support for a set of competing hypotheses both at the within-study level and across studies. This highlights inconsistencies between cohorts and allows one to address the robustness of the overall findings (see also, Zondervan-Zwijnenburg et al. 2020). Moreover, the Bayesian approach answers the focal question of which hypothesis is most plausible given the data. Furthermore, new data can be added to the analyses, because the evaluation of the hypotheses depends on posterior model probabilities, and are not affected by order of data entering. So, the results can be updated if additional data become available, facilitating the growth of knowledge by the accumulation of evidence.

A point of attention is that we only specified and tested hypotheses that were supported by literature. In theory, it is possible to specify additional (novel) hypotheses. For example, our results in some cohorts suggest that there might be no meaningful differences in self-control problem scores of mothers and fathers. In future research, we recommend including, for example, $\mu_{self} > \mu_{mother} = \mu_{father} > \mu_{teacher}$, where the ordering between the parents is not of interest.

The differences that we found between informants implies that different informants provide different information concerning self-control. One may wish to calculate self-control scores based on the ratings of all infomants (e.g., an average), but, given the differences between raters, this involves a loss of information. We note that in general one should consider the issue of measurement invariance in the comparison and interpretation of (differences in) test scores. In the present case, the interpretation of the differences between the informants in terms of differences with respect self-control on the conceptual level is based on the tacit, but testable assumption that the self-control test scores are measurement invariant with respect to informant. New datasets, preferably covering parts of the hypotheses that were underrepresented thus far, can easily be added to increase the reliability of the support and accumulate the evidence. Altogether, we feel that Bayesian evidence synthesis is a promising approach to get the most information out of the data available.

Literature

Achenbach, T. M., Ivanova, M. Y., & Rescorla, L. A. (2017). Empirically based assessment and taxonomy of psychopathology for ages 1½–90+ years: Developmental, multi-informant, and multicultural findings. *Comprehensive psychiatry*, *79*, 4-18. doi: https://doi.org/10.1016/j.comppsych.2017.03.006

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work?. *International journal of methods in psychiatric research*, *20*(1), 40-49. doi: https://doi.org/10.1002/mpr.329

Bartels, M., Beijsterveldt, C., Derks, E., Stroet, T., Polderman, T., Hudziak, J., & Boomsma, D. (2007). Young Netherlands Twin Register (Y-NTR): A Longitudinal Multiple Informant Study of Problem Behavior. *Twin Research and Human Genetics,10*(1), 3-11. doi: 10.1375/twin.10.1.3

Bartels, M., Hendriks, A., Mauri, M., Krapohl, E., Whipp, A., Bolhuis, K., ... & Roetman, P. (2018). Childhood aggression and the co-occurrence of behavioural and emotional problems: results across ages 3–16 years from multiple raters in six cohorts in the EU-ACTION project. *European child & adolescent psychiatry*, *27*(9), 1105-1121. doi: https://doi.org/10.1007/s00787-018-1169-1

Bridgett, D. J., Burt, N. M., Edwards, E. S., & Deater-Deckard, K. (2015). Intergenerational transmission of self-regulation: A multidisciplinary review and integrative conceptual framework. *Psychological bulletin*, *141*(3), 602. doi: https://doi.org/10.1037/a0038662

Friedman, N. P., & Miyake, A. (2004). The relations among inhibition and interference control functions: a latent-variable analysis. *Journal of experimental psychology: General*, *133*(1), 101. doi: https://doi.org/10.1037/0096-3445.133.1.101

Grigorenko, E. L., Geiser, C., Slobodskaya, H. R., & Francis, D. J. (2010). Cross-informant symptoms from CBCL, TRF, and YSR: Trait and method variance in a normative sample of Russian youths. *Psychological assessment*, *22*(4), 893. doi: https://doi.org/10.1037/a0020703

Gu, X., Hoijtink, H.J.A., Mulder, J. & Van Lissa, C. J. (2019). bain: Bayes Factors for Informative Hypotheses. R package version 0.2.1. https://CRAN.R-project.org/package=bain

Hoijtink, H. (2012). *Informative hypotheses: Theory and practice for behavioral and social scientists*. CRC Press. doi: https://doi.org/10.1201/b11158

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., Rosseel, Y., Miller, P., Quick, C., ... & Selig, J. (2019). Package 'semTools'.

Kooijman, M. N., Kruithof, C. J., van Duijn, C. M., Duijts, L., Franco, O. H., van IJzendoorn, M. H., ... & Moll, H. A. (2016). The Generation R Study: design and cohort update 2017. *European journal of epidemiology*, *31*(12), 1243-1264. doi: https://doi.org/10.1007/s10654-016-0224-9

Kuiper, R., Buskens, V., Raub, W., & Hoijtink, H. (2012). Combining statistical evidence from several studies: A method using Bayesian updating and an example from research on trust problems in social and economic exchange. *Sociological Methods & Research, 42*, 60–81. doi: https://doi.org/10.1177/0049124112464867

Lavine, M., & Schervish, M. J. (1999). Bayes factors: What they are and what they are not. *The American Statistician*, *53*(2), 119-122. doi: https://doi.org/10.1080/00031305.1999.10474443

Ligthart, L., van Beijsterveldt, C. E. M., Kevenaar, S. T., de Zeeuw, E., van Bergen, E., Bruins, S., … Boomsma, D.I. (2019). The Netherlands Twin Register: Longitudinal research based on twin and twin-

family designs. *Twin Research and Human Genetics*. doi: https://doi.org/10.1017/thg.2019.93

Noordhof, A., Oldehinkel, A. J., Verhulst, F. C., & Ormel, J. (2008). Optimal use of multi-informant data on co-occurrence of internalizing and externalizing problems: the TRAILS study. *International Journal of Methods in Psychiatric Research*, *17*(3), 174-183. doi: https://doi.org/10.1002/mpr.258

Nigg, J. T. (2017). Annual Research Review: On the relations among self-regulation, self- control, executive functioning, effortful control, cognitive control, impulsivity, risk- taking, and inhibition for developmental psychopathology. *Journal of child psychology and psychiatry*, *58*(4), 361-383. doi: https://doi.org/10.1111/jcpp.12675

Oldehinkel A. J., Rosmalen J. G. M., Buitelaar J. K., Hoek H. W., Ormel J., Raven, D., …, Hartman C. A. (2015). Cohort profile update. The TRacking Adolescents' Individual Lives Survey (TRAILS) . *International Journal of Epidemiology*, *44*(1), 76-76n. doi: https://doi.org/10.1093/ije/dyu225

Onland-Moret, N. C., Buizer-Voskamp, J. E., Albers, M. E., Brouwer, R. M., Buimer, E. E., Hessels, R. S., ... & Kemner, c. (2020). The YOUth study: rationale, Design, and study procedures. *Developmental cognitive neuroscience*, *46*, 100868. doi: https://doi.org/10.1016/j.dcn.2020.100868

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). doi: https://doi.org/10.1126/science.aac4716

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.r-project.org

Rescorla, L. A., Ginzburg, S., Achenbach, T. M., Ivanova, M. Y., Almqvist, F., Begovac, I., ... & Döpfner, M. (2013). Cross-informant agreement between parent-reported and adolescent self-reported problems in 25 societies. *Journal of Clinical Child & Adolescent Psychology*, *42*(2), 262-273. doi: https://doi.org/10.1080/15374416.2012.717870

Riley, R. D., Lambert, P. C., & Abo-Zaid, G. (2010). Meta-analysis of individual participant data: rationale, conduct, and reporting. *Bmj*, *340*, c22. doi: https://doi.org/10.1136/bmj.c221

Rosenthal, R., & DiMatteo, M. R. (2002). Meta-analysis. *Stevens' handbook of experimental psychology*. doi: https://doi.org/10.1002/0471214426.pas0410

Sutton, A. J., Duval, S. J., Tweedie, R. L., Abrams, K. R., & Jones, D. R. (2000). Empirical assessment of effect of publication bias on meta-analyses. *Bmj*, *320*(7249), 1574-1577. doi: https://doi.org/10.1136/bmj.320.7249.1574

Van Buuren, S.V. (2018). *Flexible Imputation of Missing Data (2nd ed.)*. CRC Press. doi: https://doi.org/10.1201/9780429492259

Van Buuren, S. & Groothuis-Oudshoorn, K. (2011) Mice: multivariate imputation by chained equations in R. *J Stat Softw*, *45*(3), 1-67. doi: https://doi.org/10.18637/jss.v045.i03

Van der Ende, J., & Verhulst, F. C. (2005). Informant, gender and age differences in ratings of adolescent problem behaviour. *European child & adolescent psychiatry*, *14*(3), 117- 126. doi: https://doi.org/10.1007/s00787-005-0438-y

Van der Ende, J., Verhulst, F. C., & Tiemeier, H. (2012). Agreement of informants on emotional and behavioral problems from childhood to adulthood. *Psychological assessment*, *24*(2), 293. doi: https://doi.org/10.1037/a0025500

Verhulst, F. C., & Van der Ende, J. (1995). The eight-year stability of problem behavior in an epidemiologic sample. *Pediatric research*, *38*(4), 612. doi: https://doi.org/10.1203/00006450-199510000-00023

Veldkamp, S.A.M. , Zondervan-Zwijnenburg, M.A.J., van Bergen, E., Barzeva, S.A., Tamayo Martinez, N., Becht, A.I., Van Beijsterveldt, C.E.M., Meeus, W., Branje, S., Hillegers, M.H.J., Oldehinkel, A.J., Hoijtink, H.J.A., Boomsma, D.I., Hartman, C. (2020). Effect of parental age on their children's neurodevelopment. doi: https://doi.org/10.1080/15374416.2020.1756298

Willems, Y. E., Dolan, C. V., van Beijsterveldt, C. E., de Zeeuw, E. L., Boomsma, D. I., Bartels, M., & Finkenauer, C. (2018). Genetic and environmental influences on self-control: Assessing self-control with the ASEBA self-control scale. *Behavior genetics*, *48*(2), 135-146. doi: https://doi.org/10.1007/s10519-018-9887-1

Zondervan-Zwijnenburg, M. A. J., Richards, J. S., Kevenaar, S. T., Becht, A. I., Hoijtink, H. J. A., Oldehinkel, A. J., ... & Boomsma, D. I. (2020). Robust longitudinal multi-cohort results: The development of self-control during adolescence. *Developmental Cognitive Neuroscience*, 100817. doi: https://doi.org/10.1016/j.dcn.2020.100817

Zondervan-Zwijnenburg, M.A.J., Veldkamp, S.A.M., Neumann, A., Barzeva, S.A., Nelemans, S.A., Van Beijsterveldt, C.E.M. Branje, S., Meeus, W.H.J., Hillegers, M.H.J., Tiemeier, H., Hoijtink, H.J.A., Oldehinkel, A.J., & Boomsma, D.I. (2019). The impact of parental age on child behavior problems: Updating evidence from multiple cohorts. *Child Development, 91*(3), 964-982. doi: https://doi.org/10.1111/cdev.13267

Supplementary Material

Inter-rater reliability

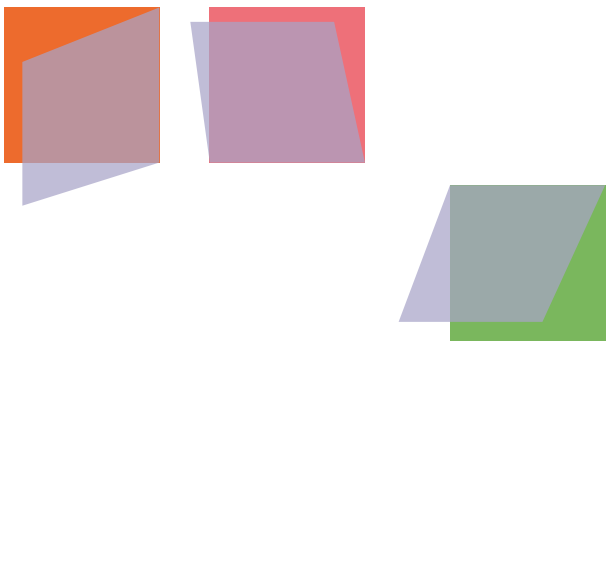
In addition to comparing mean problem levels, we evaluated whether the reported rank ordering of the children was comparable across informants. The inter-rater correlations were calculated by means of intra-class correlations (ICC 1, Shrout & Fleis, 1979). These are displayed in Supplementary Table S1.

Table S1. Intraclass correlations between the informants.

Age	Cohort	Informant	NTRS	F	T	GenRF	T	TRAILSS	YOUthT
8.5 - 10.5	NTR	M		0.70	0.38				
		F			0.37				
	GenR	M				0.66	0.44		
		F					0.44		
	TRAILS	M						0.26	
	YOUth	M							0.41
10.5 - 12.5	NTR	M	0.37	0.68	0.37				
		F	0.32		0.35				
		T	0.08						
	GenR	M				0.66			
	TRAILS	M						0.24	
	YOUth	M							0.49

Chapter 4

Self-control and grit are associated with school performance mainly because of shared genetic effects



Published as: **Kevenaar, S. T.**, Dolan, C. V., Boomsma, D. I., & van Bergen, E. (2023). Self-control and grit are associated with school performance mainly because of shared genetic effects. *JCPP Advances*, e12159.

Abstract

Background: By combining the classical twin design with regression analysis, we investigated the role of two non-cognitive factors, self-control and grit, in the prediction of school performance. We did so at the phenotypic, genetic, and environmental level. **Methods**: Teachers filled out a survey on the twins' school performance (school grades for reading, literacy, and math), self-control (ASEBA self-control scale), and grit (the perseverance aspect) for 4,891 Dutch 12-years-old twin pairs (3,837 pairs with data for both and 1,054 pairs with data for one of the twins). We employed regression analyses to first assess the contributions of self-control and grit to school performance at the phenotypic level, and next at the genetic and environmental level, while correcting for rater (teacher) effects, parental SES, and sex. **Results**: Higher SES was associated with better school performance, self-control, and grit. On average, girls had more self-control and grit than boys. Corrected for sex, SES, and teacher rater effects, genetic factors accounted for 74%, 69%, and 58% of the phenotypic variance of school performance, self-control, and grit, respectively. Phenotypically, self-control and grit explained 28.3% of the variance in school performance. We found that this phenotypic result largely reflected genetic influences. **Conclusions**: Children who have better self-control and are grittier tend to do better in school. Individual differences in these three traits are not correlated because of shared environmental influences, but mainly because of shared genetic factors.

Keywords: school performance, academic achievement, self-control, grit, heritability, non-cognitive skills

Abbreviations: SC: self-control); SP: school performance); ASEBA: Achenbach System of Empirically Based Assessment; ASCS: ASEBA self-control scale; A: Additive genetic factors; C: Common environmental effects; D: Dominant genetic factor (D); E: Unique (unshared) environmental effects; NTR: Netherlands Twin Register; SES: Socio-economic status; MZ: monozygotic; DZ: dizygotic (DZ); -2LL: -2*log likelihood (-2LL); GWAS: genome-wide association study; 95CI: 95% Confidence Interval.

Introduction

Understanding individual differences in school performance is important given the large influence they have across all domains of life. Cognitive variables, such as intelligence, are important for school success, but these only explain part of the individual differences (Kautz, Heckman, Diris, Ter Weel, & Borghans, 2014; Bartels, Rietveld, van Baal & Boomsma, 2002). Here, we considered the role of non-cognitive factors. Two such factors that have been related to school success are self-control and grit. Self-control is defined as the "capacity to resist temptation or inhibit a dominant response or activate a subdominant response" (Nigg, 2017, p. 364). Grit is defined as perseverance and passion for long-term goals (Duckworth, Peterson, Matthews, & Kelly, 2007). Grit has two aspects: consistency of interest and perseverance of effort (Duckworth et al., 2007). Of these, perseverance of effort is more strongly linked to school performance (Credé, Tynan, & Harms., 2017; Muenks, Wigfield, Yang, & O'Neal, 2017; Rimfeld, Kovas, Dale, & Plomin, 2016). The grit measure that we analyzed in this study mostly relates to this perseverance of effort aspect, especially as manifest in the classroom setting. We investigated the differential prediction by self-control and grit of individual differences in school performance of 12-year-olds in whom we collected data on these measures from their schoolteachers in a prospective study design. We analyzed the relationship between school performance and non-cognitive factors both at the phenotypic level and at the genetic and environmental levels.

Self-control and grit are distinct, but strongly correlated concepts (r ~.60; Duckworth & Gross, 2014; Duckworth et al., 2007). Grit entails persistent focused effort and long-term commitment to goals, whereas self-control encompasses the capacity to regulate attention, emotion, and behavior in the presence of distractions and temptations (Duckworth & Gross, 2014; Duckworth et al., 2007). Self-control keeps one focused on a task at hand and is required in (and outside) the school context. Grit involves making appropriate choices to reach a long-term goal. So, grit is needed to persevere in working towards a higher-order long-term goal, while self-control is needed to resist short-term distractions and temptations (Duckworth & Gross, 2014). Empirical studies have demonstrated the associations between self-control, grit, and other non-cognitive factors, like the Big Five personality trait conscientiousness (Credé et al., 2017; Muenks et al., 2017). Conscientiousness can be defined as being "self-disciplined, responsible, hardworking and thorough" (John & Srivastava, 1999). Werner, Milyavskaya, Klimo, and Levine (2019) showed that self-control, grit, and conscientiousness explained 10% of the variance in academic motivation.

Multiple studies have documented that school performance is correlated with self-control, grit, and conscientiousness. These three non-cognitive skills are overlapping constructs and poorly distinguishable (Muenks et al., 2017; Ponnock et al., 2020; Takahashi et al., 2021). Duckworth et al. (2014, 2019) showed that performance on standardized achievement tests administered at school was predicted by non-cognitive skills like self-control, motivation, and study strategies, in addition to socioeconomic status and general intelligence. Oriol, Miranda, Oyanedel, & Torres (2017) showed in primary school children that grit is related to academic self-efficacy, while self-control is related to school satisfaction. Usher, Li, Butz, and Rojas (2019) found that grit correlated modestly with self-efficacy (r ~.50), but weaker with teacher ratings in reading and math (r ~.20), and with achievement test scores (r ~.10). Self-efficacy was weakly to moderately related to all outcomes (r ~.30). Of note is that a meta-analysis confirmed that of the two facets of grit, perseverance of effort and consistency of interest, the perseverance facet is much more strongly related to academic performance (ρ = .26) than the consistency facet (ρ = .10; Credé et al., 2017). We therefore focus on perseverance.

Cognitive skills, school performance, and education-related traits are heritable, with genetic differences being the main source of individual differences. That is, for most societies that have been included in behavior genetic studies. In 12-year-old children in The Netherlands, the estimated heritability (i.e., the proportion of variance attributable to genetic influences) of standardized-test performance at the end of primary school is 74%. That is, 74% of test-score differences among children are due to genetic differences. Only 8% of individual differences were accounted for by shared-environmental influences (de Zeeuw, van Beijsterveldt, Glasner, de Geus, & Boomsma, 2016). Shared-environmental influences common to children growing up in the same family contribute to the resemblance of twins and siblings. A recent meta-analysis of twin studies found self-control to be 60% heritable (Willems, Boesen, Li, Finkenauer, & Bartels, 2019). Interestingly, the 40% environmental effects on self-control were not shared by twins. This may imply that the environmental effects do not originate in aspects of the rearing environment that are likely to be shared, such as parental upbringing or parental style, but experiences unique to each sibling, stemming from, for instance, illness, different friends, and stochastic influences (Tikhodeyev & Shcherbakova, 2019; Willems et al., 2019). The heritability of grit has been estimated at 35-61%, and like self-control, grit shows no evidence of shared environmental effects (Martinez, Holden, Hart, & Taylor, 2022; Rimfeld et al., 2016; Tucker-Drob, Briley, Engelhardt, Mann, & Harden, 2016). Martinez et al. (2022) investigated grit and

mindset in relation to reading comprehension in 422 thirteen- and fifteen-year-old twin pairs. Individuals can hold the belief that intelligence is mainly a fixed inborn trait (fixed mindset) or a malleable trait given effort and time (growth mindset). Grit and mindset were correlated with reading ability, but mindset and grit were not associated with the change in reading ability over time (Martinez et al., 2022). In a review, Malanchini, Rimfeld, Allegrini, Ritchie, and Plomin (2020) concluded that non-cognitive abilities explained genetic variance in academic performance above and beyond cognitive ability. The strong stability and heritability of academic performance appear to be partially driven by additional factors besides cognitive ability.

In conclusion, school performance, self-control, and grit are related traits that are subject to genetic influences, and self-control and grit predict school performance. Here we set out to determine the degree to which genetic and environmental factors contribute to the phenotypic relationships between these non-cognitive factors and school performance. We addressed this question by applying simultaneous regression and genetic covariance modeling, as outlined in Boomsma, van Beijsterveldt, Odintsova, Neale, and Dolan (2021).

Analyzing twin data, we can distinguish genetic and environmental sources of variation. We analyzed data that were collected from the teachers of children. Teachers assessed school performance across three domains and assessed self-control and grit. This feature of the data poses a challenge: twins in the same class were rated by the same teacher, while twins in different classes (or schools) were rated by different teachers. As teachers may have their unique views of children, and their style of rating them, in our models we included random teacher-rater effects. In so doing, we distinguished variance due to raters and variance due to child factors. We included sex and parental SES as covariates (Gil-Hernández, 2021). So, we accounted for the random effect of rater and the fixed effects of sex and parental SES.

Methods

Participants

We included data from 11.5 to 12.5-year-old twins registered in the young Netherlands Twin Register (NTR). The young NTR includes twins and multiples born in 1986, and their parents, siblings, and teachers, who participate in longitudinal research (Boomsma et al., 2006). More information about data collection, recruitment, and response rates can be found elsewhere (Van Beijsterveldt et al., 2013; Ligthart et al, 2019). After young twins are registered by their parents, usually a few weeks to months after birth, the parents are approached when the twins are 12 years old with a request for permission to approach their teachers for ratings of behavior in school and school performance. Parents, who grant permission, then provide the name of the teacher and the address of the school. Teachers are subsequently invited to complete a survey concerning the twin(s) in their class.

Our sample included 3,837 pairs with data on both twins and 1,054 incomplete pairs, that is, pairs with data on one twin member. These incomplete pairs arose because some of the twins were in different classes and rated by different teachers, so the teacher of one twin might have completed the survey, but the teacher of the other did not. There were 1,957 monozygotic and 2,934 dizygotic twin pairs with school performance, self-control, and/or grit measures. The zygosity of the same-sex twin pairs was determined by a DNA test (32.2% of the same-sex pairs) or by a questionnaire with items concerning the twin resemblance, which the parents completed. Based on this questionnaire, zygosity is correctly determined in over 95% of the cases (Ligthart et al., 2019). The data collection procedure was ethically approved by the Vaste Commissie Wetenschap en Ethiek at Vrije Universiteit Amsterdam (VCWE-2021-111).

Measures

Self-control

Self-control was assessed by the Achenbach Self-Control Scale (ASCS; Willems et al., 2018) in the ASEBA-TRF reported by teachers (Achenbach, 2001). The scale consists of 8 items, displayed in Table 1, scored on a 3-point response scale. The response options are 0 (*not true*), 1 (*somewhat or sometimes true*), and 2 (*very true or often true*). Cronbach's α of the ASCS is .82 for teacher reports at age 12. The internal inter-rater and test-retest reliability are good (Willems et

al., 2018). If three or fewer items were missing (34.5% of the sample due to ASEBA-TRF version changes over the years), the mean of the available items was substituted for the missing items to compute the sum score, as described by Willems et al. (2018). If more than three items were missing the sum scores was coded as missing. We reverse-coded the item scores so that a higher score indicated greater self-control. The total score ranged from 0-16.

Grit

The grit measure was based on teacher reports on two or three items, namely *Compared to typical pupils of the same age, 1) how hard does he/she work*; *2) how appropriately does he/she behave*, and *3) how task-oriented is he/she*. The response format was a 7-point Likert scale. Due to changes in YNTR surveys over the years, the third item was missing in 55.2%. The item scores were summed to sum scores. If more than one item was missing, the grit score was coded as missing. If a single item was missing (mostly item 3), the mean of the other two items was imputed for the missing item. The sum scores range from 1-21, with higher scores indicating more grit. The correlations among the grit items are .70 (items 1 and 2), .71 (items 2 and 3) and.82 (items 1 and 3). The correlation between the grit measure as used in the paper and the two most relevant items for grit is high (0.88; see Table 2) and justifies our use of the measure. Cronbach's α is .87.

Table 1. The Achenbach Self-Control Scale items to assess self-control problems and the items to assess grit.

Self-Control Items	Grit items
Fails to finish things he/she starts	Compared to typical pupils of the same age:
Can't concentrate, can't pay attention for long	- how hard does he/she work?
Breaks rules at home, school or elsewhere	- how appropriately does he/she behave?
Impulsive or acts without thinking	- how task-oriented is he/she?
Inattentive or easily distracted	
Stubborn, sullen or irritable	
Sudden changes in mood or feelings	
Temper tantrums or hot temper	

Note. For self-control we reversed the scores (higher scores indicating more self-control). Self-control items were scores on a 3-point scale and grit items on a 7-point scale.

School performance (grades and CITO standardized test)

Teachers reported the grades for math, reading, and literacy on 5-point scales, with scale points 1 (*fail*), 2 (*poor*), 3 (*satisfactory*), 4 (*above average*), and 5 (*good or excellent*) (de Zeeuw et al., 2014; van Bergen et al., 2018). The responses to these three items were summed, as detailed in de Zeeuw et al. (2016). School performance scores ranged from 3 to 15, with higher scores indicating better performance. If a single rating was missing (22.9% of the cases), the mean of the other two ratings was substituted. If more than one item was missing, the school performance score was coded as missing. Reading grades correlated .73 with literacy grades and .51 with math grades. Literacy grades and math grades correlated .67 (see Table 2).

In about half of the twins (N = 4,723 individual children), we had scores on a nationwide standardized educational-achievement test (i.e., CITO scores; Centraal Instituut voor Toets Ontwikkeling, 2002; de Zeeuw et al., 2020), to validate teacher-reported school performance. The CITO is a high-stakes test at the end of primary school (Grade 6; ages 11 or 12) that is taken at school over three mornings. CITO scores correlated highly with the reported school grades and with the sum score, our measure of school performance (Table 2). Both teacher reports and test scores are heritable, reliable, and predictive of future academic achievement (van Bergen et al., 2018; Rimfeld et al., 2019). When we refer to school performance in this paper, we refer to the sum of the teacher-reported school grades, because this measure was available for most children and overall and the correlation with the standardized CITO test was high (0.70; see Table 2).

Sex and socioeconomic status (SES*)*

Sex was coded 1 for males and 2 for females. SES was based on a combination of parental occupation and parental education (for details, see de Zeeuw et al., 2019), and was coded 1 (*lowest SES*) through 4 (*highest SES*).

Teacher sharing

Twins may or may not be in the same class. Twins in the same class were rated by the same teacher, while twins in different classes were rated by different teachers. This sometimes resulted in incomplete pairs, where one teacher participated in the study and the other did not. Teacher sharing was coded 1 (twins in the same class, rated by the same teacher) or 0 (different classes, different teachers).

IQ

A subsample of 421 children was assessed on full-scale IQ, by the full Dutch WISC-R (van Haasen et al., 1986). There are 12 subscales, of which half focus on verbal and the other half focus on non-verbal IQ. For a detailed description of these data, see the age-12 assessment in Bartels, Rietveld, van Baal, and Boomsma (2002). The IQ measure in the subsample allowed us to test if our non-cognitive skills predict school performance over and above IQ.

Statistical Analyses

We started with testing, in the IQ subsample, whether our non-cognitive factors explain variance in school performance over and above the cognitive factor IQ. Then we moved on to our main analyses.

To assess the differential relationship of self-control and grit with school performance, we first carried out phenotypic regression analysis, followed by genetic and environmental regression analyses (Boomsma et al., 2021) to determine the contributions of self-control and grit, and their covariance, to the variance in school performance at the phenotypic, genetic, and environmental level. The data were negatively skewed because of a ceiling effect, hence we corrected for censoring in all analyses (see de Zeeuw et al., 2019). We fitted the models using full information maximum likelihood estimation, assuming that the data follow a censored multivariate normal distribution.

