From cognitive abilities to educational outcomes

Shakeshaft, Nicholas Graham

Awarding institution:
King's College London

From cognitive abilities to educational outcomes

Nicholas Shakeshaft

**MRC Social, Genetic and Developmental Psychiatry Centre,
King's College London, University of London**

Submitted for the degree of Doctor of Philosophy in Social, Genetic and Developmental Psychiatry Research

September 2016

Abstract

This project explores the genetic and environmental underpinnings of general and specific cognitive abilities, the relationships between them, and their associations with educational outcomes. Using analyses conducted mainly within the Twins Early Development Study (TEDS), it first estimates the substantial genetic influences on outcomes at the end of compulsory education in the UK (General Certificate of Secondary Education grades; GCSEs), then examines the nature and structure of general cognitive ability (*g*) and two specific abilities, and finally uses these as predictors of the phenotypic (i.e., observed) and genetic components of educational achievement.

The specific cognitive domains examined are spatial ability (the mental manipulation of objects) and face recognition. The former has been found to be a strong predictor of educational outcomes, particularly in science, technology, engineering and mathematics (STEM) fields. However, the psychometric structure of spatial ability is highly ambiguous in the literature, reducing the clarity of its measurement and limiting its utility as a predictor; the project therefore seeks to clarify and refine it. Face recognition serves as an invaluable comparison: despite similarly being a visual perceptual ability with many of the same features as spatial ability, it appears to be highly distinct – previous research has found it to be largely unrelated to other abilities. In addition, face recognition is an important social skill; since education in practice is a highly social activity, it is also a useful predictor in its own right.

By clarifying the aetiology of these general and specific abilities, and the associations between them, the project seeks to apply the concepts with greater precision to understanding individual differences in educational outcomes. The main chapters present results indicating that i) GCSE grades are substantially heritable (58%); ii) *g* is aetiologically uniform across its whole distribution, making it suitable as a linear predictor; iii) spatial ability is phenotypically and genetically unifactorial; iv) the dissociation of face recognition from other abilities is driven by its substantial genetic component; and v) these refined measures provide useful prediction of educational outcomes, both phenotypically and genetically: spatial ability strongly predicts STEM achievement, and face recognition (as an index of social skills) is an independent predictor of non-STEM subjects such as English.

Acknowledgements

My PhD was funded in part by a UK Medical Research Council (MRC) studentship to the MRC Social, Genetic and Developmental Psychiatry Centre at the Institute of Psychiatry, Psychology and Neuroscience (IoPPN), King's College London, and in part by the IoPPN itself. The research uses data collected as part of the Twins Early Development Study (TEDS), supported by an MRC programme grant to Professor Robert Plomin (MR/M021475/1; previously G0901245 and G0500079), with additional support from the US National Institute of Health (HD044454; HD059215; NIA046938). Additional data were provided by the Swedish Multi-Generation Register.

I am deeply grateful to my supervisors, Professor Robert Plomin and Professor Francesca Happé, for their guidance, advice and patient support throughout my PhD. I would also like to thank the TEDS staff – particularly Rachel Ogden, Andy McMillan and Louise Webster – for supporting the project over many years, and without whose efforts this complex and vital study could not continue. I am grateful as ever to my friends and colleagues in TEDS and elsewhere (especially those in the Cognitive Fun Group) for their invaluable contributions, stimulation, and thoughtful maintenance of my sanity.

I would like to thank my family for their endless support and encouragement. Finally, as always, I am especially grateful to Kerry, who makes everything possible.

Author declaration

Data from the Swedish Multi-Generation Register were collected prior to the research described here, as were the educational grades and contemporaneous general cognitive ability data used for the Twins Early Development Study (TEDS). The spatial ability measures were designed in collaboration with my supervisors and other colleagues in TEDS. I developed the software used to create the stimuli for the “Bricks” spatial battery (the focus of Chapter 4) and to administer both this battery and the face recognition measures (Chapter 6) to participants, but the software used to implement and administer the “King's Challenge” spatial measures (Chapter 5) was developed in collaboration with an external development company (Helmes: http://www.helmes.ee). To the best of my knowledge, the work presented is my own in all other respects, except as acknowledged in the text.

Nicholas Shakeshaft

Index

Index of tables

Index of figures

Chapter 1:- Introduction

Background

It is virtually impossible to identify any human cognitive or behavioural trait that is not significantly heritable (Turkheimer, 2000; Plomin *et al.*, 2013). This holds true for personality (Jang *et al.*, 1996), general cognitive ability or "*g*" (Petrill and Deater-Deckard, 2004; Haworth *et al.*, 2010), thinking styles (Fletcher *et al.*, 2014), behaviour problems (Viding *et al.*, 2008; Lewis and Plomin, 2015) and every identified psychopathology from depression and anxiety (Waszczuk *et al.*, 2014) to schizophrenia (Sullivan *et al.*, 2003). For specific cognitive abilities, representing performance in particular domains, the apparent uniformity runs even deeper: not only are they all heritable, but they are all correlated with *g*, both phenotypically (the observed correlation) and genetically (Plomin and Spinath, 2002). The latter indicates substantial pleiotropy: the same "generalist" genes underpin much of the heritability of each cognitive domain (Kovas and Plomin, 2006).

g is the strongest single predictor of educational achievement (Krapohl *et al.*, 2014). Since *g* is heritable and seemingly associated strongly with every domain of ability, it is perhaps unsurprising that educational outcomes themselves are also heritable. This is true even in early indicators of literacy and numeracy (Thompson *et al.*, 1991; Kovas *et al*., 2007), and is sustained in middle childhood (Kovas *et al.*, 2013), but little work has been conducted previously to investigate whether this extends to the end of compulsory education, where examination grades heavily influence entry into the workplace or higher education. Even less is known about what exactly the genetic influence represents – intriguingly, the heritability of achievement scores throughout primary school ages is substantially greater than that of *g* at the same ages (Kovas *et al*., 2007; Kovas *et al.*, 2013); this raises the prospect that the genetic influences on education may reflect a combination of numerous, independently heritable factors. Recent findings have borne this out (Krapohl *et al.*, 2014), suggesting that *g* accounts for only half of the heritability of educational outcomes.

The work presented in this thesis extends the previous behavioural genetic research to the end of compulsory education, considers the suitability of *g* as a predictor for such outcomes, and explores part of the genetic influence on educational achievement *not* explained by *g*. For the latter, it focuses on the role of two specific cognitive abilities which are plausible candidates

to explain some of this phenotypic and genetic variation: spatial ability and face recognition.

Spatial ability has been found to be substantially associated with educational outcomes in general (Shea *et al.*, 2001), and especially with success in science, technology, engineering and mathematics (STEM) fields (Wai *et al.*, 2009). Despite being the subject of a large literature spanning many decades, however, identifying an optimal way to measure this ability has proven elusive, due to considerable uncertainty about its nature. Spatial ability is not even clearly or consistently defined (Eliot and Smith, 1983), but broadly speaking, it concerns the visualisation and manipulation of mental representations of objects and the relations between them (Carroll, 1993; Lohman, 1996). The literature proposes diverse skills and subdomains within this ability, such as mental rotation (Shepard and Metzler, 1971), visualisation (Lohman, 1979), spatial orientation (Lohman *et al.*, 1987), spatial scanning and mechanical reasoning (Carroll, 1993), among many others. Inconsistent findings have been presented about the relationships between them, and each subdomain has been subject to varied and shifting definitions among different researchers and at different times.

A huge number of tests have been developed to assess these putative factors of spatial ability, and several theories have been proposed attempting to make sense of the relationships revealed between them (Carroll, 1993; Hegarty and Waller, 2005). These invariably present spatial ability as multifactorial, encompassing numerous subdomains of abilities or processes, but there is a similar lack of consistency between the various structures these theories suggest. Even less is known about the genetic architecture of the spatial domain – it is consistently found to be heritable (Bratko, 1996; Plomin *et al.*, 2013; Tosto *et al.*, 2014), with some evidence for a substantial genetic overlap with *g* (Deary *et al.*, 2006), but no previous research has investigated any potential genetic dissociations among the apparent subdomains *within* spatial ability itself.

There is a considerable lack of clarity, therefore, about the nature, structure and aetiology of spatial ability – in fact, probably the only points of general agreement in the literature are i) that it exists as a cognitive domain partially dissociable from *g*, ii) that it is heritable, iii) that it is multifactorial, and iv) that it is a strong predictor of educational achievement in STEM fields. On the latter point, however, the inconsistencies in the literature inevitably hamper its potential as a predictor, reducing the utility and clarity of any associations found. A major focus of the present work is to develop appropriate measures, and to use behavioural genetic methods, in order to clarify the structure of spatial ability. In this way, it is hoped that the

ability may then be measured more accurately, and that its associations with educational outcomes may be investigated with greater confidence.

A previously unrelated line of research concerns social and emotional intelligence (Petrides, 2011; Mayer *et al.*, 2008b). This domain has also been found to predict educational outcomes (Nowicki and Duke, 1992; Graziano *et al.*, 2007), independently from *g* (Teo *et al.*, 1996). Various theories have been proposed regarding possible mechanisms for these associations – such as differences in social competence influencing the nature and quality of relationships with teachers (Halberstadt and Hall, 1980), or through emotion regulation influencing students' intrinsic motivation (Baumeister *et al.*, 1994). Some studies have suggested that such skills may play a more significant role in non-STEM subjects than in STEM subjects (Petrides *et al.*, 2004), perhaps by virtue of emotional intelligence being more strongly associated with verbal ability than with other cognitive abilities (Mayer *et al.*, 2008a). These dissociations in the academic subjects found to correlate with social abilities are not found consistently, however – and indeed some studies find no substantial associations with social or emotional intelligence at all (Newsome *et al.*, 2000; O'Connor and Little, 2003). As with spatial ability, the inconsistencies observed may indicate that the domain of social and emotional intelligence is not defined or understood in sufficient detail to measure the abilities and their associations reliably.

One social skill that *is* well-defined and highly specific is the ability to recognise human faces. This ability is phenotypically almost entirely dissociated from other cognitive abilities (Ishai, 2008), including the ability to recognise other types of objects (Henke et al., 1998), and is to a large extent distinct even from seemingly closely-related social skills such as interpreting emotions from facial expressions (Fitousi and Wenger, 2013). However, if the literature is correct in suggesting that social competence is predictive of educational achievement, independently from other factors such as *g*, then face recognition – to the extent that it indexes social skills – should be expected to show these relationships. No previous work has investigated this potential association.

Further, face recognition represents an ideal comparison for spatial ability. Despite the many phenotypic findings suggesting that face recognition is strongly dissociated from other cognitive abilities, both spatial ability and face recognition do (in principle) share substantial features in common – both are visual perceptual abilities, in which physical features and the relations between them are used to recognise objects from mental models. Any differences in

the associations between educational outcomes and face recognition, in comparison to those of spatial ability, may thus shed some light on *why* the latter is so predictive of achievements: are the crucial elements intrinsically spatial, or are they shared with other, similarly administered cognitive abilities? Another focus of the present work, therefore, is to compare and contrast the predictive potential of these domains. In order to establish the genetic relationships involved, as well as the phenotypic ones, it is first necessary to establish the genetic architecture between face recognition and other cognitive abilities – face recognition is known to be highly heritable (Wilmer *et al.*, 2010; Zhu *et al.*, 2010), but no multivariate studies have yet tested its genetic associations with other traits.

In summary, the work presented in this thesis aims a) to clarify the structure of spatial ability, b) to establish the degree of phenotypic and genetic dissociation between face recognition and other abilities, and then c) to use both of these specific cognitive abilities to predict educational achievement at the end of compulsory education, both together with and independently from *g*. A more detailed overview is presented below.

Methods

The research described in this thesis is centred around twin and sibling studies, using data from a variety of sources. This section serves as a general overview for context, and the methods and data specific to each study are described in detail within the chapters concerned.

Sibling and twin studies

Sibling studies are commonly used to estimate "familiality" – that is, the extent to which membership of the same family tends to promote resemblance on a trait. There are many statistical approaches of varying complexity, but the logic underlying them is straightforward: if siblings are more alike than would be expected by chance in the general population, the trait is familial. These methods may be informative to some degree about the sources of individual differences in the trait – suggesting the degree to which its development or presentation may be affected by the kinds of influences typically shared between family members – but they cannot determine the *nature* of the influences in question.

Attributing the influences on a trait to more specific aetiology requires a comparison of the degree of resemblance between multiple types or strengths of relationships. One of the most powerful such designs is the twin study, which compares the intrapair similarity of the two naturally-occurring types of twins: monozygotic (MZ) and dizygotic (DZ). MZ twins share all of their genes, while DZ twins share (on average) only half of their segregating alleles, but both (in theory; see below) share their environments to approximately the same extent. If MZs are more alike on a trait than DZs, therefore, this indicates that the familiality of the trait is to some extent genetic in origin, and further, that the phenotypic variance observed in the general population is driven in part by genetic variation. The portion of variance that is *not* attributable to genetic differences is (by definition) environmental; this is defined very broadly to include everything other than genetics, from life experiences and diet to intrauterine hormonal exposures, and may itself be further subdivided into influences "shared" and "non-shared" within families. Formally, the classic statement of these principles (Falconer's formulae) is as follows:

i) $h^2 = 2 * (rMZ – rDZ)$

ii) $c^2 = rMZ - h^2$

iii) $e^2 = 1 - rMZ$

That is: i) the difference in genetic relatedness between MZ twins (who have 100% of their genes in common) and DZ twins (50%) represents half of the maximum possible genetic relatedness, so the "heritability" of the trait (h^2) – i.e., the degree to which phenotypic variance is attributable to genetic variance – is double the difference between the MZ and DZ intrapair correlations (rMZ and rDZ, respectively). ii) Any additional similarity between MZ twins, over and above the heritability of the trait, must be due to environmental influences promoting familial resemblance: the "shared" environmental influences (c^2). iii) The residual variance (e^2) is the degree to which MZ twins are dissimilar, and therefore represents any "non-shared" environmental influences promoting differentiation even between family members, and also any error of measurement in the trait. See Plomin *et al.* (2013) for a detailed discussion of twin and other family methods.

The same principles underlying this univariate analysis, decomposing the variance in a single trait, can be extended to multivariate analyses of the covariance between traits: if twin 1's score on one trait is predictive of twin 2's score on another (a "cross-twin cross-trait" association), the traits are aetiologically related. The extent to which this cross-twin cross-trait

association is stronger for MZ than for DZ twin pairs indicates the degree to which this relationship is attributable specifically to *genetic* influences acting on both traits – in other words, the same (or perfectly correlated) genes influencing both traits. As with the univariate data, the non-genetic portion of covariance between the traits may be decomposed into shared and non-shared environmental influences ("non-shared" in this case indicating influences unique to the individual, but associated with multiple traits). Such analyses allow estimation of the structure of influences underpinning multiple traits (see, e.g., Davis *et al.*, 2009, for analyses of "generalist genes" underpinning multiple cognitive abilities), or of the stability of genetic and environmental influences on a single trait over time (Lyons *et al.*, 2009; Deary *et al.*, 2012).

In modern twin studies, the data are typically subjected to more rigorous model-fitting procedures. These allow point estimates and confidence intervals (CIs) to be derived for the genetic, shared and non-shared environmental components driving the traits and their interrelationships, and also permit the model's goodness of fit to the data to be formally tested. Complex multivariate models may be fitted to the data, allowing the structure of genetic and environmental effects on multiple traits to be explored – for example, estimating what proportion of the genetic influences on an outcome variable are in common with, or unique from, the influences on a specific predictor variable, and what proportion of those common influences can be explained by a third variable, and so on. The model-fitting procedures applicable to each analysis are described in each of the empirical chapters below, and in detail particularly in Chapters 2, 4 and 5. They are applied to continuous data in every empirical chapter, and also to dichotomous data in Chapter 3.

