**SUPPLEMENTARY ONLINE MATERIAL**

**Samples description**

**The Western Reserve Reading and Math Project** (**WRRMP)** is an ongoing longitudinal twin study, that first started in 2002 with the recruitment of 314 families of twins in Greater Cleveland, Columbus, Cincinnati metropolitan areas, and throughout Ohio and Western Pennsylvania (Hart, Petrill, Thompson, & Plomin, 2009). The families were identified through birth records, school nomination and media advertisement. The study received approval from the Office of Responsible Research Practice at the Ohio State University.

Most were two-parent households (92%) and nearly all were Caucasian (92% of mothers, 94% of fathers). The sample includes monozygotic (MZ) and same-sex fraternal dizygotic (DZ) twins. Twins' zygosity was determined using DNA markers. Where parents did not consent to DNA collection via buccal swab (n = 76), zygosity was determined using a parental twin-zygosity questionnaire (Goldsmith, 1991). Twins' individual home assessments began when children first entered school. The annual in-home assessment occurs within one month of the previous yearly assessment. Data in the present analysis were gathered in wave 8, between 2009 and 2011. Parents/caregivers' informed consent and twins' assent is obtained prior to each data collection.

**The Quebec Newborn Twin Study** (**QNTS)** children were recruited from families of twins born between April 1995 and December 1998 in the Greater Montreal area, Canada-Quebec. Of the 989 families initially contacted, 662 (67%) agreed to participate (Boivin et al., 2013). The study obtained approval from the ethical board of the Sainte-Justine Hospital Research Centre (2009-202, 2764).

The sample includes MZ twins, and both same-sex and opposite-sex DZ twin pairs. Twins’ zygosity was established at 5 and 20 months with the Zygosity Questionnaire for Young Twins (Goldsmith, 1991). For 123 twin pairs zygosity was derived from DNA samples with concordance of the two methods ranging between 91.9% and 93.8% (Forget-Dubois et al., 2003). Twins were followed from the age of 5 months and were assessed regularly on various child and family characteristics; data relevant to this study were collected in 2013. Before each data collection informed consent is obtained from parents and/or twins.

 **The Twin Early Development Study** (**TEDS)** comprises three cohorts of twins born between 1994 and 1996in England and Wales. Out of all families of twins identified through birth records, 13,694 were initially recruited into the study. The study received approval from the Ethics Committee at King’s College London, at the Institute of Psychiatry.

Throughout the years, the sample has remained reasonably representative of UK census data (Haworth, Davis, & Plomin, 2013) and although participation to the waves of data collection is entirely optional, the active twins continue to be representative of the entire sample (Tosto et al., 2017). Twins' zygosity was ascertained at the age of 12 and 18 months using a physical similarity questionnaire. This method showed over 95% accuracy when compared to DNA testing (Price et al., 2000). The sample includes MZ twins and both same-sex and opposite-sex DZ twins. Prior to the assessment relevant to this study, at age 16 (with data collected in 2010-2011), the twins have been assessed at the ages of 2, 4, 7, 8, 9, 10, 12 and 14 years. For each assessment, informed consent was obtained from parents before data collection and the twins gave their assent (Haworth, Davis, & Plomin, 2013). Two months after the main data collection at age 16, twenty-four TEDS twin-pairs matched to the entire UK twin sample on IQ and socio-economic status repeated the same tests online for test-retest purposes (Tosto et al., 2017). For this study, twins with severe medical and/or psychiatric problems and for whom English is not their first language, were excluded from the analyses.

**The Russian School Twin Registry** (**RSTR) and Russian Singletons** were recruited from standard public schools in various regions of the Russian Federation. These schools use standardised State programmes.

The families of 15 to 18 year-old twins were first contacted in 2012 and were invited to take part in the ‘Russian School Twin Registry’ (RSTR) (Kovas et al., 2013). The twins sample includes MZ twins and both same-sex and opposite-sex DZ twins. The main aim of the registry is to contribute to Progress in Education through Gene-Environment Studies (PROGRESS).

The singletons’ sample includes students of primary, secondary and high levels at the school (Grades 1 to 11). The singletons’ data for this study were collected in standard public school from the Moscow region. Data relevant to this study were collected in 2012.

Many of the tests administered to the Russian twins and singletons are the same administered to the Canadian (QNTS) and UK (TEDS) samples. The twins and singletons Russian studies received approval from the Ethics Committee of the Psychological Institute of the Russian Academy of Education. Parental and participants’ consent was obtained prior to data collection.

**Measures**

The tests used in the current study were embedded in broader batteries administered to the samples as part of their longitudinal assessments. Previous scientific manuscripts have been published on these measures in the US (Hart et al., 2016; Lukosky et al., 2014, 2017; Wang et al., 2014, 2015), and UK sample (Tosto et al., 2017), and Russian singletons (Tikhomirova, Kuzmina, & Malykh, 2018).

**Number Line** **estimation task** was used to assess estimation of numerical magnitudes. The version used in this study was adapted from Opfer and Siegler (2007). The US twins completed the test in pen and paper format. Each of the 22 trials is constituted of an A4 size sheet. A line, representing the magnitude from 0 to 1000, is in the middle and the target number is displayed on top. Participants are asked to mark the position of the target numbers on the line with a pen. Scoring was carried out according to the instruction of Opfer’s tutorial (Opfer, J.E., 2003. Analyzing the number-line task: A tutorial). Centimeters on the line were converted to the estimated numbers: one tester “graded” each number line (measured from the 0 point to the mark drawn by the participant with a ruler, and wrote that down in cm), and a second tester repeated the procedure. Discrepant measurements were discussed by the two testers and the final value reflected the settled measurement. The score for each participant (mean error) was computed averaging the absolute difference between the target number and the number estimated in each of the 22 trials. Internal validity of the pen and paper Number Line test in the US twins was excellent (α= .95).

UK, Canadian and Russian twins and the Russian singletons completed the Number Line test online. The line with magnitude 0-1000 was shown on screen and participants estimated the same 22 numbers as the US twins by releasing a cursor on the line using a computer-mouse. In this online version, the marks left by the cursor are converted to numbers based on the number of pixels between the origin of the line and the mark. The procedure to compute mean error scores was the same as described for the US twins. Internal validity of the online Number Line test was adequate: α= .69 in the Canadian twins; .63 in the UK twins; .81 in the Russian twins and .75 in the Russian singletons. The test-retest carried out in the UK subset of 45 twins was .70 (Tosto et al., 2017).

**Mathematics abilities** were measured with two tests, one assessing fluency and another assessing problem solving. Fluency and problem solving were measured respectively with the *Problem Verification* task and *Understanding Numbers* in UK, Canadian and Russian twin samples; while fluency and problem solving were measured respectively with the *Fluency* and *Applied Problems* subtests of Woodcock-Johnson III Tests of Achievement (WJ-III) in the US sample. Mathematical performance was not available for the Russian singletons.