First, we carried out phenotypic regression analyses, in which we regressed school performance (SP) on self-control (SC) and grit, and on the covariates sex, SES, and teacher sharing (t, coded 0/1).

$$SP_i = b_0 + b_{sex} * sex_i + b_{SES} * SES_i + t_i + b_{SC} * SC_i + b_{grit} * grit_i + e_i,$$

with subscript i representing individual, b_0 representing the intercept, and e_i representing prediction error. The term t_i is the random teacher effect. Conditional on sex, SES, and teacher sharing, the phenotypic school performance variance was decomposed into four parts:

$$S^2_{SP|sex,SES,teacher} = b^2_{SC} * S^2_{SC} + b^2_{grit} * S^2_{grit} + (2 * b_{SC} * b_{grit} * S_{SC,grit}) + S^2_e$$

The term $2*b_{SC}*b_{grit}*S_{SC,grit}$, due to the covariance of self-control and grit ($S_{SC,grit}$), captures variance that cannot be unambiguously attributed to either self-control or grit. We fitted the same regression model simultaneously to the data of all MZ and DZ twins, taking into account that the scores within twin pairs are dependent. The left side of Figure 1 displays the phenotypic model.

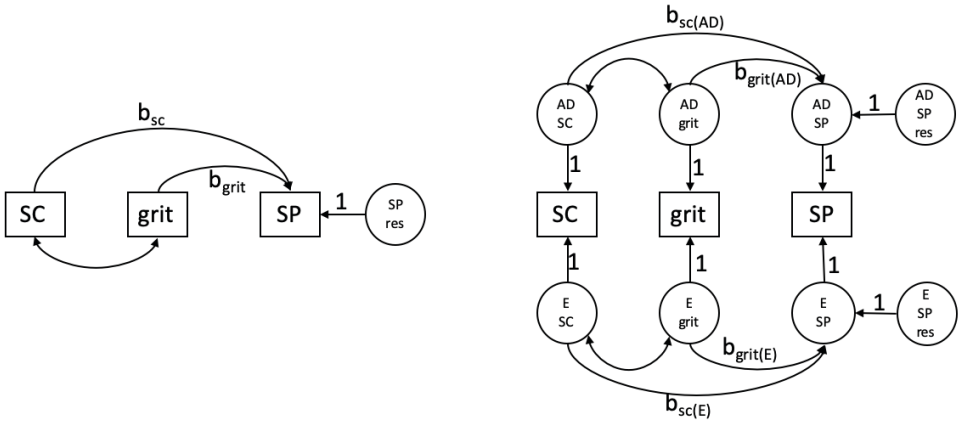


Figure 1. Figure on the left side: path diagram of the phenotypic regression model of school performance on self-control and grit, conditional on sex and SES and teacher (omitted in the figure). b_{SC} and b_{grit} represent the regression coefficients of self-control and grit respectively. Figure on the right side: path diagram of the regression of school performance on self-control and grit including the genetic (AD) and environmental (E) latent factors. The parameters $b_{SC(AD)}$, $b_{grit(AD),}$ $b_{SC(E),}$ and $b_{grit(E)}$ represent regression coefficients. AD SP res represents the residual genetic term of school performance and E SP res represents the residual environmental term of school performance.

Next, we fitted a genetic structural equation model (Figure 2). In earlier research, self-control and grit were found to be influenced by additive genetic effects and genetic dominance effects. Finding genetic dominance implies non-additive genetic effects of certain alleles (for an in-depth explanation, see Falconer & Mackay, 1983). By the common rule of thumb, we infer dominance if $r_{MZ} > 2*r_{DZ}$, where r_{MZ} and r_{DZ} are the MZ and DZ twin correlations. Given the twin correlations in Figure 4, we fitted a model including additive genetic effects (A), dominance effects (D), and unshared environmental effects (E). As shown below, we calculated the MZ and DZ covariance matrices based on the estimated additive genetic S_A and the dominance S_D covariance matrices (the dominance effects limited to self-control and grit), and the unshared environmental S_E covariance

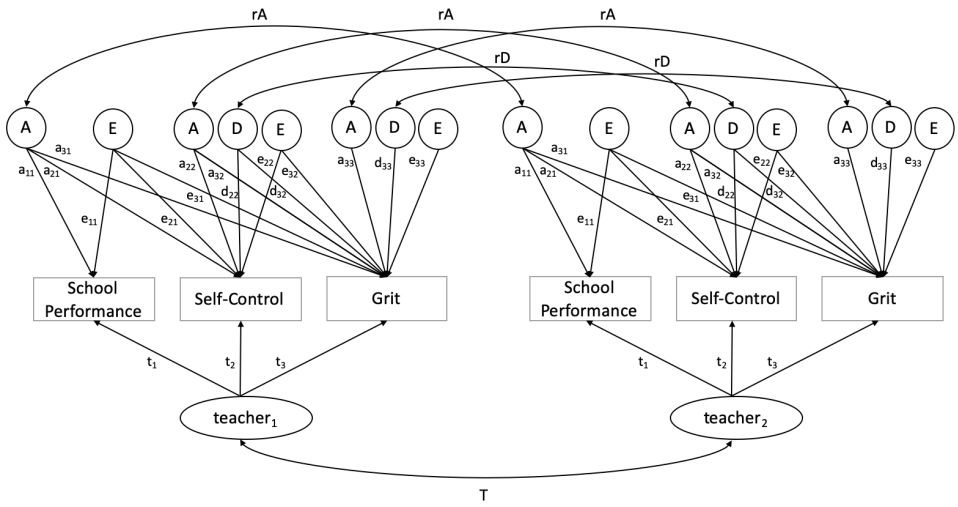


Figure 2. Path diagram of the genetical structural equation model, with twin 1 of a pair displayed on the left side of the figure and twin 2 of the same pair displayed on the right. rA denotes the correlation between the A factors in twin 1 and twin 2 and is fixed to 1 in MZ twins and to 0.5 in DZ twins. rD denotes the correlation between the D factors in twin 1 and twin 2 and is fixed to 1 in MZ twins and to 0.25 in DZ twins. T denotes teacher sharing and is fixed to 1 for twins who share a teacher and 0 for twins who do not share a teacher. The covariates SES and sex are omitted from this figure.

matrix. We included the covariance matrix S_T to accommodate possible rater (teacher) variance (see below). The 3x3 additive genetic covariance matrix S_A, the 3x3 covariance matrix S_D, and the 3x3 unshared environmental covariance matrix S_E were modeled using triangular decomposition, as $S_A = \mathbf{\Lambda}_A\mathbf{\Lambda}_A{}^t$, $\mathbf{\Lambda}_D\mathbf{\Lambda}_D{}^t$ and $S_E = \mathbf{\Lambda}_E\mathbf{\Lambda}_E{}^t$, respectively, where

$$\Lambda_A = \begin{array}{c} \\ \text{School performance} \\ \text{Self} - \text{control} \\ \text{Grit} \end{array} \begin{array}{c} \begin{array}{ccc} \text{SP} & \text{SC} & \text{Grit} \end{array} \\ \begin{bmatrix} a_{11} & 0 & 0 \\ a_{21} & a_{22} & 0 \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \end{array}$$

$$\Lambda_D = \begin{array}{c} \\ \text{School performance} \\ \text{Self} - \text{control} \\ \text{Grit} \end{array} \begin{array}{c} \begin{array}{ccc} \text{SP} & \text{SC} & \text{Grit} \end{array} \\ \begin{bmatrix} d_{11} & 0 & 0 \\ d_{21} & d_{22} & 0 \\ d_{31} & d_{32} & d_{33} \end{bmatrix} \end{array}$$

and

$$\Lambda_E = \begin{matrix} \\ \text{School performance} \\ \text{Self} - \text{control} \\ \text{Grit} \end{matrix} \begin{matrix} \begin{matrix} \text{SP} & \text{SC} & \text{Grit} \end{matrix} \\ \begin{bmatrix} e_{11} & 0 & 0 \\ e_{21} & e_{22} & 0 \\ e_{31} & e_{32} & e_{33} \end{bmatrix} \end{matrix}$$

Lastly, the random teacher-rater effect was modeled using the 3x3 covariance matrix S_T, which was modeled as $S_T = \mathbf{\Lambda}_T\mathbf{\Lambda}_T^t$, where

$$\Lambda_T = \begin{matrix} \\ \text{School performance} \\ \text{Self} - \text{control} \\ \text{Grit} \end{matrix} \begin{bmatrix} t_1 \\ t_2 \\ t_3 \end{bmatrix}$$

Here the teacher rater effect is treated as a random variable giving rise to the variances t_1^2, t_2^2, and t_3^2, and covariances among the phenotypes (t_1*t_2, t_1*t_3, t_2*t_3). We included SES and sex as fixed covariates, so the expected MZ and DZ covariance matrices conditional on sex and SES are:

$$\Sigma_{MZ} = \begin{matrix} \\ \text{MZ twin 1} \\ \text{MZ twin 2} \end{matrix} \begin{matrix} \begin{matrix} \end{matrix} \\ \begin{bmatrix} \Sigma_A + \Sigma_D + \Sigma_E + \Sigma_T & \Sigma_A + \Sigma_D + T * \Sigma_T \\ \Sigma_A + \Sigma_D + T * \Sigma_T & \Sigma_A + \Sigma_D + \Sigma_E + \Sigma_T \end{bmatrix} \end{matrix}$$

and

$$\Sigma_{DZ} = \begin{matrix} \\ \text{DZ twin 1} \\ \text{DZ twin 2} \end{matrix} \begin{matrix} \begin{matrix} \text{DZ twin 1} & \text{DZ twin 2} \end{matrix} \\ \begin{bmatrix} \Sigma_A + \Sigma_D + \Sigma_E + \Sigma_T & \frac{1}{2}\Sigma_A + \frac{1}{4}*\Sigma_D + T * \Sigma_T \\ \frac{1}{2}\Sigma_A + \frac{1}{4}*\Sigma_D + T * \Sigma_T & \Sigma_A + \Sigma_D + \Sigma_E + \Sigma_T \end{bmatrix} \end{matrix}$$

The fixed parameter T (coded 0 or 1) indicates whether the twins share the teacher (T=1) or not (T=0).

Third, we carried out the regression analysis at the level of the genetic and environmental covariance matrices to obtain which of the non-cognitive factors, self-control, grit, or their covariance, was the better predictor of school performance. We included the regression of school performance on self-control and grit at the level of the total genetic $\Sigma_{G,}$ where Σ_G equals $\Sigma_A+\Sigma_D$ and the environmental covariance matrix Σ_E,. The decomposition of the genetic variance, conditional on sex, SES, and teacher sharing, is

$$\Sigma_{G\_school\ performance|sex,SES,teacher}^2 =$$
$$b_{G\_SC}^2 * \Sigma_{G\_SC}^2 + b_{G\_grit}^2 * \Sigma_{G—grit}^2 + (2 * b_{G\_SC} * b_{G—grit} * \Sigma_{G\_SC,grit}) + \Sigma_{G\ e}^2$$

where $\Sigma_{G\ e}^2$ is the genetic prediction error variance and b_{G_SC} and b_{G_grit} are the genetic regression coefficients. The decomposition of environmental variance is

$$\Sigma_{E\_school\ performance|sex,SES,teacher}^2 =$$
$$b_{E\_SC}^2 * \Sigma_{E\_SC}^2 + b_{E\_grit}^2 * \Sigma_{E—grit}^2 + (2 * b_{E\_SC} * b_{E—grit} * \Sigma_{E\_SC,grit}) + \Sigma_{E\ e}^2$$

where $\Sigma_{E\ e}^2$ is the environmental prediction error variance and b_{E_SC} and b_{E_grit} are the environmental regression coefficients.

Statistical analyses were conducted using the OpenMx library (Neale et al., 2016) in R using full information maximum likelihood estimation. We fitted the full model with parameters accommodating the teacher-rater effect (i.e., the parameters in $\mathbf{\Lambda}_T$) estimated freely.

So, in summary, we first fitted a regression to the phenotypic data and then we fitted a regression on the (A+D) and E covariance matrices. The left side of Figure 1 represents the phenotypic regression model, in which self-control and grit predict school performance. In this model the R^2, the proportion of phenotypic school performance variance explained is decomposed into three parts: a part directly due to self-control, a part directly due to grit, and a part due to self-control and grit together. The third part involves the covariance of self-control and grit and therefore cannot be attributed to self-control or grit exclusively.

The right side of Figure 1 presents the (A+D), E regression model, in which we specify the regression relationship at the level of the total genetic covariance matrix (A+D) (comprising the additive genetic and dominance covariance). First, we calculated the R^2 of the (A+D) variance of school performance, the R^2 of the E variance of school performance. Second, we calculated the decomposition of the phenotypic school performance variance based on the A+D results and on the E results. Here we expressed the R^2 of the phenotype school performance in terms of the R^2 (with three components: a part directly due to grit, a part directly due to self-control and a part due to their covariance) based on the A+D regression and the R^2 (again with the same three components) based on the E regression. This allowed us to determine the contribution of A and D, on the one hand, and E, on the other hand, to the R^2 obtained in the phenotypic regression analyses (Figure 1, right-hand side).

Results

Descriptive Statistics

The descriptive statistics (of the raw data) are given in Table 2. School performance correlated about equally high with our non-cognitive measures (.40 with self-control and .53 with grit) as with our cognitive measure (.51 with IQ). Self-control and grit correlated .63.

For the measure of grit, we tested whether a version with just items 1 and 3 shows similar correlations with the other constructs compared to our full measure of grit. We did so, as items 1 and 3 (see Table 1) are conceptually better measures of grit than item 2. As shown in Table 2, our full measure of grit and the items-1-and-3 measure show highly similar correlations with the other construct. We continued in the following analyses with our full measure of grit to maximize the sample size.

IQ

The assessment of full-scale IQ in a subsample of 421 children allowed us to investigate if our measures of self-control and grit were associated with school performance independent of IQ. We tested if self-control and grit still predict school performance after regressing out IQ. Results indicated grit, but not self-control, still predicts school performance (b_{grit} = 0.27 [S.E. = 0.04] and ($b_{self\text{-}control}$ = 0.03 [S.E. = 0.05]). Thus, grit indeed predicts school performance above and beyond the prediction of IQ.

Table 2. Correlations among school performance, self-control, grit, CITO, and IQ.

	Self control	Grit	Grit items 1&3	School performance	Reading	Literacy	Math	CITO	IQ
Self control		0.66 (8459)	0.56 (3405)	0.40 (8087)	0.29 (6240)	0.36 (7555)	0.33 (7936)	0.29 (4642)	0.28 (417)
Grit			0.88 (3420)	0.53 (8090)	0.38 (6245)	0.46 (7564)	0.44 (7937)	0.42 (4625)	0.32 (416)
Grit items 1&3				0.48 (3393)	0.40 (3440)	0.49 (3437)	0.46 (3436)	0.43 (1061)	N.A. (0)
School performance					0.79 (6257)	0.85 (7569)	0.80 (7945)	0.70 (4381)	0.51 (377)
Reading						0.73 (6137)	0.51 (6302)	0.57 (3031)	0.43 (180)
Literacy							0.67 (7607)	0.67 (4115)	0.47 (325)
Math								0.72 (4242)	0.58 (355)
CITO									0.67 (331)
N	8521	8490	3496	8128	6401	7740	8130	4723	421
Mean	14.09	14.87	10.20	11.50	3.88	3.81	3.83	538.13	100.16
SD	2.66	4.09	2.82	2.98	1.13	1.06	1.19	8.48	13.50

***Note*.** This table is based on the raw data, uncorrected for sex, SES, teacher sharing, and censoring. The numbers between brackets refer to the number of children with overlapping data of the two constructs. Grit items 1 and 3 include the items “How hard does he/she work“ and “how task-oriented is he/she?” (so leaving out item 2). Reading, Literacy and Math are the teacher-reported school grades, and school performance is the sum of these grades. CITO is the score on the nationally-standardized school test at the end of primary school (Grade 6, ~12yo). IQ = score on the WISC-R. All correlations are significant at $p < .001$.

Table 3. Twin correlations of the raw data (uncorrected for sex, teacher sharing and censoring) for self-control, grit, and school performance, by SES and zygosity.

Zygosity	SES	Self-control	Grit	School performance
MZ	Lowest SES	.66	.67	.78
	Lower SES	.68	.67	.72
	Higher SES	.68	.71	.72
	Highest SES	.63	.64	.73
DZ	Lowest SES	.17	.15	.31
	Lower SES	.22	.24	.35
	Higher SES	.21	.23	.30
	Highest SES	.20	.22	.37

Note. All correlations are significant at $p \leq .01$.

Table 4. Parameter estimates of the effects of sex, SES and teacher sharing corrected for censoring in the saturated model, with the standard errors (se) between brackets.

	Self-control	Grit	School performance
B_0 (se)	12.05 (.05)	11.79 (.02)	8.85 (.06)
b_{sex} (se)	4.13 (.11)	4.46 (.11)	2.76 (.10)
b_{SES} (se)	0.66 (.02)	0.92 (.02)	1.07 (.03)
$b_{teacher}$ fixed (se)	0.75 (.09)	0.55 (.08)	0.62 (.10)
Teacher random (se)	0.23 (0.17)	2.11 (0.09)	0.91 (0.03)

Note. The teacher random effects equal the parameters in $\mathbf{\Lambda}_T$, where $S_T = \mathbf{\Lambda}_T\mathbf{\Lambda}_T^t$. These parameters squared equal the variance due to the teacher rater effect (i.e., .030, 4.026, and 0.928).

SES

Figure 3 displays the means of school performance, self-control, and grit for boys and girls by SES. The mean school performance, self-control, and grit vary with SES, with children with higher SES, scoring, on average, higher on school

performance, self-control, and grit. The twin correlations among these variables were highly similar across levels of SES (see Table 3 for correlational structure by SES). The main effects of sex, SES, and sharing the same teacher are displayed in Table 4.

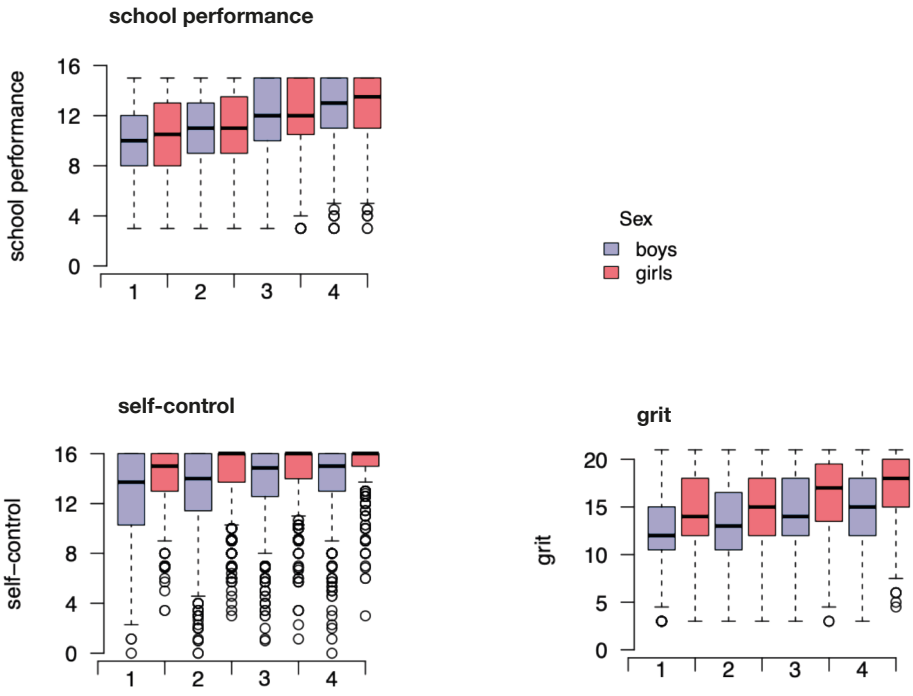


Figure 3. Boxplots of the school performance, self-control and grit scores separately for boys and girls, and for children from different socio-economic strata (SES). SES had an effect on the means of school performance, self-control and grit, but the correlational structure did not differ across SES. We included sex and SES as fixed effects in our model.

Twin Correlations

The twin correlations in Figure 4 suggest the presence of additive genetic effects on school performance (i.e., $r_{MZ} \sim 2^* r_{DZ}$) and additive genetic and as well as dominance effects for self-control and grit (i.e., $r_{MZ} > 2^* r_{DZ}$). Common environmental effects, which are suggested by $r_{MZ} < 2^* r_{DZ}$, appear to be absent.

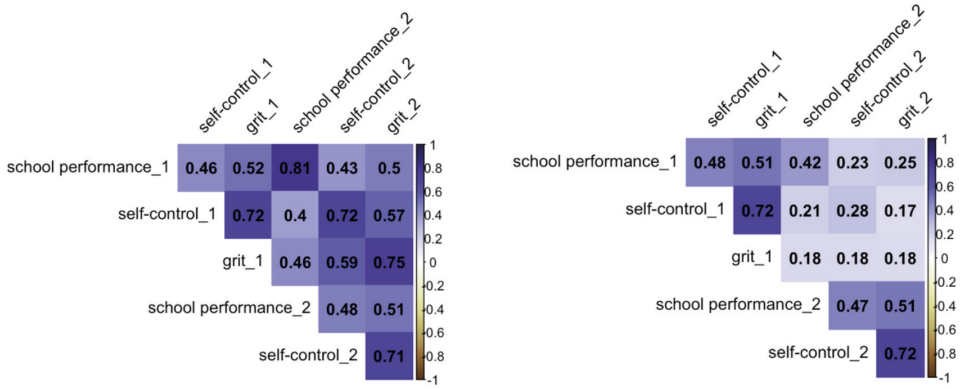


Figure 4. Twin correlations for MZ (left) and DZ (right) twins corrected for sex, SES, teacher sharing and censoring. First-born twins of a pair are indicated with _1 and second-born twins with _2. The figure includes cross-twin within-trait correlations (= the correlation between twin 1 and twin 2 for the same trait), cross-twin cross-trait correlations (= the correlation between twin 1 and twin 2 for the different traits) and within-twin cross-trait correlations (= the correlation between different traits in the same twin).

Table 5. Results of fitting the ADE mode.

ΣA			ΣD			ΣA+ΣD			ΣE			ΣT		
SP	SC	grit	SP	SC	grit	SP	SC	grit	SP	SC	grit	SP	SC	grit
9.01	4.81	4.87	0.29	0.59	1.22	9.30	5.40	6.09	2.43	0.81	0.53	0.93	0.17	1.93
4.81	6.00	2.85	0.59	4.05	5.45	5.40	10.06	8.30	0.81	4.56	2.26	0.17	0.03	0.35
4.87	2.85	2.65	1.22	5.45	8.24	6.09	8.30	10.89	0.53	2.26	3.94	1.93	0.35	4.03
standardized variances														
a^2			d^2			a^2+d^2			e^2			t^2		
.712	.410	.141	.023	.277	.437	.735	.687	.578	.192	.311	.209	.073	.002	.214

Note. Variance-covariance matrices conditional on sex and SES and corrected for censoring, and the standardized variance components attributable to additive genetic effects (A), dominance effects (D), unshared environmental effects (E), and the random teacher effect (T). The coefficient a^2 is the narrow sense heritability. The sum $a^2 + d^2$ is the broad-sense heritability. The bottom row presents the standardized proportions of variances for each of the traits explained by genetic, dominance, nonshared environmental, and teacher effects. These effects per trait add up to 1 (i.e., $a^2+d^2 e^2+t^2 = 1$).

Table 6. School performance variance explained by genetic and environmental components of self-control and grit, conditional on SES and sex and teacher sharing, and corrected for censoring. The explained variance of school performance at the level of A + D and E variance are presented in the top part, and at the level of the phenotypic variance in the bottom part. Results from the combined genetic covariance structure modeling and regression analysis. Confidence intervals (95%; 95CIs) are displayed below the estimates.

dictors	contribution to R^2	contribution to R^2	contribution to R^2	variance R^2
	At the level of A+D and E variance			
A+D	$b_{G\_SC}^2 * s_{G\_SC}^2 / s_{G\_SP}^2$ = 0.043	$b_{G\_grit}^2 * s_{G\_grit}^2 / s_{G\_SP}^2$ = 0.193	$(2 * b_{G\_SC} * s_{G\_grit} * s_{G\_SC,grit}) / s_{G\_SP}^2$ = 0.145	R^2 = 0.38 (38% of A+D variance, $s_{G\_SP}^2$)
E	$b_{E\_SC}^2 * s_{E\_SC}^2 / s_{E\_SP}^2$ = 0.045	$b_{E\_grit}^2 * s_{E\_grit}^2 / s_{E\_SP}^2$ = 0.003	$(2 * b_{E\_SC} * s_{E\_grit} * s_{E\_SC,grit}) / s_{E\_SP}^2$ = 0.013	R^2 = 0.06 (6% of E variance, $s_{E\_SP}^2$)
	At the level of the phenotypic variance			
A+D	$b_{G\_SC}^2 * s_{G\_SC}^2 / s\_{SP}^2$ = 0.034	$b_{G\_grit}^2 * s_{G\_grit}^2 / s\_{SP}^2$ = 0.153	$2 * b_{G\_SC} * s_{G\_grit} * s_{G\_SC,grit} / s\_{SP}^2$ = 0.115	R^2 = 0.303 (30.3% of Phenotypic variance, s_{SP}^2)
95CIs	[0.030 – 0.067]	[0.081– 0.213]	[0.072 – 0.121]	
E	$b_{E\_SC}^2 * s_{E\_SC}^2 / s\_{SP}^2$ = 0.0094	$b_{E\_grit}^2 * s_{E\_grit}^2 / s\_{SP}^2$ = 0.0007	$2 * b_{E\_SC} * s_{E\_grit} * s_{E\_SC,grit} / s\_{SP}^2$ = 0.0027	R^2 = 0.013 (1.3% of phenotypic variance, s_{SP}^2)
95CIs	[.003 - .011]	[.0003 - .0009]	[-.0005 - .0031]	

Phenotypic Regression Model

In the phenotypic model, the regression coefficients equal .191 (for self-control; 95%CIs: .131 - .251) and .328 (for grit; 95%CIs: .252 - .412). Self-control and grit account for 28.3% of the variance in school performance (conditional on sex, SES and teacher sharing and corrected for censoring). Of this 28.3%, self-control explained 4.4%, grit explained 13.0%, with the rest, i.e., 10.9% due to covariance between self-control and grit. So, most of the explained variance in school performance is due to grit (46%, i.e., 13.0%/28.3%) and the covariance between self-control and grit (39%, i.e., 10.9%/28.3%), while self-control accounted for 16% (i.e., 4.4%/28.3%) of the explained variance in school performance.

Genetic-and-Environmental Regression Model

Subsequently, we fitted the ADE model. In Table 5, the variance-covariance matrices are presented, with standardized variance components, based on fitting the ADE model. The standardized variance component, corrected for sex and SES, are as follows. The standardized broad-sense genetic variances (attributable to additive genetic and dominance effect) equal 73.5 (school performance), 68.7% (self-control), and 57.8% (grit); the standardized unshared environmental variances equal 19.2% (school performance), 31.1% (self-control), and 20.9% (grit), and the standardized teacher rater variances equal 7.3%, .2%, and 21.4%. Conditional on sex, SES, and teacher rater, the standardized broad-sense genetic variance components are 79% (school performance), 69% (self-control), and 73% (grit), and the standardized unshared environmental variances are 21% (school performance), 31% (self-control), and 27% (grit).

In Table 6, we present the explained variance of school performance at the level of A+D and E variance in the top part, and at the level of the phenotypic variance in the bottom part. The results in Table 6 are corrected for sex, SES, and teacher rater effects. Considering the regression as specified at the level of A+D, we found that 38% of the A+D variance of school performance is explained by the genetic (A+D) components of self-control and grit. The contributions of these genetic components are 4.3% (self-control), 19.3% (grit) and 14.5% (due to the genetic covariance of self-control and grit). Considering the regression as specified at the level of E, we found that only 6% of the E variance of school performance is explained by the unshared environmental (E) components of self-control and grit. The contributions of these unshared environmental components are 4.5% (self-control), .3% (grit) and 1.3% (due to the environmental covariance of self-control and grit). Of greater interest are the contributions to the phenotypic variance of school performance. Specifically, we know from the phenotypic regression

analyses, that self-control and grit explain about 28.3% of the phenotypic variance of school performance. In the present regression model, we explained slightly more variance, i.e., 31.6%. But of this 31.6%, 30.3% is explained by the genetic components of self-control and grit, and 1.3% is explained by the environmental components of self-control and grit. The 30.3% breaks down as follows: 3.4% (genetic component self-control), 15.3% (genetic component of grit), and 11.5% (genetic covariance of self-control and grit). The 1.3% breaks down as follows .9%, .07%, .27%. An important finding is therefore that the phenotypic regression analysis is largely a reflection of genetic influences.