Twin studies have two general limitations of note (see Plomin *et al.*, 2013). The first is generalisability: twins are an unusual subgroup, and it has been suggested on various grounds that they may not be entirely representative of the general population; however any differences appear to be minor, particularly after very early childhood. The second is that the method rests crucially on the "equal environments assumption" (EEA): that MZ and DZ twins do indeed share their environments to approximately the same extent, as the model requires. If this assumption were violated – for example, if MZ twins tended to be treated more similarly by their parents, or to experience more different intrauterine environments than DZs – then heritability estimates would be distorted: either inflated or deflated, depending on the nature of the violation of the assumption. However, the EEA has been tested in several ways and found to be approximately correct – and as *all* models are simplifications, an approximation is

sufficient. More tellingly, other methods with different assumptions (adoption studies, for example) tend to produce very similar estimates of genetic and environmental influence.

Another important assumption underpinning the genetic analyses is linearity of effect – i.e., that the same genetic influences operate continuously across the entire distribution. This would present a substantial problem in the presence of heterogeneity or discontinuity in the aetiology of the measures. With *g*, for example, one such discontinuity has indeed already been identified: severe intellectual disability is driven by different genetic influences (such as rare mutations) from the factors operating across the rest of the distribution (Ellison *et al.*, 2013; Reichenberg *et al.*, 2016) – in other words, mild and severe cognitive impairment are qualitatively distinct disorders in terms of their genetic aetiology. While this discontinuity affects only a very small minority of individuals, it has not previously been established whether there could be similar discontinuities elsewhere in the distribution, such as at the other extreme (i.e., very high intelligence). The distribution of the genetic influences on *g* is therefore a crucial methodological issue, which Chapter 3 examines in order to establish the suitability of *g* for the analyses conducted throughout this work.

In summary, the twin method is very powerful. In the absence of many replicable molecular genetic associations with complex quantitative traits – which remains true, despite recent advances – such methods have the considerable advantage of being able to estimate genetic variance and covariance, even while the specific genes involved mostly remain undiscovered.

Measures

The studies forming this thesis were mainly conducted within the longitudinal Twins Early Development Study (TEDS; Chapters 2 and 4-7), with additional data from the Swedish Multi-Generation Register and Conscription Register (Chapter 3). These samples are described in detail within the chapters concerned, as appropriate to each individual study.

Data about educational outcomes (Chapters 2 and 7) were provided by TEDS participants responding to a postal questionnaire asking them to provide the results from their General Certificate of Secondary Education (GCSE) examinations taken at the end of compulsory education, typically at age 16. IQ data for the Swedish sample (Chapter 3) were acquired as part of military conscription testing, either as pencil-and-paper or computerised tasks,

depending on the year. All other data (Chapters 4-7) were provided online, via websites created specifically for the purpose; the measures specific to each study are described in detail within their own chapters. All websites were hosted on in-house servers dedicated to TEDS web testing. The older measures used (collected prior to the present studies) assessed verbal and non-verbal cognitive abilities, administered when the twins were 16 years old; these measures were developed using the programming languages ASP and Flash, and are described in Chapters 4 and 6.

The creation and implementation of the other cognitive measures, administered after the TEDS twins reached the age of majority, was a major focus of the present work. These measures were developed with HTML, CSS and JavaScript, using *psy.js*: an open-source JavaScript library created specifically for the implementation of web-based tests and questionnaires, and freely available here: https://www.forepsyte.com/resources/public. The server-side software (i.e., the software delivering the websites to participants, processing their responses, and providing researchers with administrative controls and data access) was the *PSY framework*, a system for managing participant logins and secure data handling, written in the Python programming language. The measures are described in detail in Chapters 4-6, and some wider implications of online testing are discussed in Chapter 8.

Overview

This thesis presents work conducted i) to establish the heritability of key educational outcome indicators; ii) to test the suitability of *g* for the genetic analyses conducted; iii) to design and administer appropriate measures of the two specific cognitive domains of primary interest, spatial ability and face recognition, for analysis together with the general cognitive ability measures collected previously; and finally iv) to conduct multivariate genetic analyses with these measures, using the specific cognitive abilities to predict variation in educational outcomes, over and above the variance explained by domain-general cognitive ability (*g*).

Chapter 2 explores the genetic and environmental aetiology of GCSE grades and cross-subject composites. The previous literature is discussed, showing educational achievement to be substantially heritable at earlier ages, and then analyses are presented indicating that the same also holds true at the end of compulsory education. Possible explanations are discussed, both for the heritability being so substantial, and conversely for the finding that shared

environmental influences – which include all of the family and school environments shared between twins – are so modest. The chapter concludes with a discussion of possible implications for educational practice, arguing *inter alia* that a "personalised learning" model, adapted to each individual's strengths, may achieve better results.

Chapter 3 provides an introduction to general cognitive ability (*g*) and its aetiology, and (as noted above) presents analyses intended to establish its suitability for use as a predictor across the whole distribution of ability. The utility of *g* (and cognitive abilities generally) as a predictor for educational achievement rests in part on its normal, continuous distribution: if discontinuity or heterogeneity suggested that subgroups vary in their abilities due to different, discontinuous aetiology – such as different genetic variants operating at the extremes of the distribution – then straightforward associations with educational achievement could not be investigated meaningfully across the whole distribution. In a series of analyses comparing high-*g* individuals with the rest of the population, this chapter concludes that the same genetic and environmental influences operate continuously throughout (excepting severe intellectual disability, as noted above), suggesting that *g* is well suited to the analyses used in later chapters. In addition, a case is presented in support of "positive" genetics, arguing that genetic analyses should not focus exclusively on risk.

One of the key predictors of interest in the present research is spatial ability, which has been suggested to have strong associations with educational outcomes, but its usefulness in this regard is hampered by poor understanding of its structure. In Chapter 4, the literature on spatial ability is discussed, presenting the putative skills of "mental rotation" and "visualisation", with two- and three-dimensional stimuli, as a good starting-point to examine the inconsistent findings outlined above. The chapter argues that a key difficulty is the lack of consistent measures with which the structure of this cognitive domain can be examined with confidence. A novel battery of spatial ability tests, the "Bricks" measures, was created with the express purpose of measuring its constructs and conditions consistently, cleanly and reliably, thus allowing any genuine dissociations to emerge without introducing artifactual ones. Administering this battery to a large twin sample confirmed substantial genetic specificity for spatial ability in general, but identified no meaningful dissociations *within* the domain, either phenotypically or genetically, indicating that the true structure of spatial ability may be simpler than its complex literature suggests.

This approach is extended across the rest of the spatial domain in Chapter 5. Another battery

of tests was created and administered, sampled from across the sprawling literature. A set of 27 spatial tests was assembled, adapting existing measures where possible and creating new ones where necessary, ensuring coverage of the many putative subdomains of this ability. A series of feasibility and pilot studies, eliminating redundant and unreliable tests, reduced this to a battery of ten subtests. A large twin study conducted with this battery confirmed and extended the results in Chapter 4, finding no evidence for phenotypic or genetic dissociations among any of these spatial tests. This has very substantial implications for the literature on spatial ability, and supports the use of a single, general spatial factor as a predictor of educational achievement.

Chapter 6 turns to face recognition, the other specific cognitive ability of interest in the present work. As noted above, this ability shares some superficial features with spatial ability – both are visual skills requiring the perception of spatial relations to recognise mental models of objects, freely rotated – but has been shown to have surprisingly little in common phenotypically with any other cognitive abilities, including spatial ability. This gives it great potential as a comparison to spatial ability: does it differ in its relationship to educational (e.g., STEM) outcomes, and if so, why? While previous research has found face recognition to be highly heritable, no multivariate genetic analyses have ever been conducted to establish whether its dissociation from other abilities has a genetic basis. This chapter presents twin research comparing face recognition with general object recognition and *g*, establishing for the first time that face recognition shares almost no genetic influences in common with either of these measures, despite all three being highly heritable. If face recognition has any significant predictive potential for educational outcomes, either phenotypically or genetically, it is likely to be entirely independent from that of spatial ability.

In Chapter 7, the predictors created and described in Chapters 4-6 are put to work, predicting the key GCSE variables presented in Chapter 2, in a series of multivariate genetic analyses. The “narrow” and “broad” spatial measures (from Chapters 4 and 5, respectively) are compared, together with a “pure” face recognition measure constructed by regressing face recognition on a matched general object recognition measure (both from Chapter 6), thereby controlling for domain-general factors. Using verbal ability as a (deliberately conservative) proxy for *g*, evidence is presented indicating that the substantial association between spatial ability and STEM subjects is largely genetic in origin, as is a modest but significant association found between face recognition (a social ability) and performance in non-STEM subjects. The chapter discusses the implications of these dissociations between spatial ability

and face recognition: possible conclusions regarding the crucial features of spatial ability underpinning its strong prediction of educational outcomes; and the differential importance of social skills for different academic subjects.

Finally, an overall discussion in Chapter 8 summarises the key results and conclusions, before considering potential implications and future directions.

References

Baumeister RF, Heatherton TF, Tice DM (1994). *How and Why People Fail at Self-Regulation.* San Diego, CA: Academic Press.

Bratko D (1996). Twin study of verbal and spatial abilities. *Pers. Individ. Dif.* 21: 621–624.

Carroll JB (1993). *Human cognitive abilities: A survey of factor-analytic studies.* Cambridge: Cambridge University Press.

Davis OSP, Haworth CMA, Plomin, R (2009). Learning abilities and disabilities: Generalist genes in early adolescence. *Cogn. Neuropsychiatry* 14, 312–331.

Deary IJ, Spinath FM, Bates TC (2006). Genetics of intelligence. *Eur. J. Hum. Genet.* 14: 690–700.

Deary IJ, Yang J, Davies G, Harris SE, Tenesa A, Liewald D, Luciano M, Lopez LM, Gow AJ, Corley J, Redmond P, Fox HC, Rowe SJ, Haggarty P, McNeill G, Goddard ME, Porteous DJ, Whalley LJ, Starr JM, Visscher PM. (2012). Genetic contributions to stability and change in intelligence from childhood to old age. *Nature* 482: 212–215.

Eliot J, Smith IM (1983). *An international directory of spatial tests.* Windsor, England: NFER/Nelson; and Atlantic Highlands, NJ: Humanities Press.

Ellison JW, Rosenfeld JA, Shaffer LG (2013). Genetic basis of intellectual disability. *Annu. Rev. Med.* 64: 441–450.

Fitousi D, Wenger MJ (2013). Variants in independence in the perception of facial identity and expression. *J. Exp. Psychol. Hum. Percept. Perform.* 39: 133–155.

Fletcher JM, Marks AD, Hine DW, Coventry WL (2014). Heritability of preferred thinking styles and a genetic link to working memory capacity. *Twin Res. Hum. Genet.* 17(6): 526–534.

Graziano PA, Reavis RD, Keane SP, Calkins SD (2007). The role of emotion regulation in children's early academic success. *J. Sch. Psychol.* 45(1): 3–19.

Halberstadt AG, Hall JA (1980). Who's getting the message? Children's nonverbal skill and their evaluation by teachers. *Dev. Psychol.* 16: 564–573.

Haworth C, Wright MJ, Luciano M, Martin NG, de Geus EJC, van Beijsterveldt CEM, Bartels M, Posthuma D, Moomsma DI, Davis OSP, Kovas Y, Corley RP, DeFries JC, Hewitt JK, Olson RK, Rhea S-A, Wadsworth SJ, Iacono WG, McGue M, Thompson LA, Hart SA, Petrill SA, Lubinski D, Plomin R (2010). The Heritability of General Cognitive Ability Increases Linearly from Childhood to Young Adulthood. *Mol. Psychiatry* 15: 112–120.

Hegarty M, Waller DA (2005). Individual Differences in Spatial Abilities. *The Cambridge Handbook of Visuospatial Thinking* (eds. Shah P, Miyake A), pp 121–169. Cambridge: Cambridge University Press.

Henke K, Schweinberger SR, Grigo A, Klos T, Sommer W (1998). Specificity of Face Recognition: Recognition of Exemplars of Non-Face Objects In Prosopagnosia. *Cortex* 34(2): 289–296.

Ishai A (2008). Let's face it: It's a cortical network. *Neuroimage* 40(2): 415–419.

Jang KL, Livesley WJ, Vernon PA (1996). Heritability of the big five personality dimensions and their facets: a twin study. *J. Pers.* 64: 577–591.

Kovas Y, Haworth CM, Dale PS, Plomin R (2007). The genetic and environmental origins of learning abilities and disabilities in the early school years. *Monogr. Soc. Res. Child Dev.* 72(3)(vii): 1–144.

Kovas Y, Plomin R (2006). Generalist genes: implications for the cognitive sciences. *Trends Cogn. Sci.* 10(5): 198–203.

Kovas Y, Voronin I, Kaydalov A, Malykh SB, Dale PS, Plomin R (2013). Literacy and numeracy are more heritable than intelligence in primary school. *Psychol. Sci.* 24(10): 2048–2056.

Krapohl E, Rimfeld K, Shakeshaft NG, Trzaskowski M, McMillan A, Pingault JB, Asbury K, Harlaar N, Kovas Y, Dale PS, Plomin R (2014). The high heritability of educational achievement reflects many genetically influenced traits, not just intelligence. *PNAS* 111(42): 15273–15278.

Lewis GJ, Plomin R (2015). Heritable influences on behavioural problems from early childhood to mid-adolescence: evidence for genetic stability and innovation. *Psychol. Med.* 45(10): 2171–2179.

Lohman DF (1979). *Spatial Ability: A Review and Reanalysis of the Correlational Literature.* Stanford: School of Education, Stanford University.

Lohman DF (1996). Spatial Ability and *g*. *Human abilities: Their nature and measurement* (eds. Dennis I, Tapsfield P), pp 97–116. New Jersey, USA: Lawrence Erlbaum Associates Inc.

Lohman DF, Pellegrino JW, Alderton DL, Regian JW (1987). Spatial abilities. *Intelligence and Cognition: Contemporary Frames of Reference* (eds. Irvine SH, Newstead SE), pp. 253–312. Dordrecht: Martinus Nijhoff Publishers.

Lyons MJ, York TP, Franz CE, Grant MD, Eaves LJ, Jacobson KC, Schaie KW, Panizzon MS, Boake C, Xian H, Toomey R, Eisen SA, Kremen WS (2009). Genes determine stability and the environment determines change in cognitive ability during 35 years of adulthood. *Psychol. Sci.* 20: 1146–1152.

Mayer JD, Roberts RD, Barsade SG (2008a). Human Abilities: Emotional Intelligence. *Annu. Rev. Psychol.* 59: 507–536.

Mayer JD, Salovey P, Caruso DR (2008b). Emotional intelligence: new ability or eclectic traits? *Am. Psychol.* 63(6): 503-517.

Newsome S, Day AL, Catano VM (2000). Assessing the predictive validity of emotional intelligence. *Pers. Individ. Dif.* 29(6): 1005–1016.

Nowicki S, Duke MP (1992). The association of children's nonverbal decoding abilities with their popularity, locus of control, and academic achievement. *J. Genet. Psychol.* 153(4): 385–393.

O'Connor RM, Little IS (2003). Revisiting the predictive validity of emotional intelligence: self-report versus ability-based measures. *Pers. Individ. Dif.* 35(8): 1893–1902.

Petrides KV (2011). Social intelligence. *Encyclopedia of Adolescence* (eds. Brown BB, Prinstein MJ), pp 342–352. San Diego, CA: Academic Press.

Petrides KV, Frederickson N, Furnham A (2004). The role of trait emotional intelligence in academic performance and deviant behavior at school. *Pers. Individ. Dif.* 36(2): 277–293.

Petrill SA, Deater-Deckard K (2004). The heritability of general cognitive ability: a within-family adoption design. *Intelligence* 32: 403–409.

Plomin R, DeFries JC, Knopik VS, Neiderhiser JM (2013). *Behavioral genetics*. New York: Worth Publishers.

Plomin R, Spinath FM (2002). Genetics and general cognitive ability (*g*). *Trends Cogn. Sci.* 6(4): 169–176.

