Learning and Individual Differences 54 (2017) 196–201

Contents lists available at ScienceDirect

Learning and Individual Differences

journal homepage: www.elsevier.com/locate/lindif

Increased environmental sensitivity in high mathematics performance

I. Schwabe [a,b,*], D.I. Boomsma [c], S.M. van den Berg [a]

[a] Faculty of Behavioural, Management and Social sciences (BMS), Research Methodology, Measurement and Data Analysis (OMD), University of Twente, The Netherlands
[b] Department of Methodology and Statistics, Tilburg University, The Netherlands
[c] Department of Biological Psychology, VU University Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 29 September 2015
Received in revised form 17 October 2016
Accepted 21 January 2017
Available online xxxx

Keywords:
Mathematical ability
Genotype-environment interaction
Item response theory
Measurement error
High ability

ABSTRACT

Results of international comparisons of students in studies such as PISA (*Program for International Student Assessment*) and TIMSS (*Trends in International Mathematics and Science Study*) are often taken to indicate that mathematical education in Dutch schools is not appropriate for mathematically talented students. However, there has been no empirical study yet that investigated this hypothesis. If indeed, Dutch students with a genetic predisposition for high mathematical ability are not nurtured to their full potential, their mathematics performance should be more affected by environmental factors than that of children with a genetic predisposition for low mathematical ability. In behaviour genetics such a situation is termed *genotype-environment interaction*: the relative importance of environmental influences differs depending on students' genotypic values. To investigate genotype-environment interaction, we analyzed mathematics performance of 2110 Dutch twin pairs on a national achievement test. In the analysis we corrected for error variance heterogeneity in the measurement of mathematics performance through the application of an item response theory (IRT) measurement model. As hypothesized, results indicated that environmental influences are relatively more important in explaining individual differences in students with a genetic predisposition for high mathematical ability.

1. Introduction

While some children find it easy to solve complex mathematical equations, others are struggling to pass their math exams. Dutch teachers usually focus on the latter group: the weakest (Dekker, 2014). Often criticized as a "culture of C-grades", education in the Netherlands has the reputation of being traditionally less focused on students with high mathematics performance levels. In an ideal school system, however, the talented child should be nurtured to its full potential as well. After all, the brightest students may be the ones who make important contributions to science, find cures for diseases or invent new technologies.

International comparisons such as the Program for International Students Achievement (PISA) and the Trends in International Mathematics and Science Study (TIMSS) show that, in the Netherlands, the average mathematical performance level in primary education is relatively high. This observation can, however, be attributed mainly to the high performance in the left tail of the achievement distribution: the weakest students are performing better than the weakest students from all other countries participating in PISA and TIMMS. The variance of test scores is, however, compared to other countries, very small: the performance levels of lowest- and highest-scoring students are relatively close. In other words, whereas Netherlands' weakest students perform exceptionally well, the top students are outperformed by the brightest students from Asian and other western countries (see e.g. Meelissen et al., 2012). This appears to be a persistent phenomenon as similar patterns have been found over the years for different age groups (see e.g. Minne, Rensman, Vroomen, & Webbink, 2007). These findings are often presented as underperformance in the high-ability students (see e.g. van der Steeg, Vermeer, & Lanser, 2011) and interpreted as an indication that mathematical education in Dutch schools is better tailored to the weaker students than to the mathematically talented students. However, one cannot draw conclusions on underlying processes based on the test score distribution alone. There are alternative explanations for the relatively poor performance of the top students in the Netherlands, for one that there might indeed be different underlying distributions of talent across countries.

In this article, the underperformance of Dutch mathematically talented students was investigated from a behaviour genetics perspective. A child's genetic mathematical talent was operationalized as the *genotypic value*, a genetic concept representing the sum of the average effects of genes that influence mathematical achievement (Falconer & MacKay, 1995). If the education were ideal for every child (with or without genetic mathematical talent), this would predict that individual differences in scores are mainly explained by genetic differences rather than environmental influences (see Shakeshaft et al., 2013 for a similar argument). That is, differences in children's mathematics performance

* Corresponding author at: PO Box 90153, 5000 LE Tilburg, The Netherlands.
E-mail address: I.Schwabe@uvt.nl (I. Schwabe).

http://dx.doi.org/10.1016/j.lindif.2017.01.018
1041-6080/

can be explained solely by their different genetic talents and not by random environmental influences such as what friends or teachers they have. This line of reasoning would also imply that, if indeed, in primary education, mathematically talented children are not nurtured to their full potential, their performance should be more affected by random situational factors than the performance of average or weak students for whom the educational program is more appropriate, that is, better tailored to their personal needs. For example, talented students might be at the mercy of random events like having a teacher that is interested in their abilities or a neighbour that is willing and able to help with the more challenging homework assignments. Such a situation induces the presence of *genotype-environment interaction*: conditional on a child's *genotypic value* for mathematical ability, environmental influences can be more or less important (see e.g. Cameron, 1993), or put differently, environmental factors create more variance in test scores of the talented than the less talented children.