Figure 5 displays the proportions of phenotypic variance in school performance attributable to self-control, grit, and their covariance (conditional on sex, SES, and teacher sharing and corrected for censoring) based on the genetic covariance structure modeling. The genetic and environmental components of self-control and grit combined explained 31.6% of the variance in school performance, standardized by the total phenotypic variance. Based on the combined genetic covariance structure modeling and regression analyses (Table 5 and the right part of Figure 5), we conclude that the best predictor of school performance was the genetic (A+D) component of grit. The genetic component of grit accounted for 48.4% (i.e. 15.3%/31.6%) of the total explained phenotypic variance in school performance, and the remaining part is mostly attributable to the genetic covariance between self-control and grit (36.4%, i.e. 11.5%/31.6%). Environmental components of self-control, grit, and the covariance between self-control and grit accounted only explained 1.3% of the phenotypic variance. So, this is about 4.1% (i.e., 1.3%/31.6%) of the explained phenotypic variance.

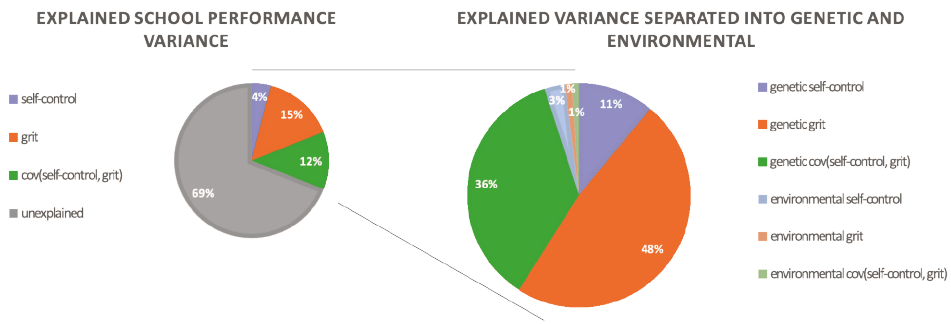


Figure 5. Explained variance in school performance in the genetically informed regression model. Left side: percentages of explained variance in school performance. Right side: percentages of explained variance in school performance by environmental and genetic components of self-control, grit, and their covariance.

Discussion

We found that self-control and grit explained 28.3% of the variance in school performance in the phenotypic model and 31.6% of the variance in school performance in the genetically informed model in 12-year-olds. Most of this 31.6% was attributable to the genetic component of grit. Because we employed twin data, we were able to use genetically-informed regression analyses to disentangle genetic and environmental contributions to the phenotypic associations. Most of the explained variance in school performance by these non-cognitive factors was accounted for by the genetic components. The best predictor of individual differences in school performance was the genetic component (A + D) of grit. About half of the 31.6% explained variance in school performance was explained by the genetic component of grit, and the remaining half was mostly explained by the genetic component of self-control and the genetic covariance between self-control and grit. A very small portion (1.3%) of individual differences in school performance was explained by unique environmental factors, mostly the environmental component of self-control.

We replicated the finding that self-control and grit are substantially heritable, with heritability estimates, conditional on sex and SES, of .69 and .58, respectively. Consistent with the results of other studies, we replicated the absence of common environmental influences (Rimfeld et al., 2016; Willems et al., 2019). The estimate of heritability of grit, conditional on SES and sex in our sample (heritability =.58) was somewhat higher than the heritability reported in a previous British study on grit (heritability = ~.4, Rimfeld et al., 2016), but the same as a previous Japanese study on grit (heritability = .59, Takahashi Zheng, Yamagata, & Ando, 2021). The estimate of heritability for school performance (heritability =.735) resembled that reported in other studies (Bartels et al., 2002; Pokropek & Sikora, 2015). Based on the twin correlations, we saw no evidence common environmental (C) influence on school performance. However, it is important to consider that we accounted for SES in our model. After correcting for SES, we found that the individual differences in the phenotypes are mainly due to genetic differences. In earlier Dutch studies that did not account for SES, C was also small (<.10) or absent (de Zeeuw et al., 2016; van Bergen et al., 2018).

Because some twins are in the same class, and shared a teacher, we modeled a random teacher-rater effect. A noteworthy finding is that sharing a class and thus being rated by the same teacher explained more variance in grit (21.4%) than in self-control (<1%) or school performance (7.3%). We hypothesized that this may be due to grit, more than self-control, being influenced by the academic climate

in the classroom (Lamb, Middeldorp, Van Beijsterveldt, & Boomsma, 2012). An Australian study that modeled the classroom effect on achievement test scores found that the variance explained by the classroom effect was only 2-3% (Grasby et al., 2020). This estimate is based on test scores, so free of a rater effect. Hence, Grasby et al.'s study suggests that the effect of the teacher and other classroom effects on school performance are small. We speculate that our effect is larger, because it includes the rater effect.

Genetic factors contribute strongly to the phenotypic correlations of non-cognitive skills. Takahashi et al. (2021) identified self-control and grit, along with conscientious and effortful control, as being part of a conscientious-related common factor. The four non-cognitive skills were strongly correlated genetically: the latent common non-cognitive factor explained 84% of the genetic variance (Takahashi et al., 2021). So, this shows that non-cognitive factors partly overlap phenotypically, mostly for genetic reasons. The current study indicates that self-control and grit have distinct aspects; they differ in their contribution to the prediction of school performance. Here, we mostly measured the perseverance aspect of grit, which is the aspect of grit found to be most related to academic outcomes in previous studies (Muenks et al., 2017; Rimfeld et al., 2016).

In a subsample we showed that school performance is similarly correlated with our non-cognitive measures as with our cognitive measure (IQ; see Table 2). Moreover, grit predicted school performance above and beyond the prediction of IQ. The finding that non-cognitive factors explain school performance over and above cognitive factors is in line with recent work at the level of measured DNA. Demange et al. (2021) operationalized a general "non-cognitive factor" by identifying genetic variants (in a genome-wide association study [GWAS] approach) that are associated with educational attainment, but not with cognition (Demange et al., 2021). Both the non-cognitive and the cognitive genetic factors predicted socioeconomic success.

A strength of the current study is its large sample of twins, which enabled us to predict children's school performance through self-control and grit at both the phenotypic, and the genetic and environmental levels. We incorporated the effects of SES, sex, and sharing the same teacher. Another strength of our study is that we corrected for censoring. Our teacher-rated measures, especially self-control, showed ceiling effects, meaning that many children scored the highest possible score.

All three main constructs were based on reports from the teacher. For validation and context, we presented in a subsample data based on individual tests (CITO

school performance and WISC-R IQ). Teachers may rate children with better academic achievement as having more self-control and being grittier, due to response bias or confirmation bias. This hypothesis fits with our observation that we find larger associations between school performance and non-cognitive skills than previously reported (meta-analyzed by Credé et al., 2017). We validated the teacher ratings of school performance: The teacher ratings correlated .75 with scores on a nationwide standardized educational-achievement test (i.e., CITO scores). In addition, the heritability estimate of teacher-rated school performance (heritability =.74; Table 5) was the same as that of the CITO scores (heritability =.74; de Zeeuw et al., 2016), though the CITO were to a small degree influenced by the shared environment (c^2 =.08; de Zeeuw et al., 2016).

A limitation of the present study concerns the measure of grit. Our measure of grit, emphasizing the perseverance of effort aspect, is weaker than the classical and validated measure, which includes items like "I finish whatever I begin" and "I am diligent" (Duckworth & Quinn, 2009). Our second item (see Table 1) theoretically seems less well related to the grit concept; however, it correlated well with the other two items. A grit measure leaving this item out correlated similarly to self-control and school performance (Table 2), thus reassuring that our findings are not driven by item 2. Our third item was missing for just over half the sample, but still leaving *N*~3,900.

Our research question concerned prediction, not causation. Accordingly, we used prediction models in cross-sectional data rather than causal models (Larson, 2021). Our findings are consistent, but do not prove, a causal effect of non-cognitive skills on school performance. Alternative explanations of the association are reverse causality (i.e., school performance influences non-cognitive factors), or a common underlying factor that influences both, without a causal association between non-cognitive factors and school performance. Future research should tackle these important but challenging research questions.

Our findings concern the status quo: we focused on (the sources of) individual differences, as they exist in the natural situation. That is, we focused on the "what is", not on the "what could be" as a consequence of intervention (van Bergen et al., 2018). Finding that individual differences in school performance can to a large extent be predicted by the genetic components of self-control and grit does not mean that these skills are immutable, but reflects that children who are performing well in school oftentimes also are genetically predisposed to be grittier and to have more self-control. In popular science, cognitive skills like IQ are sometimes thought of as innate talents that are difficult to change, while

non-cognitive skills are thought of as malleable skills that can be nurtured and taught to students (Chang, 2014; Martinez et al., 2022; Sokolowski & Ansari, 2018). Although findings from our and other heritability studies do not speak to trainability, they do show that cognitive skills and non-cognitive skills are both substantial and similarly heritable, refuting this popular distinction. The potential malleability and trainability of non-cognitive skills have been investigated with interventions (Sisk, Burgoyne, Sun, Butler, & Macnamara, 2018). For cognitive skills, Zijlstra, van Bergen, Regtvoort, de Jong, and van der Leij (2021) showed that their 2-year reading intervention was equally effective in children with and without a family risk for reading difficulties, though the family-risk group needed more intervention sessions. These findings suggest that a prolonged and tailored intervention can improve children's academic skills, also in those with a genetic predisposition for learning difficulties. Regarding non-cognitive skills, future work could investigate whether interventions targeting non-cognitive skills are equally effective in children with and without (a genetic predisposition for) learning difficulties. From our current study, we conclude that whether children do well in school can be predicted by (genetic) components of self-control and more importantly grit.

Literature

Bartels, M., Rietveld, M. J., van Baal, G. C. M., & Boomsma, D. I. (2002). Heritability of educational achievement in 12-year-olds and the overlap with cognitive ability. *Twin Research and Human Genetics*, *5*(6), 544-553.

Boomsma, D. I., De Geus, E. J., Vink, J. M., Stubbe, J. H., Distel, M. A., Hottenga, J. J., ... & Willemsen, G. (2006). Netherlands Twin Register: from twins to twin families. *Twin Research and Human Genetics*, *9*(6), 849-857.

Boomsma, D. I., Van Beijsterveldt, T. C., Odintsova, V. V., Neale, M. C., & Dolan, C. V. (2021). Genetically informed regression analysis: application to aggression prediction by inattention and hyperactivity in children and adults. *Behavior genetics*, *51*(3), 250-263

Cito (2002) Eindtoets basisonderwijs. Cito, Arnhem

Chang, W. (2014). *Grit and academic performance: Is being grittier better?*. University of Miami.

Credé, M., Tynan, M. C., & Harms, P. D. (2017). Much ado about grit: A meta-analytic synthesis of the grit literature. *Journal of Personality and social Psychology*, *113*(3), 492.

Demange, P. A., Malanchini, M., Mallard, T. T., Biroli, P., Cox, S. R., Grotzinger, A. D., ... & Nivard, M. G. (2021). Investigating the genetic architecture of noncognitive skills using GWAS-by-subtraction. *Nature Genetics*, *53*(1), 35-44.

de Zeeuw, E. L., van Beijsterveldt, C. E., Glasner, T. J., Bartels, M., de Geus, E. J., & Boomsma, D. I. (2014). Do children perform and behave better at school when taught by same-gender teachers?. Learning and Individual Differences, 36, 152-156.

de Zeeuw, E. L., van Beijsterveldt, C. E., Glasner, T. J., de Geus, E. J., & Boomsma, D. I. (2016). Arithmetic, reading and writing performance has a strong genetic component: A study in primary school children. *Learning and Individual Differences*, *47*, 156-166.

de Zeeuw, E. L., Kan, K. J., van Beijsterveldt, C. E., Mbarek, H., Hottenga, J. J., Davies, G. E., ... & Boomsma, D. I. (2019). The moderating role of SES on genetic differences in educational achievement in the Netherlands. *npj Science of Learning*, *4*(1), 1-8.

de Zeeuw, E. L., Hottenga, J. J., Ouwens, K. G., Dolan, C. V., Ehli, E. A., Davies, G. E., Boomsma, D.I., & van Bergen, E. (2020). Intergenerational transmission of education and ADHD: effects of parental genotypes. *Behavior Genetics, 50(4)*, 221-232. https://doi.org/10.1007/s10519-020-09992-w

Duckworth, A., & Gross, J. J. (2014). Self-control and grit: Related but separable determinants of success. *Current directions in psychological science*, *23*(5), 319-325.

Duckworth, A. L., Gendler, T. S., & Gross, J. J. (2014). Self-control in school-age children. *Educational Psychologist*, *49*(3), 199-217.

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: perseverance and passion for long-term goals. *Journal of personality and social psychology*, *92*(6), 1087.

Duckworth, A. L., Taxer, J. L., Eskreis-Winkler, L., Galla, B. M., & Gross, J. J. (2019). Self-control and academic achievement. *Annual review of psychology*, *70*, 373-399.

Duckworth, A. L., & Quinn, P. D. (2009). Development and validation of the Short Grit Scale (GRIT–S). *Journal of personality assessment*, *91*(2), 166-174.

Falconer, D. S., & Mackay, T. F. (1983). *Quantitative genetics*. London, UK: Longman.

Gil-Hernández, C. J. (2021). The (unequal) interplay between cognitive and noncognitive skills in early educational attainment. *American Behavioral Scientist*, *65*(11), 1577-1598.

Grasby, K. L., Little, C. W., Byrne, B., Coventry, W. L., Olson, R. K., Larsen, S., & Samuelsson, S. (2020). Estimating classroom-level influences on literacy and numeracy: A twin study. *Journal of Educational Psychology*, *112*(6), 1154.

John, O. P., & Srivastava, S. (1999). The Big-Five trait taxonomy: History, measurement, and theoretical perspectives.

J. J., Diris, R., Ter Weel, B., & Borghans, L. (2014). Fostering and measuring skills: Improving cognitive and non-cognitive skills to promote lifetime success (No. w20749) *National Bureau of Economic Research*.

Lamb, D. J., Middeldorp, C. M., Van Beijsterveldt, C. E., & Boomsma, D. I. (2012). Gene–environment interaction in teacher-rated internalizing and externalizing problem behavior in 7-to 12-year-old twins. *Journal of child psychology and psychiatry*, *53*(8), 818-825.

Larsson, H. (2021). Causation and prediction in child and adolescent mental health research. *JCPP Advances, 1*(2), e12026.

Ligthart, L., van Beijsterveldt, C. E., Kevenaar, S. T., de Zeeuw, E., van Bergen, E., Bruins, S., ... & Boomsma, D. I. (2019). The Netherlands Twin Register: longitudinal research based on twin and twin-family designs. *Twin Research and Human Genetics*, *22*(6), 623-636.

Malanchini, M., Rimfeld, K., Allegrini, A. G., Ritchie, S. J., & Plomin, R. (2020). Cognitive ability and education: How behavioural genetic research has advanced our knowledge and understanding of their association. *Neuroscience & Biobehavioral Reviews*, *111*, 229-245.

Martinez, K. M., Holden, L. R., Hart, S. A., & Taylor, J. (2022). Examining mindset and grit in concurrent and future reading comprehension: A twin study. *Developmental Psychology.* Advance online publication.

Muenks, K., Wigfield, A., Yang, J. S., & O'Neal, C. R. (2017). How true is grit? Assessing its relations to high school and college students' personality characteristics, self-regulation, engagement, and achievement. *Journal of Educational Psychology*, *109*(5), 599.

Neale, M. C., Hunter, M. D., Pritikin, J. N., Zahery, M., Brick, T. R., Kirkpatrick, R. M., Estabrook, R., Bates, T. C., Maes, H. H., & Boker, S. M. (2016). OpenMx 2.0: Extended Structural Equation and Statistical Modeling. *Psychometrika*, *81*(2), 535-549.

Nigg, J. T. (2017). Annual Research Review: On the relations among self-regulation, self-control, executive functioning, effortful control, cognitive control, impulsivity, risk-taking, and inhibition for developmental psychopathology. *Journal of child psychology and psychiatry*, *58*(4), 361-383.

Oriol, X., Miranda, R., Oyanedel, J. C., & Torres, J. (2017). The role of self-control and grit in domains of school success in students of primary and secondary school. *Frontiers in psychology*, *8*, 1716.

Ponnock, A., Muenks, K., Morell, M., Yang, J. S., Gladstone, J. R., & Wigfield, A. (2020). Grit and conscientiousness: Another jangle fallacy. *Journal of Research in Personality, 89*, 104021.

Pokropek, A., & Sikora, J. (2015). Heritability, family, school and academic achievement in adolescence. *Social Science Research*, *53*, 73-88.

Rimfeld, K., Kovas, Y., Dale, P. S., & Plomin, R. (2016). True grit and genetics: Predicting academic achievement from personality. *Journal of personality and social psychology*, *111*(5), 780.

Rimfeld, K., Malanchini, M., Hannigan, L. J., Dale, P. S., Allen, R., Hart, S. A., & Plomin, R. (2019). Teacher assessments during compulsory education are as reliable, stable and heritable as standardized test scores. *Journal of Child Psychology and Psychiatry*, *60*(12), 1278-1288.

Sisk, V. F., Burgoyne, A. P., Sun, J., Butler, J. L., & Macnamara, B. N. (2018). To what extent and under which circumstances are growth mind-sets important to academic achievement? Two meta-analyses. *Psychological science*, *29*(4), 549-571

Sokolowski, H. M., & Ansari, D. (2018). Understanding the effects of education through the lens of biology. *npj Science of Learning*, *3*(1), 1-10.

Takahashi, Y., Zheng, A., Yamagata, S., & Ando, J. (2021). Genetic and environmental architecture of conscientiousness in adolescence. *Scientific reports*, *11*(1), 1-11.

Tikhodeyev, O. N., & Shcherbakova, O. V. (2019). The problem of non-shared environment in behavioral genetics. *Behavior genetics*, *49*(3), 259-269

Tucker-Drob, E. M., Briley, D. A., Engelhardt, L. E., Mann, F. D., & Harden, K. P. (2016). Genetically-mediated associations between measures of childhood character and academic achievement. *Journal of Personality and Social Psychology, 111*(5), 790–815.

Usher, E. L., Li, C. R., Butz, A. R., & Rojas, J. P. (2019). Perseverant grit and self-efficacy: Are both essential for children's academic success?. *Journal of Educational Psychology*, *111*(5), 877.

van Beijsterveldt, C. E., Groen-Blokhuis, M., Hottenga, J. J., Franić, S., Hudziak, J. J., Lamb, D., ... & Boomsma, D. I. (2013). The Young Netherlands Twin Register (YNTR): longitudinal twin and family studies in over 70,000 children. *Twin Research and Human Genetics*, *16*(1), 252-267.

van Bergen, E., Snowling, M. J., de Zeeuw, E. L., van Beijsterveldt, C. E., Dolan, C. V., & Boomsma, D. I. (2018). Why do children read more? The influence of reading ability on voluntary reading practices. *Journal of Child Psychology and Psychiatry*, *59*(11), 1205-1214.

van Haasen, P. P., De Bruyn, E. E. J., Pijl, Y. J., Poortinga, Y. H., Lutje-Spelberg, H. C., Vander Steene, G., Coetsier, P., Spoelders-Claes, R., and Stinissen, J. (1986). *Wechsler Intelligence Scale for Children-Revised, Dutch Version*. Swets & Zetlinger B. V., Lisse, The Netherlands.

Werner, K. M., Milyavskaya, M., Klimo, R., & Levine, S. L. (2019). Examining the unique and combined effects of grit, trait self-control, and conscientiousness in predicting motivation for academic goals: A commonality analysis. *Journal of Research in Personality*, *81*, 168-175.

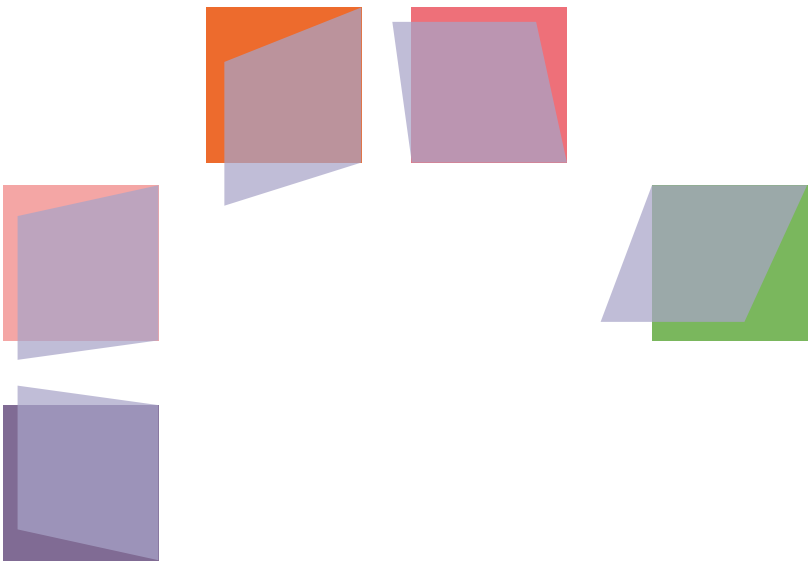
Willems, Y. E., Boesen, N., Li, J., Finkenauer, C., & Bartels, M. (2019). The heritability of self-control: A meta-analysis. *Neuroscience & Biobehavioral Reviews*, *100*, 324-334.

Willems, Y. E., Dolan, C. V., van Beijsterveldt, C. E., de Zeeuw, E. L., Boomsma, D. I., Bartels, M., & Finkenauer, C. (2018). Genetic and environmental influences on self-control: assessing self-control with the ASEBA self-control scale. *Behavior genetics*, *48*(2), 135-146.

Zijlstra, H., Van Bergen, E., Regtvoort, A., De Jong, P. F., & Van Der Leij, A. (2021). Prevention of reading difficulties in children with and without familial risk: Short-and long-term effects of an early intervention. *Journal of Educational Psychology*, *113*(2), 248.

Chapter 5

Grit and self-control predict school performance: strongly genetic, weakly causal



Submitted as: **Kevenaar, S. T.,** van Bergen, E., Oldehinkel, A. J., Boomsma, D. I. & Dolan, C. V., (Submitted for publication). Grit and self-control predict school performance: strongly genetic, weakly causal.

Abstract

Background: The non-cognitive skills self-control and grit have often been cited as predictors of school performance, but little research has investigated whether this relationship is causal. We investigated the causal nature of this association in a classical twin design, with mono- and dizygotic twin pairs. Specifically, we evaluated the direct impact of self-control and grit on school performance, while controlling for genetic or environmental influences common to all three traits (i.e., confounding). **Methods**: Teachers of 4,891 Dutch 12-year-old twin pairs (of which 3,837 were complete pairs) completed a survey about school performance (school grades), self-control (ASEBA self-control scale), and the perseverance aspect of grit. We regressed school performance on self-control and grit within the twin model to establish the phenotypic, putatively causal, regression relationship. We modelled genetic or environmental confounding (influences common to the three phenotypes) to determine their influence. In all analyses, we corrected for sex, rater effects of the teacher, and parental socioeconomic status. **Results**: Self-control and grit explained 28.4% of the school-performance variance in the phenotypic regression analysis (assuming no confounding). However, allowing for genetic confounding (due to genetic pleiotropy) revealed that the association was largely attributable to genetic influences that the three traits share. In the presence of genetic pleiotropy, the phenotypic regression of school performance on self-control and grit accounted for only 4.4%. **Conclusions**: The association between self-control and grit as predictors of school performance is attributable to both genetic pleiotropy, and to a lesser extent, direct effects of self-control and grit on school performance, which are putatively causal.

Introduction

Research has focused on grit and self-control as predictors of school and academic performance. Grit comprises consistency of interest and perseverance, and self-control is the ability to self-initiate regulation of conflicting impulses (Duckworth et al., 2016). Perseverance and self-control have generally been found to be associated with school and academic performance (Christopoulou, et al. 2018; Credé et al., 2017; De Ridder et al., 2012; Duckworth et al., 2019; Lam & Zhou, 2019; Oriol, Miranda, Oyanedel, & Torres, 2017; Wolters & Hussain, 2015). For instance, Lam & Zhou (2019) reported an average correlation of .17 between grit and school performance in school children (based on 56 correlations). The average correlation was .14 in students in higher education (based on 60 effect sizes; see also Fernandez-Martin et al, 2020). In a recent study of Czech school children, Vazsonyi, et al. (2022) found that self-control predicted school performance (both teacher-rated and grades) while controlling for motivation and intelligence. In a twin study, Kevenaar et al. (2023) found that grit and self-control together explained 28.4% of the variance in school performance. The decomposition of the phenotypic regression relationships into genetic and environmental components revealed that the phenotypic associations were mainly due to genetic factors.

It is well established that self-control and grit predict academic outcomes, however, most studies tend not to address causality (e.g. Credé et al., 2017; Eskreis-Winkler et al., 2014; Tangney et al., 2018). One of the few studies that provide a basis for a causal interpretation regarding self-control was conducted by Duckworth et al. (2010), who showed that within-individual changes in self-control over time predict changes in academic achievement, but not vice versa, which suggests a causal effect only from self-control to achievement. A small (*N* = 53) intervention study on self-regulation indicated that self-regulation training affected math performance, which is also consistent with a causal effect of self-regulation (Perels et al., 2019). Regarding grit, Jiang et al. (2019) found reciprocal effects between grit and academic achievement, consistent with a reciprocal causal relation. Postigo et al. (2021) studied a large sample of children (*N* = 5,371) longitudinally from age 10 to 14. They reported an effect of grit on school performance (grades) in a two-occasion panel model, which is consistent with a causal model. Hence, as far as it has been studied, most research suggests a causal effect of self-control and grit on school performance, rather than the other way around.

The interpretation of the effects of grit and self-control on school performance as causal is appealing, as it is plausible that these non-cognitive factors facilitate school or academic performance. However, more research, employing different designs, is needed to establish causal pathways, and to rule out possible - correlational, non-causal - sources. Such non-causal sources may be both genetic and environmental. For instance, the association may be due to common genetic influences (pleiotropy: the same genes affect multiple traits [MacKay, 2014]), or due to a rearing environment that is conducive to both cognitive and non-cognitive influences on school performance. It is also important to take note of the challenging possibility that causal and non-causal accounts of the associations are not mutually exclusive.

Our present aim is to contribute to the causal research by following up the analyses of Kevenaar et al. (2023). The classical twin design provides the means to estimate the phenotypic relationship between the non-cognitive factors and school performance while accounting for the genetic background correlation between these variables (Kohler, Behrman & Schnittker, 2011). Using this design, we investigated the causal relationship between grit, self-control, and teacher-rated school performance, while accounting for possible genetic and environmental confounding, that is, non-causal associations that are attributable to genetic and environmental influences common to the phenotypes.

The outline of this article is as follows. First, we introduce the twin model and provide a summary of the results of Kevenaar et al. (2023). Next, we present a causal twin model, which we apply to explore the putative causal influence of grit and self-control and school performance. Then, we fit the causal model by including genetic confounding, also referred to as genetic pleiotropy, and confounding by environmental influences.