Reichenberg A, Cederlöf M, McMillan A, Trzaskowski M, Kapara O, Fruchter E, Ginat K, Davidson M, Weiser M, Larsson H, Plomin R, Lichtenstein P (2016). Discontinuity in the genetic and environmental causes of the intellectual disability spectrum. *Proc. Natl. Acad. Sci. USA* 113(4): 1098–1103.

Shea DL, Lubinski D, Benbow CP (2001). Importance of assessing spatial ability in intellectually talented young adolescents: A 20-year longitudinal study. *J. Educ. Psychol.* 93(3): 604–614.

Shepard RN, Metzler J (1971). Mental rotation of three-dimensional objects. *Science* 171: 701–703.

Sullivan PF, Kendler KS, Neale MC (2003). Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Arch. Gen. Psychiatry* 60: 1187–1192.

Teo A, Carlson E, Mathieu PJ, Egeland B, Sroufe LA (1996). A prospective longitudinal study of psychosocial predictors of achievement. *J. Sch. Psychol.* 34(3): 285–306.

Thompson LA, Detterman DK, Plomin R (1991). Associations between cognitive abilities and scholastic achievement: Genetic overlap but environmental differences. *Psychol. Sci.* 2: 158–165.

Tosto MG, Hanscombe KB, Haworth CMA, Davis OSP, Petrill SA, Dale PS, Malykh S, Plomin R, Kovas Y (2014). Why do spatial abilities predict mathematical performance? *Dev. Sci.* 17(3): 462–470.

Turkheimer E (2000). Three Laws of Behavior Genetics and What They Mean. *Curr. Dir. Psychol.* 9: 160–164.

Viding E, Larsson H, Jones AP (2008). Quantitative genetic studies of antisocial behaviour. *Phil. Trans. R. Soc. B* 363(1503): 2519–2527.

Wai J, Lubinski D, Benbow CP (2009). Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *J. Educ. Psychol.* 101(4): 817–835.

Waszczuk MA, Zavos HM, Gregory AM, Eley TC (2014). The phenotypic and genetic structure of depression and anxiety disorder symptoms in childhood, adolescence, and young adulthood. *JAMA Psychiatry* 71(8): 905–16.

Wilmer JB, Germine L, Chabris CF, Chatterjee G, Williams M, Loken E, Nakayama K, Duchaine B (2010). Human face recognition ability is specific and highly heritable. *Proc. Natl. Acad. Sci. USA* 107(11): 5238–5241.

Zhu Q, Song Y, Hu S, Li X, Tian M, Zhen Z, Dong Q, Kanwisher N, Liu J (2010). Heritability of the specific cognitive ability of face perception. *Curr. Biol.* 20(2): 137–142.

Chapter 2:- Strong genetic influence on a UK nationwide test of educational achievement at the end of compulsory education at age 16

This chapter, analysing the aetiology of examination grades at the end of compulsory education, is presented as a published paper. It is an exact copy of this publication:

Shakeshaft NG, Trzaskowski M, McMillan A, Rimfeld K, Krapohl E, Haworth CMA, Dale PS, 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): e80341. doi:10.1371/journal.pone.0080341

Supplementary materials for this chapter, as detailed in the text, are attached as Appendix 1.

OPEN ACCESS Freely available online

PLOS ONE

Strong Genetic Influence on a UK Nationwide Test of Educational Achievement at the End of Compulsory Education at Age 16

Nicholas G. Shakeshaft[1]*, Maciej Trzaskowski[1], Andrew McMillan[1], Kaili Rimfeld[1], Eva Krapohl[1], Claire M. A. Haworth[2], Philip S. Dale[3], Robert Plomin[1]

1 Medical Research Council Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, King's College London, London, United Kingdom, **2** Department of Psychology, University of Warwick, Coventry, United Kingdom, **3** Department of Speech and Hearing Sciences, University of New Mexico, Albuquerque, New Mexico, United States of America

Abstract

We have previously shown that individual differences in educational achievement are highly heritable in the early and middle school years in the UK. The objective of the present study was to investigate whether similarly high heritability is found at the end of compulsory education (age 16) for the UK-wide examination, called the General Certificate of Secondary Education (GCSE). In a national twin sample of 11,117 16-year-olds, heritability was substantial for overall GCSE performance for compulsory core subjects (58%) as well as for each of them individually: English (52%), mathematics (55%) and science (58%). In contrast, the overall effects of shared environment, which includes all family and school influences shared by members of twin pairs growing up in the same family and attending the same school, accounts for about 36% of the variance of mean GCSE scores. The significance of these findings is that individual differences in educational achievement at the end of compulsory education are not primarily an index of the quality of teachers or schools: much more of the variance of GCSE scores can be attributed to genetics than to school or family environment. We suggest a model of education that recognizes the important role of genetics. Rather than a passive model of schooling as instruction (*instruere*, 'to build in'), we propose an active model of education (*educare*, 'to bring out') in which children create their own educational experiences in part on the basis of their genetic propensities, which supports the trend towards personalized learning.

Citation: Shakeshaft NG, Trzaskowski M, McMillan A, Rimfeld K, Krapohl E, et al. (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): e80341. doi:10.1371/journal.pone.0080341

Editor: Daniel Ansari, The University of Western Ontario, Canada

Received July 24, 2013; **Accepted** October 1, 2013; **Published** December 11, 2013

Funding: The Twins Early Development Study (TEDS) is supported by a program grant to RP from the UK Medical Research Council [G0901245; and previously G0500079], with additional support from the US National Institutes of Health [HD044454; HD059215]. NGS, KR, EK, and MT are supported by Medical Research Council studentships. CMAH is supported by a research fellowship from the British Academy. RP is supported by a Medical Research Council Research Professorship award [G19/2] and a European Research Council Advanced Investigator award [295366]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: nicholas.shakeshaft@kcl.ac.uk

Introduction

Children differ in their success in learning what is taught at school – skills such as reading and mathematics, and knowledge such as scientific theories and historical facts. To what extent are these individual differences in educational achievement due to nurture or nature? As academic skills and knowledge are taught at school but are seldom explicitly or systematically taught outside of school, it would be reasonable to assume that differences between students in how much they learn are due to differences in how well the educational system teaches these skills and knowledge. From this perspective, it is surprising that quantitative genetic research such as the twin method, which compares identical and fraternal twins, indicates that individual differences in educational achievement are substantially due to genetic differences (heritability) and only modestly due to differences between schools and other environmental differences [1]. For example, we have recently shown in a UK sample of 7,500 pairs of twins assessed longitudinally at ages 7, 9 and 12 that individual differences in literacy and numeracy are significantly and substantially heritable [2]. Across the three ages, the average heritability of literacy and numeracy was 68%, which means that two-thirds of the individual differences (variance) in children's performance on tests of school achievement can be ascribed to genetic differences – i.e., inherited differences in DNA sequence – between them. Remarkably, educational achievement was found to be more heritable than intelligence (68% versus 42%), even though intelligence is not taught directly in schools and is generally viewed as an aptitude of individuals rather than an outcome of schooling.

Although earlier genetic research on school achievement produced a wide range of estimates of heritability, sampling issues may have masked a more consistent pattern. For example, a classic twin study of school achievement found heritabilities of about 40% for English and mathematics in a study of more than 2000 twin pairs [3]. However, heritability estimates in this study are likely to be underestimates due to restriction of range, because the sample was restricted to the highest-achieving high-school twins in the U.S., those who had been nominated by their schools to compete for the National Merit Scholarship Qualifying Test. The wide range of heritability estimates in three other twin studies of general

educational achievement is likely to be due to their small sample sizes, which were underpowered to provide reliable point estimates of heritability: Petrill et al., 2010 (314 pairs) [4]; Thompson, Detterman, & Plomin, 1991 (278 pairs) [5]; Wainwright, Wright, Luciano, Geffen, & Martin, 2005 (390 pairs) [6].

In addition to the UK study mentioned above which showed high heritability (68%) for literacy and numeracy (Kovas et al., in press; 7,500 pairs) [2], a study of twins in Australia, the US and Scandinavia has reported high heritability (77%) for reading at age 8 (Byrne et al., 2009; 615 pairs) [7] and in the US at age 10 (Olson et al., 2011; 489 pairs) [8]. Similarly high heritability (62%) has been reported for science performance in 9-year old twins (Haworth et al., 2008; 2602 pairs) [9]. A Dutch study of 12-year-old twins reported a heritability of 60% for a national test of educational achievement (Bartels et al., 2002; 691 pairs) [10]. Another study of general educational achievement in 12-year-old twins in the Netherlands (1,178 pairs) and in the UK (3,102 pairs) did not have zygosity information (Calvin et al., 2012) [11]. However, these studies estimated identical and fraternal twin resemblance from the proportion of same-sex and opposite-sex twins, and this procedure yielded heritability estimates of about 60% in the Dutch sample and 65% in the UK sample.

The purpose of the present study was to investigate the extent to which the remarkably high heritabilities for educational achievement in the UK persist to the end of compulsory education. Unlike many countries such as the US, the UK has a nationwide examination for educational achievement, called the General Certificate of Secondary Education (GCSE), which most pupils complete at the end of compulsory education, typically at age 16. The GCSE provides a valuable test of the hypothesis of strong genetic influence on educational achievement because the GCSE is administered nationwide under standardised conditions. Furthermore, the GCSE is important for individuals, for society, and for government because it is used to make decisions about further education.

On the basis of the evidence from earlier school years – most specifically, in our research on educational achievement in the UK at ages 7, 9 and 12 – we tested the hypothesis that the high heritability of educational achievement persists to the end of compulsory education, as assessed by the GCSE at age 16. Additional support for this hypothesis comes from a recent report extending the analysis of the UK dataset described above [11] to total GCSE scores at age 16 [12]. As in the previous report for this dataset, zygosity information was not available, but estimating identical and fraternal resemblance from the proportion of same-sex and opposite-sex twins suggested substantial genetic influence on GCSE scores [12]. Although heritability was not reported because of the absence of zygosity information, the imputed correlations for identical and fraternal twins suggest a heritability of about 60%. However, a definitive estimate of the heritability of educational achievement can only be made on the basis of evidence from twins with known zygosity, which was achieved by the present study.

Materials and Methods

Participants

Twins in the Twins Early Development Study (TEDS) were recruited from birth records of twins born in England and Wales between 1994 and 1996 [13]. Their recruitment and representativeness have been described previously [14]. Children with severe medical problems or whose mothers had severe medical problems during pregnancy were excluded from the analyses. We also excluded children with uncertain or unknown zygosity, and those whose first language was not English. Zygosity was assessed through a parent questionnaire of physical similarity, which has been shown to be over 95% accurate when compared to DNA testing [15]. For cases where zygosity was unclear from this questionnaire, DNA testing was conducted. After exclusions, the total number of individuals for whom GCSE data were obtained at age 16 was 11,117, including 5,474 pairs with data for both co-twins: 2,008 pairs of monozygotic (MZ) twins, 1,730 pairs of same-sex dizygotic (DZ) twins, and 1,736 pairs of opposite-sex DZ twins. Ethical approval was provided by the King's College London ethics committee (reference: 05/Q0706/228), and the parents of the twins provided informed written consent.

Measures

The UK nationwide examination for educational achievement at the end of compulsory education is called the General Certificate of Secondary Education (GCSE). English, mathematics and science (the latter comprising physics, chemistry and biology, and taught either as a single- or double-weighted course, or as separate courses for each science) are compulsory. Many schools also require English literature and one or more modern foreign languages, among other subjects. GCSEs are typically available in a diverse range of other subjects, including history, geography, information and communications technology (ICT), music, and physical education (PE). Courses usually begin at age 14 (with some slight variations by school and subject), with exams typically being taken at age 16. There is no mandatory number of GCSEs, but students commonly take between 8–10 subjects, and receiving five or more at grades A*–C is typically a requirement for going on to further education.

Shortly after the completion of their GCSEs, each TEDS family was sent results forms by mail, (followed as necessary by telephone reminders). The forms were completed by the twins' parents, and also included results for qualifications other than GCSEs (e.g., 'Entry Level Certificates', designed to fall just below GCSE level), which were not analysed in the present study. In order to permit comparable numerical coding across different qualification types, GCSE results were coded from 11 (A*, the highest grade) to 4 (G, the lowest grade). For all analyses, outliers beyond three standard deviations from the mean were removed.

Pupils can select from a wide range of different GCSE subjects, so for many subjects the sample size is too small to analyse. The present study examined the compulsory courses, and several composites generated from the available data for individual subjects. Future papers will examine those individual subjects, including foreign languages, for which sufficient data exist.

Our main general composite was the mean GCSE grade achieved. We also calculated the number of GCSEs passed at grades A*–C, a metric commonly used for university admissions and government policies. These two composites have the advantage of including the results of all GCSE subjects in our dataset, including those taken too rarely to be analysed individually. We also created composites for the compulsory subjects: English mean grade (the mean of all English GCSEs taken; i.e., language and literature, if both were taken), a science mean composite (the mean of whichever science GCSEs were taken), and an overall 'core subjects' mean, which is the mean of the compulsory subjects (when all three were taken): the mathematics GCSE, and the English and science composites. In addition, a 'humanities' composite was generated, which is the mean of the most commonly taken humanities subjects: history, religious education (RE), media studies, music, art and drama (for those participants who took one or more of these courses); subjects such as geography are omitted, whose course content varies and which

are difficult to classify uncontroversially as either humanities or sciences. The composites are detailed further in Table S1 in File S1.

Analysis

The quantitative genetic model apportions phenotypic variance into additive genetic (A), shared or common environmental (C), and non-shared or unique environmental (E) components [16]. Figure 1 illustrates this ACE model in relation to the twin method. Within MZ twin pairs, both genetic and shared environmental effects by definition correlate 1.0, whereas within DZ twin pairs, shared environmental effects correlate 1.0 but additive genetic effects only correlate 0.5. Non-shared environmental influences are assumed to be uncorrelated for members of a twin pair and thus contribute to differences within pairs. The ACE parameters and their confidence intervals can be estimated by fitting the structural equations implied by the model to the raw data, and decomposing the phenotypic variance/covariance matrices using full-information maximum-likelihood estimation model-fitting (accounting for missing data), as described later. As is standard in twin analyses, residuals correcting for age and sex were used because the age of twins is perfectly correlated across pairs, which would otherwise be misrepresented as shared environmental influence [17]. The same applies to the sex of the twins, since MZ twins are always of the same sex.

Separately for the five twin groups (MZ male pairs and female pairs, same-sex DZ male pairs and female pairs, and opposite-sex DZ pairs), we calculated twin intraclass correlations, which index the proportion of total variance due to between-pair variance [18]. Rough ACE estimates can be calculated from these twin correlations. Heritability, the proportion of phenotypic variance ascribed to heritable genetic influences, can be estimated as twice the difference between the MZ and DZ correlations. Shared environmental influence (environmental influences that make siblings more similar to one another) is the residual familial resemblance not explained by heritability, and can be estimated by subtracting the estimate of heritability from the MZ correlation. The variance that remains is ascribed to non-shared environmental influences specific to each twin within a pair, and measurement error.

When twin correlations are compared by sex as well as zygosity, it is possible to assess quantitative and qualitative sex differences in the genetic and environmental aetiology of individual differences in GCSE scores. Quantitative sex differences refer to differences for ACE parameter estimates for male and female twin pairs. Qualitative sex differences indicate that different genes or different environmental factors influence males and females, which is suggested when the correlation for dizygotic opposite-sex (DZO) twins is less than the correlations for same-sex DZ pairs, based on the assumption that genetic or environmental influences that are specific to one sex will reduce within-pair similarity for the DZO group. It should be noted that regressing out the mean effects of sex from GCSE scores has no bearing on these analyses, which are concerned with the aetiology of variance within the sexes and covariance between the sexes, rather than the phenotypic mean difference between the sexes.