One of the methods used in behaviour genetics to estimate the relative influence of genetic and environmental factors is the twin design. Twin pairs are either identical (monozygotic, MZ) or non-identical (dizygotic, DZ). MZ twins (largely) share the same genomic sequence and the same rearing environment, including prenatal environmental conditions. DZ twins also share the same environment but on average only share half of the segregating genes. By using the twin design, the relative contributions of genetic and environmental variability can be estimated, where the heritability is defined as the ratio of genetic variance divided by total variance in a measured trait (phenotypic variance).

Although a considerable number of twin studies have studied the heritability of mathematical ability (see e.g. Alarcon, Knopik & DeFries, 2000; de Zeeuw, de Geus, & Boomsma, 2015), to our knowledge, there is only one twin study that compared the relative contributions of genetic and environmental influences in mathematically high-scoring children and children in the normal range. In a population-based sample of 10-year-old British twins, Petrill, Kovas, Hart, Thompson, and Plomin (2009) defined mathematically high-scoring twins as those who scored at or above the 85th percentile. In the top 15% of students, the heritability estimate was similar to the one obtained across the normal range of ability. Similar results were reported for high cognitive performance and high reading performance (e.g. Boada et al., 2002; Petrill et al., 1998; Ronald, Spinath, & Plomin, 2002), traits that are correlated with mathematical ability (e.g. Davis, Haworth, & Plomin, 2009). These findings seem to argue against the presence of a genotype-environment interaction, at least in the populations studied. If there were indeed genotype-environment interaction, studies focusing on the high extreme of mathematical ability should reveal that environmental influences differ in importance compared to the normal range of ability.

Although the comparison of heritability across high and normal performing twins provides a simple test for a different etiology of extreme performance scores, it does not provide information on heritability along the entire performance continuum (see also Boada et al., 2002). In addition, an often arbitrary cutoff point has to be chosen. Most importantly, comparing the heritability in two separate ranges of ability can be misleading when one does not take into account differences in measurement reliability (van den Berg, Glas, & Boomsma, 2007).

Therefore, instead, here we estimate genotype-environment interaction continuously, letting the size of environmental variance components vary as a function of the genotypic value (see below for details). Thus, rather than studying subgroups, we take advantage of the continuous nature of the scores on mathematical performance. In this approach, we also would like to correct for the increased measurement error in the upper tail of the test score distribution. While most achievement tests show little measurement error for average scores, scoring can become very unreliable for high performing students due to the small amount of information provided by only a few very difficult items, a problem that finds its most extreme form in ceiling effects. In other words: measurement error is not the same across the ability continuum (heterogeneity). The relative lack of reliability in the upper and lower tails leads to lower correlations among sum scores (attenuation), which leads to bias when estimating genetic and environmental variance components (see van den Berg et al., 2007) and furthermore can lead to the finding of spurious genotype-environment interaction effects or missing them altogether (see Molenaar & Dolan, 2014; Schwabe & van den Berg, 2014). The problem of heterogeneous measurement error can be solved by, instead of focusing on observed test scores, modelling latent variables, and using measurement models (van den Berg et al., 2007). We model genotype-environment interaction continuously, by applying a recently developed method (Molenaar & Dolan, 2014; Schwabe & van den Berg, 2014) that corrects for measurement error through the application of an item response theory (IRT) measurement model. By incorporating an IRT model into the analysis, the results regarding genotype-environment interaction presented here are free of artefacts due to heterogeneous measurement error across the performance continuum. The method was applied to data from 2110 12-year-old Dutch twin pairs on the *Eindtoets Basisonderwijs* test, a Dutch national educational achievement test that assesses what a child has learned during primary education. If the primary educational system in the Netherlands really is better suited for students without much genetic talent (i.e. low genotypic value) for mathematics than for talented students (i.e. high genotypic value), results should show more environmental variation in children genetically predisposed towards high mathematical ability than for children genetically predisposed towards low mathematical ability.