The classical twin design and model

The twin design is a genetically informative design, which is applied to decompose phenotypic variance and covariance into genetic and environmental components. With respect to the environmental components, we distinguish shared (C) and unique environmental (E) variance. The latter (E) is unique to the individual twins, not shared, and, as such, contributes to the phenotypic variance, but not the phenotypic covariance (resemblance) of the twins. Shared environmental variance originates in environmental influences that twins share and may contribute to the phenotypic covariance of the twins. With respect to the genetic components, we can distinguish additive genetic (A) variance and

dominance (D) variance, where the former is due to the additive (linear) effects of alleles, and the latter is due to non-additive effects of alleles at the relevant genetic loci on the phenotype (Falconer & MacKay, 1983). Because monozygotic (MZ) twins are genetically (nearly) identical, both A and D contribute 100% to the MZ phenotypic resemblance (i.e. covariance). Dizygotic (DZ) twins, like full sibs, on average share 50% of their alleles, as inherited from their biological parents. Based on allele sharing, we expect 50% of the additive genetic variance to contribute to the DZ phenotypic covariance. The dominance variance attributable to a given locus contributes to the phenotypic resemblance, only if the twins are genetically identical by descent at the locus. Considering a diallelic locus as an example with alleles B and b, 25% of the DZ twins are genetically identical (i.e., both BB, Bb, or bb). Therefore, we expect 25% of the dominance variance to contribute to the DZ phenotypic covariance. When fitting the classical twin model to data from MZ and DZ twin pairs to identify the variance components, we need to limit the number of components to three, that is, an ADE or an ACE model. The choice is usually based on the following rule of thumb concerning the phenotypic twin correlation, r_{mz} and r_{dz}: $r_{mz} > 2^{*}r_{dz}$ suggests an ADE model; $r_{mz} < 2^{*}r_{dz}$ suggests an ACE model (for discussion, see Keller & Coventry, 2005). Based on our earlier work, we fit an ADE model to the three phenotypes (i.e., school performance, self-control, and grit) and decompose the 3x3 covariance matrix S_{Ph} as follows: $S_{Ph} = S_A + S_D + S_E$. This decomposition is achieved by modelling the 6 x 6 MZ and DZ twin covariance matrices. The matrices are 6 x 6, because of the three phenotypes for both twin 1 and twin 2 (the first and second born, respectively), as in Table 1. The MZ and DZ twin covariance matrices $S_{Ph|MZ}$ and $S_{Ph|DZ}$ are as follows:

$$\Sigma_{Ph|MZ} = \begin{array}{c} \\ \text{MZ twin 1} \\ \text{MZ twin 2} \end{array} \begin{array}{c} \begin{array}{cc} \text{MZ twin 1} & \text{MZ twin 2} \end{array} \\ \begin{bmatrix} \Sigma_A + \Sigma_D + \Sigma_E & \Sigma_A + \Sigma_D \\ \Sigma_A + \Sigma_D & \Sigma_A + \Sigma_D + \Sigma_E \end{bmatrix} \end{array}$$

And

$$\Sigma_{Ph|DZ} = \begin{array}{c} \\ \text{DZ twin 1} \\ \text{DZ twin 2} \end{array} \begin{array}{c} \begin{array}{cc} \text{DZ twin 1} & \text{DZ twin 2} \end{array} \\ \begin{bmatrix} \Sigma_A + \Sigma_D + \Sigma_E & \frac{1}{2}\Sigma_A + \frac{1}{4}{}^{*}\Sigma_D \\ \frac{1}{2}\Sigma_A + \frac{1}{4}{}^{*}\Sigma_D & \Sigma_A + \Sigma_D + \Sigma_E \end{bmatrix} \end{array}$$

The covariance matrices S_A, S_D, and S_E may be subject to various parameterizations, depending on computational or substantive considerations (see below).

Previous Findings

As reported in Kevenaar et al. (2023), we previously analysed teacher ratings of grit, self-control, and school performance in MZ and DZ twins using the same data that were analysed for the present article. The results were obtained with a correction for the main effects of sex and SES, and a correction for the rater (i.e., the teacher of the twins). Given the ceiling effect in the distribution of the data (see below), we fitted ADE models using maximum likelihood estimation with a correction for right-censoring (see also de Zeeuw et al., 2019). First, school performance was regressed on grit and self-control (phenotypic regression). Second, the regression analyses were conducted at the broad-sense genetic level ($S_A + S_D$), and at the unshared environmental level (S_E). Because grit and self-control are correlated (about .65 in the present data), the decomposition of school performance variance (conditional on the covariates) comprised four variance components: a component due to grit, a component due to self-control, a component involving the covariance between grit and self-control, and the residual variance component. At the phenotypic level (see Figure 1 top), grit and self-control explained 28.3% of the school performance variance, with the following decomposition: 4.4% due to self-control, 13.0 % due to grit, and 10.9% involving the covariance of grit and self-control). Considering the unique contributions of grit and self-control, grit emerged as the stronger predictor (13% vs 4.4%).

Subsequently, the ADE model was fitted to the twin data, and the regression analyses were conducted twice: once at the level of $S_A + S_D$ (the broad-sense genetic covariance matrix) and once at the level of S_E (the unshared environmental covariance matrix) (see Figure 1, bottom two panels). The results showed that the phenotypic decomposition of school performance variance was largely attributable to broad-sense genetic factors. Thus, the phenotypic regression relationship between the predictors grit and self-control and school performance was largely a reflection of common genetic influences.

The results of the regression analyses, both at the phenotypic level and at the genetic and environmental level, are consistent with a causal model, but formally do not prove causality. Below we consider a causal twin model that addresses causality by fitting the phenotypic regression model, while accounting for the possibility of genetic or environmental background correlation (i.e., genetic or environmental confounding; Hart, Little & van Bergen, 2021).

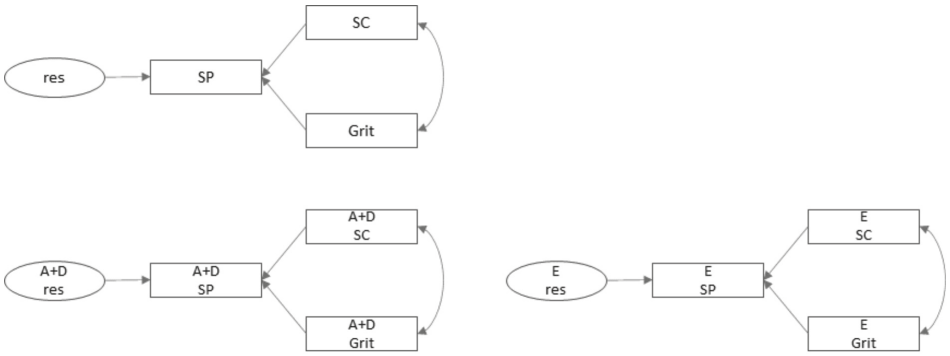


Figure 1.
Top: The phenotypic regression model. The regression residual is denoted *res*.
Bottom: The A+D regression model and the E regression model. These E and A+D models decompose the phenotypic regression results into A+D (based on $S_A + S_D$) and unshared environmental E regression results (based on S_E). SP = school performance; SC = self-control.

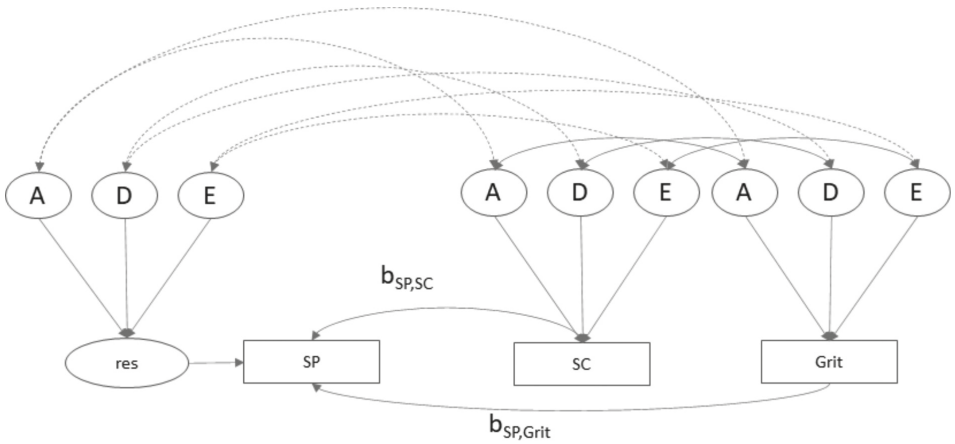


Figure 2. The causal model with background A, D, E correlation (confounding) represented by dashed double-headed arrows. The parameters $b_{SP,SC}$ and $b_{SP,Grit}$ are the causal regression coefficients. The regression residual is denoted *res*. SP = school performance; SC = self-control.

Causal twin model

The causal twin model is depicted in Figure 2. This model allows us to assess the putative causal regression relationships, while taking into account A, D, or E background correlation, i.e., A, D, or E confounding (Bruins et al., 2023; Duffy

& Martin, 1994; Heath et al., 1993; Jinks & Fulker, 1970; Kohler et al., 2011, Verhulst & Estabrook, 2012). In Figure 2, the background A, D, and E correlations are represented by the dashed double-headed arrows. In this approach, the strongest support for the causal hypothesis would be the finding that the phenotypic regression coefficients (denoted $b_{SP,SC}$ and $b_{SP,Grit}$) are significant, while the background correlations are all zero. This would support causality in that the results then demonstrate the associations between the predictors and school performance are not due to background (A, D, or E) confounding, but to the direct phenotypic, causal, relations. The causal model is refuted if the regression coefficients are zero in the presence of A, D, and/or E background correlations, as this means that the associations between grit and self-control and the dependent variable school performance are not due to direct, causal relations. Rather, they are attributable to common environmental or genetic influences. Note that the finding that $b_{SP,SC}$ and $b_{SP,Grit}$ differ significantly from zero does not rule out A, D, or E confounding. As mentioned above, the direct (phenotypic) causal effects and confounding are not mutually exclusive. In this causal twin model, we explore this possibility by fitting the phenotypic regression model, while allowing for A, D, or E confounding. We do this by including the dashed double-headed arrow in the model (Figure 2).

Methods

Participants

The sample consisted of children registered in the Netherlands Twin Register (NTR). The NTR collects data from twins, their parents, and their siblings. The data of the children include self-ratings and parental and teacher ratings (Boomsma et al., 2006; van Beijsterveldt et al., 2013; Ligthart et al, 2019). The data for this study are teacher ratings of the grit, self-control and school performance in 11.5 - 12.5-year-old twins. First, the parents of these twins were asked for permission to contact the teachers. Twins could be either in the same class and share a teacher, or be different classes and be rated by different teacher. The sample included 3,837 complete pairs and 1,054 incomplete pairs (i.e., data missing on one member). The sample consisted of 1,957 monozygotic and 2,934 dizygotic twin pairs. To ascertain the zygosity of the same-sex twin pairs, a DNA or blood test was conducted for 32.2% of the pairs, while for the remainder, parents completed a questionnaire that contained items related to the twins'

resemblance. Based on this questionnaire, zygosity is correctly determined in more than 96% of cases (Ligthart et al., 2019)

Materials

Self-control

The measure of self-control was based on the teacher ratings. The teachers completed the 8 items of the Achenbach Self-Control Scale (ASCS; Willems et al., 2018) in the ASEBA-TRF (Achenbach & Rescorla, 2000). The response options of each item are 0 (*not true*), 1 (*somewhat or sometimes t*rue), and 2 (*very true or often true*). If more than three items were missing the sum scores was coded as missing. If three or fewer items were missing, the missing items were imputed by the person's mean of the available items (Willems et al. 2018). The scores were reverse-coded, so the total score ranged from 0-16, with higher scores indicating better self-control. The Cronbach's alpha of the ASCS in teachers is 0.82 (Willems et al., 2018).

Grit

The measure of grit was based on the teacher ratings, based on the following three items relating to the perseverance aspect of grit: *Compared to typical pupils of the same age, 1) how hard does he/she work*; *2) how appropriately does he/she behave*, and *3) how task-oriented is he*/she was available. The response scale was a 7-point Likert scale and item scores were summed to sum scores. If only a single item was missing, the mean of the other two items was used to calculate the sum score. If more than one item was missing, the sum score was coded as missing. The grit sum scores ranged from 1-21, with higher scores indicating more grit. Cronbach's alpha of the grit measure is 0.87.

School performance

The measure of school performance was the sum score of teacher reports on math, reading, and literacy on 5-point scales (de Zeeuw et al., 2014; van Bergen et al., 2018). School performance scores ranged from 3 to 15, with higher scores indicating better school performance. If a single rating was missing the mean of the other two ratings was used for the missing value and the sum score was coded as missing in case more than one score was missing.

Sex and socioeconomic status (SES)

In the study, sex was denoted by 0 for males and by 1 for females. The socio-economic status (SES) of the participants was determined by a combination of

their parents' occupation and education, as described in de Zeeuw et al. (2019). The SES variable was coded on a scale of 1 to 4, with 1 indicating low SES and 4 indicating high SES.

Same/different class

Twins in the same class were rated by the same teacher, while twins in different classes were rated by different teachers. This may result in data for one twin, and not for their cotwin (the incomplete pairs), where one teacher of one of the twins participated in the study, and the teacher of the other twin did not participate. 'Class' was coded 1 (twins rated by the same teacher) or 0 (twins rated by different teachers).

Statistical modelling

We modelled the data in R using the OpenMx library (Neale et al., 2016). As shown below, the distributions of all three phenotypes display excessive negative skewness, as a consequence of ceiling effects. This is most notable in the distributions of self-control and grit. We fitted the models using full information maximum likelihood estimation, assuming that the data follow a right-censored multivariate normal distribution (as in de Zeeuw et al., 2019). We took the censoring into account explicitly to avoid bias stemming from the ceiling effects.

In addition to the saturated model (a baseline model that has a perfect fit to the data because it contains as many parameters as there are observations in the data), fitted to obtain the 6x6 MZ and DZ correlation matrices, we fitted the following six models: 1) the standard phenotypic regression model (taking into account the clustering of twins in families); 2) the trivariate ADE model to estimate the 3x3 covariance matrices S_A, S_D, and S_E; 3) the causal regression model as depicted in Figure 2, without A, D, or E confounding (i.e., the model with the correlations associated with the dashed double-headed arrows fixed to zero); and models 4, 5, and 6, i.e., the causal regression model with A confounding (model 4), D confounding (model 5), or E confounding (model 6). Model 1 produces results based on the regression of school performance of grit and self-control, as one would obtain them in a sample of unrelated children. Model 2 is a standard trivariate ADE twin model. This model does not include any regression analyses, it provides estimates of the 3x3 covariance matrices S_A, S_D, and S_E, and serves as a baseline model to evaluate the fit of model 3, and model 4, 5, and 6, as these models are nested under model 2. If there is no confounding and if the regression relations are truly causal, we expect model 2 to produce

regression results comparable to those of model 3, and we expect model 3 to fit well (compare to model 2). In case of A confounding, we expect model 3 to fit poorly (compared to model 2), and we expect model 4 (causal regression with A confounding to fit well (compared to model 2).

We conducted a total of six statistical tests, based on the likelihood ratio: the comparison of the causal regression model without confounding with the ADE model (one test with 4 degrees of freedom [df]); the comparison of the causal regression model with A, D, or E confounding with the ADE model (three tests, each with 2 df); and the test of the causal regression parameter (two tests, each 1 df) in the ultimate model of choice. As we conducted 6 likelihood ratio tests (LRTs), we corrected our family-wise alpha level of 0.05 using the Bonferroni correction, resulting in an alpha of 0.05/6 = ~0.008 for each LRT.

In all models, sex and SES were included as fixed covariates, and teacher was included as a random covariate, allowing for the possibility that the rater variance is shared (i.e., contributes to the phenotypic twin covariance), if the twins are rated by the same teacher.

Results

Descriptives

Histograms of the raw data are given in Figure 3. The right censoring (ceiling effects) is evident in all three phenotypes. In the MZ and DZ twin 1 members, the skewnesses equal -.63 (school performance), -1.63 (self-control), -.31 (grit); in the MZ and DZ twin 2 members, these equal -.57, -1.93, and -.48, respectively. The estimates of the MZ and DZ correlation matrices, based on the saturated model, are given in Table 1 (these are conditional on the covariates sex, SES, and rater, and corrected for censoring).

The MZ twin correlations for school performance, self-control, and grit equal .809 (95% CIs: .788 - .928), .715 (95% CIs: .685 - .716), and .751 (95% CIs: .722 - .778), respectively. The DZ twin correlations equal .419 (95% CIs: .376 - .460), .276 (95% CIs: .250 - .289), and .176 (95% CIs: .114 – .234), respectively. The twin correlations of self-control and grit suggest an ADE model ($r_{MZ} > 2*r_{DZ}$). The twin correlations of school performance suggest an AE model, but the 95% CIs do not rule out the possibility of an ADE model (see Keller and Coventry, 2005).

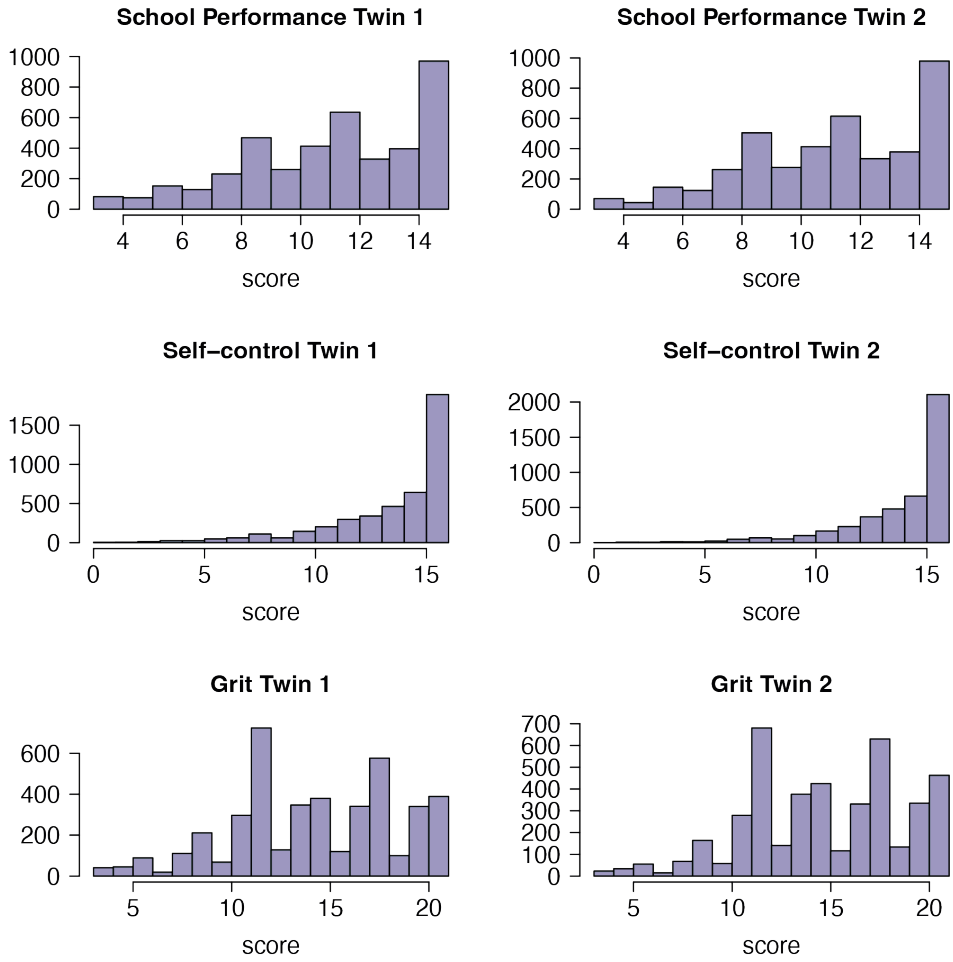


Figure 3. Histograms of the raw data of twin 1 (firstborn) and twin 2 (second born) for the total sample of MZ and DZ twins

The correlations between grit and self-control are about .72. The correlations between school performance on the one hand and grit or self-control on the other hand range from .458 to .523.

Phenotypic regression model (model 1)

We fitted the phenotypic regression model in OpenMx, correcting for the family clustering of the data (i.e., MZ and DZ twins in pairs). The aim of this is to obtain regression results at the population level. We decomposed the proportion of explained variance of school performance (i.e., R^2 statistic), into the part due

Table 1. MZ and DZ correlation matrices (conditional on fixed sex and SES effects, and random rater effects; corrected for censoring).

MZ	SP1	SC1	GRIT1	SP2	SC2	GRIT2
SP1	1.000					
SC1	0.458	1.000				
GRIT1	0.523	0.721	1.000			
SP2	**0.809**	0.395	0.458	1.000		
SC2	0.428	**0.715**	0.587	0.480	1.000	
GRIT2	0.499	0.566	**0.751**	0.510	0.710	1.000

DZ	SP1	SC1	GRIT1	SP2	SC2	GRIT2
SP1	1.000					
SC1	0.478	1.000				
GRIT1	0.511	0.717	1.000			
SP2	**0.419**	0.214	0.181	1.000		
SC2	0.226	**0.276**	0.184	0.474	1.000	
GRIT2	0.248	0.166	**0.176**	0.506	0.723	1.000

Note. The correlations shown in dark blue represent the within-person correlations between traits, which are expected to be similar in MZ and DZ. The correlations shown in bold represent the within-trait twin correlations. These are higher in MZ than DZ, suggesting genetic influences on the traits. The correlations shown in light blue represent the cross-trait, cross-twin correlations (e.g., the correlation between school performance of one twin and self-control of the cotwin). These are higher in MZ than DZ, suggesting genetic correlations between the traits. 1 = MZ = monozygotic; DZ = dizygotic; 1 = firstborn 2 = second born; SP = school performance; SC = self-control

to self-control, the part due to grit, and the part that involves the covariance of grit and self-control. Because the third part involves the covariance, it cannot unambiguously be attributed to either grit or self-control. As reported before (Kevenaar et al., 2023), we found that self-control and grit explained 28.4% of the variance in school performance (R^2=.284). The unique contributions of self-control and grit equaled 4.4% (95% CIs: 2.07% – 7.91%) and 13.0% (95% CIs: 8.03% - 19.48%), respectively. The remaining 10.9% was a function of the covariance of the predictors (95% CIs: 9.08% - 12.43%). The regression coefficients equaled $b_{SP,SC}$ = .191 (95% CIs: .132 - .251) and $b_{SP,Grit}$ =.331 (95% CIs: .252 - .412). From these results, grit emerges as the stronger predictor.

ADE twin model (model 2)

The 3x3 covariance matrices Σ_A, Σ_D, and Σ_E were parameterized using lower triangle matrices (i.e., the Cholesky decomposition, e.g., Bruins, Franic, Dolan, Borsboom, & Boomsma, 2023): $\Sigma_A = D_A D_A^t$, $\Sigma_D = D_D D_D^t$, $\Sigma_E = D_E D_E^t$, where D_A, D_D,

Table 2: Covariance (cov) matrices, correlation (cor) matrices; and proportions (conditional on fixed sex and SES effects, and random rater effects) based on the ADE twin model.

	Phenotypic cov matrix			Phenotypic cor matrix					
Σ_{Ph}	SP	SC	GRIT	SP	SC	GRIT	SP	SC	GRIT
SP	11.725	6.207	6.615	1.000	0.474	0.502	-		
SC	6.207	14.616	10.565	0.474	1.000	0.718	-	-	
GRIT	6.615	10.565	14.824	0.502	0.718	1.000	-	-	-
	A cov matrix			A cor matrix			proportions Σ_A/Σ_{Ph}		
Σ_A	SP	SC	GRIT	SP	SC	GRIT	SP	SC	GRIT
SP	9.009	4.806	4.868	1.000	0.653	0.996	**0.768**	0.774	0.736
SC	4.806	6.006	2.852	0.653	1.000	0.715	0.774	**0.411**	0.270
GRIT	4.868	2.852	2.649	0.996	0.715	1.000	0.736	0.270	**0.179**
	D cov matrix			D cor matrix			proportions Σ_D/Σ_{Ph}		
Σ_D	SP	SC	GRIT	SP	SC	GRIT	SP	SC	GRIT
SP	0.288	0.587	1.220	1.000	0.543	0.792	**0.025**	0.095	0.184
SC	0.587	4.054	5.449	0.543	1.000	0.943	0.095	**0.277**	0.516
GRIT	1.220	5.449	8.239	0.792	0.943	1.000	0.184	0.516	**0.556**
	E cov matrix			E cor matrix			proportions Σ_E/Σ_{Ph}		
Σ_E	SP	SC	GRIT	SP	SC	GRIT	SP	SC	GRIT
SP	2.428	0.813	0.527	1.000	0.244	0.170	**0.207**	0.131	0.080
SC	0.813	4.557	2.264	0.244	1.000	0.535	0.131	**0.312**	0.214
GRIT	0.527	2.264	3.936	0.170	0.535	1.000	0.080	0.214	**0.265**

and D_D are 3x3 lower triangular matrices. The results of fitting the ADE twin model are given in Table 2.

Table 3 includes the estimates of the covariance matrices Σ_A, Σ_D, and Σ_E, the phenotypic covariance matrix Σ_{Ph} (i.e., $\Sigma_A + \Sigma_D + \Sigma_E$), and the associated correlation matrices (i.e., Σ_A, Σ_D, Σ_E, and Σ_{Ph} standardized). The right columns of Table 2 show the proportions of the phenotypic variances and covariances attributable to A, D, and E factors. These proportions provide an interpretable decomposition of phenotypic (co)variance. For instance, the standardized variance of grit, conditional on the covariates (SES, sex, and rater), is expressed in proportions as follows.179 (A), .556 (D), and .265 (E). So, we know that about 73% of the phenotypic variance is due to genetic effects (17.9%+55.6%). The phenotypic correlation between

Table 3. ADE standardized variance components (corrected for sex and SES), including the variance attributable to rater (95% CIs in parentheses). The standardized A component gives narrow-sense heritability and the standardized A component + the standardized D component gives the broad-sense heritability.

	A	D	E	Teacher
School Performance	.712	.022	.192	.073
	(.609-.757)	(.003 - .105)	(.183-.221)	(.044 -.011)
Self-control	.410	.276	.311	.002
	(.261-.538)	(.207-.374)	(.286 - .315)	(.000-.113)
Grit	.139	.437	.209	.212
	(.093-.226)	(.371-.512)	(.194-.231)	(.183-.250)

Note. The four variance components are standardized, so add up to 1.

school performance and grit is .502. This correlation is expressed as proportions .736 (A), .184 (D), and .080 (E). So, 8% of the phenotypic correlation is attributable to E, unshared environmental factors, and 92% (73.6%+18.4%) is attributable to genetic factors). The correlation matrices are displayed in the middle columns. So, for example, the additive genetic correlation between self-control and grit is 0.715. The results in Table 2 are conditional on the covariates sex, SES, and rater (teacher). As mentioned above, the rater effect was modelled as a random effect, i.e., part of the covariance structure. Table 3 contains the standardized variance components including the proportion attributable to the rater effect.

So, the standardized variance of grit, conditional on the covariates (SES, sex, and rater), is expressed as proportions as follows: .139 (A), .437 (D), .209 (E), and .212 (rater). We note that the rater (teacher) effects, in terms of standardized variance are quite variable, ranging from 21.2% (grit) to .2% (self-control).

Causal regression model without confounding (model 3)

The causal regression model is depicted in Figure 2. In this model, the background correlations (associated with the dashed double-headed arrows in Figure 2) are fixed to zero, meaning that there is no background correlation due to common A, D, or E influences (i.e., no confounding). As such, this is model is consistent with the strong causal hypothesis that self-control and grit are causes of school performance. The LRT of this model relative to the ADE model equals LRT = 155.8, df = 4 ($p < 0.008$). The test has 4 degrees of freedom, because the ADE model includes six parameters to model the phenotypic covariance between self-control

and school performance and grit and school performance (two A covariances, two D covariances and two E covariance). But the causal model includes two parameters to model these covariances (i.e., the regression coefficients $b_{SP, SC}$ and $b_{SP, Grit}$). The difference in the number of parameters, which equals the degrees of freedom, is four. The LRT (155.8, df = 4, p<.008) clearly indicates that the causal model, without confounding, does not fit well, relative to the ADE model. This suggests at least that the effects of the predictors grit and self-control on the outcome school performance are not purely causal.