In the *Problem Verification*task (fluency) UK, Canadian and Russian twins were presented with arithmetic problems such as 76÷7=10. The task requires to judge, as fast and as accurately as possible without calculating, whether the proposed answer is correct or not, within a limited time. Participants were administered an online adaptation of the Problem Verification Task described in Murphy & Mazzocco (2008). The test includes 48 trials: 24 arithmetic and 24 fraction problems. The 24-sets contain 6 items each of addition, subtraction, multiplication and division. Participants were instructed to respond as quickly as possible within 10 seconds by pressing the keys corresponding to 'correct', 'incorrect', and 'don’t know'. A time bar on the top-left corner of the screen reminded participants of the elapsing time. The program records accuracy and reaction time of response. Number of correct response was used in the analyses. The Problem Verification task revealed good Cronbach’s alpha (α = .83. in Canadian twins, .85 in UK twins, and .91 in Russian twins) and good test re-test in the UK twins, r = .78, n = 48 (Tosto et al., 2017).

In *Fluency WJ-III,* the fluency subtest from Woodcock-Johnson III Tests of Achievement(Woodcock, McGraw, & Mather, 2001) the US twins were asked to solve addition, subtraction, and multiplication problems in a 3-minute time limit. The test was administered in pen and paper format. Cronbach’s alpha in this sample was adequate: α = .66 (Hart et al., 2016).

The test *Understanding Numbers* (problem solving) used in UK, Canadian and Russian online batteries, evaluates mathematical achievement expected at age 15-16 according to levels set by the UK National Curriculum. Eighteen items such as: “Find two odd numbers that add up to 8 and type the answers in the boxes” were selected from the understanding numbers component of the National Foundation for Educational Research booklets level 1 to 8 (nferNelson, 1994, 1999, 2001). This task requires an understanding of numerical operations (e.g., division is the inverse of multiplication) and of patterns of numbers. Number of correct response was used in the analyses. Cronbach’s alpha was .87 in the Canadian data (n = 318), .90 in UK data (n = 2153) and .92 in the Russian data (n = 178). Test re-test correlation was .67 (n = 48) (Tosto et al., 2017).

*Applied Problems*(problem solving) subtest from the WJ-III (Woodcock et al., 2001) was used to assess problem-solving skills involving understanding numbers in the US twins; the test was administered in pen and paper format. The initial items require application of simple number concepts while most items require a student to listen to the problem, recognize the mathematical procedure that must be followed, and perform the appropriate calculations. For example, looking at a picture of a thermometer and reading the temperature. Cronbach’s alpha in this sample was .83 (Hart et al., 2016).

**Analyses**

**The twin method** is based on comparison of the observed resemblance (intra-class correlation, ICC) (Shorut & Fleiss, 1979) between monozygotic (MZ, or identical twins) and dizygotic (DZ, or fraternal twins). Observed twins' similarity stems from common genetic and common environmental factors (Plomin et al., 2013). MZ twins share the same genotype, in most cases, therefore their genetic similarity is assumed to be 100%; in DZ twins genetic similarity is on average 50%, like in siblings. It is assumed that MZ and DZ twins living in the same family are influenced by shared elements of the environments that contribute to their similarity in a trait.

**Univariate** **genetic analyses** examine the etiology of individual differences - the extent to which genetic and environmental influences contribute to variation in a trait. Heritability (h2), the sum of all genetic influences contributing to inter-individual variation in a trait, is estimated as twice the difference between the MZ and DZ ICCs. Shared environmental influences (c2) are computed as the difference between MZ ICCs and h2; these are environments experienced by both twins in a pair contributing to their similarity on a trait. Factors such as socio-economic status, home environment, and school are often thought to play a role in similarities among family members. Conversely, non-shared environmental influences (e2) include individual unique experiences, such as different friends and classmates, differential (either objective or perceived) treatment by their parents and teachers (Kovas et al., 2007b). These are estimated by subtracting MZ ICCs from 100%; the estimates also incorporate measurement error (Plomin et al., 2013).

**Multivariate genetic analyses** allow examining the etiology of co-occurrence. These analyses estimate genetic correlation (ra), the extent to which the same genetic factors influence two measures; and shared and non-shared environmental correlations (rc and re respectively) indexing shared and non-shared environmental factors common to both traits. Another outcome of these analyses is the bivariate heritability (the proportion of phenotypic correlation between the two measures that can be explained by common genetic factors) and bivariate shared environmentalities (the proportion of phenotypic correlation explained by shared and non-shared environments affecting both measures).

**Sex-limitation models** are conducted to examine the etiology of sex differences in traits. From a behavioural genetic perspective, qualitative sex differences exist if the genetic and/or environmental causes of individual variation for that trait are (partly) different for males and females. Genetic and environmental causes of individual variation could be the same for males and females but influencing them to a different extent, generating quantitative sex-differences. Alternatively, there could be no qualitative or quantitative differences in the etiology of individual differences between males and females.

In this study, univariate and multivariate (bivariate) twin analyses were conducted using Cholesky decompositions (see Appendix in Plomin et al., 2013) run with OpenMx software (Boker et al., 2011).

**Number line estimation and mathematics across countries: ANOVAs results**

Differences across samples were examined with ANOVAs using country as factor (fix effects). Analyses were carried out on both uncorrected variable and variables corrected for age and sex. These analyses were conducted on the pooled samples subsequently divided by age to generate age-homogenous groups with reasonable sample sizes for the comparisons (Table S2). In the comparison of Number Line estimation, a cut-off with age < 12 years generated only two groups including US twins and Russian singletons. The second cut-off with age between 12 and 16 years included participants from all samples, generating five comparison groups (UK, US, Canadian and Russian twins, and Russian singletons). The third cut-off with age >16 years included Russian singletons, Russian and UK twins.

No cut-offs were applied when comparing Canadian, UK and Russian twins in mathematics (fluency and the problem solving tests) as these samples had very close mean ages with small age range. When the US sample was included in the mathematics comparisons, the age cut-off was set <16 years, as US participants’ age ranged between 8 and 15 years.

Table S2 summarizes the ANOVAs results. For example, using the above described cut-offs resulted in two groups with Number Line data with participants' age below 12 years, 109 US twins and 773 Russian singletons. Their comparison on Number Line performance showed no significant differences in the uncorrected (F(1,880) = .569, p = .451) and age-sex corrected measure (F(1,876) = .469, p=.494). Participants from the five samples with ages between 12 and 16, and with ages greater than 16 years (UK and Russian twins, and Russian singletons) showed some significant differences in uncorrected Number Line variable. However, when comparisons were performed on the age-sex corrected Number Line variable, no differences were found among the groups.

Similarly, for both Problem Verification (fluency) and Understanding Numbers (problem solving), the groups showed significant difference on the uncorrected variables, but these differences disappeared when controlling for age and sex. The effect of age and sex on the uncorrected variables was small as they accounted between 0.1% and 5.8% (Table S2). This consistent pattern suggests that variance in age and sex, rather than variance stemming from countries or test, drives the small differences observed.