2. Method

2.1. Data

The sample of twins for this study comes from the Netherlands Twin Register (NTR, Boomsma et al., 2002). Data on the *Eindtoets Basisonderwijs* test of 12-year-old twins from birth cohorts 1998–2000 were analyzed to study genotype-environment interaction in mathematical achievement on the *Eindtoets Basisonderwijs* test. Conducted and analyzed by the testing company Cito, this test consists of 290 multiple choice items in four different subjects (language, arithmetic/mathematics, study skills and world orientation [optional]). For this paper, the 60 dichotomous item scores (coded as 0 = incorrect, 1 = correct) of the mathematics subscale of this test were analyzed. The methods used in this study required item data, whereas at the NTR only total test scores were available. The NTR data on twins for whom signed informed consent forms for database linking were available were therefore linked to item data available at Cito. This was done by an ICT employee at Cito who was not involved in the study. Linking was based on name, sex, birth year, name of the school, and total Cito score, if available, for 7031 twins. The first step was to link the NTR data to a BRIN code, a 6-digit number that is given to educational institutes by the Dutch ministry. Then 12 different queries with a different combination of the BRIN code, birth year, sex, surname and initials of a twin were used to identify the item data associated with an individual. 1017 twins had more than one unique match and 2427 twins could not be matched at all, reducing the dataset to 3587 twins consisting of 2149 families. To link twins with item scores to the NTR data of their co-twin, a unique family ID was used. Excluding triplets (N = 63 individuals), this led to a dataset of 4238 twins (2119 twin pairs). Twin pairs with unknown zygosity (N pairs = 9) were excluded from the analysis, leading to a total of 4220 twins, forming 581 MZ pairs and 1529 DZ pairs. Of the monozygotic twins, 282 pairs were male and 299 were female; of the dizygotic twins, 360 pairs were male, 309 were female, and 860 were of opposite sex. For 711 twins, item scores were unknown. Scores were missing either because the child had not reached final grade yet (N twins = 52), the child was attending special education (N twins = 34), a different test was used at the school the twin was attending (N

twins = 13), the child (N twins = 2) or the whole school (N twins = 1) did not attend the test or the reason was unknown (N twins = 609).

The *Eindtoets Basisonderwijs* was administered using different test versions. In each year a regular test (paper-based) and an anchor test (paper-based) was used. The anchor test is an adapted version of the regular test in which 20 items are replaced by anchor items that are common between the years. This creates an internal anchor with which the tests from different years can be linked. Thus, 40 of the 60 items were the same in the regular test version and the anchor test version of the particular year (2010, 2011 & 2012) whereas 20 items were unique in the regular and anchor test version of a particular year. The 20 unique items were the same every year in the anchor test version but these unique items differed from year to year in the regular test version. Furthermore, there were three different digital test versions (computer-based tests) of which two shared 45 items while the rest of the items were unique. In our study the different combination of items from the regular test, the anchor test and the digital test led to nine different test versions.

These different versions may differ with respect to their overall difficulty and, as a result, the sum scores are not comparable across versions. To make them comparable, measures needed to be harmonized such that data from twins assessed by a different test version could be compared meaningfully. Instead of equating sum scores from different versions to make them comparable, we equated item parameters across versions by an IRT model. To link the different regular and anchor test versions, the psychometric group at Cito made a concurrent estimate of all the item parameters in the twin data. Note that the data allowed for a test linking design that is a combination of *common item equating* and *common person equating*. That is, the regular test and the anchor test item parameters were linked via common item blocks and the anchor item blocks and the digital tests were linked via common persons. The resulting item difficulty and discrimination parameters for all items were imputed in the measurement model which is described in further detail below.

2.2. Genetic models

With twin data, we can fit different genetic models to the data. The most commonly used model is the ACE model, which decomposes phenotypic variance, σ^2_P, into variance due to additive genetic influences (denoted as σ^2_A), variance due to common-environmental influences (denoted as σ^2_C), and variance due to unique-environmental influences (denoted as σ^2_E). Common-environmental influences are influences that lead to resemblance between twins and cannot be attributed to their genetic resemblance. They are parameterized to be perfectly correlated within a twin pair. Unique- or non-shared environmental influences are not shared within pairs and are parameterized to be uncorrelated for members of a twin pair. It is also possible to fit an AE model or an ADE model in which D represents dominance effects (non-additive genetic variance). All models, with and without genotype-environment interaction, were fitted to find the genetic model that fitted the data well, while, at the same time, being parsimonious. All genetic models were fitted simultaneously with a measurement model (IRT model). In the following, the modelling of genotype-environment interaction (based on the ACE model) and the IRT model will be discussed separately.