To test the observations derived from the intraclass correlations and to derive ACE estimates and confidence intervals, data for each of the five zygosity-sex groups were analysed in a series of models using the structural equation program OpenMx [19]. These models are based on the standard univariate twin model shown in Figure 1 but extended to a so-called sex-limitation model with the inclusion of DZO twin pairs [20]. Within same-sex twin pairs, the correlation between additive genetic influences on Twin 1 and Twin 2 was fixed at 1.0 for MZ and 0.5 for DZ twin pairs. The correlation between shared environmental influences was fixed at 1.0 for both zygosity groups. Within DZO pairs, in contrast, the genetic and shared environmental correlations may be less than the expected values of 0.5 and 1.0, respectively, if there are significant sex-specific genetic or environmental influences.

The full model allows all parameters to vary across sex: the genetic (or shared environmental) correlation in DZO twins; A, C, and E parameters for boys and girls; and variances for boys and girls. Sex-limitation model fitting involves a series of models that are hierarchically related (nested), which makes it possible to test

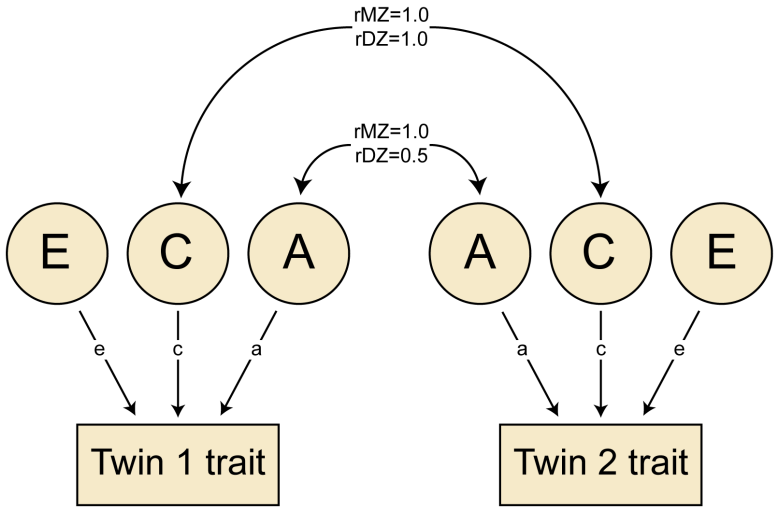


Figure 1. Path diagram representing the basic twin model. A = additive genetic influence; C = shared environmental influence; E = non-shared (unique) environmental influence. Paths a, c and e = effects of A, C and E on the trait. rMZ and rDZ = genetic or shared environmental correlations for monozygotic and dizygotic twins, respectively.
doi:10.1371/journal.pone.0080341.g001

the relative fit of each alternative model using standard chi-squared difference tests with degrees of freedom equal to the difference in degrees of freedom between the two models [20]. As a test of qualitative sex differences, the fit of the full model was compared to a nested model in which either the genetic or shared environmental correlation was fixed at the expected values of 0.5 and 1.0, respectively (common effects model). As it is not possible to estimate the genetic and shared environmental correlations for DZO twins simultaneously, we cannot ascertain whether any qualitative sex differences are genetic or environmental in origin. As a test of quantitative sex differences, a further nested model (called a scalar model) constrained all ACE parameter estimates to be equal for boys and girls, as well as constraining the genetic correlation to 0.5 in DZO twins; this model is called scalar because it allows differences in phenotypic variance between boys and girls [21]. The third nested model, called the null model, tests for variance differences between boys and girls by constraining all parameters including variances to be equal for males and females. AE, CE and E sub-models within the null model were also tested, fixing the missing ACE parameter(s) to zero in each case. More parsimonious models are typically considered preferable unless a significant deterioration in fit is observed, with ACE estimates being derived from the best-fitting sex-limitation model. Greater detail about sex-limitation modelling in TEDS is available [14]. The model-fitting analyses assume equality of shared environmental effects across MZ and DZ twin pairs, the absence of assortative mating, and independence and additivity of the A, C, and E components [16].

Results

Descriptive statistics

Table 1 presents unadjusted raw score means and standard deviations for GCSE scores for the total sample, for all boys and all girls, and for each of the five twin groups. Comparing our results to normative results for GCSE (https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/167426/sfr25-2012.pdf) indicates that our sample is reasonably representative of the UK population: for example, the number of students who receive 5 or more GCSEs with grades of A* to C, an index often used in government policy analyses, is 81.1% nationally and 83.6% in our sample. Mean sex differences can be seen for English, with girls scoring about one-third of a standard deviation higher than boys, and for mathematics, with boys scoring about one-tenth of a standard deviation higher than girls. No significant mean sex differences were found for science. Analysis of variance (ANOVA) was performed on each GCSE score in order to assess the mean effects of sex and zygosity and their interaction. It can be seen from Table 1 that, although significant mean differences emerged for sex and zygosity, they explain less than 3% of the variance. Because GCSE scores are negatively skewed, which is generally interpreted as a ceiling effect, in subsequent analyses, we applied a van der Waerden transformation to all GCSE scores, which normalized the distribution.

We also note that the GCSE scores are for the most part highly correlated: .56 on average, excluding subjects with sample sizes too small to analyse individually. Very high correlations were found between English language and English literature (.80), the science subjects (.83 on average), and the 'core' subjects of English, science and mathematics (.70 on average); the high phenotypic correlations led us to create composite scores for English, science, the three 'core' GCSE subjects (comprising the English and science composites and the mathematics GCSE), and for the overall mean of all subjects. The correlation for the subjects included in the 'humanities' composite (history, religious education (RE), media studies, music, art and drama) is somewhat lower on average (.51), but we argue that the traditional division between 'sciences' and 'humanities' justifies the creation of this composite in order to compare heritability between these areas. Correlation matrices are included in Tables S2 and S3 in File S1, for all subjects with sufficient data, and also for the subset of subjects included in our composites.

For subsequent analyses, the data were age- and sex-regressed as described above.

Twin correlations

Table 2 presents intraclass twin correlations for all MZ and same-sex DZ twins as well as separately for the five twin groups. Looking first at the twin correlations for all MZ and same-sex DZ twins, the GCSE scores yield MZ correlations that are greater than DZ correlations, suggesting genetic influence. The non-overlapping confidence intervals between the MZ and DZ correlations indicate that the differences are significant. Table 2 includes rough estimates of heritability based on doubling the differences between the MZ and DZ correlations. The average heritability estimate is 53% across the GCSE scores and composites, similar to the mean GCSE score heritability estimate of 52%. Shared environmental influence, estimated as the difference between the MZ correlation and heritability, is 29% on average across the GCSE scores and 36% for the mean GCSE score. A remarkable finding is that the estimates of heritability and shared environmental influence do not differ substantially across diverse subjects. The humanities subjects have the lowest estimate (40%), and science subjects the highest (60%).

The twin correlations are suggestive of sex differences. Looking at the intraclass correlations for the five sex and zygosity twin groups, quantitative sex differences are apparent across most subjects, in that heritabilities are somewhat greater for boys than for girls and shared environmental influences are greater for girls than for boys. There is much less evidence for qualitative sex differences (indicated by lower correlations for opposite-sex DZ twins as compared to same-sex DZ twins), but the correlations are suggestive of such effects for some subjects. These questions are addressed more precisely by the model-fitting results below.

Model-fitting results

The results seen in the basic twin correlations can be tested more rigorously using model fitting. For all variables, the comparison between nested sex-limitation models described above indicated the presence of significant quantitative sex differences. No qualitative sex differences of any kind were found for any subject.

The finding of quantitative sex differences would suggest that the full sex-limitation model should be used to derive ACE estimates – i.e., separately for males and females. However, the differences between the heritability estimates for males and females are small (e.g., 57% vs. 47%, respectively, for the overall mean GCSE grade), with overlapping confidence intervals for all our measures (see Table S4 in File S1). Despite being statistically significant, therefore, the quantitative sex differences observed are minor, and would probably not be significant for smaller samples (indeed they are not significant for those individual GCSE subjects with small samples in our data). For this reason, the most informative (and parsimonious) model is the null model, with ACE parameter estimates and variances equated between males and females. The AE, CE and E sub-models all resulted in a significant deterioration in fit when compared with the null model, indicating that all the ACE parameters are required. The full sex-limitation

Table 1. GCSE grade means (standard deviations).

	N	Whole sample	Male	Female	MZm	DZm	MZf	DZf	DZos	Sex	Zyg	Sex×Zyg	R^2
Mean grade for GCSE passes	11011	8.89 (1.14)	8.77 (1.15)	9.00 (1.12)	8.72 (1.16)	8.83 (1.12)	9.00 (1.12)	8.97 (1.14)	8.90 (1.15)	68.85**	1.47	0.35	0.01
Number of GCSE passes at grade A*–C	11117	8.09 (3.16)	7.81 (3.26)	8.34 (3.04)	7.67 (3.29)	7.96 (3.17)	8.38 (2.97)	8.20 (3.11)	8.12 (3.19)	51.28**	0.90	1.13	<.01
GCSE English mean grade	10928	8.93 (1.17)	8.72 (1.19)	9.11 (1.12)	8.66 (1.20)	8.77 (1.17)	9.10 (1.11)	9.07 (1.14)	8.95 (1.18)	166.47**	4.14*	0.46	0.03
GCSE science mean grade	10166	9.03 (1.25)	9.03 (1.24)	9.03 (1.27)	9.02 (1.23)	9.06 (1.22)	9.03 (1.26)	9.01 (1.29)	9.04 (1.26)	1.77	0.01	0.07	<.01
Mathematics	10852	8.96 (1.40)	9.02 (1.39)	8.90 (1.40)	8.95 (1.41)	9.09 (1.35)	8.91 (1.38)	8.87 (1.42)	8.97 (1.41)	4.75*	1.02	0.29	<.01
GCSE core subjects mean grade	10037	9.05 (1.13)	9.00 (1.13)	9.09 (1.13)	8.96 (1.13)	9.03 (1.11)	9.07 (1.13)	9.07 (1.13)	9.07 (1.14)	15.67**	2.38	0.00	<.01
GCSE humanities mean grade	9349	9.03 (1.33)	8.82 (1.39)	9.20 (1.27)	8.76 (1.39)	8.91 (1.35)	9.19 (1.27)	9.18 (1.30)	9.02 (1.33)	106.51**	1.82	2.16	0.02

Scores for composite means and mathematics GCSE have a maximum of 11 and a minimum of 4, representing grades A* to G. N = sample size after exclusions (individuals); MZ = monozygotic; DZ = dizygotic; m = male; f = female; os = opposite sex. ANOVA performed (on cleaned, normality-transformed data from one randomly-selected twin per pair) to test effects of sex and zygosity: results = F statistic; * = $p<.05$; ** = $p<.01$; R^2 = proportion of variance explained by sex, zygosity and their interaction. All variables except for mathematics are composites.

doi:10.1371/journal.pone.0080341.t001

model results are available in Table S4 in File S1, together with a comparison of the nested sub-models (Tables S5–11 in File S1).

The null model results are summarised in Table 3; in each case the best-fitting model was an ACE model that included additive genetic effects (A) and shared environmental effects (C), in addition to residual variance (E) not accounted for by A or C.

These model-fitting results confirm the major conclusions gleaned from the twin correlations. First, heritability is substantial across all GCSE scores. The average heritability is 53%, similar to the heritability of 52% for the mean GCSE score. Second, shared environmental influence is significant for all GCSE scores, but these shared environment estimates are much lower than the heritability. The average shared environment estimate is 30%, and 36% for the mean GCSE score. Third, these estimates do not vary much across most GCSE scores, with heritability estimates for the core subjects all falling into the 52–58% range, and shared environmental variance for these subjects ranging from 24–31%.

One striking finding, closely echoing the estimates derived from the twin correlations in Table 2, is the apparent distinction between the subjects loosely termed as 'sciences' or 'humanities': the science subjects, on average, are the most heritable (58%), and the humanities the least (42%). The non-overlapping confidence intervals for the heritability estimates suggest that this difference is significant.

Discussion

Our results indicate that individual differences in educational achievement are just as strong at the end of compulsory education at age 16 as they are in the earlier school years. Heritability is substantial not only for the core subjects of English (52%), mathematics (55%) and science (58%), but also for the (usually optional) humanities subjects in our dataset (42%). We discuss below the implications of finding that GCSE scores are highly heritable.

Also important is the finding that shared environment accounts for much less variance than does genetics. On average, genetics accounts for almost twice as much of the variance of GCSE scores (53%) as does shared environment (30%), even though shared environmental influences include all family, neighbourhood, and school influences that are shared by members of twin pairs growing up together and attending the same school. In addition, estimates of shared environment are also similar across subjects: English (31%), mathematics (26%), science (24%), and the humanities (32%).

Quantitative sex differences emerged for most subjects, with heritability generally greater for boys and shared environmental influence greater for girls (see Table S4 in File S1). Despite the small effect sizes, it is interesting to speculate about how such a pattern of results could occur; for example, girls might be more susceptible to the shared environmental influences of schools or peers. However, we prefer merely to note these significant sex differences in our sample and to defer speculation about their origins until these results are replicated, for reasons discussed later.

We discuss each of these three topics, acknowledge limitations of our study, and conclude by discussing the policy implications of finding such strong genetic influence and moderate shared environmental influences on educational achievement at the end of compulsory education.

Why is there such strong genetic influence for all GCSE subjects?

It was surprising to us to find such strong genetic influence on educational achievement in the early school years, and now, as seen in the present results, at the end of the compulsory school years as well. The surprise stems from thinking that, as these subjects are taught at school, differences in educational achievement are primarily due to differences in teaching. This thinking is not entirely wrong-headed: differences between schools account for about a third of the variance in educational achievement [22]. However, most of the variance in achievement lies within schools: that is, children within a school differ widely in their performance. Teachers within a school account for some variance, but children in the same classroom also differ widely in their achievement [14].

Table 2. Intraclass twin correlations (with 95% confidence intervals) and approximate variance component estimates for monozygotic (MZ) and same-sex dizygotic (DZss) twins, and separately for MZ and DZ males (m) and females (f), and opposite-sex (os) DZs.

	INTRACLASS CORRELATIONS (95% CONFIDENCE INTERVALS)							VARIANCE COMPONENTS (estimated from MZ vs DZss twin correlations)		
	MZ	DZss	MZm	DZm	MZf	DZf	DZos	h^2	c^2	e^2
Mean grade for GCSE passes	0.88 (0.87–0.89)	0.62 (0.60–0.64)	0.87 (0.85–0.88)	0.57 (0.53–0.62)	0.90 (0.89–0.91)	0.66 (0.63–0.70)	0.55 (0.51–0.58)	0.52	0.36	0.12
Number of GCSE passes at grade A*–C	0.82 (0.81–0.84)	0.57 (0.55–0.60)	0.81 (0.79–0.83)	0.53 (0.48–0.58)	0.83 (0.81–0.85)	0.61 (0.57–0.65)	0.50 (0.47–0.54)	0.50	0.32	0.18
GCSE English mean grade	0.82 (0.80–0.83)	0.56 (0.54–0.58)	0.80 (0.77–0.82)	0.52 (0.47–0.57)	0.83 (0.81–0.85)	0.60 (0.56–0.64)	0.48 (0.44–0.52)	0.52	0.30	0.18
GCSE science mean grade	0.82 (0.80–0.83)	0.52 (0.50–0.55)	0.81 (0.79–0.83)	0.46 (0.40–0.51)	0.83 (0.81–0.85)	0.58 (0.53–0.62)	0.51 (0.47–0.54)	0.60	0.22	0.18
Mathematics	0.82 (0.80–0.83)	0.53 (0.51–0.56)	0.80 (0.78–0.82)	0.50 (0.45–0.55)	0.83 (0.81–0.85)	0.56 (0.51–0.60)	0.45 (0.41–0.49)	0.58	0.24	0.18
GCSE core subjects mean grade	0.87 (0.86–0.88)	0.58 (0.55–0.60)	0.85 (0.83–0.86)	0.53 (0.48–0.58)	0.89 (0.88–0.90)	0.62 (0.58–0.65)	0.53 (0.49–0.56)	0.58	0.29	0.13
GCSE humanities mean grade	0.73 (0.71–0.75)	0.53 (0.50–0.55)	0.71 (0.67–0.74)	0.53 (0.48–0.58)	0.76 (0.73–0.78)	0.52 (0.47–0.57)	0.42 (0.38–0.46)	0.40	0.33	0.27

Variance component estimates are heritability (h^2: double the difference between MZ and DZSS), shared environment (c^2: the MZ correlation minus h^2), and unique environment including error (e^2: 1 - h^2 - c^2). All variables except for mathematics are composites.

doi:10.1371/journal.pone.0080341.t002

Neighbourhoods within a school district account for perhaps 10–15% of the variance, but at least half of this variance can be attributed to differences between families [12].