 The cut-offs, reduced the overall age gap in the compared groups, however the groups still showed great variability, as homogeneity of variance in age was not met in any of the comparisons. From Table S2, it can also be observed that the cut-offs generated groups of different sample-size. Further, the slight difference of tests needs to be considered when the comparisons include the US sample. These factors may have introduced biases.

|  |
| --- |
| Table S1. Means and standard deviation on all measures for each sample |
|    | Age range | Age (years) | Raw data | Raw data  | Raw data males | Raw data females | MZm | MZf | DZm | DZf | DZos |
|  | (years) | M (SD) | Mediann | M (SD)n | M (SD)n | M (SD)n | M (SD)n | M (SD)n | M (SD)n | M (SD)n | M (SD)n |
| **WRRMP**    | 8.75-15.33 | 12.27 (1.20) |  |  |  |  |  |  |  |  |  |
| Number Line estimation |  |  | 55.31 n=246 | 81.53 (78.21) n= 246 | 66.15 (55.29) n= 106 | 93.17 (90.21) n= 140 | -.14 (.85) n= 41 | 0 (1.02) n= 53 | .09 (.84) n= 65 | -.02 (1.13) n= 87 | -- |
| Number Line (Lambda) |  |  | .09 n=246 | .19 (.25) n= 246 | .13 (.19) n= 106 | .23 (.28) n= 140 | -.04 (.74) n= 41 | -.03 (1.02) n= 53 | .03 (.79) n= 65 | .02 (1.22) n= 87.5 | -- |
| Fluency WJ-III |  |  | 100 n=242 | 100.36 (16.30) n= 242 | 101.52 (19.38) n= 103 | 99.49 (13.54) n= 139 | 0 (1.45) n= 39 | .13 (.96) n = 53 | -.01 (1.02) n = 64 | -.05 (.74)n=86 | -- |
| Applied Problems WJ-III |  |  | 107 n=245 | 106.4 (11.78) n= 245 | 108.25 (11.12) n= 105 | 105.02 (12.08) n= 140 | -.05 (.85) n= 41 | .09 (1.10) n= 53 | .02 (1.02) n= 64 | -.05 (.98) n= 87 | -- |
| Mathematics Composite  |  |  | - | - |   | - | -.03 (1.13) n= 39 | .12 (1.07) n= 53 | .02 (1.05) n= 63 | -.08 (.84) n= 86 | -- |
| **QNTS**   | 14.50-16.08 | 15.17 (.29) |  |  |  |  |  |  |  |  |  |
| Number Line estimation |  |  | 41.00n=332 | 45.68 (25.33)n=332 | 44.82 (24.26)n=158 | 46.46 (26.31)n=174 | -.10 (.90)n=56 | -.02 (1.02)n=76 | -.15 (.80)n=49 | .04 (.98)n=49 | -.12 (1.03)n=92 |
| Number Line (Lambda) |  |  | .00n=331 | .022 (.07)n=331 | .017 (.06)n=157 | .027 (.07)n=174 | .017 (.07)n=56 | .027 (.09)n=76 | .011 (.03)n=49 | .032 (.07)n=49 | .022 (.06)n=91 |
| Problem. Verif. (Fluency) |  |  | 35.00n=331 | 33.86 (7.14)n=331 | 34.37 (7.96)n=157 | 33.40 (6.30)n=174 | -.03 (1.14)n=56 | -.10 (.91)n=76 | .12 (1.23)n=49 | .02 (1.03)n=49 | .17 (.84)n=91 |
| Understanding Numbers |  |  | 11.00n=318 | 10.57 (3.86)n=318 | 10.55 (4.19)n=149 | 10.58 (3.55)n=169 | .12 (.94)n=53 | .008 (.98)n=73 | -.009 (1.18)n=49 | -.01 (.87)n=47 | -.04 (.89)n=88 |
| Mathematics Composite  |  |  |  | -- | -- | -- | .05 (1.11)n=56 | -.04 (.89)n=76 | .07 (1.25)n=49 | .02 (.88)n=49 | .07 (.86)n=91 |
| **TEDS**  | 15.82-17.28 | 16.48 (.27) |  |  |  |  |  |  |  |  |  |
| Number Line estimation |  |  | 33.88n=2555 | 36.25 (15.18)n=2555 | 35.51 (15.02)n=1070  | 36.78 (15.28)n=1485 | .03 (.93)n=363 | .03 (.94)n=585 | .03 (.98)n=336 | -.10 (1.00)n=480 | .00 (1.00)n=783 |
| Number Line (Lambda) |  |  | .00n=2555 | .01 (.02)n=2555 | .01 (.02)n=1070 | .01 (.02)n=1485 | .01 (.02)n=363 | .01 (.02)n=585 | .01 (.03)n=336 | .01 (.02)n=480 | .01 (02)n=783 |
| Problem. Verif. (Fluency) |  |  | 37.00n=2238 | 36.08 (6.68)n=2238 | 37.66 (6.68)n=931 | 34.96 (6.35)n=1307 | -.03 (1.02)n=321 | .00 (.99)n=523 | -.07 (1.05)n=295 | .07 (1.01)n=418 | .08 (.63)n=677 |
| Understanding Numbers |  |  | 12.25n=2153 | 11.94 (4.06)n=2153 | 12.58 (3.81)n=885 | 11.48 (4.16)n=1268 | -.02 (.90)n=306 | -.01(1.04)n=516 | .08 (.96)n=2888 | .06 (1.04)n=415 | .03 (.98)n=628 |
| Mathematics Composite  |  |  |  | -- | -- | -- | -.02 (.96)n=327 | -.01 (1.01)n=542 | -.01 (1.02)n=305 | .07 (1.02)n=437 | .08 (.96)n=687 |
| **RSTR**  | 14.54-18.84 | 16.44 (.91) |  |  |  |  |  |  |  |  |  |
| Number Line estimation  |  |  | 40.50n=165 | 47.94 (28.27)n=165 | 47.35 (22.49)n=68 | 48.76 (31.99)n=95  | -- | -- | -- | -- | -- |
| Number Line (Lambda) |  |  | .00n=166 | .05 (.00)n=166 | .05 (.13)n=70 | .04 (.10)n=96 | -- | -- | -- | -- | -- |
| Problem. Verif. (Fluency) |  |  | 35.00n=171 | 33.95 (8.88)n=171 | 33.29 (9.77)n=75 | 34.65 (8.00)n=93 | -- | -- | -- | -- | -- |
| Understanding Numbers |  |  | 9.00n=173 | 8.86 (4.79)n=173 | 7.93 (5.07)n=76 | 9.19 (4.53)n=94 | -- | -- | -- | -- | -- |
| Mathematics Composite  |  |  |  | -- | -- | -- | -- | -- | -- | -- | -- |
| **Singletons Younger age < 16 years** | 7.51-15.99 | 11.42 (2.59) |  |  |  |  |  |  |  |  |  |
| Number Line estimation |  |  | 61.55n=1232 | 85.87 (65.05)n=1232 | 80.13 (60.30)n=657  | 92.42 (69.55)n=575  | -- | -- | -- | -- | -- |
| Number Line (Lambda) |  |  | .04n=1231 | .12 (.17)n=1231 | .09 (.14)n=656 | .14 (.19)n=575 | -- | -- | -- | -- | -- |
| **Singletons Older age > 16 years** | 16.00-18.85 | 17.14 (.74) |  |  |  |  |  |  |  |  |  |
| Number Line estimation |  |  | 41.10n=211 | 43.95 (16.96)n=211 | 41.07 (15.12)n=84  | 45.85 (17.88)n=127  | -- | -- | -- | -- | -- |
| Number Line (Lambda) |  |  | .00n=211 | .02 (.04)n=211 | .01 (.03)n=84 | .02 (.04)n=127 | -- | -- | -- | -- | -- |
| MZm = Monozygotic males; MZf = Monozygotic females; DZm = Dizygotic same-sex males; DZf = Dizygotic same-sex females; DZos = Dizygotic opposite-sex. For independence of data, mean (M) and standard deviation (SD) are reported on one twin randomly selected in each pair. Descriptive statistics for the twins divided by zygosity are reported on standardized variables corrected for age and sex. |