2.2.1. Genotype-environment interaction

In case of genotype-environment interaction, the amount of variance due to environmental influences varies systematically with genotypic value *A*, parametrized as a latent (e.g., unobserved) variable. Thus, environmental variance components are larger at either higher or lower levels of the genotypic value – in this application, this means that environmental influences are more important either for children with a genetic talent for mathematics or for children without such a talent.

When considering genotype-environment interactions, we can distinguish between two different types of interaction effects: There can be an interaction with unique-environmental influences (henceforth referred to as AxE) and there can be an interaction with common-environmental influences (henceforth referred to as AxC).

In case of AxE, we partition variance due to unique-environmental influences into an intercept (estimating environmental variance when $A = 0$) and a part that is a function of *A*. This makes the unique-environmental variance component different for each individual *j* with genotypic value A_j:

$$\sigma^2_{Ej} = \exp(\beta_0 + \beta_1 A_j) \quad (1)$$

where β_0 denotes the intercept (i.e., unique-environmental variance when $A_j = 0$) and β_1 is a slope parameter that represents AxE. Likewise, to model AxC, we portion variance due to common-environmental influences into an intercept (i.e., common-environmental variance when $A = 0$) and a part that is a function of *A*:

$$\sigma^2_{Cj} = \exp(\gamma_0 + \gamma_1 A_j) \quad (2)$$

where γ_0 denotes the intercept and γ_1 represents AxC.

Both interaction effects are modelled here as (log)linear effects, meaning that environmental variance is larger at either higher or lower levels of the genotypic value (i.e., larger differences among individuals with similar genotypic value). The sign of the slope determines the direction of the interaction effect. The exponential function is used to avoid negative variances (see e.g. SanChristobal-Gaudy, Elsen, Bodin, & Chevalet, 1998).

2.3. Measurement model

An IRT approach uses properties of each item as information to be incorporated into the scaling of individual test performance. The probability of a correct answer of an individual *j* on a test item *k* is modelled using both the latent trait θ_j (e.g., mathematical ability) and the item difficulty, b_k. The item difficulty represents the trait level associated with a 50% chance of endorsing an item. Furthermore, a discrimination parameter, a_k, can be incorporated into the IRT model. The discrimination parameter indicates how rapidly the probability of giving a correct answer changes with varying levels of the latent trait (e.g., factor loading in a factor analysis). Latent trait value θ_j can be interpreted as the true theoretical performance level for individual *j* that is corrected for the difficulty levels of the items that were in a student's test version (for further reading see Embretson & Reise, 2009). Here the one parameter logistic model (OPLM, Verhelst, Glas, & Verstralen, 1995) version of an IRT model was used that is suitable for dichotomous data where item responses are scored as correct/false.

In the context of IRT, the *test information curve* can be calculated, which represents the amount of psychometric information that a test contains for all points along the continuum of latent traits, θ_j. Furthermore, the standard error of measurement can be calculated for all latent traits.

2.4. Incorporating biometric and measurement models

Van den Berg et al. (2007) showed that, to take full advantage of the IRT approach, the genetic model and the IRT model have to be fitted concurrently – which we do here using Bayesian statistical modelling. In a Bayesian analysis, statistical inference is based on the joint posterior density of the model parameters, which is proportional to the product of a prior probability and the likelihood function of the data (for further reading, see e.g. Box & Tiao, 1992). We use a Markov chain Monte Carlo (MCMC) algorithm called Gibbs sampling (Gelfand & Smith, 1990; Geman & Geman, 1984) to obtain this joint posterior density. For a detailed description of the specification of the ACE model in this

Table 1
Model fit (DIC) for all biometric models.

Biometric model	DIC
I. AE	204356
a) With AxE	204337
II. ACE	204357
a) With AxE	204338
b) With AxC	204344
c) With AxE + AxC	204334
III. ADE	204356
a) With GxE	204337

Note. DIC = deviance information criterion.

context, the reader is referred to Schwabe and van den Berg (2014); Molenaar and Dolan (2014); Schwabe, Jonker, and van den Berg (2015) or Molenaar, Middeldorp, van Beijsterveldt, and Boomsma (2015). In addition to the ACE model, an ADE model was fitted. The detailed specification of this model is described in the online supplementary material.