Differences between families could be due to nature or nurture, but the present results indicate that familial resemblance for educational achievement is primarily due to nature rather than nurture. Paradoxically, individual differences in educational achievement may be highly heritable precisely because these subjects are taught at school. To the extent that children receive the same education, which is the goal of a one-size-fits-all national curriculum, this potential source of environmental differences between children's educational achievement is attenuated. As a result, the individual differences that remain will be due to genetic differences to a greater extent. This line of thinking leads to what may be an uncomfortable realisation: success in achieving widely accepted educational goals such as educational equity, social mobility, and personalised learning will all increase heritability. Indeed, heritability could be viewed as an index of equity in educational opportunities.

For this reason, one might predict that countries with a tightly prescribed national curriculum, such as the UK, might yield higher heritability estimates than countries with decentralized educational systems, such as the US. Although cross-country comparisons of twin results have reported such differences, the studies were too small to provide adequate tests of cross-country differences in heritability [7][23]. One argument against this environmental explanation for the high heritability of educational achievement is that it seems odd, perhaps, that the effect of universal education would emerge full blown in the earliest school years [14]. It also seems odd that the effect does not diminish during the school years as education moves beyond teaching basic skills such as literacy and numeracy. For example, after children learn to read, they read to learn, which might weaken the impact of universal education as children educate themselves to a greater extent; this could be seen as an example of a gene-environment correlation (discussed below), which would have the effect of increasing the heritability estimate beyond the level produced by genes alone.

Another possibility is that educational achievement shows strong genetic influence because it taps into many genetically influenced traits, not just aptitudes of cognition but also appetites of personality and motivation which also have genetic influences. Multivariate genetic analysis, which addresses the genetic and environmental origins of covariance among traits [16], can be used to investigate why educational achievement is so heritable, by identifying the genetic correlates of educational achievement. In other words, multivariate genetic analysis can be used to investigate the extent to which the high heritability of educational achievement is due to the genetic influence of traits such as cognitive abilities, personality, motivation, and adjustment. It can also be used to examine two additional features of the present results: all GCSE scores intercorrelate substantially, 0.56 on average, and all GCSE scores are substantially heritable, 0.53 on average. Although these two findings might suggest that some common genetic mechanisms affect all GCSE scores, it is also possible that each GCSE score could be heritable for different genetic reasons. Multivariate genetic analysis can estimate the extent to which the same genes affect different GCSE scores. Such analyses into genetic correlates of GCSE scores, and genetic intercorrelations among GCSE scores, are the focus of our ongoing analyses, which will be presented in a future paper.

We noted that one possible exception to the finding that all GCSE subjects show strong genetic influence is that subjects loosely termed as 'sciences' are more heritable (58%) than

Table 3. Model fitting results for additive genetic (A), shared environmental (C) and residual (E; i.e., non-shared environment and error) components of variance, with 95% confidence intervals.

	Variance components (95% confidence intervals)			Sample (numbers of pairs)				
	A	C	E	MZm	DZm	MZf	DZf	DZos
Mean grade for GCSE passes	0.52 (0.47–0.58)	0.36 (0.31–0.41)	0.11 (0.11–0.12)	891	820	1108	935	1743
Number of GCSE passes at grade A*–C	0.51 (0.45–0.57)	0.32 (0.26–0.37)	0.17 (0.16–0.19)	898	824	1114	940	1759
GCSE English mean grade	0.52 (0.46–0.58)	0.31 (0.24–0.36)	0.18 (0.17–0.19)	881	812	1104	928	1728
GCSE science mean grade	0.58 (0.52–0.66)	0.24 (0.17–0.30)	0.18 (0.16–0.19)	831	770	1018	865	1598
Mathematics	0.55 (0.49–0.62)	0.26 (0.20–0.32)	0.18 (0.17–0.20)	879	799	1085	928	1719
GCSE core subjects mean grade	0.58 (0.52–0.64)	0.29 (0.23–0.35)	0.13 (0.12–0.14)	819	753	1007	856	1573
GCSE humanities mean grade	0.42 (0.35–0.51)	0.32 (0.24–0.39)	0.26 (0.24–0.28)	715	670	974	811	1492

Numbers of pairs are shown for male (m), female (f) and opposite sex (os) monozygotic (MZ) and dizygotic (DZ) twins; figures include incomplete pairs (i.e., those with missing data for one twin). All variables except for mathematics are composites.

doi:10.1371/journal.pone.0080341.t003

'humanities' (42%). This finding is interesting because it is contrary to the 'folk psychology' view that science is something you learn from teaching (i.e., environment) but abilities in the humanities are 'gifts' (i.e., genetics). Multivariate genetic analyses might help to explain this heritability difference if different patterns of genetic correlates are found for sciences and humanities.

Why is shared environmental influence so modest for all GCSE subjects?

Just as important as the finding of high heritability is the finding that shared (as opposed to non-shared) environmental influence accounts for 30% of the variance of GCSE scores on average, compared to the 53% accounted for by genetics. On the one hand, it is interesting that so much of the variance is due to shared environment because it often has negligible influence on behavioural traits [24]. This estimate of 30% of the variance of GCSE scores being due to shared environment is greater than what we have found at earlier ages, where the average estimate of shared environmental influence for National Curriculum scores for literacy and numeracy across ages 7, 9 and 10 is 12% [14]. It would be interesting if this jump in shared environmental influence at the end of secondary school proved to be replicable, as it would suggest that secondary schools have more of an impact than primary schools. We are currently obtaining data on school quality to test the hypothesis that the quality of secondary schools mediates this effect.

On the other hand, it is remarkable that only 30% of the variance is due to shared environment for GCSE scores because familial resemblance is indexed in our study using siblings who have grown up in the same family, lived in the same neighbourhood, attended the same school, and perhaps even studied and revised together during their education. In comparison, resemblance between parents and offspring is more limited environmentally because parents and offspring grow up at least two decades apart, and in different homes; their resemblance is also limited genetically because different genes can affect adults (parents) and children (offspring). Moreover, the siblings in our study are twins, which means that they also lived together prenatally in the same womb and grew up together at exactly the same age. In other words, twin siblings maximally share their environments, and yet our results indicate that their resemblance owes substantially more to genetics than to shared environment.

It should be mentioned that even this modest estimate of shared environmental influence might be inflated. Twins have been reported to have twin-specific shared environmental effects – that is, environmental effects that are shared by twins but not by other siblings – such as the extra resemblance that might be derived from growing up together at exactly the same age [25]. Data from the recent sibling study of GCSE scores [12] appear to provide at most modest support for this hypothesis, because correlations for DZ twins are only slightly greater than correlations for non-twin siblings: the GCSE correlation for DZ brothers was 0.62, as compared to 0.59 for non-twin brothers; for DZ sisters and non-twin sisters, the correlations were 0.64 and 0.62, respectively. However, the study did not assess zygosity, so the same-sex DZ correlations may not be accurate.

It should also be noted that the term 'shared environment' is shorthand for 'shared environmental effects', not 'shared environmental events'. That is, twins manifestly share environmental events such as the same parents, the same home, and the same school. However, quantitative genetic analyses such as the twin method address the genetic and environmental sources of individual differences, that is, genetic and environmental factors that make a difference. In the case of shared environment, this refers to the influence of environmental factors that contribute to the covariance of siblings after controlling for the genetic contribution to their covariance. In other words, shared environments such as shared families and schools might not have shared environmental effects.

Does finding only modest shared environmental influence mean that schools do not matter? Of course not: schools systematically teach children basic skills such as reading, writing and arithmetic, and basic cultural knowledge. Although the difference in educational achievement between the best schools and the worst schools might not be great compared to the wide range of individual differences within schools, the difference between going to school and not going to school would be enormous. Moreover, shared environmental influence refers to only one specific type of environmental influence: for example, the extent to which children attending the same school are similar in their educational achievement after controlling for genetic influence. Controlling for genetic influence is important: differences between schools cannot be safely assumed to be entirely environmental in origin,

because families are not assigned randomly to schools. Genetic factors are likely to contribute to this non-random assortment of children to schools – including the parents' own educational achievement, as discussed later.

Some of the clearest evidence for the impact of schools on intelligence and cognitive development comes from studies which have used the school cut-off method [26]. Children who have just missed the cut-off date for entering school are compared at later times with those who just made the cut-off. The groups are nearly identical in age and many other characteristics, but differ by one year's schooling. Not only does the additional year of schooling have a significant effect on IQ and a range of cognitive tasks, a year of schooling generally has at least twice as much of an effect as does a year of additional age without an additional year of schooling. Thus schooling has a very substantial mean impact, but – based on studies such as the present one – relatively little impact on the relative differences between children.

Environmental effects that are not shared by family members are called non-shared environmental influences [24]. While non-shared environment accounts for only a modest proportion of variance in our sample (very modest, considering that measurement error is included in this estimate), it is still significant. One direction for research is to attempt to identify these non-shared environmental influences on educational achievement. What environmental factors could be responsible for making children in the same classroom in the same school differ so much in their educational achievement? For example, are teachers differentially effective in teaching some children more than others? The difficulty in investigating non-shared environmental influences is to disentangle them from genetic influences. That is, teachers might respond differently to some children on the basis of the children's genetically driven differences. Identical twins are a powerful tool for studying non-shared environment while controlling for genetics. Since members of identical twin pairs are identical in terms of inherited DNA sequences, differences within pairs of identical twins can only be due to non-shared environmental influences. Nonetheless, in general it has proven difficult to identify specific factors that account for non-shared environment [24]. However, some positive results were found in a study of non-shared classroom experiences of MZ twins who were in the same classrooms and were assessed every school day for two weeks. MZ twins experienced their teachers, classrooms, and peers somewhat differently, and these experiential differences within MZ twin pairs were significantly associated with differences in educational achievement, especially in mathematics and science [27]. In relation to our finding that science subjects may be more heritable than humanities subjects, it is interesting that we find *less* non-shared environmental influence for sciences than humanities. Since estimates of non-shared environmental effects include measurement error, one possibility is that humanities are less reliably measured than sciences.

Sex differences?

When examining the phenotypic variance difference between sexes, we found that individual differences within sex are far greater than average differences between boys and girls. An important point is that the description and causes of individual differences are not necessarily related to the description and causes of average differences between groups. That is, regardless of whether there are mean sex differences, sex differences at the level of individual differences can still be found. Genetic analyses focus on the origins of individual differences for boys and girls, not mean differences. Therefore, the mean differences were regressed prior to model fitting analyses.

For several of our measures, we found significant quantitative (but no qualitative) sex differences: greater heritability for boys, and greater shared environment for girls. However, these differences were small for all measures, with overlapping confidence intervals (Table S4 in File S1). Moreover, we had not anticipated these findings because our research on the same sample in the earlier school years did not find significant quantitative sex differences. For example, at ages 7, 9 and 10, we found similar estimates of heritability and shared environment for boys and girls [14]. Indeed, when quantitative sex differences were found, they were in the opposite direction from those in the present study: heritability was slightly lower for boys, and shared environment slightly lower for girls. It is noteworthy that our finding of quantitative sex differences cannot be tested by comparing correlations for non-twin siblings, because sibling studies cannot separate genetic and environmental influences. If heritability is greater for boys and shared environment is greater for girls, these quantitative sex differences would be counterbalanced; that is, heritability would contribute to a higher correlation for brothers and shared environment would contribute to a higher correlation for sisters. Although our results suggest that the magnitude of these counterbalancing effects is similar – heritability is about 10% greater for boys and shared environment is about 10% greater for girls – in fact, we find a slightly lower average correlation for DZ boys (0.52) than for DZ girls (0.59). In this context, it is noteworthy that in the recent paper on GCSE scores mentioned in the Introduction [12], correlations for non-twin siblings were in a similar direction although the difference was even smaller: 0.59 for brothers and 0.62 for sisters.

For these reasons, although there is some support in the literature for our findings of quantitative sex differences, we suggest caution in accepting and interpreting these results until they are replicated in independent studies.

Limitations

Limitations of the present study include general limitations of the twin method, most notably the equal environments assumption – that environmentally-caused similarity is equal for MZ and DZ twins – and the assumption that results for twins generalize to non-twin populations [16]. The equal environments assumption has survived several tests of its validity, but the most persuasive evidence is that similar results are found using two other methods with different assumptions: the adoption method and a quantitative genetic method based on DNA alone [28][29]. In terms of the generalization from twin to non-twin samples, GCSE scores for twins and non-twin siblings have been shown to be very similar in means and variances [12].

Specific limitations involve aspects of the sample and measures. As mentioned earlier, although our sample was relatively large, the sex differences that emerged from our sex-limitation model fitting were so small that caution is warranted in interpreting these results until they are replicated in other studies. In terms of the measure, although the GCSE may not be the best or most thorough test of educational achievement, it is important because it is a nationwide test that is used to make decisions about further education and employment. Moreover, our results for the GCSE at age 16 are comparable to those we obtained using web-based tests of reading and mathematics at age 12 [30]. Our sample tended to score more highly than the national average, and our dataset does not contain information about failed exams (i.e., below grade G), but these account for only around 1.5% of exams nationally (https://docs.google.com/spreadsheet/pub?key = 0AoEZjwuqFS2PdEZfSVpFd0UwdExROXlQbHR4d2laUHc). A possible specific limitation of our study is that GCSE scores were reported by parents. However,

for 7,367 of the twins, we were able to obtain official GCSE scores from the UK National Pupil Database (http://www.education.gov.uk/researchandstatistics/national-pupil-database); the correlations between parent-reported scores and official scores were 0.98 for English, 0.99 for mathematics, and >0.95 for all science subjects, so obtaining GCSE results from parents was not problematic. Another limitation is that the present analyses are univariate; as mentioned earlier, multivariate genetic analyses are in progress that address the genetic and environmental origins of the phenotypic correlations among GCSE subjects, and those between GCSE scores and other traits.

A genetic model of education

Education has been slow to take on board the importance of genetics for educational achievement [31][32][33]. Some of this reluctance comes from general misconceptions of what it means to say that genetics influences educational achievement. One major misconception is that finding genetic influence diminishes the importance of schools: even if the heritability of educational achievement were 100%, this means that the differences in achievement between pupils are due to genetic differences between them but it would not mean that schools are unimportant. As noted earlier in relation to the modest impact of schools on shared environmental influence, the differential impact of good and bad schools is not great, but the difference between schools and no schools is likely to be enormous. Without educational curricula, whether taught in schools or homes, children would not systematically learn basic skills such as literacy and numeracy or basic knowledge such as history and science. In addition, there is a more subtle way in which schools could be important even if heritability were 100%: heritability of 100% means that inequalities of educational opportunity do not exist. In this counter-intuitive sense, heritability can be considered as an index of equality.

Rather than a universal, one-size-fits-all approach to educational curricula, a more individually tailored approach is needed that recognizes the strong genetic contribution to individual differences in educational achievement. Education is not imposed on a passive organism. When a universal educational curriculum is imposed on children, children differ in their response to it, in large part for genetic reasons [1]. In quantitative genetics, this process is known as genotype-environment interaction, in which the effect of an imposed environment differs as a function of individuals' genetic propensities. However, a farther-reaching view of the interface between the environment and genes is genotype-environment correlation, which denotes genetic influence on exposure to environments. Genotype-environment correlation involves choice of environments rather than the imposition of an environment: children select, modify and create environments in part for genetic reasons [34]. There are three types of genotype-environment correlation: passive, evocative, and active. The passive type occurs because children passively receive environments correlated with their genotypes when they are reared by their genetic parents. For example, parents whose genetic propensities lead them to read more are also likely to read more to their children. Evocative genotype-environment correlation occurs when children, on the basis of their genetic propensities, evoke reactions from other people, such as teachers noticing a child who loves to read and then encouraging that propensity. Active genotype-environment correlation occurs when children select, modify, and construct or re-construct experiences that are correlated with their genetic propensities. For example, children who like to read can cultivate their own reading in the library, on the internet, and via friends.