|  |
| --- |
| Table S2: ANOVAs comparison of Number Line estimation and Mathematics means in groups with restricted age range |
| Comparison variable | age cut-offs | Sample size in each comparison | F-value  | ƞ2 | F-value  | ƞ2 |
| WRRMPUS | QNTSCA | TEDSUK | RSTRRU | R.SinglRU | Uncorrected scores |  | Corrected scores |  |
| Number Line estimation  | age < 12 | 109 | -- | -- | -- | 773 | F(1, 880) = .57, p = .451  | .001 | F(1, 876) = .47, p = .494  | .001 |
| 12<age≤16 | 134 | 323 | 93 | 60 | 458 | F(4, 1063) = 12.93, p < .001  | .046 | F(4, 1054) = 1.32, p = .260 | .005 |
| age > 16 | -- | -- | 2449 | 101 | 222 | F(2, 2769) = 87.76, p <.001  | .058 | F(2, 2754) = 2.10, p = .123 | .002 |
| Homogeneity of variance for age was not met in any of the age cut-offs. Effect size (ƞ2) of age on Number Line estimation ranged between .195 at age <12 cut-off and .441 with age between 12 and 16. |
| Understanding Numbers/Applied Problems WJ-III | No age cut off | -- | 313 | 2153 | 169 | N/A | F(2, 2632) = 43.86, p < .001 | .032 | F(2, 2632) = .09, p = .913 | .000 |
| age < 16 | 247 | 309 | 87 | 63 | N/A | N/A |  | F(3, 702) = .12, p = .949 | .001 |
| Homogeneity of variance for age was not met in any of the age cut-offs. Effect size (ƞ2) of age was .596 when the 3 samples were compared on Understanding Numbers test with no age cut-off, and .815 when the US twins tested on Applied Problems WJ-III were included. |
| Problem Verification/ Fluency WJ-III | No age cut off | -- | 326 | 2238 | 167 | N/A | F(2, 2798) = 20.38, p < .001 | .015 | F(2, 2798) = .14, p = .868 | .000 |
| age < 16 | 243 | 322 | 89 | 63 | N/A | N/A |  | F(3, 713) = .22, p = .882 | .001 |
| Homogeneity of variance for age was not met in any of the age cut-offs. Effect size (ƞ2) of age was .601 when the 3 samples were compared on fluency Problem Verification, with no age cut-off, and .810 when the US twins tested on Fluency WJ-III were included. |
| Groups for the means comparison were generated pooling together all participants and applying 3 age cut-offs: age <12 years, age between 12 and 16 years, and age >16 years. These age cut-offs allow groups of reasonable sample size with a more restricted age range. For Number Line estimation, the 4 twin samples and the singletons were pooled together. No data on mathematics was available for the Russian singletons, therefore, only the twin samples were pooled together for the comparison of means in mathematics. UK, Canadian and Russian twins were pooled together for the comparison of Understanding Numbers and Problem Verification with no age cut-off as these groups had a restricted age range of ~16 years. When the US sample was included in the comparison, the age cut-off was set <16 years to exclude the few older (~18 years old) twins in this pooled data. Comparisons were conducted on the uncorrected and age-sex corrected scores of Number Line estimation, Number Line Lambda, Problem Verification and Understanding Numbers. Results for Mathematics Composite (age-sex corrected scores) means comparison are not reported in this table as no differences were detected. Results for Lambda parameters are also not reported in this table as they follow the same pattern as Number Line estimation. The only exception is that for age > 16, significant differences in Lambda were found in both uncorrected and corrected variable. |

CI = 95% confidence intervals; MZm = Monozygotic males; DZm = Dizygotic same-sex males; MZf = Monozygotic females; DZf = Dizygotic same-sex females; DZss = Dizygotic same-sex; DZos = Dizygotic opposite-sex