The causal model with confounding (models 4, 5, 6)

We added A, D, and E confounding to the model by including the relevant background A, D, and E correlations (dashed double-headed arrows in Figure 2). We considered A, D, and E confounding consecutively. We did not consider more than one source of confounding, as this, in combination with the phenotypic regression coefficients, renders the model equivalent to the ADE model in terms of the number of parameters used to model the associations. The LRT statistics, based on the comparison of the ADE model with the causal model

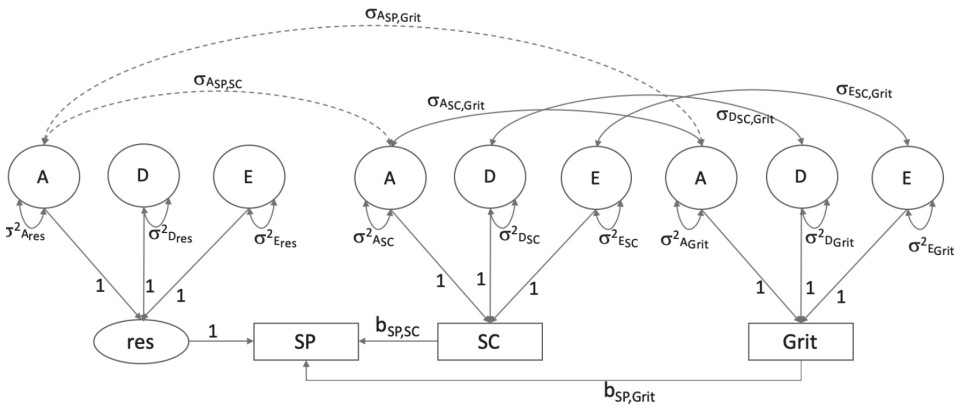


Figure 4. The model of choice: direct causal effects (parameters $b_{SP,SC}$ and $b_{SP,Grit}$) and A confounding (dashed double-headed arrows with covariances $s_{ASP,Grit}$, $s_{ASP,SC}$).

with confounding, are LRT = 0.407, df = 2, p = .815, LRT = 24.8, df = 2, p <.008, and LRT = 38.4, df = 2, p <.008, given A, D, and E confounding, respectively. The tests have two degrees of freedom, because the ADE model includes six parameters to

model the phenotypic covariance between self-control and school performance and grit and school performance (two A covariances, two D covariances and two E covariances). The causal regression model with confounding does this with four parameters: the regression coefficients and two A, D, or E covariances. These results suggest that the causal model with A confounding fits the data well, relative to the ADE model, but the other models clearly do not. The model of choice with direct causal effects and A confounding is shown in Figure 4.

In the causal model with A confounding, the estimates of the causal regression parameters $b_{SP,SC}$ and $b_{SP,Grit}$ are .151 (s.e. .027) and .047 (s.e. .033), respectively. The LRT statistics of the tests of $b_{SP,SC} = 0$ and $b_{SP,Grit} = 0$ are 2.44, df = 1, p = .118 ($b_{SP,SC}$) and 29.18, df = 1, p<.008 ($b_{Grit,SC}$). While the test of $b_{SP,SC}$ is not statistically significant, we conservatively retained this parameter in the model, and in the calculation of components of variance of school performance. The decomposition of the variance of school performance in raw and standardized variance components is given in Table 4.

Table 4. Decomposition of the school performance variance in raw and standardized estimates with 95% confidence intervals (95% CIs) in the causal regression model with A confounding.

variance components of school performance		raw estimate	proportion of variance (95% CIs)
causal due to self-control (SC)	$\beta_{SP,SC}^{2*}(\sigma^2_{Asc} + \sigma^2_{Dsc} + \sigma^2_{Esc})$	0.333	0.0284 (.019 - .051)
causal due to grit	$\beta_{SP,Grit}^{2*}(\sigma^2_{AGrit} + \sigma^2_{DGrit} + \sigma^2_{EGrit})$	0.032	0.0027 (.001 - .011)
causal due to covariance SC-grit	$2*\beta_{SP,SC}*\beta_{SP,Grit}*(\sigma_{Asc,Grit} + \sigma_{Dsc,Grit} + \sigma_{Esc,Grit})$	0.148	0.0126 (.007-.208)
confounding due to A	$2*\beta_{SP,SC}*\sigma_{ASC,SP} + 2*\beta_{SP,Grit}* \sigma_{AGrit,SP}$	1.457	0.124 (.101 - .134)
residual (res) variance	$\sigma^2_{Ares} + \sigma^2_{Dres} + \sigma^2_{Eres}$	9.769	0.832 (.816-.868)
total	σ^2_{SP}	11.739	1

The total explained variance of school performance is 16.8%, but by far the largest part of this 16.8% (12.4%) is due to genetic confounding. The causal effects account for 4.4% (i.e., see Table 4: 2.84% +.27%+1.26%) of the school performance variance. The decomposition of the 4.4% reveals that self-control (2.84% of the 4.4%) is a stronger predictor than grit (.27% of the 4.4%).

The results based on the causal regression model with A confounding differ appreciably from the phenotypic regression results both in terms of explained variance and in terms of the relative contributions of grit and self-control. In the phenotypic regression analyses, we found that grit and self-control accounted for 28.4% of the school performance variance, and we found that grit was the stronger predictor in terms of unique contributions (grit contributed 13%, self-control contributed 4.4% to the total of 28.4%). In the causal regression model with A confounding, we found that the total explained variance is lower at 16.8%: 4.4% due to the causal effects of self-control and grit, and 12.4% due to additive genetic confounding. In contrast to the phenotypic regression, the stronger predictor here is self-control. Grit is the stronger predictor in the phenotypic regression model because of the very high genetic correlation between grit and school performance (.996, see Table 2). This is largely due to A confounding. However, in the phenotypic regression model, which does not correct for A confounding, this A confounding contributes to the regression of school performance on grit. Consequently, grit emerges as the stronger predictor. Once we account for A confounding, the predictive value of grit is greatly reduced, and self-control emerges as the stronger predictor. The difference in total explained variance (28.4% in model 1 vs 16.8% in model 4) is a consequence of the influence of A confounding on the regression coefficients. The regression coefficients in the phenotypic regression model (model 1) are $b_{SP,SC}$ = .191 and $b_{SP,Grit}$ = .331, compared to $b_{SP,SC}$ = .151 and $b_{SP,Grit}$ = .047 in the causal regression model with A confounding (model 4). The bias in the regression model is due to confounding, see Supplementary Material for a detailed explanation.

Discussion

We investigated the direct causal pathway of two non-cognitive skills, grit and self-control, to school performance, while considering the role of additive genetic (A), dominant genetic (D), and unique environmental (E) confounding. Our results supported a model in which school performance is causally dependent on grit and self-control, in the presence of additive genetic (A) correlations (= additive genetic confounding). This means the association between grit, self-control,

and school performance is partially attributable to the direct causal effects of grit and self-control on school performance, and partly attributable to genetic correlations. The genetic correlations arise from the correlated effects of genes contributing to the variance of school performance, self-control and grit. This is called pleiotropic genetic effects. In this model, 83.2% was residual variance and 16.8% of the school performance variance was explained. Of this 16.8%, the genetic pleiotropy accounted for 12.4% and the causal effects accounted for 4.4%.

Our results support that self-control and grit predict school performance for of two reasons. Firstly, we found that self-control and grit had a direct impact on school performance, explaining some of its variance. Secondly, our results indicate that a portion of the association between self-control, grit, and academic performance was attributable to genetic confounding, or pleiotropy. As a result, children who struggle in school may face a double disadvantage. They are likely to have inherited genetic variants associated with lower self-control and grit, which are also associated with lower school performance. In contrast, children who excel in school are likely to have inherited genetic variants that increase their likelihood of exhibiting both better self-control and grit, as well as performing well academically.

The effectiveness of interventions designed to improve self-control and grit in enhancing school performance cannot be accurately predicted using phenotypic regression analysis. Specifically, it is not possible to predict the efficacy of a given intervention based on phenotypic regression analysis. Our model 1 results showed a fairly strong (multiple) correlation between self-control and grit with school performance. About 28% of the school-performance variance was accounted for by grit and self-control, which implies a fairly large multiple correlation of ($\sqrt{.28}$=).53. However, it is important to note that the phenotypic regression analysis does not correct for confounding, which could lead to greatly overestimating the direct causal effects of self-control and grit: the explained variance attributable to the causal effects of self-control and grit (model 4) was 4.4%, implying a multiple correlation of about ($\sqrt{.044}$=) .21. So, the phenotypic correlation is not a sound basis to predict the expected effect of an intervention addressing self-control and grit. Furthermore, we saw that the degree of additive genetic confounding may differ for self-control and grit. Specifically, when conducting the regression analysis while correcting for confounding (model 4), we found that self-control was a stronger causal predictor than grit, while in the standard phenotypic regression (model 1), grit was the stronger predictor. This

discrepancy was due to the high genetic correlation between grit and school performance (see Table 2). This stresses the importance of accounting for possible confounding. In the presence of confounding, the regression coefficients in the phenotypic regression model are biased estimates of the true causal regression coefficients. It is evident that the regression coefficients were affected by the inclusion of confounding: self-control appeared to be the better predictor in the causal model with confounding, while in the phenotypic regression model, grit was the most predictive factor.

The 4.4% causal effect of grit and self-control on school performance, as revealed by our study, suggests that an intervention targeting these non-cognitive skills could produce a modest yet meaningful improvement in academic performance. This is arguably a small effect, at least relative to the 28% as suggested by the standard phenotypic regression analysis. Previous studies on interventions on non-cognitive skills have shown to impact academic outcomes nevertheless. For example, Yeager et al. (2016) studied a growth-mindset intervention in high school students aimed to improve their academic outcomes. The intervention involved a series of online activities designed to help students adopt a growth mindset, which emphasizes the belief that intelligence is not fixed and can be developed through effort. Picking more challenging math problems correlated with self-control (.14) and grit (.16). Results showed that adolescents who adopted a growth mindset chose difficult problems more often. Results also indicated that the intervention significantly improved students' grades. Another intervention, the "tools of the mind" curriculum, focused on improving executive functioning and self-regulation skills in kindergarten children. Results indicated that this intervention improved academic outcomes (Diamond et al., 2019). Some of the effects were specific to high-poverty schools, suggesting that an intervention on executive functions and self-regulation in early elementary education might be an effective approach to reduce the achievement gap (Blair & Raver, 2014).

We acknowledge that this study has some limitations. We obtained our results with a causal model, in which self-control and grit influence school performance, and did not test direct causal pathways from school performance to grit and self-control, i.e. we have not ruled out the possibility of reverse causation. Jiang et al. (2019) found indications of a bidirectional effect between academic achievement and grit, suggesting that reverse causal effects play a role.

Our measures had limitations too, and we made our best efforts to properly address these in the analyses. To start with, all three measures displayed ceiling effects. As their distributions were skewed, we corrected for this by fitting the

models using maximum likelihood estimation, assuming that the data followed a multivariate right-censored distribution. Another challenge was that all measures were obtained by teacher ratings, and these teachers could be either the same teacher or different teachers for both twins in a pair. To address this challenge, we included the teacher as a random variable in our model and modelled its contribution. As a result, we were able to quantify the variance due to teacher sharing. Teachers are reliable reporters of school performance: our teacher ratings correlate .75 with scores on a nationally-standardised test of educational achievement (CITO scores). Both indicators of school performance had an estimated heritability of 74% (Kevenaar et al., 2023). A final limitation is that our grit measure mostly captured the perseverance of effort aspect of grit, not the consistency of interest. Previous studies have shown that the perseverance of effort aspect of grit is more related to academic outcomes than the consistency of interest (Credé, et al., 2017; Muenks, Wigfield, Yang, & O'Neal, 2017; Rimfeld, Kovas, Dale, & Plomin, 2016). Our second item of grit (on appropriate behaviour) relates less well to the grit concept. In Kevenaar et al. (2023) we demonstrate that this item nevertheless highly relates to the other items, and that the predictive value for school performance is not driven by this one item.

Some other points worth discussing concern the assumptions underlying the classical twin design. These include the assumption of non-assortative mating, the equal environment assumption, the representativeness of twins and the equality of variance across zygosity groups. Assortative mating refers to the tendency of individuals to seek out partners who resemble them. In our genetic structural equation models, we assumed random mating with regard to the traits under study in the parents of twins. Assortative mating can introduce biases that affect the estimates of the genetic and environmental influences on a trait (Kempthorne, 1959; Cavalli-Sforza & Bodmer, 1999). In the classical twin model, assortative mating will lead to an increase in the resemblance of DZ pairs, thereby inflating the shared environmental variance component. The assumption of random mating does not hold true for education, and we do not know if it holds for self-control and grit. We did not obtain evidence for the presence of C in our analyses, but it is possible that it was concealed by D. Another assumption of the classical twin design is the assumption of equal environment (EEA). The EEA states that environmental effects do not depend on the zygosity of the twins. Such dependence may arise due to the manifest physical similarity of MZ twins or similarity in the environmental response that MZ twins elicit. However, Evans and Martin (2000), Derks et al. (2006) and others found no detectable violation of the EEA. With respect to the equal variance assumption, i.e. the assumption

that the variance in the trait of interest is the same in MZ and DZ twins, we saw that the variances are similar in the MZ and the DZ twins, indicating that social interactions between twins are not likely to play a role (Martin et al., 1978). Regarding the representativeness of twins compared to the population, most of the differences found between twins and singletons are found in young children, but these differences usually disappear in adulthood. Concerning cognitive abilities specifically, no difference is found between twins and their singleton siblings (Evans and Martin, 2000; Willemsen et al. 2021).

In summary, our study sheds light on the relationship between non-cognitive traits and school performance, showing that while genetic factors play a significant role, self-control and grit also have a modest direct effect. The majority of the association is due to genetic influences shared by all three traits (12.4% due to pleiotropy), but importantly, self-control and grit do have a small direct effect on school performance (4.4%). This putatively causal effect is driven by self-control. So, interventions might want to target self-control to increase school performance, but the effects are expected to be small. Our findings have important implications for intervention researchers, educators, and policymakers seeking to improve student outcomes, suggesting that efforts to promote self-control may be a worthwhile strategy.

Literature

Achenbach, T. M., & Rescorla, L. A. (2000). *Manual for the ASEBA preschool forms and profiles* (Vol. 30). Burlington, VT: University of Vermont, Research center for children, youth, & families.

Blair, C., & Raver, C. C. (2014). Closing the achievement gap through modification of neurocognitive and neuroendocrine function: Results from a cluster randomized controlled trial of an innovative approach to the education of children in kindergarten. PloS one, 9(11), e112393.

Boomsma, D. I., De Geus, E. J., Vink, J. M., Stubbe, J. H., Distel, M. A., Hottenga, J. J., ... & Willemsen, G. (2006). Netherlands Twin Register: from twins to twin families. *Twin Research and Human Genetics*, *9*(6), 849-857.

Bruins S., Franić S., Dolan C.V., Borsboom, D., & Boomsma, D.I. (2023). Structural Equation Modeling in Genetics. In: R. Hoyle (ed.), *Handbook of Structural Equation Modeling,* Second Edition (pp. 646-663). The Guilford Press, New York.

Cavalli-Sforza, L. L., & Bodmer, W. F. (1999). *The genetics of human populations*. Courier Corporation.

Christopoulou, M., Lakioti, A., Pezirkianidis, C., Karakasidou, E., & Stalikas, A. (2018). The role of grit in education: A systematic review. *Psychology*, *9*(15), 2951-2971.

Credé, M., Tynan, M. C., & Harms, P. D. (2017). Much ado about grit: A meta-analytic synthesis of the grit literature. *Journal of Personality and social Psychology*, *113*(3), 492.

de Ridder, D. T., Lensvelt-Mulders, G., Finkenauer, C., Stok, F. M., & Baumeister, R. F. (2012). Taking stock of self-control: A meta-analysis of how trait self-control relates to a wide range of behaviors. *Personality and Social Psychology Review*, *16*(1), 76-99.

Derks, E. M., Dolan, C. V., & Boomsma, D. I. (2006). A test of the equal environment assumption (EEA) in multivariate twin studies. *Twin Research and Human Genetics*, *9*(3), 403-411.

de Zeeuw, E. L., van Beijsterveldt, C. E., Glasner, T. J., Bartels, M., de Geus, E. J., & Boomsma, D. I. (2014). Do children perform and behave better at school when taught by same-gender teachers?. Learning and Individual Differences, 36, 152-156.

de Zeeuw, E. L., van Beijsterveldt, C. E., Glasner, T. J., de Geus, E. J., & Boomsma, D. I. (2016). Arithmetic, reading and writing performance has a strong genetic component: A study in primary school children. *Learning and Individual Differences*, *47*, 156-166.

Diamond, A., Lee, C., Senften, P., Lam, A., & Abbott, D. (2019). Randomized control trial of Tools of the Mind: Marked benefits to kindergarten children and their teachers. PloS one, 14(9), e0222447.

Duckworth, A. L., Gendler, T. S., & Gross, J. J. (2016). Situational strategies for self-control. *Perspectives on Psychological Science*, *11*(1), 35-55.

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: perseverance and passion for long-term goals. *Journal of personality and social psychology*, *92*(6), 1087.

Duckworth, A. L., Taxer, J. L., Eskreis-Winkler, L., Galla, B. M., & Gross, J. J. (2019). Self-control and academic achievement. *Annual review of psychology*, *70*, 373-399.

Duckworth, A. L., Tsukayama, E., & May, H. (2010). Establishing causality using longitudinal hierarchical linear modeling: An illustration predicting achievement from self-control. *Social psychological and personality science*, *1*(4), 311-317.

Duffy, D. L., & Martin, N. G. (1994). Inferring the direction of causation in cross-sectional twin data: Theoretical and empirical considerations. *Genetic epidemiology*, *11*(6), 483-502.

Eskreis-Winkler, L., Shulman, E. P., Beal, S. A., & Duckworth, A. L. (2014). The grit effect: Predicting retention in the military, the workplace, school and marriage. *Frontiers in psychology*, *5*, 36.

Evans, D. M., & Martin, N. G. (2000). The validity of twin studies. *GeneScreen*, *1*(2), 77-79.

Hart, S. A., Little, C., & van Bergen, E. (2021). Nurture might be nature: Cautionary tales and proposed solutions. *NPJ science of learning*, *6*(1), 2.

Heath, A. C., Kessler, R. C., Neale, M. C., Hewitt, J. K., Eaves, L. J., & Kendler, K. S. (1993). Testing hypotheses about direction of causation using cross-sectional family data. *Behavior Genetics*, *23*, 29-50.

Henningsen, A. (2010). Estimating censored regression models in R using the censReg Package. *R package vignettes*, *5*, 12

Falconer, D. S., & Mackay, T. F. (1983). *Quantitative genetics*. London, UK: Longman.

Fernández Martín, F. D., Arco Tirado, J. L., & Hervás Torres, M. (2020). Grit as a predictor and outcome of educational, professional, and personal success: A systematic review.

Jiang, W., Xiao, Z., Liu, Y., Guo, K., Jiang, J., & Du, X. (2019). Reciprocal relations between grit and academic achievement: A longitudinal study. *Learning and Individual Differences*, *71*, 13-22.

Jinks, J. L., & Fulker, D. W. (1970). Comparison of the biometrical genetical, MAVA, and classical approaches to the analysis of the human behavior. *Psychological bulletin*, *73*(5), 311.

Keller, M., & Coventry, W. (2005). Quantifying and Addressing Parameter Indeterminacy in the Classical Twin Design. *Twin Research and Human Genetics, 8*(3), 201-213. doi:10.1375/twin.8.3.201

Kempthorne, O. (1957). *An introduction to genetic statistics*. Wiley, New York

Kevenaar, S. T., Dolan, C. V., Boomsma, D. I., & van Bergen, E. (2023). Self-control and grit are associated with school performance mainly because of shared genetic effects. JCPP Advances, e12159. Kohler, H. P., Behrman, J. R., & Schnittker, J. (2011). Social science methods for twins data: Integrating causality, endowments, and heritability. Biodemography and social biology, 57(1), 88-141.

Lam, K. K. L., & Zhou, M. (2019). Examining the relationship between grit and academic achievement within K-12 and higher education: A systematic review. *Psychology in the Schools*, *56*(10), 1654-1686.

Ligthart, L., van Beijsterveldt, C. E., Kevenaar, S. T., de Zeeuw, E., van Bergen, E., Bruins, S., ... & Boomsma, D. I. (2019). The Netherlands Twin Register: longitudinal research based on twin and twin-family designs. *Twin Research and Human Genetics*, *22*(6), 623-636.

Mackay, T. F. (2014). Epistasis and quantitative traits: using model organisms to study gene–gene interactions. *Nature Reviews Genetics*, *15*(1), 22-33.

Martin, N. G., Eaves, L. J., Kearsey, M. J., & Davies, P. (1978). The power of the classical twin study. *Heredity*, 40(1), 97-116.

Muenks, K., Wigfield, A., Yang, J. S., & O'Neal, C. R. (2017). How true is grit? Assessing its relations to high school and college students' personality characteristics, self-regulation, engagement, and achievement. *Journal of Educational Psychology*, *109*(5), 599.

Neale, M. C., Eaves, L. J., Kendler, K. S., & Hewitt, J. K. (1989). Bias in correlations from selected samples of relatives: The effects of soft selection. *Behavior Genetics*, *19*(2), 163-169

Neale, M. C., Hunter, M. D., Pritikin, J. N., Zahery, M., Brick, T. R., Kirkpatrick, R. M., Estabrook, R., Bates, T. C., Maes, H. H., & Boker, S. M. (2016). OpenMx 2.0: Extended Structural Equation and Statistical Modeling. *Psychometrika*, *81*(2), 535-549.

Oriol, X., Miranda, R., Oyanedel, J. C., & Torres, J. (2017). The role of self-control and grit in domains of school success in students of primary and secondary school. *Frontiers in psychology*, *8*, 1716.

Perels, F., Dignath, C., & Schmitz, B. (2009). Is it possible to improve mathematical achievement by means of self-regulation strategies? Evaluation of an intervention in regular math classes. *European Journal of Psychology of Education*, *24*(1), 17-31.

Postigo Gutiérrez, Á., Cuesta Izquierdo, M., Fernández Alonso, R., García Cueto, E., & Muñiz, J. (2021). Temporal stability of grit and school performance in adolescents: A longitudinal perspective. *Psicología Educativa, 27(1),* 77-84.

Rimfeld, K., Kovas, Y., Dale, P. S., & Plomin, R. (2016). True grit and genetics: Predicting academic achievement from personality. *Journal of personality and social psychology*, *111*(5), 780.

Tangney, J. P., Boone, A. L., & Baumeister, R. F. (2018). High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. In *Self-regulation and self-control* (pp. 173-212). Routledge.

van Beijsterveldt, C. E., Groen-Blokhuis, M., Hottenga, J. J., Franić, S., Hudziak, J. J., Lamb, D., ... & Boomsma, D. I. (2013). The Young Netherlands Twin Register (YNTR): longitudinal twin and family studies in over 70,000 children. *Twin Research and Human Genetics*, *16*(1), 252-267.

van Bergen, E., Snowling, M. J., de Zeeuw, E. L., van Beijsterveldt, C. E., Dolan, C. V., & Boomsma, D. I. (2018). Why do children read more? The influence of reading ability on voluntary reading practices. *Journal of Child Psychology and Psychiatry*, *59*(11), 1205-1214. https://doi.org/10.1111/jcpp.12910

Vazsonyi, A. T., Javakhishvili, M., & Blatny, M. (2022). Does self-control outdo IQ in predicting academic performance?. *Journal of Youth and Adolescence*, *51*(3), 499-508.

Verhulst, B., & Estabrook, R. (2012). Using genetic information to test causal relationships in cross-sectional data. *Journal of theoretical politics*, *24*(3), 328-344.

Willems, Y. E., Dolan, C. V., van Beijsterveldt, C. E., de Zeeuw, E. L., Boomsma, D. I., Bartels, M., & Finkenauer, C. (2018). Genetic and environmental influences on self-control: assessing self-control with the ASEBA self-control scale. *Behavior genetics*, *48*(2), 135-146.

Willemsen, G., Odintsova, V., de Geus, E., & Boomsma, D. I. (2021). Twin-singleton comparisons across multiple domains of life. *Twin and Higher-order Pregnancies*, 51-71.

Wolters, C. A., & Hussain, M. (2015). Investigating grit and its relations with college students' self-regulated learning and academic achievement. *Metacognition and Learning*, *10*, 293-311.

Yeager, D. S., Romero, C., Paunesku, D., Hulleman, C. S., Schneider, B., Hinojosa, C., ... & Dweck, C. S. (2016). Using design thinking to improve psychological interventions: The case of the growth mindset during the transition to high school. *Journal of Educational Psychology*, *108*(3), 374.

Supplementary Material

Abbreviations:

SP School performance, a phenotypic observed variable, conveyed in a square (Fig 1)

SC Self-control, a phenotypic observed variable, conveyed in a square (Fig 1)

Grit Grit, a phenotypic observed variable , conveyed in a square (Fig 1)

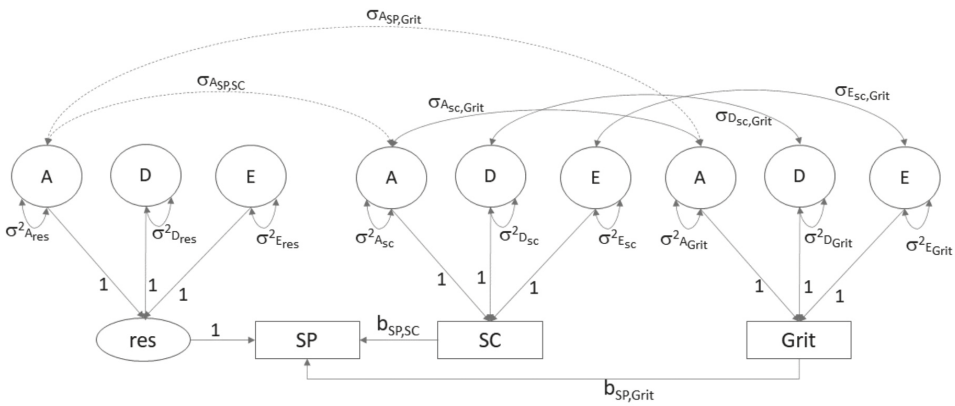
A Additive genetic variable, a latent variable, conveyed in a circle (Fig 1)

D Dominance variable, a latent variable, conveyed in a circle in the path diagram (Fig 1)

E Unshared environmental variable, a latent variable, conveyed in a circle in the path diagram (Fig 1)

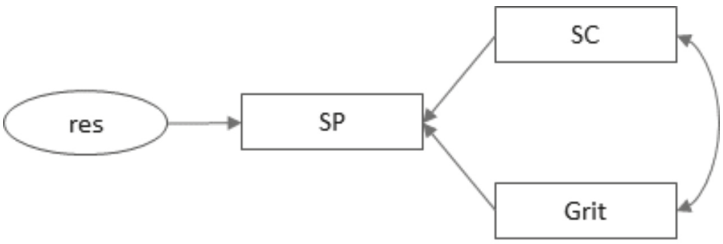
This supplement provides an account of the bias in the regression model originating in additive genetic confounding. The model is depicted in Figure 1. This model includes the phenotypic regression of SP on SC and Grit. The additive genetic confounding is represented by the dashed double headed arrows. The confounding represents a source of association between SP and SC and between SP and Grit, which is independent of the direct, phenotypic regression relations (coefficients $b_{SP,SC}$ and $b_{SP,Grit}$).

Figure 1



Given the model in Figure 1, we determine how the A confounding affects the phenotypic regression relationship given the following regression model (Figure 2):

Figure 2



The model in Figure 1 can be fitted given twin data; the model in Figure 2 can be fitted given data obtained in unrelated individuals, as this is the straightforward regression of SP on SC and Grit. The results of fitting the modelling in Figure 2 are informative from a predictive point of view, but the relations cannot be interpreted causally. Any confounding will affect the explained variance of SP (R^2) in the phenotypic regression model. This R^2 could be completely due to confounding, in the most extreme case.