The passive type of genotype-environment correlation is one reason why it is unsafe to assume that correlations between family background and educational achievement are mediated environmentally. The evocative type occurs to the extent that parents and teachers recognize and foster genetically driven aptitudes and appetites among children. Active genotype-environment correlation has the broadest ramifications for education because it suggests an active model of education in which children actively select, modify and create their own environments, even within an ostensibly 'universal' curriculum. Using reading again as an example, children with reading problems will benefit from increased reading instruction but because reading is difficult they are less likely to be motivated to read on their own.

Active genotype-environment correlation may be the most general process by which genotypes develop into phenotypes, in education as well as other developmental domains. The distinction between the prevailing passive model of imposed environments and this active model of education can be captured by the contrast between the word 'instruction', which is derived from the Latin word *instruere* meaning 'to build in', and the word 'education', which is derived from *educare* meaning 'to bring out'. The instruction model of imposed environments is consistent with a one-size-fits-all national curriculum approach, whereas the education model of active experiences fits the trend towards adaptive learning systems tailored to each pupil [35]. For example, there is increasing evidence that individualized reading instruction is more effective than instruction of similar quality that is not individualized [36]. Genetics will become more specifically useful in such personalized learning programs as specific genes responsible for the high heritability of educational achievement are identified, and the dynamic interplay of genetic and environmental factors, e.g., genotype-environment correlation, is better understood. However, as is the case for complex traits in all of the life sciences, progress has been slow in identifying genes responsible for heritability [37].

In closing, we note that accepting the evidence for strong genetic influence on individual differences in educational achievement has no necessary implications for educational policy, because policy depends on values as well as knowledge. For example, a deep-seated fear is that accepting the importance of genetics justifies inequities – educating the best and forgetting the rest. However, depending on one's values, the opposite position could be taken, such as putting more educational resources into the lower end of the distribution to guarantee that all children reach minimal standards of literacy and numeracy, so that they are not excluded from our increasingly technological societies. It is to be hoped that better policy decisions will be made with knowledge than without. Part of that knowledge is the strong genetic contribution to individual differences in educational achievement.

Supporting Information

File S1 Supporting information tables. Table S1. Construction of composites. **Table S2.** Correlation matrix for all GCSE subjects. **Table S3.** Correlation matrix for subjects included in composites. **Table S4.** Sex limitation A, C and E estimates. **Table S5.** Sex limitation sub-model comparisons: Mean grade for GCSE passes. **Table S6.** Sex limitation sub-model comparisons: Number of GCSE passes at grade A*–C. **Table S7.** Sex limitation sub-model comparisons: GCSE English mean grade. **Table S8.** Sex limitation sub-model comparisons: GCSE science mean grade. **Table S9.** Sex limitation sub-model comparisons: Mathematics. **Table S10.** Sex limitation sub-model comparisons: GCSE core subjects mean grade. **Table S11.** Sex limitation sub-model comparisons: GCSE humanities mean grade. (PDF)

Acknowledgments

We gratefully acknowledge the ongoing contribution of the twins and their families in the Twins Early Development Study (TEDS).

Author Contributions

Conceived and designed the experiments: NGS AM RP. Analyzed the data: NGS MT AM KR EK. Contributed reagents/materials/analysis tools: MT AM. Wrote the paper: NGS MT AM KR EK CMAH PSD RP.

References

1. Haworth CMA, Asbury K, Dale PS, Plomin R (2011) Added value measures in education show genetic as well as environmental influence. PLoS ONE 6: e16006. doi: 10.1371/journal.pone.0016006.t004.
2. Kovas Y, Voronin I, Kaydalov A, Malykh SB, Dale PS, et al. (2013) Literacy and numeracy are more heritable than intelligence in primary school. Psychol Sci. doi: 10.1177/0956797613486982.
3. Loehlin JC, Nichols J (1976) Heredity, environment and personality. Austin: University of Texas.
4. Petrill SA, Hart SA, Harlaar N, Logan J, Justice LM, et al. (2010) Genetic and environmental influences on the growth of early reading skills. J Child Psychol Psychiatry 51: 660–667. doi: 10.1111/j.1469-7610.2009.02204.x.
5. Thompson LA, Detterman DK, Plomin R (1991) Associations between cognitive abilities and scholastic achievement: Genetic overlap but environmental differences. Psychol Sci 2: 158–165. doi: 10.1111/j.1467-9280.1991.tb00124.x.
6. Wainwright MA, Wright MJ, Luciano M, Geffen GM, Martin NG (2005) Multivariate genetic analysis of academic skills of the Queensland core skills test and IQ highlight the importance of genetic g. Twin Res Hum Genet 8: 602–608. doi: 10.1375/twin.8.6.602.
7. Byrne B, Coventry WL, Olson RK, Samuelsson S, Corley R, et al. (2009) Genetic and environmental influences on aspects of literacy and language in early childhood: Continuity and change from preschool to Grade 2. J Neurolinguistics 22: 219–236. doi: 10.1016/j.jneuroling.2008.09.003.
8. Olson RK, Keenan JM, Byrne B, Samuelsson S, Coventry WL, et al. (2011) Genetic and environmental influences on vocabulary and reading development. Sci Stud Read 15: 26–46. doi: 10.1080/10888438.2011.536128.
9. Haworth CMA, Dale PS, Plomin R (2008) A twin study into the genetic and environmental influences on academic performance in science in nine-year-old boys and girls. Int J Sci Educ 30: 1003–1025. doi: 10.1080/0950069070 1324190.
10. Bartels M, Rietveld MJ, van Baal GC, Boomsma DI (2002) Heritability of educational achievement in 12-year-olds and the overlap with cognitive ability. Twin Res 5: 544–553. doi: 10.1375/136905202762342017.
11. Calvin CM, Deary IJ, Webbink D, Smith P, Fernandes C, et al. (2012) Multivariate genetic analyses of cognition and academic achievement from two population samples of 174,000 and 166,000 school children. Behav Genet 42: 699–710. doi: 10.1007/s10519-012-9549-7.
12. Nicoletti C, Rabe B (2013) Inequality in pupils' test scores: How much do family, sibling type and neighbourhood matter? Economica 80: 197–218. doi: 10.1111/ecca.12010.
13. Haworth CMA, Davis OSP, Plomin R (2013) Twins Early Development Study (TEDS): A genetically sensitive investigation of cognitive and behavioral development from childhood to young adulthood. Twin Res Hum Genet 16: 117–125. doi: 10.1017/thg.2012.91.
14. Kovas Y, Haworth CMA, Dale PS, Plomin R (2007) The genetic and environmental origins of learning abilities and disabilities in the early school years. Monogr Soc Res Child Dev 72: 1–144. doi: 10.1111/j.1540-5834.2007.00453.x.
15. Price TS, Freeman B, Craig IW, Petrill SA, Ebersole L, et al. (2000) Infant zygosity can be assigned by parental report questionnaire data. Twin Res 3: 129–133. doi: 10.1375/136905200320565391.
16. Plomin R, DeFries JC, Knopik VS, Neiderhiser JM (2013) Behavioral genetics. New York: Worth Publishers.
17. McGue M, Bouchard TJ Jr (1984) Adjustment of twin data for the effects of age and sex. Behav Genet 14: 325–343. doi: 10.1007/BF01080045.
18. Shrout PE, Fleiss J (1979) Intraclass correlations: Uses in assessing rater reliability. Psychol Bull 86: 420–428. doi: 10.1037/0033-2909.86.2.420.
19. Boker S, Neale M, Maes H, Wilde M, Spiegel M, et al. (2011) OpenMx: An open source extended structural equation modeling framework. Psychometrika 76: 306–317. doi: 10.1007/s11336-010-9200-6.
20. Neale MC, Maes HHM (1999) Methodology for genetic studies of twins and families. Dordrecht, Netherlands: Kluwer.
21. Medland SE (2004) Alternate parameterization for scalar and non-scalar sex-limitation models in Mx. Twin Res 7: 299–305. doi: 10.1375/136905204774200587.
22. OECD (2004) Learning for tomorrow's world: First results from PISA 2003. Paris: OECD.
23. Samuelsson S, Byrne B, Olson RK, Hulslander J, Wadsworth S, et al. (2008) Response to early literacy instruction in the United States, Australia, and Scandinavia: A behavioral-genetic analysis. Learn Individ Differ 18: 289–295. doi: 10.1016/j.lindif.2008.03.004.
24. Plomin R (2011) Commentary: Why are children in the same family so different? Non-shared environment three decades later. Int J Epidemiol 40: 582–592. doi: 10.1093/ije/dyq144.
25. Koeppen-Schomerus G, Spinath FM, Plomin R (2003) Twins and non-twin siblings: Different estimates of shared environmental influence in early childhood. Twin Res 6: 97–105. doi: 10.1375/136905203321536227.
26. Morrison FJ, Griffith EM, Frazier JA (1996) Schooling and the 5–7 shift: A natural experiment. In: Sameroff A, Naith MN, editors. Reason and responsibility: The passage through childhood. Chicago: University of Chicago Press. pp. 161–186.
27. Asbury K, Almeida D, Hibel J, Harlaar N, Plomin R (2008) Clones in the classroom: A daily diary study of the nonshared environmental relationship between monozygotic twin differences in school experience and achievement. Twin Res Hum Genet 11: 586–595. doi: 10.1375/twin.11.6.586.
28. Plomin R, Haworth CMA, Meaburn EL, Price T, Wellcome Trust Case Control Consortium, et al. (2013) Common DNA markers can account for more than half of the genetic influence on cognitive abilities. Psychol Sci 24: 562–568. doi: 10.1177/0956797612457952.
29. Trzaskowski M, Davis OP, DeFries J, Yang J, Visscher P, et al. (2013) DNA evidence for strong genome-wide pleiotropy of cognitive and learning abilities. Behav Genet 43: 267–273. doi: 10.1007/s10519-013-9594-x.
30. Davis OSP, Haworth CMA, Plomin R (2009) Learning abilities and disabilities: Generalist genes in early adolescence. Cogn Neuropsychiatry 14: 312–331. doi: 10.1080/13546800902797106.
31. Haworth CMA, Plomin R (2011) Genetics and education: Towards a genetically sensitive classroom. In Harris KR, Graham S, Urdan T, editors. American Psychological Association Handbook of Educational Psychology. Washington, DC: APA. pp. 529–559.
32. Wooldridge A (1994) Measuring the Mind: Education and psychology in England, c. 1860–c. 1990. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511659997.
33. Asbury K, Plomin R (2013) G is for genes: The impact of genetics on education and achievement. London: Wiley-Blackwell.
34. Hanscombe KB, Haworth CMA, Davis OSP, Jaffee SR, Plomin R (2011) Chaotic homes and school achievement: a twin study. J Child Psychol Psychiatry 52: 1212–1220. doi: 10.1111/j.1469-7610.2011.02421.x.
35. Tseng JCR, Chu H-C, Hwang G-J, Tsai C-C (2008) Development of an adaptive learning system with two sources of personalization information. Comput Educ 51: 776–786. doi: 10.1016/j.compedu.2007.08.002.
36. Connor CM, Morrison FJ, Fishman B, Crowe EC, Al Otaiba S, et al. (2013) A longitudinal cluster-randomized controlled study on the accumulating effects of individualized literacy instruction on students' reading from first through third grade. Psychol Sci. Advanced online publication. doi: 10.1177/0956797612472204.
37. Plomin R, Simpson MA (2013) The future of genomics for developmentalists. Dev Psychopathol. In press.

Chapter 3:- Thinking positively: The genetics of high intelligence

This chapter, examining the homogeneity of genetic influences on general cognitive ability, is presented as a published paper. It is an exact copy of this publication:

Shakeshaft NG, Trzaskowski M, McMillan A, Krapohl E, Simpson MA, Reichenberg A, Cederlöf M, Larsson H, Lichtenstein P, Plomin R (2015). Thinking positively: The genetics of high intelligence. *Intelligence* 48: 123–132. doi:10.1016/j.intell.2014.11.005

Intelligence 48 (2015) 123–132

Contents lists available at ScienceDirect

Intelligence

Thinking positively: The genetics of high intelligence

Nicholas G. Shakeshaft [a], Maciej Trzaskowski [a], Andrew McMillan [a], Eva Krapohl [a], Michael A. Simpson [a], Avi Reichenberg [a,b], Martin Cederlöf [c], Henrik Larsson [c], Paul Lichtenstein [c], Robert Plomin [a,*]

[a] King's College London, MRC Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology & Neuroscience, London, SE5 8AF, United Kingdom
[b] Department of Psychiatry, Mount Sinai School of Medicine, NY, 10029, USA
[c] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Box 281, 17177 Stockholm, Sweden

ARTICLE INFO

Article history:
Received 3 June 2014
Received in revised form 24 October 2014
Accepted 11 November 2014
Available online xxxx

Keywords:
Intelligence
Human genetics
Twins
Siblings
Positive genetics

ABSTRACT

High intelligence (general cognitive ability) is fundamental to the human capital that drives societies in the information age. Understanding the origins of this intellectual capital is important for government policy, for neuroscience, and for genetics. For genetics, a key question is whether the genetic causes of high intelligence are qualitatively or quantitatively different from the normal distribution of intelligence. We report results from a sibling and twin study of high intelligence and its links with the normal distribution. We identified 360,000 sibling pairs and 9000 twin pairs from 3 million 18-year-old males with cognitive assessments administered as part of conscription to military service in Sweden between 1968 and 2010. We found that high intelligence is familial, heritable, and caused by the same genetic and environmental factors responsible for the normal distribution of intelligence. High intelligence is a good candidate for "positive genetics" — going beyond the negative effects of DNA sequence variation on disease and disorders to consider the positive end of the distribution of genetic effects.

1. Introduction

High intelligence is precious human capital for advancing and maintaining society in the information age, as documented in studies that demonstrate that high intelligence is responsible for exceptional performance in many societally-valued outcomes (Kell, Lubinski, & Benbow, 2013; Lubinski, Benbow, Webb, & Bleske-Rechek, 2006; Rindermann & Thompson, 2011). Understanding the genetic and environmental origins of high intelligence is crucial for government policy (for example, for education in the STEM subjects of science, technology, engineering and mathematics), for neuroscience (for investigating the high-performance brain), and for genetics. A key question for genetic research is the extent to which the aetiology of high intelligence differs from the aetiology of the normal distribution of intelligence. More specifically, do the same genes affect both high intelligence and the rest of the distribution to the same extent? It cannot be assumed that the aetiology of high intelligence is the same. For example, very low intelligence (severe intellectual disability) differs aetiologically from the normal distribution, as proposed initially by Lionel Penrose (1938). In quantitative genetic studies (Nichols, 1984; Reichenberg et al., in preparation), a critical piece of evidence is that siblings of individuals with severe intellectual disability have an average IQ near 100, whereas siblings of those with mild intellectual disability have an average IQ of around 85, about one standard deviation below the population mean. In recent molecular genetic studies, rare non-inherited mutations appear to be a major source of severe intellectual disability (Ellison, Rosenfeld, & Shaffer, 2013).