|  |
| --- |
| Table S3. Intraclass correlation (ICC). Twins similarity in Number Line estimationand mathematics |
|   | MZ (CI) n | DZ (CI) n | MZm (CI) n | DZm (CI) n | MZf (CI) n | DZf (CI) n  | DZss (CI) n | DZos (CI) n |
| **WRRMP**  |   |   |   |   |   |   |   |   |
| Number Line estimation | .47 (.30; .62) n=94 | .43 (.30; .56) n=152 | .42 (.13; .64) n=41 | .15 (-.10; .37) n=65 | .50 (.27; .68) n=53 | .56 (.39; .69) n=87 |  -- |  -- |
| Fluency WJ-III | .87 (.80; .91) n=94 | .54 (.41; .64) n=152 | .89 (.8; .94) n=41 | .64 (.47; .76) n=65 | .82 (.71; .89) n=53 | .40 (.21; .56) n=87 | -- | -- |
| Applied Problems WJ-III | .76 (.66; .83) n=94 | .56 (.45; .66) n=152 | .64 (.42; .79) n=41 | .66 (.49; .78) n=65 | .81 (.70; .89) n=53 | .49 (.32; .64) n=87 | -- | -- |
| Mathematics Composite | .88 (.83; .92) n=94 | .59 (.48; .69) n=152 | .85 (.73; .92) n=41 | .69 (.54; .80) n=65 | .91 (.85; .95) n=53 | .48 (.30; .63) n=87 | -- | -- |
| **QNTS**  |  |  |  |  |  |  |  |  |
| Number Line estimation | .41 (.25; .54) n=130 | .26 (.12; .39) n=189 | .29 (.03; .51) n=55 | .36 (.09; .58) n=49 | .47 (.28; .63) n=75 | .35 (.08; .57) n=48 | .35 (.16; .51) n=97 | .19 (-.02; .38) n=92 |
| Problem Verification (Fluency) | .46 (.32; .59) n=130 | .19 (.05; .32) n=187 | .49 (.26; .67) n=55 | .23 (-.05; .48) n=49 | .44 (.24; .60) n=75 | .18 (-.11; .44) n=48 | .20 (.00; .39) n=97 | .17 (-.04; .36) n=90 |
| Understanding Numbers | .49 (.34; .61) n=122 | .28 (.14; .41) n=177 | .46 (.22; .65) n=53 | .39 (.13; .60) n=49 | .51 (.31; .67) n=69 | .25 (-.04; .51) n=45 | .33 (.14; .50) n=94 | .19 (-.02; .39) n=83 |
| Mathematics Composite | .58 (.45; .68) n=130 | .28 (.14; .41) n=187 | .59 (.38; .74) n=55 | .34 (.07; .57) n=49 | .58 (.41; .71) n=75 | .21 (-.07; .46) n=48 | .28 (.09; .46) n=97 | .28 (.07; .46) n=90 |
| **TEDS**  |  |  |  |  |  |  |  |  |
| Number Line estimation | .26 (.20; .32) n=872 | .14 (.08; .19) n=1412 | .33 (.23; .42) n=328 | .09 (-.02; .20) n=304 | .22 (.14; .30) n=544 | .18 (.09; .27) n=443 | .14 (.07; .21) n=746 | .13 (.05; .20) n=765 |
| Problem Verification (Fluency) | .54 (.49; .59) n=757 | .25 (.20; .30) n=1206 | .57 (.48; .64) n=282 | .34 (.23; .44) n=262 | .54 (.46; .59) n=475 | .22 (.12; .31) n=436 | .27 (.19; .34) n=641 | .23 (.15; .31) n=565 |
| Understanding Numbers | .64 (.59; .68) n=738 | .34 (.49; .54) n=3273 | .61 (.54; .68) n=277 | .45 (.35; .54) n=257 | .65 (.59; .70) n=461 | .36 (.27; .45) n=374 | .39 (.32; .45) n=631 | .28 (.20; .35) n=535 |
| Mathematics Composite | .68 (.64; .72) n=799 | .36 (.32; .41) n=1278 | .67 (.60; .73) n=295 | .44 (.34; .53) n=277 | .69 (.64; .73) n=504 | .37 (.28; .45) n=414 | .39 (.33; .45) n=691 | .33 (.25; .40) n=587 |

|  |
| --- |
| Table S4. Heritability estimates from twin model-fitting for Number Line estimation and mathematics measures |
|   | h2 | c2 | e2 |
| **WRRMP**  |   |   |   |
| Number Line estimation  | .16 (.00; .52) | .34 (.05; .53) | .50 (.37; .65) |
| Number Line (Lambda) | .26 (.00; .59) | .31 (.05; .55) | .42 (.31; .58) |
| Fluency WJ-III | .39 (.24; .64) | .43 (.19; .60) | .18 (.14; .25) |
| Applied Problems WJ-III | .39 (.16; .64) | .37 (.14; .56) | .24 (.18; .33) |
| Mathematics Composite | .46 (.28; .69) | .40 (.18; .56) | .14 (.10; .19) |
| **QNTS**  |  |  |  |
| Number Line estimation  | .35 (.00, .55) | .08 (.00, .36) | .57 (.45, .73) |
| Number Line (Lambda) | .40 (.10; .55) | .05 (.00; .29) | .55 (.44; .67) |
| Problem Verification. (Fluency) | .46 (.19, .57) | .00 (.00, .20) | .54 (.43, .68) |
| Understanding Numbers | .35 (.00, .58) | .12 (.00, .40) | .53 (.42, .67) |
| Mathematics Composite | .59 (.28, .68) | .00 (.00, .24) | .41 (.32, .53) |
| **TEDS**  |  |  |  |
| Number Line estimation  | .27 (.12; .32)  | .00 (.00; .11)  | .73 (.68; .78)  |
| Number Line (Lambda) | N/A | N/A | N/A |
| Problem Verification (Fluency) | .56 (.52; .60)  | .00 (.00; .08)  | .44 (.40; .48)  |
| Understanding Numbers | .59 (.47; .68)  | .06 (.00; .16)  | .35 (.32; .39)  |
| Mathematics Composite | .64 (.64; .72)  | .05 (.00; .14)  | .31 (.27; .34)  |
| h2 = heritability, proportion of variance explained by genetic influences; c2 = proportion of variance explained by shared-environmental influences; e2 = proportion of variance explained by non shared-environmental influences. The 95% confidence intervals, in brackets, are wider in the smaller samples. The results show about similar genetic heritability, shared and nonshared environmental factors to individual variation of the Number Line (linear – mean error) and of the logarithmic (non-linear) representation of numbers (Lambda) in both US and Canadian twins. |