To determine how the A confounding (as shown in Fig 1) affects the results of the phenotypic regression analysis, we first determine the expected phenotypic covariance matrix associated with the "true" model (Figure 1), S. This 3x3 covariance matrix S can be expressed as follows:

$\Sigma = (\mathbf{I}\text{-}\mathbf{B})^{-1}(\Sigma_A + \Sigma_D + \Sigma_E)\,(\mathbf{I}\text{-}\mathbf{B})^{-1t}$, where

$$\Sigma_A = \begin{matrix} a11 & \mathbf{a21} & \mathbf{a31} \\ \mathbf{a21} & a22 & a32 \\ \mathbf{a31} & a32 & a33 \end{matrix} = \begin{matrix} \sigma^2_{Ares} & \sigma_{ASP,SC} & \sigma_{ASP,Grit} \\ \sigma_{ASP,SC} & \sigma^2_{ASC} & \sigma_{ASC,Grit} \\ \sigma_{ASP,Grit} & \sigma_{ASC,Grit} & \sigma^2_{AGrit} \end{matrix}$$

where a21 and a31 are the source of confounding, i.e., a21 = $\sigma_{SP,SC}$ and a31 = $s_{SP,Grit}$.

$$\Sigma_D = \begin{matrix} d11 & 0 & 0 \\ 0 & d22 & d32 \\ 0 & d32 & d33 \end{matrix} = \begin{matrix} \sigma^2_{Dres} & 0 & 0 \\ 0 & \sigma^2_{DSC} & \sigma_{DSC,Grit} \\ 0 & \sigma_{DSC,Grit} & \sigma^2_{DGrit} \end{matrix}$$

$$\Sigma_E = \begin{matrix} e11 & 0 & 0 \\ 0 & a22 & e32 \\ 0 & e32 & e33 \end{matrix} = \begin{matrix} \sigma^2_{Eres} & 0 & 0 \\ 0 & \sigma^2_{ESC} & \sigma_{ESC,Grit} \\ 0 & \sigma_{ESC,Grit} & \sigma^2_{EGrit} \end{matrix}$$

$$(\mathbf{I}\text{-}\mathbf{B}) = \begin{matrix} 1 & -bs & -bg \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{matrix}$$

$$(\mathbf{I}\text{-}\mathbf{B})^{-1} = \begin{matrix} 1 & bs & bg \\ 0 & 1 & 0 \\ 0 & 0 & 1, \end{matrix}$$

where bs and bg are the regression coefficients (i.e., $b_{SP,SC}$ and $b_{SP,Grit}$, in Figure 2)

We represent the 3x3 covariance matrix $Y = \mathbf{\Sigma}_A + \mathbf{\Sigma}_D + \mathbf{\Sigma}_E$ as follows:

$$\begin{matrix} \alpha & \beta & \chi \\ \beta & \delta & \varepsilon \\ \chi & \varepsilon & \phi \end{matrix} = \begin{matrix} a11+d11+e11 & a21 & a31 \\ a21 & a22+d22+e22 & a32+d32+e32 \\ a31 & a32+d32+e32 & a33+d33+e33 \end{matrix}$$

The expected covariance matrix is $S = (I\text{-}B)^{-1} Y (I\text{-}B)^{-1t} =$

α + bs*β+bg*χ + bs*(β + bs*δ+bg* ε) + bg*(χ + bs*ε+bg*ϕ)	β + bs*δ + bg*ε	χ + bs*ε + bg*ϕ
β + bs*δ + bg*ε	δ	ε
χ + bs*ε + bg*ϕ	ε	ϕ

In fitting the linear regression model (Figure 2), we estimate the regression coefficients ***b*** (2x1) as follows:

$\boldsymbol{b} = \Sigma_X^{-1}\Sigma_{XY}$, where

$$\Sigma_X = \begin{matrix} \delta & \varepsilon \\ \varepsilon & \phi \end{matrix} = \begin{matrix} a22+d22+e22 & a32+d32+e32 \\ a32+d32+e32 & a33+d33+e33 \end{matrix}$$

and

$$\Sigma_{xy} = \begin{matrix} \beta + bs^{*}\delta + bg^{*}\varepsilon \\ \chi + bs^{*}E + bg^{*}\phi \end{matrix}$$

The inverse Σ_x^{-1} equals

$$\begin{matrix} \phi/(\delta^{*}\phi - \varepsilon^{2}) & -\varepsilon/(\delta^{*}\phi - \varepsilon^{2}) \\ -\varepsilon/(\delta^{*}\phi - \varepsilon^{2}) & \phi/(\delta^{*}\phi - \varepsilon^{2}), \end{matrix}$$

where $(\delta^{*}\phi - \varepsilon^{2})$ is the determinant of Σ_x. Let Σ_x^{-1} equal

$$\begin{matrix} \phi/(\delta^{*}\phi - \varepsilon^{2}) & -\varepsilon/(\delta^{*}\phi - \varepsilon^{2}) \\ -\varepsilon/(\delta^{*}\phi - \varepsilon^{2}) & \phi/(\delta^{*}\phi - \varepsilon^{2}) \end{matrix} = \begin{matrix} x11 & x21 \\ x21 & x22 \end{matrix}$$

So, we can express the regression coefficients in the phenotypic regression model (Fig 2) as follows:

b =

$$\begin{matrix} x11^{*}(\beta + bs^{*}\delta + bg^{*}\varepsilon) + x21^{*}(\chi + bs^{*}\varepsilon + bg^{*}\phi) \\ x21^{*}(\beta + bs^{*}\delta + bg^{*}\varepsilon) + x22^{*}(\chi + bs^{*}\varepsilon + bg^{*}\phi) \end{matrix}$$

b =

$$\begin{matrix} \phi/(\delta^{*}\phi - \varepsilon^{2})^{*}(\beta + bs^{*}\delta + bg^{*}\varepsilon) + (-\varepsilon/(\delta^{*}\phi - \varepsilon^{2}))^{*}(\chi + bs^{*}\varepsilon + bg^{*}\phi) \\ (-\varepsilon/(\delta^{*}\phi - \varepsilon^{2}))^{*}(\beta + bs^{*}\delta + bg^{*}\varepsilon) + \phi/(\delta^{*}\phi - \varepsilon^{2})^{*}(\chi + bs^{*}\varepsilon + bg^{*}\phi) \end{matrix}$$

We know that the A confounding is due to β =a21 and χ =a31. So, making this substitution, we have (conveying a21 and a31 in green to highly these parameters):

b =

x11*(a21+bs*δ + bg*ε) + x21*(a31+bs*ε + bg*ϕ)

x21*(a21+bs*δ + bg*ε) + x11*(a31+bs*ε + bg*ϕ)

The a21 and a31, being positive (see main article), result in an upwards bias in the regression estimates. Given a21=a31=0, we have

x11*(bs*δ + bg*ε) + x21*(bs*ε + bg*ϕ)

x21*(bs*δ + bg*ε) + x11*(bs*ε + bg*ϕ)

So, the regression coefficients in the phenotypic regression analysis (Figure 2) are overestimated by

x11*a21 + x21*a31 = ϕ /(δ *ϕ - ε^2)*a21 + (-ε /(δ *ϕ - ε^2))*a31

and

x21*a21 + x11*a31 = (-*e* /(*d* **f* - *e*^{2}))*a21 + *f*/(*d* **f* - *e*^{2})*a31

Numerical results

The values of bs and bg in the causal + A confounding model (Figure 1):

```
> print(c(bs, bg))
[1] 0.151 0.047
```

These values maximum likelihood estimates taken from the OpenMx output. We calculate these values based on the above, and obtain about the same results:

```
> print(c(f1_,f2_))
[1] 0.15100000 0.05110704
```

So ~.15 (self-control) and ~.05 (grit) are the values of the regression coefficients in the true model (Figure 1)

The values as calculated in the linear regression model (Figure 2) are shown below (based on OpenMx output):

```
> print(c(f1,f2))
[1] 0.1906227 0.3307272
```

The values based on the above are about the same:


```
> print(c(g1,g2))
[1] 0.2067383 0.3123943
```

The bias in the regression parameters due to confounding.

```
> print(c(bias1, bias2))
[1] 0.05573831 0.26128727
```

This corresponds to

$$x11*a21 + x21*a31 = \phi /(\delta *\phi - \varepsilon^2)*a21 + (-\varepsilon /(\delta *\phi - \varepsilon^2))*a31$$

and

$$x21*a21 + x11*a31 = (-\varepsilon /(\delta *\phi - \varepsilon^2))*a21 + \phi/(\delta *\phi - \varepsilon^2)*a31$$

The confounding is the source of the bias. In the true model (Figure 1), the regression coefficients are about .151 (SC) and .05 (Grit). In the regression model (Figure 2), we obtain about .20 (SC) and .32 (Grit). The differences about .151 vs .20 and .05 vs .32 is due to confounding (i.e., parameters a21 and a32, or, as conveyed in Figure 1, the covariances $s_{ASP,SC}$ and $s_{ASP,Grit}$).

#Numerical check fitted model

```
bs= 0.1510
bg= 0.0470
a11= 2.7348
a21= 1.2714
a31= 1.5858
a22= 2.1455
a32= 0.2914
a33= 0.0003
d11= 0.0023
d22= 1.9542
d32= 2.8777
d33= 0.0021
e11= 1.5123
e22= 2.1378
e32= 1.0477
```

```
e33= 1.6795


DA=matrix(
c(a11,0,0,
a21,a22,0,
a31,a32,a33),3,3,byrow=T)
A=DA%*%t(DA)
a11=A[1,1]
a21=A[2,1]
a31=A[3,1]
a22=A[2,2]
a32=A[3,2]
a33=A[3,3]

DD=matrix(
c(d11,0,0,
0,d22,0,
0,d32,d33) ,3,3,byrow=T)
D=DD%*%t(DD)
d11=D[1,1]
d21=D[2,1]
d31=D[3,1]
d22=D[2,2]
d32=D[3,2]
d33=D[3,3]
```

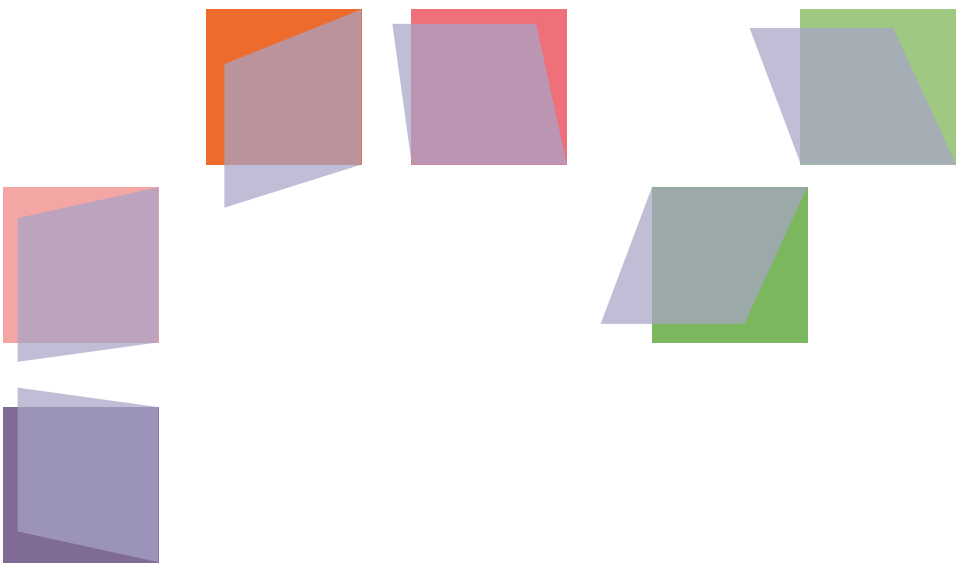
```
DE=matrix(
c(e11,0,0,
0,e22,0,
0,e32,e33) ,3,3,byrow=T)
E=DE%*%t(DE)
e11=E[1,1]
e21=E[2,1]
e31=E[3,1]
e22=E[2,2]
e32=E[3,2]
e33=E[3,3]


#The linear regression model.

f1=0.1906227
f2=0.3307272

#

A=a11+d11+e11;B=a21;C=a31
B=a21;D=a22+d22+e22;E=a32+d32+e32
C=a31;E=a32+d32+e32;F=a33+d33+e33
#
x11=F/(D*F-E^2); x21=-E/(D*F-E^2)
x21=-E/(D*F-E^2); x11=F/(D*F-E^2)
#
```

```
g1=x11*(a21+bs*D+bg*E)  + x21*(a31+bs*E+bg*F)
g2=x21*(a21+bs*D+bg*E) + x11*(a31+bs*E+bg*F)
#

f1_= x11*(bs*D+bg*E)  + x21*(bs*E+bg*F)
f2_= x21*( bs*D+bg*E) + x11*( bs*E+bg*F)

bias1=x11*a21 + x21*a31
bias2=x21*a21+ x11*a31



print(c(bs, bg))
print(c(f1_,f2_))
#
print(c(f1,f2))
print(c(g1,g2))
#
print(c(bias1, bias2))
#
#
SX=matrix(c(
a22+d22+e22, a32+d32+e32,
a32+d32+e32, a33+d33+e33), 2,2)
resV=a11+d11+e11
#
G=matrix(c(g1,g2),2,1) #  biased
(t(G)%*%SX%*%G) / (t(G)%*%SX%*%G+resV)
```

```
#
(t(G)%*%SX%*%G)
B=matrix(c(bs,bg),2,1)
t(B)%*%SX%*%B
```

Chapter 6

Summary and general discussion



Summary

The aim of the research in this dissertation was to disentangle the source of individual differences in a range of childhood traits and characteristics, namely height, self-control, grit, and school performance, employing a variety of models. I first summarize the findings of the research in this dissertation below and next offer a general discussion and reflection on the research presented.

In chapter 2, I investigated the role of geographical location as a third-level variable in a twin design, where the first level is the child level and the second level is the family level. The reparameterization of the classical twin model into an equivalent multilevel model provided the possibility to include higher-level variables in which the lower-level variables are nested, in this case, geographical location as determined by postal codes. I demonstrated the use of a 3-level multilevel model to analyse twin data and investigate the regional clustering of 7-year-old children's height in the Netherlands. Our research revealed that ~2% of the phenotypic variance in children's height can be attributed to regional clustering, which accounts for 7% of the variance explained by common environmental components between families. On average, children in the North of the Netherlands are taller than children in the South, and boys are taller than girls. I also examined the potential effect of genetic ancestry on regional clustering by assessing the impact of genetic principal components in a subset of participants with genome-wide single nucleotide polymorphism (SNP) data. Principal component analysis of the covariance matrix of the SNP data allows us to identify genetic principal components, which reflect allele frequency gradients. In the Netherlands, the first genetic principal component reflects a north-south gradient, the second an east-west division, and the third the more central regions of the country. Our results indicated that, after accounting for genetic factors, region no longer had a significant effect on height variation. These findings suggest that the phenotypic variance explained by regional clustering are attributable to ancestry effects on height.

In chapter 3, the power of multi-study collaboration and Bayesian evidence synthesis was demonstrated. In this study, I showcase the utilization of Bayesian evidence synthesis to quantify and compare the level of support for different competing hypotheses, and to consolidate this support across various studies. I implemented this approach to investigate the ranking of multi-informant scores on the ASEBA Self Control Scale (ASCS) in a multi-cohort design with information from four Dutch cohorts. As the set of available reporters on children's self-

control varied across the cohorts (e.g., parents, teachers, self-reports), each cohort evaluated different aspects of the overall competing hypotheses. Our findings consistently provided evidence for the partial hypothesis that parents reported more self-control issues than teachers. The aggregated results showed most support for the hypothesis that children report the highest number of self-control problems, followed by their mothers and fathers, while teachers report the fewest problems. However, there were inconsistencies in the ordering of self-reported self-control problems. This chapter illustrated the importance of taking the informant into account, and the potential of combining results from different studies with Bayesian evidence synthesis.

In chapter 4, I combined genetic covariance structure modelling with regression to investigate the prediction of school performance by self-control and grit. The design allowed for disentangling genetic and environmental sources of variation. The results showed that a large portion of the individual differences in school performance, namely 28.4%, can be explained by self-control, grit, and their covariance. Zooming in on the aetiology, results showed that most of the explained variance in school performance was attributable to the genetic components of grit and self-control. After correcting for SES, sex, and rater (teachers), only 1.3% of the explained variance in school performance was attributable to environmental factors. In this study, I addressed two data issues. First, I had to correct for censoring in the analyses, because the distribution of the variables, especially that of self-control, was skewed. This was the consequence of a ceiling effect, meaning that a lot of children scored the highest possible score. I handled this by fitting the models using maximum likelihood estimation, subject to the assumption that the data followed a multivariate, right-censored distribution. Second, I recognized that the teachers, who rated the twins with respect to the phenotype, were a source of systematic variance. A strength of this study is that I could estimate the teacher (rater) variance because some twins were in the same class and hence shared the teacher, while others did not. This allowed to estimate the rater variance.

In chapter 5 I further explored the relationship between self-control, grit, and school performance, by applying an explicit causal model. That is, I considered the phenotypic regression of school performance on grit and self-control, while taking into account possible confounding due to background genetic or environmental influences (influences common to all three phenotypes). Demonstrating the phenotypic regression relationship, while taking into account confounding, supports the interpretation of the influence of grit and self-control on school

performance as causal. The results showed that most of the relationship between self-control, grit, and school performance was due to genetic confounding, which is likely to reflect genetic pleiotropy: i.e., the genes that explained individual differences in self-control and grit also explained individual differences in school performance. I did observe direct (phenotypic) effects of self-control and grit on school performance as well, but these direct effects accounted for only 4.4% of the variation, while 12.4% was accounted for by pleiotropy.

General Discussion

In this general discussion, I aim to address some of the methodological issues I have encountered while working on this dissertation and some future directions for investigating childhood individual differences.

In the general introduction of my dissertation, I raised the topic of how to measure phenotypes, including traits, behaviours, and skills, in children. The phenotypes that feature in this dissertation are important in multiple contexts in the life of a child, both at school and in the home environment. I had the opportunity to investigate these phenotypes across multiple contexts, by analysing measures obtained from different informants, namely the fathers, mothers, and teachers of children, and at different ages of the children. All these informants see children in different contexts. In chapter 3, I used Bayesian evidence synthesis to compare self-control scores rated by different informants in multiple different datasets, collected by large cohort studies across the Netherlands. This study showed that it is important to consider who provided information about the children, because informants rate children's behaviour differently.

In the projects that constitute this dissertation, I had the opportunity of working with different types of data. These included easily measurable data such as children's height using a measuring rod, but also school grades reported by teachers, and standardized tests developed by CITO. Additionally, I analysed non-cognitive skills data of children collected in large-scale surveys including questionnaires such as the ASEBA instruments, which were completed by parents, teachers, or the children themselves. While the psychometric properties of these questionnaires are good, there were still challenges in using them in the present projects. One of these challenges was the presence of ceiling effects. In the ASCS (ASEBA Self-Control Scale, Willems et al., 2019) that was used in chapters 3, 4, and 5, items are scores on a three-point scale indicating problems with self-control, where 0 = not true, 1 = somewhat or sometimes true, and 2 = very true or often true. I reverse-coded these scores in chapters 4 and 5 so that

higher scores indicated better self-control. A lot of children obtained the maximal score, so there was a ceiling effect (in the original scale used in chapter 3, the minimal score, so a floor effect, was observed), resulting in a skewed distribution. This can lead to biased estimates of the underlying statistical distribution or relationships between variables. To analyse the skewed (censored) data optimally, a correction was applied in chapters 4 and 5. This entailed maximum likelihood estimation assuming that the data followed a censored multivariate normal distribution. Note that in chapter 3, the raw scores were used instead of the scores that were corrected for censoring, because the goal in chapter 3 was to investigate and report mean informant differences when using this scale, so original scale was used. By adjusting for censoring, researchers can obtain more accurate estimates of parameters and relationships, leading to better insights and conclusions. However, the impact of the censoring correction on the results and conclusions in chapter 4 and 5 appeared to be small.

The research in this dissertation also illustrates the power of the classical twin design. The classical twin design is in principle a simple, intuitive, and therefore attractive study design. It entails a comparison of the resemblance in monozygotic (MZ) and dizygotic (DZ) twins. This requires, of course, a proper operationalization of the phenotype, and the means to diagnose zygosity (MZ vs DZ). As described in Appendix 1 of this dissertation, determining the zygosity of same-sex twin pairs is feasible with a few items about the twins' resemblance. I assessed the accuracy of this manner of diagnosing zygosity by a series of discriminant analyses, in which I compared zygosity classification by survey items on twin resemblance to the zygosity determination based on blood group or DNA polymorphism information as an index of 'true zygosity'. The accuracy of the zygosity determination procedure by the resemblance items was as high as 96.8% in children, meaning that this is a very useful tool to determine if twin pairs are monozygotic or dizygotic.

Children's height, which features in chapter 2 of this dissertation, is relatively simple to measure accurately, shows a normal distribution in the population, and has been studied well. Therefore, analyses of height often serve as a 'golden standard' in genetic studies. Together with postal code information about the clustering of families in geographical regions, the data on height created the opportunity to explore a multilevel implementation of the twin design. I also incorporated information about genetic variants into a twin model by inclusion of the Principal Components (PCs; an abbreviation that, by the way, has turned out to be very prominent in my PhD trajectory. Depending on the context it could

mean Principal Component, Postal Code, Parental Consent, Personal Computer, or Politically Correct). This study illustrated the possibilities of twin designs perfectly: adding additional clustering information to the model results in a new realm of questions that can be answered.

However, in some situations, the implementation of the classical twin design turns out to be challenging. A challenge of working with teacher ratings in twin data is that one has to account for the fact that some twin pairs are in the same class with the same teacher, and some twin pairs are in different classes with different teachers. This implies that, when teachers rate twins with respect to given phenotypes, the members of some twin pairs are assessed by the same person, whereas the members of other twin pairs are not. This may be problematic because the assignment of twins to one class or to different classes may not be a random process. Two examples of a non-random process are that, first, behavioural problems may be a reason to separate twins, and second, smaller schools (in rural areas) may only have one class for every year grade, which effectively rules out separating the twins into different classes. In addition to this classroom sharing not being random, teacher ratings can be influenced by twin zygosity. When rated by the same teacher, a teacher might be more prone to focus on the similarities in monozygotic twins and the differences in dizygotic twins.

In all analyses, the fact that some twins shared a teacher and some did not, was accounted for by adding an extra “latent teacher variable”, which was correlated 1 between twins sharing a class and teacher and 0 between twins in different classes with different teachers. Addressing teacher sharing in this way provided an interesting additional research opportunity, because it allowed for quantification of individual differences that could be attributed to being rated by the same or by different teachers. Part of the variance in children’s scores is due to child factors and part is due to informant or rater effects. Twin studies that include both monozygotic (MZ) and dizygotic (DZ) twin pairs who either share a class or not can help researchers separate variation in childhood traits due to child factors (such as genetics and personality) from informant or rater effects. Child factors are correlated between twins, and the difference in correlation between MZ and DZ twins can provide information on the proportion of variance attributable to genetic effects, shared environmental effects, and unique environmental effects. With twin data, we are able to quantify the teacher effects and estimate the heritability of the child part. We found that genetic factors explain 58%, 69%, and 74% of the individual differences in grit, self-control and school performance respectively.

The teacher effect consists of two parts, the rater part and the classroom part. The rater part is due to the child being evaluated by the same teacher. This part of the teacher effect reflects how much teachers' ratings of the child are influenced by their own personal biases or preferences, rather than being a true reflection of the child's abilities or performance. The classroom part of the teacher effect reflects the influence that teachers and the rest of the class have on the child. This part of the teacher effect can be seen as a measure of the overall quality of the learning environment created by the teacher and the group dynamic within the class. If a measure is an objective test, like a standardized performance test, you only capture the classroom part. If a measure is a teacher rating, like ratings of school grades, you capture both the rater part and the classroom part. Lamb et al. (2012) demonstrated the interaction between genetics and environment in relation to teacher-rated internalizing and externalizing problem behaviours, and argued that the classroom environment, including the teacher and peer dynamics, plays a role in the manifestation of problem behaviour. As this study made use of teacher reports, it was not possible to disentangle the rater part and the classroom part. The research in this dissertation indicated that the teacher effects (including both the classroom and rater part) appeared to vary considerably, ranging from almost no variance explained by teacher in self-control to as much as 21% of the variance explained by teacher in grit. A way to investigate purely the classroom part, is to work with standardized tests instead of teacher ratings. Grasby et al. (2020) studied the impact of the classroom environment on achievement test scores. Their findings indicate that the classroom effect explains 2-3% of the variance in test scores. Stienstra et al. (2022) made use of the fact that some twins are in the same class and some twins are in different classrooms in studying if classrooms increase or decrease educational inequality. The influence of the classroom appeared to be larger for children with lower socioeconomic backgrounds and smaller for children from higher socioeconomic backgrounds, indicating a compensatory effect of classroom. The effect of the classroom on children's traits and skills is part of environmental effects. Overall, twin studies can provide valuable insights into the complex interplay between genes and environment in shaping childhood individual differences.

Prediction with correlated predictor variables requires special attention. Chapters 4 and 5 illustrate that the covariance between self-control and grit, hence variance that cannot be unambiguously ascribed to either self-control or grit, explains a significant portion of the variance in school performance. Thus, when in a study only one of these constructs is available for prediction, one automatically also captures part of the variance explained by the other ones. Consequently, as long

non-cognitive factors are covered in a study it might not matter that much what specific non-cognitive factors are included.

Confounding, both within and outside of the genetic field, is a common concept that relates to the adage "correlation does not imply causation". Confounding refers to the phenomenon that a putatively causal relationship between two variables is wholly or partly attributable to a third variable that is associated with both variables. A confounding variable that is not properly accounted for in the analysis can lead to inaccurate conclusions about the true relationship between the variables of interest. When comparing the results reported in chapters 4 and 5, the parameter estimates of the predictors showed different patterns for the same data, pointing to the importance of model choice and the inclusion of confounders. In chapter 4, genetic grit appeared to be the best predictor of school performance. However, when I tested a causal model and accounted for genetic confounding (multiple traits that are influenced by the same genes), the direct causal path to school performance appeared to be stronger for self-control than for grit. Moreover, the direct causal effect of self-control and grit on school performance appeared to be small, with most of the association between self-control, grit, and school performance due to genetic pleiotropy, i.e. genes that affect self-control and grit as well as school performance.

I presented Bayesian evidence synthesis as a useful approach for combining results from different data sources, even when the datasets each provide information about only a specific part of the hypotheses. This method offers several benefits. One benefit is that each dataset can be used to test partial hypotheses, and with Bayesian evidence synthesis, this information can be combined to obtain the evidence for the full set of hypotheses, as illustrated in chapter 3. Bayesian evidence synthesis can also be used to combine information obtained using different measurement instruments. This enables researchers to combine evidence collected in a wide range of studies. A benefit of Bayesian evidence synthesis over frequentist approaches concerns the interpretation of the evidence, namely as the degree of support for competing hypotheses instead of the rejection (or not) of a single hypothesis. Furthermore, Bayesian techniques are not subject to the effects of optional stopping, meaning that the results can be updated when new data become available. However, contrary to meta-analysis, Bayesian evidence synthesis requires access to the raw data. In the Consortium of Individual Development (CID), this prerequisite was satisfied, and we could analyse raw data collected at four different sites. Each site provided information to test different parts of the hypotheses, resulting in a comprehensive evaluation of the entire set of hypotheses.

Big consortium collaborations, like the Consortium of Individual Development (CID), have proven to be very fruitful. These consortia enable researchers to combine knowledge and data sources to undertake studies that are feasible due to the collaboration and aggregation of data resources. Because of the collaborations in the Consortium of Individual Differences, self-control got to play an important role within the Netherlands Twin Register. Willems et al. (2018) developed a self-control measure based on the ASEBA scales (Child Behaviour Checklist (CBCL), Teacher Report Form (TRF), Youth Self-Report (YSR); Achenbach & Rescorla, 2014), which are administered in many (longitudinal) cohort studies. This provided the opportunity to study self-control in large, established datasets like the Netherlands Twin Register (NTR), TRacking Adolescents' Individual Lives Survey (TRAILS), Generation R, and YOUth. A valuable way for data enrichment, relevant to future projects, is record linkage, where participants of a study consent to their data being linked to other data sources in a safe environment (for example, data from Statistics Netherlands; CBS). Currently, research infrastructures like ODISSEI (Open Data Infrastructure for Social Science and Economic Innovations) make it possible to undertake new and interdisciplinary research questions. For example, an ODISSEI-supported study by de Zeeuw et al. (2021) described and utilized record linkage between the NTR, which includes genotype data, and register data from Statistics Netherlands (CBS) to conduct a Genomewide Association Study on healthcare expenditure.