One of the earliest studies in behavioural genetics was Galton's *Hereditary Genius* (1869), an analysis of family pedigrees for brains as well as beauty and brawn. Since there

* Corresponding author at: King's College London, MRC Social, Genetic & Developmental Psychiatry Centre, Institute of Psychiatry, Psychology & Neuroscience, DeCrespigny Park, Denmark Hill, London, SE5 8AF, United Kingdom. Tel.: +44 20 7848 0985; fax: +44 20 7848 0092.
E-mail address: robert.plomin@kcl.ac.uk (R. Plomin).

http://dx.doi.org/10.1016/j.intell.2014.11.005
0160-2896/

was no satisfactory way at the time to measure intelligence, Galton had to rely on reputation as an index of eminence, which he found to be highly familial. Since Spearman's (1904) seminal work on general cognitive ability (*g*) over a century ago, research has focused on intelligence as a general factor that indexes what diverse tests of cognitive abilities have in common (Jensen, 1998). Intelligence was the target of the first twin and adoption studies in the 1920s (Burks, 1928; Freeman, Holzinger, & Mitchell, 1928; Merriman, 1924; Theis, 1924), and continues to be among the most studied traits in behavioural genetics (Plomin, DeFries, Knopik, & Neiderhiser, 2013).

For these reasons, it is surprising that few behavioural genetic studies have focused on high intelligence (Plomin & Haworth, 2009). We review these studies below, but we begin with hypotheses about why genetic and environmental factors might differ for high intelligence (the Discontinuity Hypothesis), and why the results might be similar (the Continuity Hypothesis).

2. The Discontinuity Hypothesis

The Discontinuity Hypothesis posits different environmental and genetic aetiologies for high intelligence in contrast to the rest of the distribution (Petrill, Kovas, Hart, Thompson, & Plomin, 2009). Although the evidence showing substantial heritability for the normal distribution of intelligence is one of the most consistently documented findings in the behavioural sciences (Deary, Johnson, & Houlihan, 2009), researchers in the field of expert training have argued that "differences in early experiences, preferences, opportunities, habits, training, and practice are the real determinants of excellence" (Howe, Davidson, & Sloboda, 1998, p. 403). A recent special issue of the journal Intelligence examines this environmental view of the acquisition of expertise (Detterman, 2014), including its relationship to genetic research (Plomin, Shakeshaft, McMillan, & Trzaskowski, 2014). Although the critical importance of deliberate practice is most often considered in the domain of specialist skills such as games, arts and sports, intelligence is also sometimes viewed as acquired expertise rather than inherited talent (Sternberg, 1999). If one accepts the overwhelming evidence showing substantial heritability for variation in the normal range of intelligence, the expert training position would suggest a discontinuity in the sense that it assumes that excellence is primarily due to environmental factors. Quantitative genetic research such as the twin method can test this hypothesis by investigating whether environmental influence is more important for high intelligence as compared to the rest of the distribution. Another more subtle environmental source of discontinuity can also be tested: the hypothesis that "differences in early experiences" are especially important for excellence would lead to the prediction that shared environment – environmental factors that make family members similar – should be greater for high intelligence.

Genetic reasons for discontinuity are also plausible, beginning with the folk wisdom that there could be "genes for genius." The most persuasive case for genetic discontinuity for genius has been made by David Lykken (1998). He notes that a key problem of genius is "its mysterious irrepressibility and its ability to arise from the most unpromising of lineages and to flourish even in the meanest of circumstances" (p. 29). He proposed that genius emerges from unique combinations of genes; he referred to these higher-order nonadditive (epistatic) interactions as emergenic (Lykken, 1982, 2006). The emergenesis hypothesis does not necessarily predict that different genes affect high intelligence, but it does predict that genetic effects are nonadditive for high intelligence. The hallmark of an epistatic trait is one for which identical twins, who share all their genes, are more than twice as similar as fraternal twins and other first-degree relatives, who share on average 50% of their segregating genes. The twin design can test this hypothesis that nonadditive genetic effects are greater for high intelligence as well as testing the "genes for genius" hypothesis that different genes are responsible for high intelligence.

For both environmental and genetic discontinuity hypotheses, a crucial issue is the cut-off used to define high ability. If the cut-off is extremely high, scientific research gives way to case studies, as has been recently avowed by a leader in research on expert training, who advocated case studies of the "less than a handful of individuals… with the very highest levels of performance" (Ericsson, 2014). In genetics, too, there is interest in the very highest levels of performance. For example, Galton benchmarked the top 1 in a million (.0001%) as "illustrious" and the top 250 in a million (.025%) as "eminent" (Galton, 1869), and Lykken referred to "genius" although he did not suggest a specific cut-off. Such extreme cut-offs are beyond the reach of quantitative genetics research or gene-hunting research, both of which require large sample sizes. However, once genes accounting for at least a few percent of the variance at any level of performance are identified, they can be used with adequate power as a polygenic score in research on even "a handful of individuals with the very highest levels of performance" (Plomin & Deary, 2014). This is beginning to happen in the world of elite athletic performance where, contrary to the Discontinuity Hypothesis, the same genes appear to be associated additively with both ordinary and extraordinary performance (Epstein, 2013).

3. The Continuity Hypothesis

The Continuity Hypothesis posits that high performance is the quantitative extreme of the same environmental and genetic factors responsible for the rest of the normal distribution. From an environmental perspective, the prodigious practice and concentrated effort of high performers might be only quantitatively (e.g., number of hours of deliberate practice) but not qualitatively different from the factors responsible for the rest of the distribution. In terms of genetics, the Continuity Hypothesis is the foundation for quantitative genetic theory (Fisher, 1918). If multiple genes affect a trait, their joint effects are distributed as a normal bell-shaped curve, which means that the same genes affect the low and high extremes of such polygenic traits. Molecular genetic research has begun to confirm this polygenic prediction as genes are identified that contribute to the heritability of complex dimensions and disorders (Plomin, Haworth, & Davis, 2009). For example, genes identified by their association with obesity are associated with body weight throughout the distribution of weight (Speliotes et al., 2010).

4. Quantitative genetic analysis of high intelligence

When genes associated with intelligence are identified, they will provide a strong competitive test of these two hypotheses by assessing the extent to which genes

associated with normal variation in intelligence are also associated with high intelligence and vice versa. Until that time, quantitative genetic methods such as the twin design can be used to compare the hypotheses. Quantitative genetic analyses have an advantage over molecular genetic approaches in terms of investigating environmental as well as genetic sources of continuity and discontinuity. For example, a twin study can test whether shared environmental influence is greater for high intelligence.

There are several ways that the twin method can be used to investigate whether genetic and environmental influences differ for high intelligence as compared to the rest of the distribution. These methods are described in greater detail in Methods section, but we introduce them here because of their relevance for reviewing previous studies of high intelligence. One set of methods uses a dichotomous "diagnosis" of high intelligence (case) or not (control). Monozygotic (MZ) and dizygotic (DZ) twin concordances can be compared to estimate genetic and environmental influence on high intelligence. Such dichotomous data are often analysed using a liability–threshold model, which assumes that liability is distributed normally until a threshold is exceeded, even though the analysis is based on dichotomous data (Rijsdijk & Sham, 2002). If the only available data were a "diagnosis" of high intelligence, the liability–threshold model is a useful way of assuming an underlying continuous liability despite having assessed a dichotomy.

Analysing high intelligence as a dichotomy loses much information when intelligence in the "cases" and "controls" has been assessed as a continuum. A method called DeFries–Fulker (DF) extremes analysis (DeFries & Fulker, 1985, 1988; DeFries, Fulker, & LaBuda, 1987) makes use of such quantitative trait data in estimating the genetic and environmental origins of the mean difference between the high intelligence group and the rest of the population. For this reason, heritability from DF extremes analysis is called group heritability to distinguish it from the usual estimate of heritability, which could be called individual differences heritability because it refers to genetic influence on individual differences throughout the distribution. Importantly, DF extremes analysis broaches the issue of the extent to which the same genes affect high intelligence and the rest of the distribution, as explained in Methods section.

5. Previous studies of high intelligence

Twin studies of high intelligence in childhood (Petrill et al., 1997; Plomin & Thompson, 1993; Ronald, Spinath, & Plomin, 2002) and in adulthood (Saudino, Plomin, Pedersen, & McClearn, 1994) have generally used DF extremes analysis and reported results consistent with the Continuity Hypothesis, in that group heritability was similar to individual differences heritability. However, the high-intelligence groups in these studies were small, just a few dozen pairs of twins, with the exception of one study (Ronald et al., 2002) which was limited by the age of the sample (2–4 years) and the measure (ratings of intelligence by parents). Low power to detect differences in heritability biases results in favour of the Continuity Hypothesis. Other studies have investigated the heritability of individual differences within high-intelligence groups, or asked more generally whether heritability differs across the population as a function of level of intelligence (Thompson, Detterman, & Plomin, 1993). However, such analyses address why one highly intelligent person is slightly more or less intelligent than another highly intelligent person, rather than asking why highly intelligent individuals as a group differ from the rest of the population.

In response to the neglect of research on high intelligence, the Genetics of High Cognitive Abilities (GHCA) Consortium was formed to bring together intelligence data on 11,000 twin pairs for the purpose of enabling an adequately powered comparison between high intelligence and the normal distribution. Liability–threshold model-fitting yielded evidence supporting the Continuity Hypothesis because estimates of genetic influence did not differ for high intelligence (0.50 with a 95% confidence interval of 0.41 to 0.60) and the entire sample (0.55; 0.51–0.59) (Haworth, Dale, & Plomin, 2009; Haworth et al., 2009). The overlapping confidence intervals suggest that heritability from the liability–threshold model in the high-intelligence group does not differ significantly from individual differences heritability. Estimates of shared environmental influence were also similar: 0.28 (0.19–0.37) for high intelligence and 0.21 (0.17–0.25) for the entire sample. However, the large confidence intervals for the high-intelligence group indicate that replication is needed to confirm the Continuity Hypothesis.

Finding similar heritabilities for high intelligence and the rest of the distribution does not confirm that the same genes are involved, which is the strength of DF extremes analysis. Moreover, in the GHCA study, only the top 15% were selected and the sample came from six twin studies each using different measures, in four countries, with a wide age range (6–71 years).

6. The present study

In contrast to the GHCA study, the present study used a higher cut-off (5%). It included non-twin siblings as well as twins. The sample was drawn from a single population and was assessed at the same age (18 years) on the same battery of cognitive measures, and the data were analysed with multiple methods including DF extremes analysis. Using a general factor from cognitive assessments of 3 million 18-year-old males administered as part of compulsory military service in Sweden between 1968 and 2010, we identified 370,000 sibling pairs and 9000 twin pairs. We selected the highest-scoring (top 5%) non-twin siblings and twins in order to investigate the familiality and heritability of high intelligence and its links to the normal distribution.

7. Methods

7.1. Sample

We tested the Continuity Hypothesis using cognitive assessments administered as part of military service in Sweden, from 1968 to 2010. Conscription was compulsory for males in Sweden until 2009, excluding those with severely disabling physical or psychiatric disorders, and achieved approximately 98% participation: 3 million 18-year-old males. From these, 363,905 families were identified containing at least two conscripted male siblings born in Sweden. From each family, we selected one twin pair if present (the youngest, if the family contained more than one pair); if there were no twins,

we selected the two male siblings closest to one another in age (the youngest, again, if two such pairs had the same age difference). These selections were made using the Swedish Multi-Generation Register, which includes all individuals born in Sweden since 1932 or living in Sweden since 1961. The resulting data set comprised 3039 monozygotic (MZ) and 3196 dizygotic (DZ) twin pairs, 2780 twin pairs of unknown zygosity, and 354,890 pairs of non-twin brothers. The vast majority (96.7%) of the non-twin sibling pairs were separated in age by less than 2 years.

7.2. Measures

General cognitive ability was assessed with the Swedish Enlistment Battery (SEB), administered as part of the military conscription testing. Three different versions of the SEB were used during the 40-year period for which cognitive data were available: the SEB67 during the years 1970–1979, the SEB80 during 1980–1993, and the CAT-SEB during 1994–2009 (Carlstedt, 2000). The SEB67 and SEB80 were paper and pencil tests consisting of four subtests assessing verbal, visuospatial, technical and inductive abilities, which were summed to derive a measure for general cognitive ability. High internal consistency for the SEB80 has been reported (coefficient $\alpha = .79$–$.91$) (Carlstedt & Mårdberg, 1993). Due to theoretical and methodological developments in intelligence research and the advent of the personal computer, a new version of the SEB (CAT-SEB), utilising computer-aided testing, was launched in 1994. The CAT-SEB was based on a three-level hierarchical model of cognitive abilities and included 12 tests, of which 10 were used to form the latent general ability factor, plus secondary factors of crystallised intelligence and general visualisation. The reliability of the CAT-SEB tests is also good (coefficient $\alpha = .70$–$.85$) (Mårdberg & Carlstedt, 1998). The general cognitive ability variable, available from the Conscription Register and based on the different versions of the SEB, was measured on a stanine scale, i.e., a normally-distributed variable divided into nine levels (higher scores indicating greater ability), with a mean of 5 and standard deviation of 2.

7.3. Analyses

In addition to traditional individual differences analyses of the entire sample of twins and non-twin siblings (Plomin et al., 2013), two types of analysis were used for high intelligence: liability–threshold model-fitting using dichotomous data (high intelligence versus normal-range intelligence), and DeFries–Fulker (DF) extremes analysis, in which an "extreme" (or proband) group is selected (high-intelligence individuals, in this case), and quantitative variation in their siblings or co-twins is analysed. We begin with a brief description of other ways that have been used to analyse data of this type.

One general approach is to test for an interaction across the population between heritability (and environmental parameter estimates) and level of intelligence (Cherny, Cardon, Fulker, & DeFries, 1992; Logan et al., 2012). However, because there are relatively few individuals of high intelligence in the population, testing for an interaction throughout the entire population has little power to detect a difference in heritability specifically for high intelligence. Low power to detect interactions biases this approach in favour of the Continuity Hypothesis.

A more focused approach is to compare heritability for a high-intelligence group and an unselected group. A methodological problem with this apparently straightforward approach is that the variance of a high-intelligence group is restricted because they are highly selected, and this is likely to affect twin correlations. An important conceptual problem is that the focus of traditional heritability estimates is on individual differences. For understanding the origins of high intelligence, the issue is not whether one highly intelligent person is slightly more or less intelligent than another highly intelligent person, which is what is assessed in traditional heritability estimates. Instead, we are interested in the genetic and environmental causes of high intelligence — why highly intelligent individuals as a group differ from the rest of the population.

7.4. Liability–threshold model-fitting

The dichotomous data – high intelligence versus the rest of the distribution – can be analysed by comparing the degree of concordance for MZ and DZ twins, and for non-twin siblings. Here, we used probandwise concordance: the proportion of "affected" individuals (i.e., those with a stanine score of 9, in this case) who have a twin or sibling who is also affected. This method indicates morbidity risk, i.e., the probability that a sibling or co-twin of someone in the high-intelligence group will also be in that group.

Liability–threshold models assume that liability is normally distributed, but with the "disorder" (membership of the high-intelligence group, in this case) occurring only when a certain threshold is reached. Tetrachoric twin correlations and thresholds were calculated from our dichotomous data (Falconer, 1965; Smith, 1974), and liability–threshold (and all other) model-fitting analyses were conducted using OpenMx (Boker et al., 2011). This model-fitting produces ACE estimates analogous to those produced by twin model-fitting for continuous data, but the heritability estimate is the heritability of a hypothetical continuous liability construct, derived from the dichotomous data.

7.5. DeFries–Fulker (DF) extremes analysis

Analysing continuous data as dichotomous loses a great deal of information. Here, intelligence is assessed as a continuous stanine (standardised, nine-point) score, so much more information is available than the dichotomised "diagnosis" of high intelligence assessed by liability–threshold modelling.

We can use these continuous data to estimate the genetic and environmental origins of the mean difference between the high-intelligence group and the rest of the distribution, using DF extremes analysis (DeFries & Fulker, 1985, 1988; DeFries et al., 1987). This technique assesses the degree to which the co-twins or siblings of the extreme (high intelligence) group regress to the population mean. If the co-twin/sibling mean differs from the population mean, the trait is familial. Further, if the mean for MZ twins regresses less than that for DZ twins (and non-twin siblings), this indicates genetic influence on the mean difference between the high-intelligence group and the rest of the population.