|  |
| --- |
| Table S5. Sex limitation model fitting and parameter estimates for males and females separately (UK TEDS sample only) |
| Variable | -2LL | df | Δ-2LL | AIC | BIC | p-val | ep | a (95%CI) | c (95%CI) | e (95%CI) |
| **Number Line estimation** |  |  |  |  |  |  |  | males | males | males |
| Full Sex-Limitation | 12053.30 | 4320 |  | 3413.30 | -25242.88 |  | 9 | .31 (.15; .40) | .00 (.00; .11) | .69 (.60; .78) |
| Common Effects Model | 12053.30 | 4321 | .00 | 3411.30 | -25251.51 | 1.00 | 8 | females | females | females |
| Scalar Effects Model | 12056.52 | 4323 | 3.22 | 3410.53 | -25265.56 | .19 | 6 | .12 (.00; .30) | .11 (.00; .24) | .80 (.70; .84) |
| **Null Model** | **12056.74** | **4324** | **3.43** | **3408.74** | **-25273.97** | **.48** | **5** | all | all | all |
|  |  |  |  |  |  |  |  | .27 (.11; .32) | .00 (.00; .11) | .73 (.68; .79) |
| **Problem Verification (Fluency)** |  |  |  |  |  |  |  | males | males | males |
| Full Sex-Limitation | 10409.00 | 3788 |  | 2833.00 | -22294.22 |  | 9 | .53 (.30; .66) | .06(.00; .26) | .40 (.34; .43) |
| Common Effects Model | 10409.00 | 3789 | .00 | 2831.00 | -22302.86 | 1.00 | 8 | females | females | females |
| Scalar Effects Model | 10410.74 | 3791 | 1.74 | 2828.74 | -22318.38 | .63 | 6 | .55 (.41; .51) | .00 (.00; .12) | .45 (.39; .51) |
| **Null Model** | **10410.74** | **3792** | **1.74** | **2826.74** | **-22327.02** | **.78** | **5** | all | all | all |
|  |  |  |  |  |  |  |  | .57 (.43; .61) | .00 (.00; .11) | .43 (.39; .47) |
| **Understanding Numbers** |  |  |  |  |  |  |  | males | males | males |
| Full Sex-Limitation | 9866.97 | 3646 |  | 2574.97 | -21610.31 |  | 9 | .37 (.16; .66) | .27 (.01; .44) | .36 (.30; .43) |
| **Common Effects Model** | **9867.47** | **3649** | **.50** | **2573.47** | **-21618.45** | **.47** | **8** | females | females | females |
| Scalar Effects Model | 9880.03 | 3649 | 12.56 | 2582.03 | -21623.16 | .002 | 6 | .59 (.41; .70) | .06 (.00; .22) | .34 (.30; .39) |
| Null Model | 9884.31 | 3650 | 17.34 | 2584.31 | -21627.51 | .002 | 5 | all | all | all |
|  |  |  |  |  |  |  |  | 58 (.45; .68) | .07 (.00; .18) | .35 (.31; .39) |
| **Mathematics Composite** |  |  |  |  |  |  |  | males | males | males |
| Full Sex-Limitation | 10420.62 | 3909 |  | 2602.62 | -23327.25 |  | 9 | .51 (.32; .71) | .18 (.00; .35) | .31 (.26; .37) |
| Common Effects Model | 10420.62 | 3910 | .00 | 2600.62 | -23335.88 | 1 | 8 | females | females | females |
| Scalar Effects Model | 10425.05 | 3912 | 4.43 | 2601.05 | -23348.71 | .11 | 6 | .68 (.52; .73) | .03 (.00; .18) | .29 (.26; .34) |
| **Null Model** | **10425.18** | **3913** | **4.57** | **2599.18** | **-23357.21** | **.33** | **5** | all | all | all |
|  |  |  |  |  |  |  |  | .63 (.51; .72) | .07 (.00; .20) | .30 (.27; .34) |
| -2LL = -2 Loglikelihood; df = degrees of freedom; Δ-2LL = differences in likelihood between compared models; AIC = Akaike Information Criterion; BIC= Bayesian Information Criterion; p-val=significance in change of likelihood between compared models. ep = estimated parameters. The best fitting model (bold characters) is inferred when the change in likelihood associated to the drop of parameter estimated does not produce a worsening in fit of the nested model. Comparisons are conducted between the Full and Common models to test for qualitative sex-differences; between the Common and Scalar for quantitative sex-differences; between the Null and Scalar for variance differences. The table reports the fit comparison between the Null and Full models. For Number Line, Fluency and the Mathematics Composite the best model fitting is the Null Model, which indicates no qualitative, quantitative or variance sex differences. For Understanding Numbers, the best fitting model is the Common Effects Model, which allows for quantitative, but not qualitative sex differences. The fit of this model however, is not significantly better than the full model (p=.47), indicating no significant qualitative sex differences. The model testing for quantitative sex differences (Scalar Effects Model) has a significantly different fit than the Common model (p=.002). This fit is better only according BIC, but not AIC, suggesting that quantitative sex differences may exist for this measure. Such differences may be small as the discrepancies between the two fit indexes suggest a weak fit of this model. Finally, the difference in fit between the Null and Scalar model is significant (non reported in table: p=0.04). This suggest variance differences, however, the fit if the Null model is better according BIC but not AIC, suggesting that if there are variance differences these must be small. The difference in fit between the Null and Full model is significant. The Full model is a better fit only by BIC and not AIC, suggesting small, non-qualitative sex differences in this measure. |

|  |
| --- |
| Table S6. Cross-trait twin correlations between Number Line estimation and mathematics measures |
|   | MZ ICC (95%CI) N | DZ ICC (95%CI) N |
| **WRRMP**  |  |  |
| NL tw1 & Fluency (WJ-III) tw2 | -.13 (-.33; .07) n=94 | -.27 (-.41; -.12) n=152 |
| NL tw1 & Applied Problems (WJ-III) tw2 | -.48 (-.62; -.3) n=94 | -.42 (-.55; -.28) n=152 |
| NL tw1 & Mathematics Composite tw2 | -.34 (-.51; -.14) n=94 | -.42 (-.54; -.27) n=152 |
| **QNTS**  |  |  |
| NL tw1 & Problem Verification (Fluency) tw2 | -.44 (-.56; -.29) n=130 | -.16 (-.29; -.01) n=188 |
| NL tw1 & Understanding Numbers tw2 | -.38 (-.52; -.22) n=125 | -.18 (-.32; -.04) n=183 |
| NL tw1 & Mathematics Composite tw2 | -.46 (-.59; -.31) n=130 | -.19 (-.32; -.05) n=188 |
| **TEDS**  |  |  |
| NL tw1 & Problem Verification (Fluency) tw2 | -.29 (-.36; -.23) n=792 | -.16 (-.22; -.11) n=1294 |
| NL tw1 & Understanding Numbers tw2 | -.28 (-.35; -.22) n=762 | -.19 (-.24; -.13) n=1253 |
| NL tw1 & Mathematics Composite tw2 | -.32 (-.38; -.26) n=810 | -.21 (-.26; -.16) n=1337 |
| NL = Number Line estimation; tw1 = twin 1; tw2 = twin 2; MZ = monozygotic twin pairs; DZ = dizygotic twin pairs; ICC = intra-class correlations. Cross-trait twin correlations are computed using scores of 1 twin in the pair in a measure (Number Fine for example) and scores of twin 2 in the pair in the other measure (Problem Verification). |