2.4.1. Prior distributions

As a Bayesian approach was used, prior distributions had to be specified. We used a uniform prior distribution for all variance components ($\sigma^2_A, \sigma^2_D \sim U(0,100)$ and $\sigma^2_E, \sigma^2_C \sim U(0,100)$. In the biometric models with interaction effects, the prior for the intercepts and the slope parameters was normal and relatively non-informative ($\beta_0, \gamma_0 \sim N(-1,2)$ and β_1, $\gamma_1 \sim N(0,10)$). A normal prior distribution was placed on the phenotypic population mean ($\mu \sim N(0,10)$).

2.5. Analysis

For the ACE model we considered either AxE or AxC interaction effects, or both, and similarly for the ADE model. To assess model fit, the deviance information criterion (DIC, Spiegelhalter, Best, Carlin, & van der Linde, 2002) was calculated for each model. The DIC is a measure that estimates the amount of information that is lost when a given model is used to represent the process that generates the data. It takes account of both the goodness of fit and the complexity of a model.

For the MCMC estimation, we used the freely obtainable MCMC software package JAGS (Plummer, 2003). For further data handling, the statistical programming language R was used (R Development Core Team, 2008). As an interface from R to JAGS, we used the rjags package (Plummer, 2013).

After a burn-in phase of 12,000 iterations for each separate Markov chain, the characterization of the posterior distribution was based on a total of 75,000 iterations from five different Markov chains. The mean and standard deviation of the posterior point estimates was calculated for each parameter as was the 95% highest posterior density (HPD, see e.g. Box & Tiao, 1992) interval, which can be interpreted as the Bayesian analog of a confidence interval (CI). The influence of model parameters can be regarded as significant when the respective HPD interval does not contain zero. This does not hold for the variance components, as these are bounded at zero. For all test versions, the test information and standard error of measurement were calculated for a range of latent trait θ values. Furthermore, the effect size, defined as the factor with which the environmental variance component increases for an individual with an additive genetic effect of $A_i = \sigma_a$ (for technical details see Schwabe & van den Berg, 2014), was determined for both (AxE and AxC) interaction parameters.

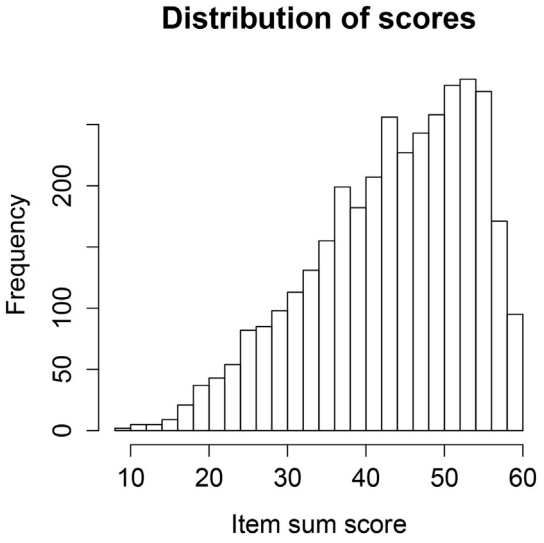

Fig. 2. Histogram of the raw sum scores (not corrected for different test versions).

3. Results

Table 1 presents the DIC for all fitted biometric models. The ACE model with AxE and AxC showed the lowest DIC and was therefore chosen as the preferred model.

Fig. 1 shows the test information curve and standard errors of measurement for each different test version for a range of latent trait values ([−0.20; 0.90]). This range of values for latent traits was chosen based on the 95% HPD interval of the θ values that occurred in the posterior based on the ACE model with AxE and AxC. It can be seen that all test