Research on individual differences in children is valuable to advance our understanding of why children differ in skills, behaviour, health, and physical traits, why some children do better than others, and why some children thrive, whereas other do not. It might seem straightforward to point to environmental factors to explain differences between children, but research in behaviour genetics indicates that genetic factors play an important role as well, at as early as age 3 (Bartels et al., 2004; Van den Oord, Verhulst & Boomsma, 1997). A role for genetics implies that we need to consider the gene sharing between parents and children when examining areas such as parenting or the home environment. An association does not necessarily imply a causal mechanism (Hart, Little & Van Bergen, 2021). Assuming a causal mechanism where none exists can be harmful.

Conclusion

In analysing childhood individual differences, the aim is to identify how children differ from each other, and what factors underlie these differences. This can be done by (combining) data from large samples and adopting a variety of methods and models. In childhood, survey reports are often an important data source. In analysing such survey data, it is important to take into account who provided the information about the children and to recognise potential differences between father and mothers, teachers, and the children themselves. Children grow up in groups, i.e., families, classes, schools, neighbourhoods, and countries. In statistics, we refer to this as clustering. Accounting for clustering offers additional research opportunities, e.g., investigating the aetiology of a trait, because we know the family clustering. Because we can identify the zygosity of twin pairs well, we can use the difference between groups of monozygotic and dizygotic twin pairs to disentangle genetic and environmental sources of individual differences. Another utilization of clustering involves geographical location. Knowing which children cluster in the same neighbourhoods and combining this with family clustering, zygosity information and genetic principal components enables the investigation of the role of geographical location and genetic ancestry. Furthermore, when examining individual differences, it is crucial to consider confounding, particularly if the goal is to make any inferences about causality. This dissertation shows that childhood individual differences as they arise in the world around us are to a large extent due to genetic differences between children, but the environment also plays a role in why children differ from each other.

Literature

Achenbach, T. M., & Rescorla, L. A. (2014). The Achenbach system of empirically based assessment (ASEBA) for ages 1.5 to 18 years. In The use of psychological testing for treatment planning and outcomes assessment (pp. 179-214). Routledge.

Bartels, M., Van den Oord, E. J. C. G., Hudziak, J. J., Rietveld, M. J. H., Van Beijsterveldt, C. E. M., & Boomsma, D. I. (2004). Genetic and environmental mechanisms underlying stability and change in problem behaviors at ages 3, 7, 10, and 12. *Developmental psychology*, *40*(5), 852.

de Zeeuw, E. L., Voort, L., Schoonhoven, R., Nivard, M. G., Emery, T., Hottenga, J. J., ... & Boomsma, D. I. (2021). Safe Linkage of Cohort and Population-Based Register Data in a Genomewide Association Study on Health Care Expenditure. *Twin Research and Human Genetics*, *24*(2), 103-109.

Emery, T., Braukmann, R., Wittenberg, M., van Ossenbruggen, J., Siebes, R., & van de Meer, L. (2020). The ODISSEI Portal: Linking Survey and Administrative Data

Grasby, K. L., Little, C. W., Byrne, B., Coventry, W. L., Olson, R. K., Larsen, S., & Samuelsson, S. (2020). Estimating classroom-level influences on literacy and numeracy: A twin study. *Journal of Educational Psychology, 112(6),* 1154

Hart, S. A., Little, C., & van Bergen, E. (2021). Nurture might be nature: Cautionary tales and proposed solutions. *NPJ Science of Learning, 6(1)*, 2.

Lamb, D. J., Middeldorp, C. M., Van Beijsterveldt, C. E., & Boomsma, D. I. (2012). Gene–environment interaction in teacher-rated internalizing and externalizing problem behavior in 7-to 12-year-old twins. *Journal of Child Psychology and Psychiatry, 53(8),* 818-825

Stienstra, K., Knigge, A., Maas, I., de Zeeuw, E. L., & Boomsma, D. I. (2022). Are classrooms equalizers or amplifiers of inequality? A genetically informative investigation of educational performance. *European Sociological Review*.

van den Oord, E. J., Verhulst, F. C., & Boomsma, D. I. (1996). A genetic study of maternal and paternal ratings of problem behaviors in 3-year-old twins. *Journal of Abnormal Psychology*, *105*(3), 349.

Willems, Y. E., Dolan, C. V., van Beijsterveldt, C. E., de Zeeuw, E. L., Boomsma, D. I., Bartels, M., & Finkenauer, C. (2018). Genetic and environmental influences on self-control: Assessing self-control with the ASEBA self-control scale. *Behavior Genetics*, *48*, 135-146.

Appendix

Appendix 1

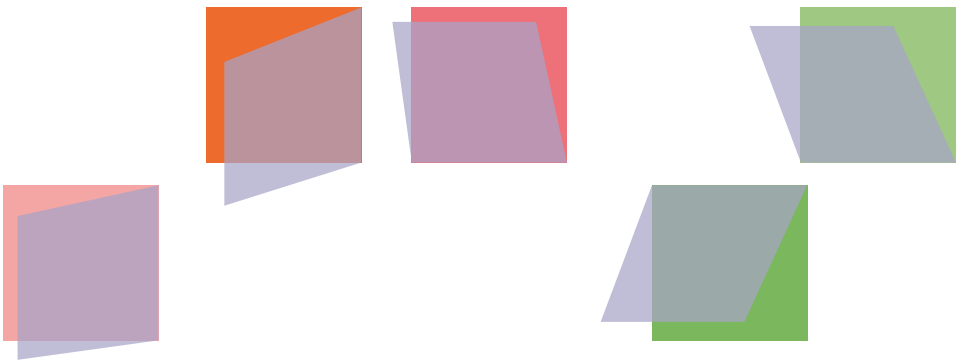
Zygosity assessment by survey items: agreement with zygosity based on a blood group or DNA tests

Appendix 2

Data collection in teachers of twins and their siblings in the Netherlands Twin Register

Appendix 3

Attachments of data collection in teachers of twins and their siblings in the Netherlands Twin Register



This appendix was published online as a supplement to: Ligthart, L., van Beijsterveldt, C. E., **Kevenaar, S. T.,** de Zeeuw, E., van Bergen, E., Bruins, S., ... & Boomsma, D. I. (2019). The Netherlands Twin Register: longitudinal research based on twin and twin-family designs. *Twin Research and Human Genetics*, *22*(6), 623-636.

Appendix 1

Zygosity assessment by survey items: agreement with zygosity based on a blood group or DNA tests.

To determine the accuracy of zygosity determination in children aged 0-16, the agreement between zygosity based on a blood group or DNA tests and zygosity based on discriminant analysis of survey items on resemblance was investigated. In earlier research, prediction accuracy of zygosity was around 93% in children (Rietveld et al., 2000). Since then, an additional item was added to the NTR questionnaires and substantially more DNA data have become available. We re-evaluated zygosity assignments, based on surveys obtained at age 3, 5, 7, 10, 12, 14, and 16 years. These contain 10 zygosity items about resemblance between the twins (6 items about physical similarities and 4 items about confusion by parents and others). At ages 3 to 12 years, mothers and fathers and at ages 14 to 16 the twins themselves filled out the questions. In adults aged 18-99 years, multiple surveys contain 8 zygosity items (5 items about resemblance and 3 items about confusion by parents by parents and others). Here, the questions were answered by the twins themselves.

Because knowledge of the result of a zygosity test may affect responses, we only included data from same-sex twins whose survey information had been completed before they, or their parents, received the results of the DNA tests. For the children, this resulted in a sample of 5,776 twins (individuals) and for adults in a sample of 3,512 twins (individuals).

For children, the data were randomly divided into a training set (60%) and a testing set (40%). In the training set, linear discriminant analysis was applied to survey data from each informant (mother, father, and self) and ages 3 to 16. Linear discriminant analysis determines the axes that maximize the separation of different classes, in our case mono- and dizygotic (MZ and DZ). This analysis generated a linear function of the weighted sum of the items, in which the weights were optimized to distinguish between MZ and DZ twin pairs. The outcomes then were applied to the data from the testing set (N = 2,260). If multiple surveys were available (either multiple ages, multiple informants or both), the most frequently predicted zygosity was chosen as the assigned zygosity. If an equal number of MZ and DZ outcomes was observed, the mean probability of being MZ determined

by the discriminant analyses was used (MZ: probability > .5 and DZ: probability < .5). To determine the accuracy of our zygosity classification by the questionnaire items, we determined the proportion in which the true zygosity and the assigned zygosity corresponded. If there was no childhood survey after age 2 years, the item about resemblance from the questionnaire collected at age 2 was used (MZ: “yes, but barely different” or “yes, but well distinguishable” and DZ: “no, not a lot” or “no, not at all”). If there also was no information at age 2, the item from the questionnaire at age 1 was used (MZ: “MZ twins” and DZ: “DZ twins”).

For adults a similar scheme was used: 60% of the data was randomly assigned to be in the training set and 40% to be in the testing set. The first available survey with zygosity items was analyzed. The outcomes of the linear discriminant analysis of the training set were used to predict zygosity in the testing set (N = 1,362).

Results

In children, results indicate that the entire procedure of zygosity prediction by the 10-item zygosity questionnaires correctly classified zygosity in 96.8%. Prediction accuracy for all YNTR questionnaires separately can be found in Table 1. For 95.8% of this sample, data on the 10-item zygosity list included in the questionnaires at age 3 to 16 was available. If only these participants were considered, the accuracy of zygosity classification was 97.2%. In the remaining sample, only surveys at earlier ages, when twins are 1 and 2 years, were available. These were completed by mothers and included one question at age 1 (“According to you, the twins are”, with answer options “DZ twins” and “MZ twins”) and one question at age 2 (“Do the children resemble each other” - with answer options “yes, they are barely different”, “yes, but well distinguishable”, “no, not a lot” and “no, not at all”). If only the questionnaire at age 2 is used, zygosity prediction is accurate in 93.8% of the cases. When only the questionnaire at age 1 is used, prediction accuracy of zygosity drops to 78.9%. Participants with data at age 1 or 2 only are underrepresented in the sample, (N = 96).

In adults, zygosity prediction based on survey items was accurate in 95.9%. In Table 2, the results of the linear discriminant analysis are displayed. In adults, the item that distinguishes best between MZ and DZ twins was “Were you each other’s spitting image as children?”, whereas the item “Did mother and father mix you up when you were young” does not distinguish well. In parent-reports on children, “confusion by other family members than the parents” was a well-distinguishing item.

Table 1: Zygosity prediction accuracy of YNTR questionnaire items compared to DNA.

YNTR questionnaire	Correctly classified
Overall (multiple questionnaires)	96.8%
Age 1	78.9%
Age 2	93.8%
Age 3 mother	94.9%
Age 3 father	94.1%
Age 5 mother	97.5%
Age 7 mother	95.9%
Age 7 father	96.5%
Age 10 mother	96.1%
Age 10 father	96.3%
Age 12 mother	96.6%
Age 12 father	95.9%
Age 14 self	91.8%
Age 16 self	94.9%

Table 2: Ranks for the predictive value of the zygosity items in YNTR and ANTR surveys.

Item	Age 3 mother -rated	Age 3 father-rated	Age 12 mother - rated	Age 12 father-rated	Age 16 self-rated	Adults self-rated
Facial appearance	7	3	6	5	5	7
Hair colour	3	8	5	4	4	5
Face colour	4	6	4	6	7	6
Eye colour	5	4	7	7	2	4
Hair structure	2	2	3	3	3	-
Spitting image	10	9	9	9	1	1
Confusion mom/dad	8	5	10	8	6	8
Confusion other family members	1	1	2	2	8	3
Confusion strangers	6	7	1	1	9	2
Confusion on photos	9	10	8	10	-	-

1 = most predictive item; 10 = least predictive item.

Literature

Rietveld, M. J. H., van Der Valk, J. C., Bongers, I. L., Stroet, T. M., Slagboom, P. E., & Boomsma, D. I. (2000). Zygosity diagnosis in young twins by parental report. *Twin Research and Human Genetics*, *3*(3), 134-141.

Appendix 2

Data collection in teachers of twins and their siblings in the Netherlands Twin Register.

This document describes the steps that NTR undertakes to collect data from twins, and their siblings, from their elementary school teachers. The first step is approaching the parents for consent and ask them for information about the school and teacher. If the parents return the informed consent forms, the second step is the to approach the teachers.

Parental consent (PC)

Every year, parents of twins aged 7, 9 and 12 are asked by NTR if they are willing to provide consent to approach the teacher of their children, the twins and any other siblings within the appropriate age range, to provide information about the children's development from the teacher's point of view. The parents receive an email with information about the survey for the teachers and a personal link to a survey to fill in if they want to give consent or not and to fill in the contact information about the teachers and the schools, as well as if the twins are going to the same school and are in the same class with the same teacher.

Parents are also asked if they have other children in aged 6-12 that go to primary school and if so, if they give permission and the contact information about their teacher. Families with multiple twin pairs in the nuclear family were not asked about additional siblings in primary school.

Parents are also asked if they give permission to link data to external registers (e.g. CBS: Statistics Netherlands) for research purposes. Some of the families are registered in the Netherlands Twin Register, but rarely participated in earlier surveys. For these twin children, we do not know their zygosity, which is essential information to have in twin model research. Therefore, the parents with twins with an unknown zygosity also received the question if the zygosity was ever determined by a DNA or blood test as well as 10 items about the resemblance of the twins. Zygosity determination by analysing responses these items has proven to have be accurate (with parents reporting on twin children there is correct classification in 96.8% of the cases), for more information see Appendix 1.

Teacher Survey

The teacher survey is sent out via email to the teachers of the children whose parents gave consent to contact them. The teachers are free to decide if they participate or not. The survey contains questions about the background of the school (e.g. school type, socio-economic status of parents), children's school results as well as the children's behaviour at school (including the complete ASEBA Teacher Report Form (TRF), Conners Teacher Rating Scale and Social Skills Rating System (SSRS)).

Teachers also receive an invitation to upload pdf files containing the pupil monitor system (Leerling Volg Systeem). Since 2022, this is done via the protected SURF environment. The teachers were invited by a separate email to upload the pupil monitor system pdf and the encrypt it with a password in the SURF environment, where it could be retrieved by the researchers. This safe environment eliminated the need for pseudonymization by the teachers.

Ethics assessment

The data collection procedure was re-approved by the ethics committee (Vaste Commissie Wetenschap en Ethiek (VCWE) of the Vrije Universiteit Amsterdam in 2021 (VCWE number 2021-111).

Overview of the data collection for 2018/19, 2019/20, 2020/21 and 2021/2022

Year 2018/2019

We approached parents with twins born between 01-10-2011 and 30-09-2012, 01-10-2009 and 30-09-2010 and 01-10-2006 and 30-09-2007. In cases where mother's email address was known, we sent the consent form to mother's email address. If only father's email address was known, we sent the consent form to him. In families with multiple mothers who had email addresses available, we randomly selected one of the mother's to receive the parental consent form. This resulted in a total of 2571 parents who were asked to give parental consent by email. The parents were reminded by email and a subset of the parents also were reminded by telephone. Of the approached parents, 1137 parents (44.23%) gave permission to approach the teacher.

	Parents selected	Parents Replied
Age 7	876	399
Age 9	913	335
Age 12	782	403

Year 2019/2020

We approached parents with twins born between 01-10-2012 and 30-09-2013, 01-10-2010 and 30-09-2011 and 01-10-2007 and 30-09-2008. If mother's email address was known, we sent the consent form to mother's email address and if only father's email address was known, we sent the consent form to the father. In families with multiple mothers with known email addresses, we randomly selected one of them. This selection resulted in a total of families. Of these, 749 parents (27.3%) returned the parental consent form. We obtained zygosity data on 46 twin pairs that before had an unknown zygosity.

Due to the outbreak COVID-19 pandemic, the teacher survey was not sent out. Parents were informed about this change in data collection by email.

	Parents selected	Parents Replied
Age 7	934	232
Age 9	968	242
Age 12	839	273

Year 2020/2021

We approached parents with twins born between 01-10-2013 and 30-09-2014, 01-10-2011 and 30-09-2012 and 01-10-2008 and 30-09-2009 (total of 2741 parents, consisting of mostly mothers but if mother's email address was missing, the father was approached, and in families with multiple mothers with known email addresses, we randomly selected one of the mothers). 516 parents (20.5%) filled in the parental consent form. Of the parents who responded, 406 parents gave permission to approach the teacher and 110 did not give permission. We obtained zygosity data on 4 twin pairs that before had an unknown zygosity.

Because of the COVID-19 pandemic, the teacher survey was not sent out. Parents were informed about this change in data collection by email.

	Parents selected	Parents Replied
Age 7	805	134
Age 9	914	142
Age 12	798	240

Year 2021/2022

We approached parents with twins born between 01-10-2014 and 30-09-2015, 01-10-2012 and 30-09-2013 and 01-10-2009 and 30-09-2010 (total of 2632 parents). The parents mostly consisted of mother', but if mother's email address was unknown and the father's email address was available, the father was approached, and in families with multiple mothers with known email addresses, we randomly selected one of the mothers. They were reminded by email and a subset also by telephone. Of the approached parents, 614 parents (23.3%) filled in the parental consent form. Of these 614 parents, 501 parents gave permission to approach the teachers and 113 did not. In this data collection effort, we collected the items about zygosity to parents of twins who never provided information about zygosity before as well as to parents of twins who only provided information about zygosity when the twins were 1 or 2 years old.

	Parents selected	Parents Replied
Age 7	814	183
Age 9	878	171
Age 12	940	260

Literature

Ligthart, L., van Beijsterveldt, C. E., Kevenaar, S. T., de Zeeuw, E., van Bergen, E., Bruins, S., ... & Boomsma, D. I. (2019). The Netherlands Twin Register: longitudinal research based on twin and twin-family designs. *Twin Research and Human Genetics*, *22*(6), 623-636.

Appendix 3

Appendix 3. Attachments of data collection in teachers in the Netherlands Twin Register.

A. Uitnodigen Parental Consent email

B. Herinnering parental consent email

C. Uitnodigen leerkrachten email

D. Herinnering leerkrachten email

E. Mail aan ouders die toestemming hadden gegeven om leerkracht te benaderen tijdens COVID-19

F. Email bij de uitnodigingslink om LVS pdf te uploaden via SurfFileSender

A. Uitnodigen Parental Consent email

Beste ouder/verzorger van {{member.naam_1}} en {{member.naam_2}},

U doet mee aan onderzoek van het Nederlands Tweelingen Register (NTR) naar de ontwikkeling van tweelingen en andere meerlingen. Dankzij uw deelname en die van vele andere meerlingfamilies heeft dit onderzoek geleid tot belangrijke inzichten. Om een nog vollediger beeld te krijgen van de ontwikkeling van opgroeiende meerlingen betrekken we ook graag hun leerkrachten bij het onderzoek. Na uw toestemming, sturen wij de leerkracht(en) een uitnodiging om een vragenlijst in te vullen over gedrag en schoolprestaties en vragen wij de leerkrachten om resultaten uit het leerlingvolgsysteem te delen. Voor de leerkrachten duurt het invullen van de vragenlijst ongeveer 20-25 minuten per kind.

Om leerkrachten te kunnen benaderen, hebben wij uw toestemming en de gegevens van de school nodig. Wilt u op onderstaande website aangeven of u hiervoor toestemming geeft?

Volg deze link om de enquête te starten als het via bovenstaande knop niet lukt: {{survey.personal_link}}

Uw toestemming en de deelname van leerkrachten aan dit onderzoek is geheel vrijwillig. Alle gegevens worden vertrouwelijk behandeld. Dit betekent onder andere dat ouders, kinderen en leerkrachten geen inzage krijgen in elkaars antwoorden. Als u niet wilt dat wij de leerkracht van uw kinderen benaderen dan vragen wij u dit ook aan te geven, zodat u hierover geen berichten meer ontvangt. We realiseren ons dat leerkrachten het erg druk hebben, maar zeker in deze vreemde tijden is de blik van de leerkracht op het kind zeer waardevol.

De afgelopen jaren hebben al veel leerkrachten meegewerkt aan ons onderzoek. Eén van de bevindingen is dat het voor de sociale ontwikkeling van tweelingen geen verschil maakt of de kinderen wel of niet bij elkaar in de klas zitten. U kunt hierover meer lezen op: https://tweelingenregister.vu.nl/deelnemers/onderzoek/lopend_onderzoek/vragenlijst-voor-leerkrachten

Als u vragen heeft dan kunt u contact met opnemen met Sofieke Kevenaar per e-mail (ntr.leerkrachten@vu.nl) of per telefoon (020-5982458, laat bij geen gehoor een voicemail achter, dan bellen wij u zo snel mogelijk terug).

Wij stellen uw medewerking bijzonder op prijs!

Met vriendelijke groet
Prof. dr. Dorret Boomsma & Sofieke Kevenaar
Het Nederlands Tweelingen Register

B. Herinnering parental consent email

Beste ouder/verzorger van {{member.naam_1}} en {{member.naam_2}},

Enige tijd geleden heeft u van ons een mail gekregen over het geven van toestemming voor het benaderen van de leerkracht(en) van uw kinderen. Volgens onze administratie heeft u deze nog niet (helemaal) ingevuld. Vandaar dat u nu deze herinneringsmail ontvangt.

U doet mee aan onderzoek van het Nederlands Tweelingen Register (NTR) naar de ontwikkeling van tweelingen en andere meerlingen. Dankzij uw deelname en die van vele andere meerlingfamilies heeft dit onderzoek geleid tot belangrijke inzichten. Om een nog vollediger beeld te krijgen van de ontwikkeling van opgroeiende meerlingen betrekken we ook graag hun leerkrachten bij het onderzoek. Na uw toestemming, sturen wij de leerkracht(en) in de loop van het schooljaar (rond maart) een uitnodiging om een vragenlijst in te vullen over gedrag en schoolprestaties. Voor de leerkrachten duurt het invullen van deze vragenlijst ongeveer 30 minuten per kind.

Om leerkrachten te kunnen benaderen, hebben wij uw toestemming en de gegevens van de school nodig. Wilt u op onderstaande website aangeven of u hiervoor toestemming geeft?

ENQUÊTE STARTEN

Volg deze link om de enquête te starten als het via bovenstaande knop niet lukt:
{{survey.personal_link}}
Uw toestemming en de deelname van leerkrachten aan dit onderzoek is geheel vrijwillig. Alle gegevens worden vertrouwelijk behandeld. Dit betekent onder andere dat ouders, kinderen en leerkrachten geen inzage krijgen in elkaars antwoorden. Als u niet wilt dat wij de leerkracht van uw kinderen benaderen dan vragen wij u dit ook aan te geven, zodat u hierover geen berichten meer ontvangt. We realiseren ons dat leerkrachten het erg druk hebben, maar zeker in deze vreemde tijden is de blik van de leerkracht op het kind zeer waardevol.

De afgelopen jaren hebben al veel leerkrachten meegewerkt aan ons

onderzoek. Eén van de bevindingen is dat het voor de sociale ontwikkeling van tweelingen geen verschil maakt of de kinderen wel of niet bij elkaar in de klas zitten. U kunt hierover meer lezen op: https://tweelingenregister.vu.nl/deelnemers/onderzoek/lopend_onderzoek/vragenlijst-voor-leerkrachten

Als u vragen heeft dan kunt u contact met opnemen met Sofieke Kevenaar per e-mail (ntr.leerkrachten@vu.nl) of per telefoon (020-5982458, laat bij geen gehoor een voicemail achter, dan bellen wij u zo snel mogelijk terug).

Wij stellen uw medewerking bijzonder op prijs!

Prof. dr. Dorret Boomsma & Sofieke Kevenaar
Het Nederlands Tweelingen Register

C. Uitnodigen leerkrachten email

Beste {{member.Voornaam_Leerkracht}} {{member.Achternaam_Leerkracht}},

Deze mail gaat over {{member.first_name}}.
Het Nederlands Tweelingen Register (NTR) volgt de ontwikkeling van twee- en meerlingen en hun broertjes en zusjes. Hun ouders vullen vaak al sinds de geboorte van de kinderen vragenlijsten in over hun ontwikkeling. Een belangrijk aspect van het onderzoek is het gedrag van kinderen op school. De afgelopen jaren deden veel leerkrachten mee. Achtergronden en resultaten vindt u op onze website). Bij u in de klas zitten één of meerdere kinderen uit een meerlinggezin. Hun ouders hebben toestemming gegeven om u aan te schrijven. Wij willen u vragen om een vragenlijst in te vullen over **{{member.first_name}}**. Dit duurt ongeveer 20-25 minuten. Over een eventuele andere leerling ontvangt u een aparte mail. U opent de vragenlijst via “Enquête starten”:

ENQUÊTE STARTEN

Alle gegevens worden vertrouwelijk behandeld.
Dit betekent onder meer dat ouders geen inzage krijgen in de antwoorden van leerkrachten en dat de leerkracht geen inzage krijgt in de antwoorden van ouders.

NTR onderzoekt gedrag en de schoolprestaties, en ook hun samenhang. Daarom vragen we u ook om een uitdraai van het leerlingvolgsysteem (LVS) van de leerling met NTR te delen. Dit kan in de beveiligde omgeving van SURF

(Samenwerkende Universitaire Reken Faciliteiten). Hiervoor ontvangt u een aparte mail van FileSurfSender. In deze mail kunt u via de link de LVS uitdraai uploaden door op “select files” te klikken. Daarna selecteert u het bestand met de LVS uitdraai en vinkt u het vakje “Send file(s) more secure by enabling File Encryption” aan. Uw wachtwoord is:

<wachtwoord>

Daarna vinkt u “I accept the following code of conduct when using this service” aan en klikt u op “Send”.

We realiseren ons dat leerkrachten het erg druk hebben, en stellen uw medewerking bijzonder op prijs! Zeker in deze vreemde tijden is de blik van de leerkracht op het kind zeer waardevol. Mocht u vragen hebben, dan kunt u contact met Sofieke Kevenaar opnemen via e-mail: ntr.leerkrachten@vu.nl of via telefoon: 020 - 5982458 (laat bij geen gehoor een voicemail achter, wij bellen u zo snel mogelijk terug).

Met vriendelijke groet en bij voorbaat heel hartelijk dank voor uw medewerking,

Prof. dr. Dorret Boomsma
Sofieke Kevenaar

D. Herinnering leerkrachten email

Beste <naam leerkracht> ,

Enige tijd geleden ontving u van Nederlands Tweelingen Register (NTR) een uitnodiging om een vragenlijst in te vullen over <**naam kind**>.

Het NTR volgt de ontwikkeling van twee- en meerlingen en hun broers/zusjes. Hun ouders verschaffen informatie over hun ontwikkeling, maar een belangrijk aspect is het gedrag van kinderen op school. Achtergronden van dit onderzoek vindt u hier: website.

De ouders hebben toestemming gegeven om u aan te schrijven met het verzoek een lijst in te vullen. Dit duurt ongeveer 20-25 minuten. Over een eventuele andere leerling ontvangt u een aparte mail. U opent de vragenlijst via “Enquête starten”:

ENQUÊTE STARTEN

Alle gegevens worden vertrouwelijk behandeld. Dit betekent onder meer dat ouders en leerkrachten geen inzage krijgen in elkaars antwoorden.

We vragen u ook, als mogelijk, om een uitdraai van het leerlingvolgsysteem (LVS). Dit kan in een beveiligde omgeving. Hiervoor ontvangt u een aparte mail van FileSurfSender. Via de link kunt u de LVS uitdraai uploaden door op “select files” te klikken. Daarna selecteert u het bestand met de LVS uitdraai en vinkt u het vakje “Send file(s) more secure by enabling File Encryption” aan. Uw wachtwoord is:

<Wachtwoord>

Daarna vinkt u “I accept the following code of conduct” aan en klikt u op “Send”.

We realiseren ons dat leerkrachten het erg druk hebben, en stellen heel veel prijs op uw medewerking! De blik van de leerkracht op het kind is zeer waardevol. Het invullen van de vragenlijst en het uploaden van de LVS, staan los van elkaar. Ook als we een van beide ontvangen, is dit waardevol.

Als u vragen heeft, kunt u contact opnemen met Sofieke Kevenaar [020 - 5982458; bij geen gehoor een bericht inspeken, wij bellen u terug of] of via: ntr.leerkrachten@vu.nl.