|  |
| --- |
| Table S7. Genetic and environmental correlations; bivariate heritabilities; phenotypic correlations |
|   | ra | rc  | re | h2-bivariate | c2-bivariate  | e2-bivariate | ph. corr. |
| **WRRMP**  |  |  |  |  |  |  |  |
| NL & Fluency WJ-III | -  | **-.28 (-.44; -.11)** | -.11 (-.23; .02) | -  | **.79 (.30; 1.0)** | .21 (-.05; .46) | **-.20 (-.30; -.10)** |
| L & Fluency WJ-III | -  | **-.26 (-.42; -.09)** | -.08 (-.20; .05) |  - | **.84 (.31; 1.0)** | .16 (-.11; .42) | **-.18 (-.28; -.08)** |
| NL & Applied Problems WJ-III |  -.93 (-1.0; -1.0) | **-.68 (-1.0; -.68)** | .01 (-.19; .18) | **.53 (.13; .93)** | **.47 (.14; .85)** | .00 (-.15; .12) | **-.49 (-.56; -.41)** |
| L & Applied Problems | -.49 (-1.0; 1.0) | **-.88 (-1.0; -.88)** | -.05 (-.25; .15) | .33 (-.09; .69) | **.63 (.31; 1.0)** | .04 (-.12; .17) | **-.47 (-.55; -.39)** |
| NL & Mathematics Composite | -.37 (-1.0; 1.0) | **-.75 (-1.0; -.75)** | -.12 (-.31; .08) | .29 (-.17; .72) | **.63 (.24; 1.0)** | .08 (-.07; .20) | **-.40 (-.49; -.32)** |
| L & Mathematics Composite | -.18 (-1.0; 1.0) | **-.85 (-1.0; 1.0)** | -.14 (-.33; .06) | .17 (-.29; .61) | **.75 (.37; 1.0)** | .09 (-.06; .21) | **-.39 (-.47; -.30)** |
| **QNTS**  |  |  |  |  |  |  |  |
| NL & Problem Verif. (Fluency) | **-1.0 (-1.0; -.67)** | 1.0 (-1.0; 1.0) | -.08 (-.23; .05) | **.95 (.53; 1.00)** | -.06 (-.16; .27) | .11 (-.06; .30) | **-.44 (-.50; -.37)** |
| L & Problem Verif. (Fluency) | **-.61 (-1.0; -.20)** | 1.0 (-1.0; 1.0) | -.20 (-.34; .04) | **.42 (.13; .57)** | **.03 (.00; .23**) | .**55 (.42; .69)** | **-.32 (-.39; -.25)** |
| NL & Understanding Numbers | **-.78 (-1.0; -.26)** | **-.73 (-1.0; -.42**) | -.03 (-.19; .13) | **.77 (.13; 1.00)** | .19 (-.22; .69) | .04 (-.20; .29) | **-.37 (-.44; -.30)** |
| L & Understanding Numbers | **-.98 (-1.0; -.37)** | 1.0 (-1.0; 1.0) | -.05 (-.23; .11) | **.32 (.06; .57)** | .**16 (.00; .37)** | **.52 (.41; .65)** | **-.32 (-.39; -.23)** |
| NL & Mathematics Composite | **-1.0 (-1.0; -.64)** | 1.0 (-1.0; 1.0) | -.05 (-.20; .09) | **1.0 (.51; 1.00)** | -.06 (-.17; .35) | .04 (-.10; .22)  | **-.45 (-.52; -.39)** |
| L & Mathematics Composite | **-.77 (-1.0; -.38)** | 1.0 (-1.0; 1.0) | **-.17 (-.32; -.02**) | **.51 (.24; .66)** | **.07 (.00; .27)** | **.42 (.33; .54)** | **-.36 (-.43; -.29)** |
| **TEDS**  |  |  |  |  |  |  |  |
| NL & Problem Verif. (Fluency) | **-.84 (-1.0; -.6**7) | 1.0 (-1.0; 1.0) | **-.13 (-.19; -.06)** | **.82 (.66; .99**) | .00 (.00; -.01) | **.18 (.09; .28)** | **-.38 (-.41; -.35)** |
| L & Problem Verif. (Fluency) | N/A | N/A | N/A | N/A | N/A | N/A | **-.06 (-.10; -.01)** |
| NL & Understanding Numbers | **-.76 (-1.0; -.65)** | 1.0 (-1.0; 1.0) | **-.10 (-.13; -.03)** | **.74 (.51; .96)** | **.12 (.00; .29)** | **.14 (.01; .22)** | **-.38 (-.40; -.35)** |
| L & Understanding Numbers | N/A | N/A | N/A | N/A | N/A | N/A | **-.08 (-.13; -.04)** |
| NL & Mathematics Composite | **-.80 (-1.0; -.69)** | 1.0 (-1.0; 1.0) | **-.14 (-.21; -.07)** | **.80 (.59; .96)** | **.04(.00; .20)** | .**16 (.08; .23)** | **-.42 (-.44; -.39)** |
| L & Mathematics Composite | N/A | N/A | N/A | N/A | N/A | N/A | **-.08 (-.12; -.04)** |
| **RSTR**  |  |  |  |  |  |  |  |
| NL & Problem Verif. (Fluency) | N/A | N/A | N/A | N/A | N/A | N/A | **-.43 (-.55; -.30)** |
| L & Problem Verif. (Fluency) | N/A | N/A | N/A | N/A | N/A | N/A | **-.35 (-.48; -.21)** |
| NL & Understanding Numbers | N/A | N/A | N/A | N/A | N/A | N/A | **-.36 (-.49; -.22)** |
| L & Understanding Numbers | N/A | N/A | N/A | N/A | N/A | N/A | **-.30 (-.44; -.15)** |
| NL & Mathematics Composite | N/A | N/A | N/A | N/A | N/A | N/A | **-.45 (-.57; -.32)** |
| L & Mathematics Composite | N/A | N/A | N/A | N/A | N/A | N/A | **-.37 (-.50; -.23)** |
| NL = Number Line estimation (mean error); L = Number Line Lambda; ra, rc, re = genetic, shared and non-shared environmental correlations respectively. They represent the probability that genetic, shared and non-shared environmental factors influencing trait 1 (Number Line, for example) also influence trait 2 (Problem Verification). h2, c2, e2-bivariate = bivariate heritability, bivariate shared and nonshared environmentalities. They index the proportion of phenotypic correlation explained by genetic (A), shared (C) and non-shared (E) environmental factors. ph. corr. = phenotypic correlations Significant estimates are in bold characters. |