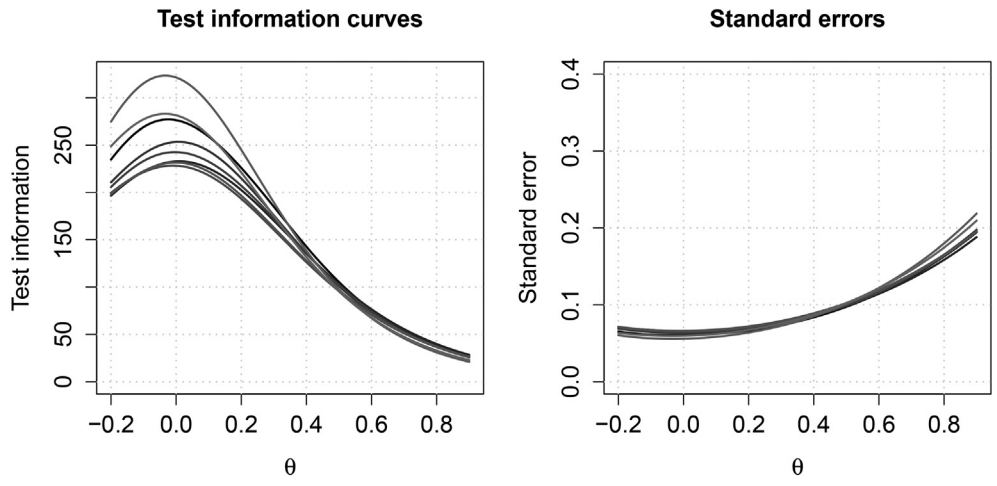

Fig. 1. Test information curves (left) and standard errors (right) for all different test versions for a representative part of the latent trait continuum.

versions provide substantial information for the average and even low-performing range of trait values, but that there is less information for the very-high performing students. This is also reflected in the standard errors of measurement. Although standard errors are generally low, they are higher for high-performing students. The distribution of the sum scores (see Fig. 2) shows that, although only 35 (<1%) students got a perfect score of 60 correct items, only 828 students (20%) scored above the mode of 52 correct item answers.

Based on the ACE model with AxE and AxC, the posterior means and standard deviations of all parameters and narrow-sense heritability can be found in Table 2. The 95% credibility region of the AxE interaction effect is displayed for the entire range of estimated genotypic values in Fig. 3.

The results suggest that most of the phenotypic variance can be explained by genetic influences, resulting in a narrow-sense heritability h^2 of 0.7286. Narrow-sense heritability is the proportion of the additive genetic variance of the phenotypic variance and was defined here as $\frac{\sigma_A^2}{\sigma_A^2 + \exp(\gamma_0) + \exp(\beta_0)}$. A substantial part of the phenotypic variance could be explained by unique-environmental influences while common-environmental influences were negligibly small. The results showed a positive AxE interaction effect such that individuals having high genotypic values show more variance due to unique-environmental influences than individuals with lower genotypic values. The 95% HPD interval shows that this effect was significant. Furthermore, the results suggest that there is a positive AxC effect such that individuals having high genotypic values for mathematical ability show more variance due to common-environmental influences than individuals with lower genotypic values. The 95% HPD interval however shows that this interaction effect was not significant. Effect sizes of the interaction effects were 1.63 (AxE) and 1.85 (AxC).

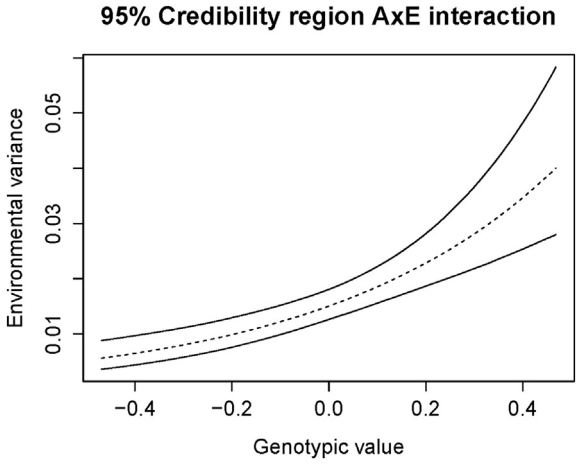

Fig. 3. 95% credibility region of the AxE interaction effect for the entire range of estimated genotypic values.

4. Discussion

The results of international comparisons are often interpreted as an indication that education in Dutch schools is more appropriate for the weakest students than for the mathematically talented students. However, until now there has been no empirical study that tested this hypothesis. From a behaviour genetics perspective, we can translate this hypothesis into the presence of *genotype-environment interaction*: if children with a talent for mathematics are not nurtured to their full (genetic) potential, we expect their academic performance to show more environmental variability, that is, achievement levels should depend more on random chance in talented children than in less talented children.

Genotype-environment interaction was investigated in Dutch students on the continuous dimension of mathematical performance. As hypothesized, we found significant genotype-environment interaction (AxE): the unique environmental variance component was larger in children with a genetic predisposition towards high mathematical ability than in children with a genetic predisposition towards low mathematical ability.