Met vriendelijke groet en bij voorbaat heel hartelijk dank voor uw medewerking,

Prof. dr. Dorret Boomsma
Sofieke Kevenaar

E. Mail aan ouders die toestemming hadden gegeven om leerkracht te benaderen tijdens COVID-19

Beste ouders van {{member.naam_1}} en {{member.naam_2}},

Bij het Nederlands Tweelingen Register (NTR) hopen we van harte dat het goed gaat met u en uw kinderen, dat al uw dierbaren in goede gezondheid verkeren en dat dit ook zo zal blijven in deze ongekende tijd.

Een poosje geleden heeft u het NTR toestemming gegeven voor het benaderen van de leerkracht(en) van uw kinderen. Normaal gesproken nodigen wij de leerkracht rond deze tijd in het schooljaar uit om de vragenlijst(en) in te vullen. We willen u hierbij laten weten dat we dat in deze huidige situatie vanzelfsprekend niet zullen gaan doen.

We wensen u heel veel sterkte en (als dat voor u geldt) succes met werk en het geven van thuisonderwijs!

Met vriendelijke groet,

namens het Nederlands Tweelingen Register, Dorret Boomsma, Eveline de Zeeuw en Sofieke Kevenaar

---

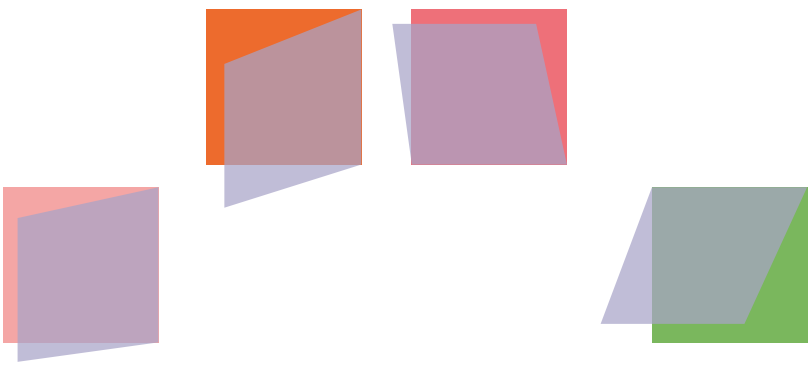
F. Email bij de uitnodigingslink om LVS pdf te uploaden via SurfFileSender

Zoals in de eerder verstuurde mail aangekondigd volgt hierbij de uitnodiging voor het beveiligd uploaden van de leerlingvolgsysteem informatie. NTR onderzoekt gedrag en de schoolprestaties, en ook hun samenhang. Daarom vragen we u ook om een uitdraai van het leerlingvolgsysteem (LVS) van de leerling met NTR te delen. Dit kan in de beveiligde omgeving van SURF (Samenwerkende Universitaire Reken Faciliteiten). In deze mail kunt u via de link de LVS uitdraai uploaden door op de voucher link te klikken en de voorwaarden te accepteren. Vervolgens klikt u op “select files”. Daar selecteert u het bestand met de LVS uitdraai (of uitdraaien als u uitdraaien van meerdere kinderen wil uploaden) en vinkt u het vakje “Send file(s) more secure by enabling File Encryption” aan. Uw wachtwoord is hier: <wachtwoord>.
Daarna vinkt u “I accept the following code of conduct when using this service” aan en klikt u op “Send”. We realiseren ons dat leerkrachten het erg druk hebben, en stellen uw medewerking bijzonder op prijs! Zeker in deze vreemde tijden is de blik van de leerkracht op het kind zeer waardevol. Mocht u vragen hebben, dan kunt u contact met Sofieke Kevenaar opnemen via e-mail: ntr.leerkrachten@vu.nl of via telefoon: 020 - 5982458 (laat bij geen gehoor een voicemail achter, wij bellen u zo snel mogelijk terug).

Met vriendelijke groet en bij voorbaat heel hartelijk dank voor uw medewerking,

Prof. dr. Dorret Boomsma
Sofieke Kevenaar, MSc

Summary of author contributions



Summary of author contributions

Chapter 2

Chapter 2 was based on Tamimy*, Z., Kevenaar*, S. T., Hottenga, J. J., Hunter, M. D., de Zeeuw, E. L., Neale, M. C., van Beijsterveldt, C. E. M., Dolan, C. V., van Bergen, E., & Boomsma, D. I. (2021). Multilevel twin models: geographical region as a third level variable. *Behavior Genetics*, *51*(3), 319-330.

The study was designed by S.T. Kevenaar, Z. Tamimy, E. van Bergen and D.I. Boomsma. The data were prepared by C.E.M. van Beijsterveldt (phenotypic data) and J.J. Hottenga (genetic data). The analyses were performed by S.T. Kevenaar and Z. Tamimy. The paper was written by S.T. Kevenaar and Z. Tamimy and all co-authors reviewed and edited the manuscript.

Chapter 3

Chapter 3 was based on Kevenaar, S. T., Zondervan-Zwijnenburg, M. A., Blok, E., Schmengler, H., Fakkel, M. T., De Zeeuw, E. L., van Bergen, E., Onland-Moret, N. C., Peters, M., Hillegers, M. H. J., Boomsma, D. I., & Oldehinkel, A. J. (2021). Bayesian evidence synthesis in case of multi-cohort datasets: An illustration by multi-informant differences in self-control. Developmental Cognitive *Neuroscience, 47*, 100904.

The study was designed by S.T. Kevenaar, A.J. Oldehinkel, D.I. Boomsma, E. van Bergen and E. de Zeeuw. The scripts for all the analyses were prepared by S.T. Kevenaar in collaboration with M. A. Zondervan-Zwijnenburg. The local analyses were performed by E. Blok, H. Schmengler and M.T. Fakkel. The paper was written by S.T. Kevenaar and all co-authors reviewed and edited the manuscript

Chapter 4

Chapter 4 was based on Kevenaar, S. T., Dolan, C. V., Boomsma, D. I., & van Bergen, E. (2023). Self-control and grit are associated with school performance mainly because of shared genetic effects. *JCPP Advances*, e12159.

The study was designed by S.T. Kevenaar, C.V. Dolan, D.I. Boomsma and E. van Bergen. The analyses were performed by S.T. Kevenaar and C.V. Dolan. The paper was written by S.T. Kevenaar and all co-authors actively reviewed and edited the manuscript.

Chapter 5

Chapter 5 was based on Kevenaar, S. T., van Bergen, E., Oldehinkel, A. J., Boomsma, D. I. & Dolan, C. V., (Submitted for publication). Grit and self-control predict school performance: strongly genetic, weakly causal.

The study was designed by S.T. Kevenaar, C.V. Dolan, D.I. Boomsma and E. van Bergen.. The analyses were performed by S.T. Kevenaar and C.V. Dolan. The paper was written by S.T. Kevenaar and all co-authors actively reviewed and edited the manuscript.

Nederlandse samenvatting

Nederlandse samenvatting

Het doel van het onderzoek in dit proefschrift was om de bron van individuele verschillen in een reeks kenmerken en eigenschappen van kinderen te analyseren met behulp van verschillende modellen. In dit proefschrift komen de fenotypes lengte, zelfcontrole, doorzettingsvermogen/vastberadenheid (het Engelse *grit*) en schoolprestaties aan bod. Hieronder vat ik de bevindingen van het onderzoek in dit proefschrift samen.

In hoofdstuk 2 onderzocht ik de rol van geografische locatie als een derde-niveau variabele in een tweelingmodel, waarbij het eerste niveau het kinderniveau is en het tweede niveau het familieniveau. Door de herparameterisatie van het klassieke tweelingmodel tot een equivalent multilevel model ontstond de mogelijkheid om een hogere niveau variabele, in dit geval geografische locatie zoals bepaald door postcodegebieden, op te nemen, waarin de lagere niveau variabelen zijn genest. We illustreerden hier het gebruik van een multilevel model met drie niveaus door het model toe te passen op data van tweelingen en de regionale clustering van de lengte van 7-jarige kinderen in Nederland te onderzoeken. Ons onderzoek onthulde dat ongeveer 2% van de fenotypische variantie in de lengte van kinderen kan worden toegeschreven aan regionale clustering. Deze 2% correspondeert met 7% van de variantie verklaard door gemeenschappelijke omgevingscomponenten tussen families. Gemiddeld genomen zijn kinderen in het noorden van Nederland langer dan kinderen in het zuiden, en jongens zijn langer dan meisjes. Ik onderzocht ook het mogelijke effect van genetische afkomst op regionale clustering door de impact van genetische principale componenten te beoordelen in een subset van deelnemers met genoom-brede enkelvoudige nucleotide polymorfisme (SNP) gegevens. De principale componentenanalyse van de covariantiematrix van de SNP-gegevens stelt ons in staat om genetische principale componenten te identificeren, die allelfrequentiegradiënten weerspiegelen. In Nederland weerspiegelt de eerste genetische principale component een noord-zuid gradiënt, de tweede een oost-west verdeling en de derde de meer centrale regio's van het land. Onze resultaten gaven aan dat na correctie voor genetische factoren de regio geen significant effect meer had op de variatie in lengte. Deze bevindingen suggereren dat de fenotypische variantie in lengte die wordt verklaard door regionale clustering toe te schrijven is aan genetische afkomsteffecten.

In hoofdstuk 3 werd de kracht van multi-cohort samenwerking en *'Bayesian evidence synthesis*' gedemonstreerd. In deze studie toon ik het gebruik

van '*Bayesian evidence synthesis*' aan om de mate van ondersteuning voor verschillende concurrerende hypotheses te kwantificeren en te vergelijken, en om deze ondersteuning over verschillende studies te combineren. Ik paste deze aanpak toe om de rangorde van multi-informant scores op de ASEBA Zelfcontroleschaal (ASCS) te onderzoeken in een multi-cohort ontwerp met informatie van vier Nederlandse cohorten. Omdat de beschikbare rapporteurs over zelfbeheersing van kinderen varieerden tussen de cohorten (bijvoorbeeld ouders, leraren, zelfrapportages), evalueerde elk cohort verschillende aspecten van de concurrerende hypotheses. Onze bevindingen leverden consistent bewijs voor de gedeeltelijke hypothese dat ouders meer zelfbeheersingsproblemen rapporteerden dan leraren. De geaggregeerde resultaten toonden de meeste ondersteuning voor de hypothese dat kinderen het hoogste aantal zelfbeheersingsproblemen rapporteren, gevolgd door hun moeders en vaders, terwijl leraren de minste problemen rapporteerden. In de positie van zelfrapportages van de zelfbeheersingsproblemen waren er inconsistenties. Dit hoofdstuk illustreert het belang van rekening houden met de informant, en het potentieel van het combineren van resultaten uit verschillende studies met '*Bayesian evidence synthesis*'.

In hoofdstuk 4 combineerde ik genetische covariantiestructuur modellering met regressie om de voorspelling van schoolprestaties door zelfcontrole en doorzettingsvermogen/vastberadenheid (*grit*) te onderzoeken. Deze analysemethode maakte het mogelijk om genetische en omgevingsbronnen van variantie te onderscheiden. De resultaten toonden aan dat een groot deel van de individuele verschillen in schoolprestaties, namelijk 28,4%, kan worden verklaard door zelfcontrole, doorzettingsvermogen/vastberadenheid en hun covariantie. Bij het verder inzoomen op de etiologie bleek dat het grootste deel van de verklaarde variantie in schoolprestaties toe te schrijven was aan de genetische componenten van doorzettingsvermogen/vastberadenheid en zelfcontrole. Na correctie voor sociaal-economische status, geslacht en al dan niet delen van dezelfde beoordelaar (in dit geval de leraar), was slechts 1,3% van de verklaarde variantie in schoolprestaties toe te schrijven aan omgevingsfactoren. In deze studie heb ik twee dataproblemen aangepakt. Ten eerste moest ik corrigeren voor *censoring* in de analyses, omdat de verdeling van de variabelen, vooral die van zelfcontrole, scheef was. Dit was het gevolg van een plafondeffect, wat betekent dat veel kinderen de hoogst mogelijke score behaalden. Dit heb ik opgelost door de modellen te fitten met *maximum likelihood estimation*, onder de aanname dat de gegevens een multivariate, rechts-gecensureerde verdeling volgden. Ten tweede vormden de leraren een bron van systematische variantie,

omdat de fenotypes van een deel van de tweelingen beoordeeld werden door dezelfde leerkracht. Een kracht van deze studie is dat de beoordelaar (leraar) variantie geschat kon worden, omdat sommige tweelingen in dezelfde klas zaten en dus dezelfde leraar deelden, terwijl andere tweelingen in verschillende klassen met verschillende leraren zaten. Dit maakte het mogelijk om het deel van de variantie dat toe te schrijven is aan de beoordelaar te kwantificeren.

In hoofdstuk 5 heb ik de relatie tussen zelfcontrole, doorzettingsvermogen/ vastberadenheid en schoolprestaties verder onderzocht door een expliciet causaal model toe te passen. Het toegepaste model combineert fenotypische regressie van schoolprestaties op doorzettingsvermogen en zelfcontrole met het schatten van mogelijke *confounding* door genetische of omgevingsinvloeden (invloeden die gemeenschappelijk zijn voor alle drie de fenotypen). Door gebruik van dit model is het mogelijk om te onderzoeken of er ondersteuning wordt gevonden voor een causale relatie. De resultaten toonden aan dat het grootste deel van de relatie tussen zelfcontrole, doorzettingsvermogen/vastberadenheid en schoolprestaties toe te schrijven was aan genetische *confounding*, wat hier waarschijnlijk genetische pleiotropie weerspiegelt. Dat wil zeggen dat de genen die individuele verschillen in zelfcontrole en doorzettingsvermogen/ vastberadenheid verklaarden, ook individuele verschillen in schoolprestaties verklaarden. Er bleken ook directe (fenotypische) effecten van zelfcontrole en doorzettingsvermogen/vastberadenheid op schoolprestaties te zijn, maar deze directe effecten verklaarden slechts 4,4% van variantie, terwijl 12,4% werd verklaard door pleiotropie.

Conclusie

Bij het analyseren van individuele verschillen in de kindertijd is het doel om te identificeren hoe kinderen van elkaar verschillen en welke factoren aan deze verschillen ten grondslag liggen. Dit kan worden gedaan door (het combineren van) gegevens uit grote datasets en het gebruik van verschillende methoden en modellen. In de kindertijd zijn vragenlijsten vaak een belangrijke bron van gegevens. Bij het analyseren van dergelijke vragenlijsten is het belangrijk om rekening te houden met wie de informatie over de kinderen heeft verstrekt en om mogelijke verschillen tussen vaders en moeders, leraren en de kinderen zelf te erkennen en hier rekening mee te houden. Kinderen groeien op in groepen, zoals gezinnen, klassen, scholen, buurten en landen. In statistische termen noemen we dit clustering. Rekening houden met clustering biedt extra onderzoeksmogelijkheden, bijvoorbeeld het onderzoeken van de etiologie van

een eigenschap, omdat we de familieclustering kennen. Omdat we de zygositeit van tweelingparen goed kunnen identificeren, kunnen we het verschil tussen groepen monozygote en dizygote tweelingparen gebruiken om genetische en omgevingsbronnen van individuele verschillen te onderscheiden. Een andere toepassing van clustering heeft betrekking op geografische locatie. Door rekening te houden met welke kinderen zich in dezelfde buurten clusteren en deze gegevens te combineren met familieclustering, zygositeitsinformatie en genetische principale componenten, ontstaat de mogelijkheid om de rol van geografische locatie en genetische afkomst te onderzoeken. Verder is het bij het onderzoeken van individuele verschillen cruciaal om rekening te houden met *confounding*, vooral als het doel is om enige inferenties over causaliteit te maken. Dit proefschrift toont aan dat individuele verschillen in de kindertijd zoals ze zich in de wereld om ons heen manifesteren grotendeels te wijten zijn aan genetische verschillen tussen kinderen, maar de omgeving speelt ook een rol in waarom kinderen van elkaar verschillen.

List of publications

List of publications

Bignardi, G., Chamberlain, R., **Kevenaar, S. T.**, Tamimy, Z., & Boomsma, D. I. (2022). On the etiology of aesthetic chills: a behavioral genetic study. *Scientific Reports*, *12*(1), 1-11.

Derks, K., Burger, J., van Doorn, J., Kossakowski, J. J., Matzke, D., Atticciati, L., ... **Kevenaar, S. T.** …& Wagenmakers, E. J. (2018). Network models to organize a dispersed literature: the case of misunderstanding analysis of covariance. *Journal of European Psychology Students*, *9*(1).

Fakkel, M., Peeters, M., Lugtig, P., Zondervan-Zwijnenburg, M. A. J., Blok, E., White, T., ... **Kevenaar, S. T.** …& Vollebergh, W. A. M. (2020). Testing sampling bias in estimates of adolescent social competence and behavioral control. *Developmental cognitive neuroscience*, *46*, 100872.

Kevenaar, S. T., Dolan, C. V., Boomsma, D. I., & van Bergen, E. (2023). Self-control and grit are associated with school performance mainly because of shared genetic effects. *JCPP Advances*, e12159.

Kevenaar, S. T., van Bergen, E., Oldehinkel, A. J., Boomsma, D. I. & Dolan, C. V., (Submitted for publication). Grit and self-control predict school performance: strongly genetic, weakly causal.

Kevenaar, S. T., Zondervan-Zwijnenburg, M. A., Blok, E., Schmengler, H., Fakkel, M. T., De Zeeuw, E. L., ... & Oldehinkel, A. J. (2021). Bayesian evidence synthesis in case of multi-cohort datasets: An illustration by multi-informant differences in self-control. *Developmental cognitive neuroscience*, *47*, 100904

Ligthart, L., van Beijsterveldt, C. E., **Kevenaar, S. T.**, de Zeeuw, E., van Bergen, E., Bruins, S., ... & Boomsma, D. I. (2019). The Netherlands Twin Register: longitudinal research based on twin and twin-family designs. *Twin Research and Human Genetics*, *22*(6), 623-636.

Tamimy*, Z**., Kevenaar*, S. T**., Hottenga, J. J., Hunter, M. D., de Zeeuw, E. L., Neale, M. C., & Boomsma, D. I. (2021). Multilevel twin models: geographical region as a third level variable. *Behavior Genetics*, *51*(3), 319-330
*shared first author

Zondervan-Zwijnenburg, M. A. J., Richards, J. S., **Kevenaar, S. T.**, Becht, A. I., Hoijtink, H. J. A., Oldehinkel, A. J., ... & Boomsma, D. I. (2020). Robust longitudinal multi-cohort results: The development of self-control during adolescence. *Developmental cognitive neuroscience*, *45*, 100817.

Dankwoord

Dankwoord

Dankwoord

Many people contributed to this dissertation and were a significant part of my PhD trajectory. I want to thank all my colleagues, collaborators, co-authors, friends and family. I want to thank some people explicitly, and depending on who I am addressing, I will do so in English or Dutch. So, sorry for the chaotic mix of English and Dutch in this section.

Allereerst wil ik wil graag alle deelnemers van het Nederlands Tweelingen Register, TRAILS, Generation R en YOUth bedanken. Zonder hen was het onderzoek in dit proefschrift niet mogelijk. Ik wil de leescommissie, prof. dr. Maartje Raaijmakers, prof. dr. Peter de Jong, prof. dr. Tina Kretschmer, dr. Michelle Achterberg, dr. Stéphanie van den Berg en dr. Camiel van der Laan bedanken voor de tijd en aandacht die zij hebben besteed aan het lezen en beoordelen van mijn proefschrift.

Verder wil ik natuurlijk graag mijn promotoren en copromotoren bedanken voor hun begeleiding en adviezen en het vertrouwen dat ze in me hebben gehad. Dorret, bedankt voor de zorgvuldige feedback en alle kennis die je hebt gedeeld de afgelopen jaren. Ik sta echt versteld van de snelheid en grondigheid waarmee je alles bekijkt en de parate kennis die je zo gemakkelijk oprakelt. Conor, van jou heb ik zo veel geleerd. Jouw methodologische en statistische kennis en kunde is ongelofelijk. Bedankt dat je dit met mij hebt willen delen. Elsje, bedankt voor het delen van je inhoudelijke kennis, voor je betrokkenheid en voor ‘checking in’. Vooral tijdens de Corona-periode was dit heel fijn. Tineke, bedankt voor de waardevolle samenwerking. Ook al was het vaak van een afstand, jouw bijdrages aan de papers waren altijd een zeer welkome aanvulling op ons Biologische Psychologie perspectief.

I want to thank all (former) colleagues of the department of Biological Psychology. You all made a difference. Thank you all for being there and for keeping up with my distractive behavior. I want to thank some people in particular. I want to thank the current and former PhD students: Anne L., Anne H., Bodine, Camiel, Denise, Eshim, Fiona, Floris, Hekmat, Hill, Lianne, Margot, Matthijs, Nicole, Nikki, Perline, Saskia, Selim, Susanne, Sjors, Veronika, Wonu, Yahua, Yayouk and Zenab. Denise, thank you for your openness and honesty. Eshim, thanks for all the fun we had on our ‘island’ and for all the nice dinner nights. Nicole, thank you for all the nice coffee dates. Lianne, thanks for all the fun we had at the department and in Boulder. Nikki, thank you for sharing my ticket buying stress and concert craziness. I am so glad I’ve had our PhD community. Michiel and Natascha,

thanks for all you do for the department. Cyrina and Toos, thank you for your amazing help with the data collection. Eco, Meike, Lannie, Jouke-Jan and Quinta, thank you for sharing your expert knowledge with me. I want to thank Eveline for all the help with starting up my PhD. Martin, thanks for alle 'gezelligheid' and for introducting me to your Ajax group. Michel, thank you for all the coffee moments and our conversations about everything and nothing.

What I never could have imagined when I started this PhD, is how many real and life changing friendships I would gain. Floris, thank you for all the very open and honest conversations we shared over dinner. Your interesting perspectives always make me reconsider mine. I am very glad that I've got to know you. Anne H., thanks for your friendship and all your great advice. I always leave reassured after we spent time togheter.

My lovely TTT girls: my paranymphs Wonu and Zenab, and Perline and Margot, I cannot express in words how important and valuable it was that I have had you during this PhD trajectory. I have way too many amazing memories with you to share here, but please know that I cherish them all. My dear Wonu, where do I even begin? I really cannot imagine my life without you anymore. You are an amazing scientist, human and friend. Thank you for allowing me to be my insane, unhinged self with you, it means the world. Whether we go on a crazy adventure in Italy (or New York or London) or just chill at home, I always have the best time with you. Our conversation never ends. I love you so much. Zenab, we were friends long before our time at the VU. We have been through so much together, and I am so grateful that you were by my side for everything. You are so smart, fun and engaged, I really admire that. Thanks for being my friend. Perline, you are truly inspiring and incredibly funny. It was a privilege to share a room with you, you always make me laugh when I least expect it. You have such a strong scientific vision and have been a great support throughout my academic journey. Margot, thanks for always organizing the most amazing and special get-togethers (the personalized game!), for everything you did for PhD students, and for always being there for me.

Ik wil ook heel graag al mijn andere lieve vrienden bedanken. Ik heb zoveel aan jullie te danken. Bedankt dat jullie er voor me waren. Ik heb echt de leukste vrienden van de wereld!

Mijn PML-guapas Zenab, Koen, Ria, Jessica, Rens, Frank, Femke, Gaby en Leonie. Onze tijd bij psychologische methodenleer was fantastisch. Ondanks en misschien juist wel dankzij alle wodka-woensdagen, uitgaansavonden en biertjes bij de Roeter is mijn enthousiasme voor de wetenschap echt aangewakkerd in

deze tijd. Jullie zijn stuk voor stuk heel slimme, inspirerende en bovenal super gezellige vrienden.

Inge, bedankt voor je trouwe vriendschap. Je staat altijd voor anderen klaar. Je ambitie en zorgzaamheid zijn heel inspirerend. Of we nou naar de film gaan, koffie drinken, samen eten of net iets teveel cocktails drinken, het is altijd gezellig. Amber, we go way back. Er zijn veel periodes geweest waarin we elkaar bijna elke dag zagen (vroeger op school of toen we huisgenoten waren) en ook periodes dat we elkaar bijna niet zagen (omdat jij op reis of geëmigreerd was naar een ver oord). Bedankt dat je er altijd voor me bent geweest, of we nou samen in één huis woonden of dat jij ergens aan de andere kant van de wereld zat. Als we elkaar zien of spreken is het altijd fijn. Eefje en Sofie, bedankt voor de fantastische tijd in de Hacquartstraat.

Heel erg veel dank aan mijn fantastische vriendinnen Saar, Roos en Rosanne, die me altijd enorm hebben gesteund. De gesprekken met jullie, ooit begonnen in de late (of beter gezegd, vroege) uurtjes aan de bar van de komedie, hebben me gevormd. Jullie zijn alle drie zo slim, cultureel onderlegd en maatschappelijk betrokken. Jullie zijn echt ongelofelijk belangrijk voor me. Rosanne, bedankt voor je eerlijkheid en je vragen die altijd een nieuw perspectief op de dingen werpen. Roos, bedankt dat je er altijd voor me bent, voor alle serieuze gesprekken maar ook zeker voor alle uit de hand gelopen kroeg- en uitgaansavonden. Ik vertrouw blind op alles wat jij me aanraadt, er is nog nooit één misser tussen geweest. Saar, je bent een van de sociaalste mensen die ik ken, en toch heb je altijd alle tijd en aandacht voor mij. Je zit vol goede ideeën en je kan zo goed delen. Bedankt dat ik altijd bij je terecht kan. Ik leer altijd heel veel van (de gespreken met) jullie. Ik voel me bij jullie echt thuis. Daarnaast doen jullie het écht belangrijke werk, want jullie staan elke dag voor de klas, met veel passie voor jullie vak en vooral heel veel betrokkenheid bij jullie leerlingen en hun omgeving. Ik heb hier enorm veel bewondering voor. En dan jullie hebben ook nog eens de leukste kinderen ooit gekregen die mijn leven heel erg hebben verrijkt.

Lieve Zuzu, je bent een super lief en gezellig meisje. Ik ben heel blij dat ik je nog beter heb leren kennen op onze vakantie. Je bent eigenlijk bijna altijd vrolijk, vanaf dat je wakker wordt tot je naar bed moet. Ik vind het zo leuk om te zien dat je steeds meer dingen kan en om te zien dat je je nu al zo goed kan concentreren op waar je mee bezig bent. Ik kan niet wachten om je verder te zien opgroeien.

Allerliefste Eli, jij brengt zoveel geluk en liefde in mijn leven. Fantastisch om te zien hoe sommige aspecten van je karakter er al vanaf het eerste moment in zaten, maar ook om te zien ook hoe je je ontwikkelt en wat je allemaal leert, het

gaat zo snel en elke dag word je nóg leuker! Je bent zo'n ongelofelijk lief, vrolijk en grappig kind. Ik kan uren kijken naar hoe lief je speelt. Het is heel speciaal om je van zo dichtbij te mogen zien opgroeien en ik ben je geweldige moeder hier voor eeuwig dankbaar voor. Het is echt altijd het leukste gedeelte van m'n dag als ik je zie of als we 'lellen', met als absolute hoogtepunt de gezellige vakanties. Ik hou van je.

Dank aan mijn lieve familie, mijn opa en oma, mijn ooms en tantes en neven en nichten. Natuurlijk wil ik graag mijn lieve ouders in het bijzonder bedanken voor alle liefde, interesse, steun en betrokkenheid, niet alleen tijdens mijn promotietraject, maar tijdens mijn hele leven. Jullie zijn fantastische ouders en ik heb super veel geluk met het warme nest wat jullie me hebben gegeven. Mam, ik bewonder je eerlijkheid en oprechtheid. Je bent een van de grappigste mensen die ik ken en een heel liefdevolle moeder. Pap, jouw loyaliteit en vastberadenheid inspireren me. Je staat altijd voor ons klaar. Bedankt dat jullie me altijd de ruimte hebben gegeven om mezelf te zijn. Als laatste wil ik graag mijn allerliefste zusje Iesje bedanken. Jij begrijpt me als geen ander, ondanks dat we zo verschillend zijn. Je bent ongelofelijk zorgzaam en betrokken. Ik ben heel trots op hoe goed je kan doorzetten als je iets echt wil. Je bent mijn steun en toeverlaat, ik hou heel veel van jou.