|  |
| --- |
| Table S8. Bivariate model fitting between Number Line (NL) and mathematics measures  |
| Measures | Models | -2LL | DF | (Δ-2LL) | AIC | BIC | p-value | ep | Obs. Stat. |
| **WRRMP**  | Saturated | 2465.54 | 947 |  | 571.54 | -2748.01 |  | 28 | 975 |
| NL & Fluency WJ-III | ACE | 2503.204 | 964 | 37.66 | 575.2 | -2803.94 | 0.003 | 11 | 975 |
|  | **AE** | 2520.08 | 967 | 54.54 | 586.08 | -2803.58 | <.001 | 8 | 975 |
| NL & Applied | Saturated | 2455.05 | 953 |  | 549.05 | -2791.53 |  | 28 | 981 |
| Problems WJ-III | **CE** | 2465.76 | 970 | 10.72 | 525.76 | -2874.41 | 0.87 | 11 | 981 |
|  | AE | 2479.4 | 973 | 24.35 | 533.4 | -2877.29 | 0.23 | 8 | 981 |
| NL & Mathematics | Saturated | 2398.49 | 944 |  | 510.49 | -2798.55 |  | 28 | 972 |
| Composite | **CE** | 2419.54 | 961 | 21.05 | 497.54 | -2871.09 | 0.22 | 11 | 972 |
|  | AE | 2446.8 | 964 | 48.31 | 518.8 |  | <.001 | 8 | 972 |
| **QNTS**  | Saturated | 3345.38 | 1261 |  | 823.38 | -3955.75 |  | 28 | 1289 |
| NL & Problem Verif. (Fluency) | ACE | 3364.70 | 1278 | 19.32 | 808.70 | -4034.86 | .31 | 11 | 1289 |
|  | **AE** | **3366.15** | **1281** | **1.45** | **804.15** | **-4050.78** | **.69** | **8** | **1289** |
| NL & Understanding | Saturated | 3315.81 | 1239 |  | 837.81 | -3857.94 |  | 28 | 1267 |
| Numbers | ACE | 3337.36 | 1256 | 21.54 | 825.36 | -3934.83 | .20 | 11 | 1267 |
|  | **AE** | **3338.16** | **1259** | **.80** | **820.16** | **-3951.40** | **.85** | **8** | 1267 |
| NL & Mathematics | Saturated | 16538.99 | 1280 |  | 13978.99 | 9127.84 |  | 28 | 1308 |
| Composite | ACE | 16701.91 | 1297 | 162.92 | 14107.91 | 9192.33 | .00 | 11 | 1308 |
|  | **AE** | 16702.14 | 1300 | .23 | 14102.14 | 9175.18 | .97 | 8 | 1308 |
| **TEDS**  | Saturated | 25808.42 | 9544 |  | 6720.42 | -58257.23 |  | 28 | 9572 |
| NL & Problem Verif. (Fluency) | ACE | 25833.65 | 9561 | 25.24 | 6711.65 | -58381.73 | .09 | 11 | 9572 |
|  | **AE** | **25834.07** | **9564** | **25.65** | **6706.07** | **-58407.74** | **.18** | **8** | **9572** |
| NL & Understanding | Saturated | 25228.89 | 9362 |  | 6504.89 | -57233.66 |  | 28 | 9390 |
| Numbers | ACE | 25253.61 | 9379 | 24.72 | 6495.61 | -57358.68 | .10 | 11 | 9390 |
|  | **AE** | **25255.49** | **9382** | **26.60** | **6491.49** | **-57383.23** | **.15** | **8** | **9390** |
| NL & Mathematics | Saturated | 25665.10 | 9676 |  | 6313.10 | -59563.24 |  | 28 | 9704 |
| Composite | ACE | 25690.60 | 9693 | 25.51 | 6304.60 | -59687.47 | .08 | 11 | 9704 |
|  | **AE** | **25691.92** | **9696** | **26.82** | **6299.92** | **-59712.58** | **.14** | **8** | **9704** |
| -2LL = -2 Log likelihood; df = degrees of freedom; Δ-2LL = differences in likelihood between compared models; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; p-value = significance in change of likelihood between compared models. ep = estimated parameters; Obs. Stat. = Observed statistics. The fit of the nested-alternated models is compared with the fit of the Saturated model. Lower AIC and BIC index the better fit. The best fitting model is marked with bold characters. In WRRMP, the best bivariate fitting model between Number Line and Fluency WJ-III was an AE model, suggesting that only genetic and non-shared environmental factors are important to their covariation. The covariation between Number Line with Applied Problems and the Mathematics Composite is best described as a contribution of genetic, shared and non-shared environmental factors. For UK and Canadian samples the best bivariate model fitting was an AE model for all measures, suggesting that shared environmental factors are not important to the covariation of Number Line and mathematics in these two samples. CE models fitting are not reported as these were not a good it in any model for these 3 samples. The bivariate model fitting for Lambda parameter of the US and Canadian samples are available from the authors. |

**Number Line: Lambda parameter analyses**

We conducted another set of analyses on Number Line scores using Lambda. Lambda is a continuous parameter ranging from 0 to 1 indicating logarithmic nonlinearity. Lambda values of 0 indicate linear responding and Lambda values of 1 indicate logarithmic, less accurate, responding. Younger children are found to have a strong logarithmic component in approximation, so children’s Lambda scores are closer to 1 than older children and adults’ Lambda scores. Lambda scores were computed using model fitting using the log-liner regression model:

$$R=a\left(\left(1-λ\right)N+λ\frac{1000}{ln⁡(1000)}ln⁡(N)\right)$$

where R is the subject’s response; ‘*a*’ is a scaling factor; ‘$λ$’is the Lambda parameter; ‘*N*’ represents the number to be estimated by participants; 1000 is the maximum number in the range to be estimated - the line ranged from 0 to 1000; ‘$ln$’ is the natural logarithm (Cicchini, Anobile, & Burr, 2014). The advantage of Lambda scores is that they account for the shape of individual response, whereas the absolute mean error does not.

The oldest samples, UK and the Older Russian singletons, show an overall linear trend. Lambda parameter in both samples is very close to zero and the correlation with the number line mean error is negligent in the UK twins (.06, Figure S1c) and very modest in the singletons (.15, Figure S2b). In younger samples there is an increasing logarithmic trend with the decrease of age and in the youngest samples we observe the highest correlation between Lambda and the number line mean error: .89 in US twins (Figure S1a) and .77 in the Younger Russian singletons (Figure S2a).

Because of the small proportion of logarithmic nonlinear responding in the UK twins (16% of Lambda > 0), the analyses carried out using Lambda parameters were non-significant in this sample. In US, Canadian and Russian twins, participants with logarithmic trends (Lambda >0) were 73.8%, 30% and 34.5% respectively. It can be observed that in these 3 samples the phenotypic correlation of Lambda with mathematics is very similar to the correlation of the Number Line estimation mean error and mathematics (Table S7). The univariate genetic analyses conducted on US and Canadian samples show heritability estimates of Lambda similar to the Number Line estimation mean error (Tables S4). The multivariate genetic analyses show more reliable estimates in the US than in the Canadian twins due to the stronger cross-trait twin correlations in the former sample. However, in both samples, the genetic correlations and bivariate heritabilities are similar in Lambda and Number Line estimation mean error (Table S7).

Figure S1. Phenotypic correlations between Lambda scores and the absolute mean error scores of Number Line estimation task for the twin samples

a) WRRMP twins  b) QNTS twins  ****

N=246, age range years 8.75-15.33; Mage = 12.27 (1.20) N=333; age range years14.50-16.08; Mage = 15.17 (.29)

**R2 = .89**

**R2 = .65**

c) TEDS twins **** d) RSTR twins ****

N=2555; age range years 15.82-17.28; Mage = 16.55 (.27) N=164; age range years 14.54-18.84; Mage = 16.44 (.91)

**R2 = .06**

**R2 = .42**

Note. In the US-WRRMP, 73.8% of the twins have a non-linear logarithmic trend (Lambda > 0), while only 34.5% of the RU-RSTR twins, 30% of the CA-QNTS twins, and 16% of the UK-TEDS twins have a non-linear logarithmic trend. Mean error is associated with the degree of logarithmic nonlinear responding (Lambda) in early ages (8 year-olds in US-WRRMP), but the association declines with acquisition of linearity stemming from the age increase (15 year-olds in CA-QNTS, 16 years-old in UK-TEDS, and 16.5 years-old in RU-RSTR).

Figure S2. Phenotypic correlations between Lambda scores and the absolute mean error scores of Number Line estimation task

for the Russian singletons

a) Younger Russian Singletons (age < 16 years)  b) Older Russian Singletons (age > 16 years) 

N=164; age range years: 7.51-15.99; Mage = 11.42 (2.59) N=1232; age range years: 16.00-18.85; Mage = 17.14 (.74)



**R2 = 0.77**