We have, however, to be cautious in drawing conclusions. The distribution of the twin's mathematics sum scores was negatively skewed. Although only 1% obtained a perfect score of 60 correct items, a relatively low percentage scored higher than the mode. Thus, there was a clear ceiling effect, with relatively little information on individual differences in performance levels on the high end of the scale. With such a limited reliability in the right tail of the distribution, even correcting for attenuation through the IRT model does not completely solve the problem of lack of reliability. If the test does not discriminate well in the right tail of the distribution, then the estimates of genotype-environment interaction effects are based foremost on the information on the rest of the population. Therefore, to draw valid conclusions on children with extreme high ability, the current study should at one time be replicated using a mathematics achievement scale with less pronounced ceiling (and floor) effects.

Recent years have seen a remarkable increase in attention for excellence in the Netherlands. The government aims to encourage excellent performance by offering talented students education tailored to their individual needs (Dekker, 2014). Twin studies can be a valuable complement to the findings of educational research, because the twin design makes it possible to correct for genetic influences when the effect of specific environmental influences is investigated. The results of present twin study add to our understanding on the issue, but drawing conclusions for policy measures requires further research. The method that we used here to model genotype-environment interaction was parametrized such that both, genetic as well as environmental influences were modelled as latent (i.e., unmeasured) variables. Therefore, no conclusions can be drawn on the nature and importance of *specific* environmental influences that are more important for students genetically predisposed towards high mathematical ability than for students with a genetic predisposition towards low or average mathematical ability. Using a population of children where the most talented ones receive tests with very difficult items, future research should focus on the exact nature of the genotype-environment interaction, for example by using the parametrization introduced by Purcell (2002) that regresses *measured* environmental moderators directly on the genotypic value. There is a broad range of influences that can contribute to differences in twin pairs, ranging from prenatal differences to different perceptions of the environment to subtle differences in brain structure. Future research should first focus on variables that have proven to be important for talented students, such as peer influences (Austin & Draper, 1981), personality characteristics (Ackerman, 1997) and motivation (Vallerand, 1994).

Table 2
Estimates of all parameters and narrow-sense heritability, based on the ACE model with AxC and AxE interactions. See Section 2.2.1 for a detailed interpretation of the parameter estimates.

	Posterior mean (SD)	HPD
σ_A^2	0.0552 (0.0035)	[0.0480; 0.0617]
$\exp(\gamma_0)$	0.0054 (0.0024)	[0.0012; 0.0099]
$\exp(\beta_0)$	0.0151 (0.0016)	[0.0120; 0.0183]
β_1 (AxE)	2.0837 (0.4939)	[1.0886; 3.0344]
γ_1(AxC)	2.6159 (1.3503)	[−0.1358; 5.2189]
h^2	0.7286 (0.0396)	[0.6448; 0.7988]

Note. Total phenotypic variance, defined as $\sigma_A^2 + \exp(\gamma_0) + \exp(\beta_0)$, was 0.0757. HPD refers to the 95% highest posterior density interval.

Acknowledgements

This study was funded by the Netherlands Scientific Organization (NWO-PROO 411-12-623).

References

Ackerman, C. (1997). Identifying gifted adolescents using personality characteristics: Dabrowski's overexcitabilities. *Roeper Review*, *19*(4), 229–236.
Alarcon, M., Knopik, V., & DeFries, J. (2000). Covariation of mathematics achievement and general cognitive ability in twins. *Journal of School Psychology*, *28*, 63–77.
Austin, A., & Draper, D. (1981). Peer relationships of the academically gifted: A review. *The Gifted Child Quarterly*, *25*, 129–133.
Boada, R., Willcutt, E., Tunick, R., Chabildas, N., Olson, R., DeFries, J., & Pennington, B. (2002). A twin study of the etiology of high reading ability. *Reading and Writing*, *15*, 683–707.
Boomsma, D., Vink, J., Beijsterveldt, C., de Geus, E., Beem, A., Mulder, E., & van Baal, G. (2002). Netherlands twin register: A focus on longitudinal research. *Twin Research and Human Genetics*, *5*, 401–406.
Box, G., & Tiao, G. (1992). *Bayesian inference in statistical analysis.* New York: John Wiley & Sons.
Cameron, N. D. (1993). Methodologies for estimation of genotype with environment interaction. *Livestock Production Science*, *35*(3–4), 237–249.
Davis, O., Haworth, C., & Plomin, R. (2009). Learning abilities and disabilities: Generalist genes in early adolescence. *Cognitive Neuropsychiatry*, *14*(4–5), 312–331.
de Zeeuw, E. L., de Geus, E. J. C., & Boomsma, D. I. (2015). Meta-analysis of twin studies highlights the importance of genetic variation in primary school educational achievement. *Trends in Neuroscience and Education*, *4*, 69–76.
Dekker, S. (2014). *Plan van aanpak toptalenten 2014–2018 (Kamerbrief).* Den Haag: Ministerie van Onderwijs, Cultuur en Wetenschap.
Embretson, S., & Reise, S. (2009). *Item response theory for psychologists.* New Jersey: Psychology Press.
Falconer, D. S., & MacKay, T. F. C. (1995). *Introduction to quantitative genetics.* Essex, UK: Pearson Education Limited.
Gelfand, A., & Smith, A. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*(410), 398–409.
Geman, S., & Geman, D. (1984). Stochastic relaxation, gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*(6), 721–741.
Meelissen, M., Netten, A., Drent, M., Punter, R., Droop, M., & Verhoeven, L. (2012). *Pirls-en timss-2011. trends in leerprestaties in lezen, rekenen ennatuuronderwijs.* Nijmegen & Enschede, Netherlands: Radbout Universiteit & Universiteit Twente.
Minne, B., Rensman, M., Vroomen, B., & Webbink, D. (2007). *Excellence for productivity? Bijzondere publicatie 69.* Den Haag, Netherlands: Centraal Planbureau.
Molenaar, D., & Dolan, C. (2014). Testing systematic genotype by environment interactions using item level data. *Behavior Genetics*, *44*(3), 212–231.
Molenaar, D., Middeldorp, C., van Beijsterveldt, T., & Boomsma, D. I. (2015). Analysis of behavioral and emotional problems in children highlights the role of genotype × environment interaction. *Child Development*, *86*(6), 1999–2016.
Petrill, S., Saudino, K., Cherny, S., Emde, R., Fulker, D., Hewitt, J., & Plomin, R. (1998). Exploring the genetic and environmental etiology of high general cognitive ability in fourteen- to thirty-six month-old twins. *Child Development*, *69*, 68–74.
Petrill, S., Kovas, Y., Hart, S., Thompson, L., & Plomin, R. (2009). The genetic and environmental etiology of high math performance in 10-year-old twins. *Behavior Genetics*, *39*(4), 371–379.
Plummer, M. (2003). *Jags: A program for analysis of bayesian graphical models using gibbs sampling.*
Plummer, M. (2013). *rjags: Bayesian graphical models using mcmc [Computer software manual].* (Retrieved from http://CRAN.R-project.org/package=jags (R package version 3 10)).
Purcell, S. (2002). Variance components models for gene-environment interaction in twin analysis. *Twin Research and Human Genetics*, *5*(6), 554–571.
R Development Core Team (2008). *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)).
Ronald, A., Spinath, F., & Plomin, R. (2002). The etiology of high cognitive ability in early childhood. *High Ability Studies*, *13*, 103–114.
SanChristobal-Gaudy, M., Elsen, J., Bodin, L., & Chevalet, C. (1998). Prediciton of the response to a selection for canalisation of a continous trait in animal breeding. *Genetics Selection Evolution*, *30*, 423–451.
Schwabe, I., & van den Berg, S. M. (2014). Assessing genotype by environment interaction in case of heterogeneous measurement error. *Behavior Genetics*, *44*(4), 394–406.
Schwabe, I., Jonker, W., & van den Berg, S. M. (2015). Genes, culture and conservatism – A psychometric-genetic approach. *Behavior Genetics*. http://dx.doi.org/10.1007/s10519-015 9768-9 (In press).
Shakeshaft, N., Trzaskowski, M., McMilan, A., Rimfeld, K., Krapohl, E., Haworth, C., & Plomin, R. (2013). Strong genetic influence on a UK nationwide test of educational achievement at the end of compulsory education at age 16. *PloS One*, *8*(12).
Spiegelhalter, D., Best, N., Carlin, B., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society*, *64*, 583–639.
Vallerand, R. (1994). A comparison of the school intrinsic motivation and perceived competence of gifted and regular students. *The Gifted Child Quarterly*, *38*(4), 172–175.
van den Berg, S., Glas, C., & Boomsma, D. (2007). Variance decomposition using an irt measurement model. *Behavior Genetics*, *37*, 604–616.
van der Steeg, M., Vermeer, N., & Lanser, D. (2011). *Nederlandse onderwijsprestaties in perspectief.* Den Haag: Centraal Planbureau.
Verhelst, N., Glas, C., & Verstralen, H. (1995). *One-parameter logistic model oplm. CITO.*