In DF extremes analysis, the trait scores are standardised and transformed to account for the mean differences between the MZ and DZ groups, then fitted to the regression equation: $C = \beta_1 P + \beta_2 R + A$. C is the predicted score for the co-twin; P, the proband score; R, the coefficient of genetic relatedness (1.0 for MZ twins, 0.5 for DZ twins and non-twin siblings) and A, the regression constant. β_1 is the partial regression of the co-twin score on the proband score, and represents the average twin resemblance, independent of β_2. β_2 is the partial regression of the co-twin score on R independent of β_1, and is equal to double the difference between the MZ and DZ co-twin means (adjusted for any differences between MZ and DZ probands). Dividing β_2 by the difference between the proband and population means provides the "group heritability," the proportion of the difference between the proband and population phenotypic means that is genetic in origin. (This should be contrasted against the usual heritability estimates produced by traditional twin model fitting analyses, which represent the genetic influence on individual differences, rather than the influence on the mean difference between probands and the rest of the population.)

A finding of group heritability indicates that both the extreme trait and the rest of the distribution are heritable. Importantly, however, it also indicates that the genetic contributions in both cases are not independent from one another: the group heritability for two heritable but unrelated traits would be zero (Plomin & Kovas, 2005). In effect, DF extremes is a bivariate analysis: in this case, between the extreme score and the rest of the quantitative dimension. Finding substantial group heritability thus indicates not only that both the dimensional trait and its quantitative extreme are heritable, but also that they are influenced in part by the same genes: extreme scores are not qualitatively distinct from the rest of the distribution.

8. Results

This study assessed the genetic architecture of high intelligence. We present the results of classical twin model-fitting for the whole distribution of intelligence. For high intelligence vs. the rest of the distribution, we present twin and sibling pair concordances, liability–threshold model-fitting analyses of dichotomous twin data, and DF extremes analyses incorporating quantitative twin data. First, we provide descriptive results and a simple representation of the familiality of high intelligence.

8.1. Descriptive statistics

Fig. 1 shows the distribution of stanine scores for intelligence for the total sample described above, selecting one sibling at random from each pair. As indicated, the data are normally distributed (mean = 5.16, SD = 1.94), with 5% of individuals achieving the highest possible stanine score (9), corresponding to an IQ above 125. The siblings of these probands were selected, comprising a sample of 185 MZ twins, 196 DZ twins, and 28,339 non-twin siblings.

The aim of this study is to estimate the genetic and environmental influences accounting for the difference (amounting to 1.98 standard deviations) between the highest scoring individuals and the population mean. As explained above, we are not concerned with the individual differences between the highest scoring individuals themselves (which cannot be assessed in any case, as only stanine scores are available for this sample), but rather with the differences between this group as a whole and the rest of the population.

For subsequent analyses, to account for any changes in population mean intelligence over time, the raw stanine scores were regressed on year of birth, and standardised.

8.2. Individual differences (whole twin sample)

Before exploring the differences between the high-intelligence group and the rest of the population, the twin sample as a whole was analysed to confirm the validity and representativeness of these data and to provide a comparison for the analysis of high intelligence. The correlation between scores for MZ twins was 0.80, and for DZ twins 0.51, which suggests heritability of 0.58 for intelligence in this sample (by doubling the difference between these correlations to produce a rough estimate) and is inconsistent with nonadditive genetic effects (as the MZ correlation is less than double the DZ

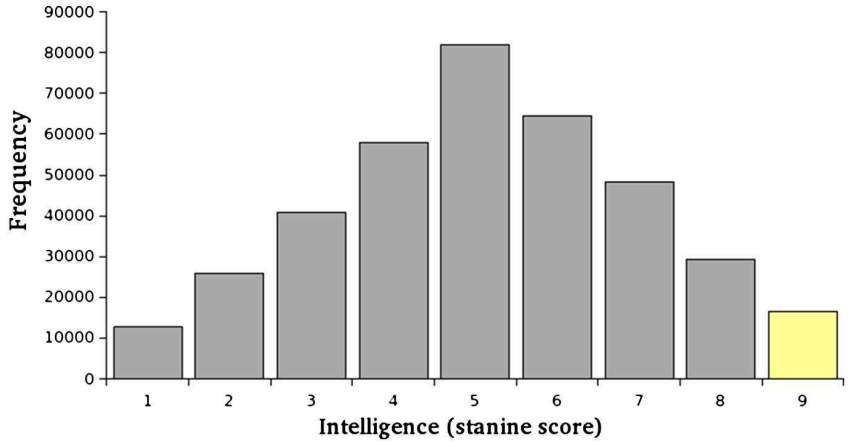


Fig. 1. Distribution of intelligence scores. $N = 363{,}905$, mean = 5.16, SD = 1.94. Data shown include one randomly-selected individual per sibling pair. The highest-scoring individuals (stanine 9) are highlighted ($N = 16{,}058$).

Table 1
Model-fitting results for whole twin sample. Results are additive genetic (A), shared environmental (C) and residual (E; i.e., non-shared environment and error) components of variance, with 95% confidence intervals.

Variance components (95% confidence intervals)			Sample (numbers of pairs)	
A	C	E	MZ	DZ
0.58 (0.53–0.63)	0.22 (0.17–0.27)	0.20 (0.19–0.21)	3039	3196

correlation). More rigorous estimates, produced by univariate ACE twin model-fitting, are presented in Table 1. This analysis partitions variance in the sample's scores into additive genetic (A), shared environmental (C) and non-shared environmental (E) components.

These results, suggesting substantial genetic influence, with environmental influences evenly divided between shared and non-shared effects, correspond very closely to those typically found in the literature for participants of this age (Haworth et al., 2010). This suggests that these data are in line with those obtained by other studies.

8.3. Familiality of high intelligence

The familiality of high intelligence can be observed simply by comparing the mean scores of the non-twin siblings of high-intelligence probands to the population mean.

As shown in Fig. 2, high intelligence, defined as the highest 5% of scores, is highly familial. For siblings of probands (i.e., those with a standardised score of 1.98, equivalent to a raw stanine score of 9), the distribution of intelligence is shifted sharply to the right of that of the rest of the population, with a mean score (0.81) approximately halfway between the proband score and the population mean (0). These results suggest a sibling "group correlation" (the ratio between the siblings' deviation from the population mean to the probands' deviation from the population mean) of 0.41. In other words, almost half of the difference between high intelligence and the rest of the population is familial in origin.

Familiality could be due to genetic or environmental influences. However, the twin data presented in Fig. 3 indicates that the familial effect is substantially genetic in origin. The mean for DZ twins of probands (0.95) does not differ substantially from that of non-twin siblings, as shown in Fig. 1. In contrast, MZ co-twins have a substantially higher mean score (1.39) than that of DZs, suggesting a strong genetic association.

More rigorous and specific results can be obtained, as described below.

8.4. Dichotomous data: Concordances

Table 2 presents twin/sibling concordances for the high-intelligence MZ twin, DZ twin and non-twin sibling groups.

For the 28,339 selected non-twin sibling pairs, there were 6604 individuals in 3302 concordant pairs, and 50,074 individuals in 25,037 discordant pairs. Simple (pairwise) concordance is thus 12% (i.e., 3302/28339, the proportion of pairs that are concordant). However, probandwise concordance is a better measure, since it indicates morbidity risk. For these siblings, probandwise concordance is 21% (i.e., (2*3302)/((2*3302) + 25037)), which is the probability that the twin or sibling of a proband will also be a proband. These results indicate substantial familiality for high intelligence.

For twins, the same calculations indicate probandwise concordance of 45% for MZ twins, and 25% for DZ twins. In other words, there is a 45% probability that the MZ twin of an individual in the high intelligence group will also be in that

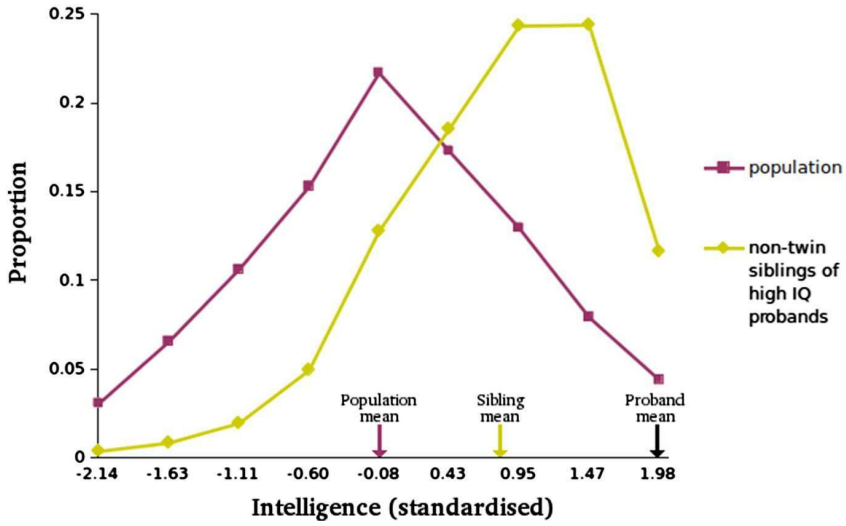


Fig. 2. Familiality of high intelligence. Male siblings of high-intelligence probands (with a standardised score of 1.98) have significantly and substantially higher intelligence (mean = 0.81, SD = 0.81, N = 28,339) than the population (mean = 0, SD = 1, N = 727,810).

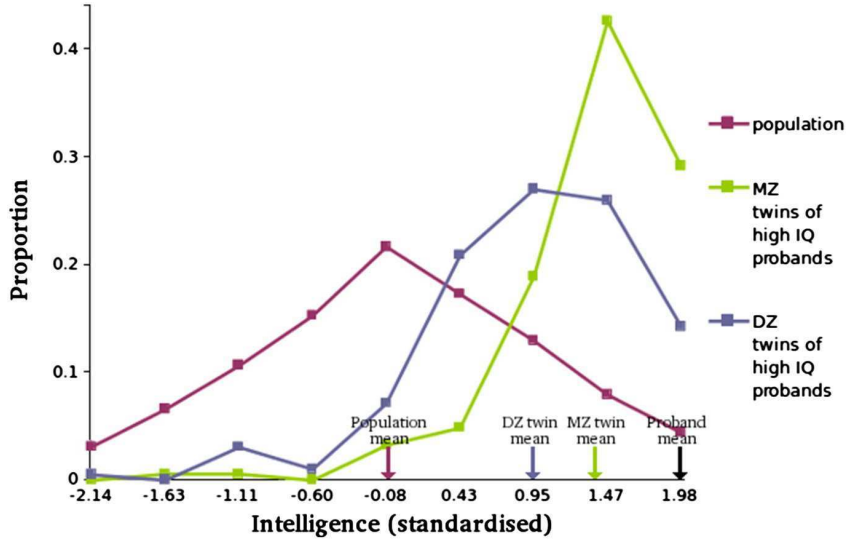


Fig. 3. Heritability of high intelligence. Male MZ co-twins of high-intelligence probands (with a standardised score of 1.98) have significantly and substantially higher intelligence (mean = 1.39, SD = 0.58, N = 185) than DZ co-twins (mean = 0.95, SD = 0.75, N = 196), who in turn score significantly and substantially higher than the population (mean = 0, SD = 1, N = 727,810).

group (and 25% for a DZ twin). Doubling the difference between the MZ and DZ concordances would suggest heritability of 0.40 — but as these concordances do not take account of population base rates, this is not entirely appropriate statistically. Tetrachoric and group correlations (presented below) are preferable for this reason.

All subsequent analyses (tetrachoric correlations, liability–threshold model-fitting and DF extremes analysis) were conducted using the full twin sample of 6235 pairs.

8.5. Dichotomous data: Liability–threshold model-fitting

As discussed in Methods, dichotomous data liability–threshold modelling may be used to analyse dichotomous data, assuming that liability (i.e., the "risk" of high intelligence, in this case) is normally distributed, but a certain threshold must be exceeded for an individual to become a proband. The liability–threshold model is based on twin tetrachoric correlations, which are presented in Table 3.

These tetrachoric correlations, derived from dichotomous data, may be analysed in the same way as twin correlations from continuous data. For example, doubling the difference between the MZ and DZ correlations suggests heritability of 0.44 for high intelligence. As with the twin correlations for the whole distribution, these results do not suggest the existence of nonadditive genetic effects. Liability–threshold model-fitting provides a more rigorous analysis. Results are presented in Table 4.

These results suggest substantial heritability (0.42), with environmental influences divided between shared and non-shared effects. All of these variance components were significant, and an analysis of sub-models (eliminating variance components and testing the decrease in fit to the data) indicated that this full model best fit the data. As noted in Methods, however, these results refer to the variance of a hypothetical construct of continuous liability for high intelligence, derived from the dichotomous data, rather than that of a quantitative, continuous measure of intelligence.

8.6. Continuous data: DeFries–Fulker (DF) extremes analysis

As shown in Fig. 3, MZ co-twins of those in the high intelligence group regress to the population mean to a much smaller extent than do DZ co-twins, suggesting genetic influence. As discussed in Methods, DF extremes analysis uses continuous data, and can estimate the genetic and environmental factors influencing the difference in mean intelligence between the two intelligence groups (high intelligence vs. the rest of the population), by quantifying the differential regression to the mean for MZ and DZ co-twins of probands.

Table 2
Concordances. Concordance is shown both pairwise (the proportion of concordant pairs) and probandwise (the proportion of probands whose twin/sibling is also a proband).

	Number of pairs			Concordance	
	Total	Concordant	Discordant	Pairwise	Probandwise
MZ twins	185	54	131	0.29	0.45
DZ twins	196	28	168	0.14	0.25
Non-twin siblings	28,339	3302	25,037	0.12	0.21

Table 3
Tetrachoric correlations. N = 6235 twin pairs.

	Tetrachoric correlation (95% confidence interval)	Std. error
MZ twins	0.78 (0.71–0.84)	0.05
DZ twins	0.56 (0.45–0.66)	0.08

Table 4
Liability–threshold model-fitting results. Results are additive genetic (A), shared environmental (C) and residual (E; i.e., non-shared environment and error) components of variance. $N = 6235$ twin pairs.

Variance components (95% confidence intervals)		
A	C	E
0.42 (0.17–0.68)	0.36 (0.12–0.57)	0.22 (0.16–0.30)

The process can be illustrated without using model-fitting, as shown above for non-twin siblings ("Familiality of High Intelligence"). Whereas a conventional twin correlation refers to individual differences on a trait, a "group" correlation quantifies the mean difference between the extreme group (i.e., the high intelligence group, here) and the rest of the population (Plomin, 1991). This may be calculated as the ratio between the two groups' differences from the mean, i.e., that of the probands and that of their co-twins. (In animal selection studies, these are known as the "selection differential" and "response to selection," respectively; Plomin et al., 2014.) For these data, this yields group correlations of 0.70 for MZ twins, and 0.48 for DZ twins. Doubling the difference between these group correlations estimates group heritability at 0.44, suggesting that almost half of the mean difference between the high intelligence group and the rest of the population is explained genetically.

DF extremes model-fitting is preferable, because it uses the full twin data set (6235 pairs), and does not rely on randomly selecting one member of each concordant pair. It would also take into account any mean differences between MZ and DZ probands, although there are none with these stanine data. The DF extremes model-fitting results are presented in Table 5.

The DF extremes group heritability estimate (0.40) is similar to that estimated using the simpler group method above, and to the liability–threshold model-fitting results. The close approximation between the DF extremes and liability–threshold model-fitting results suggests that the assumptions of the latter are correct (Plomin & Kovas, 2005).

As with the previous analyses using dichotomous data, the DF extremes results indicate that just under half of the mean difference between the high intelligence group and the rest of the population is explained genetically, with the remaining variance divided between shared and non-shared environmental influences.

9. Discussion

These results provide strong support for the Continuity Hypothesis. Familial resemblance from non-twin sibling analyses and heritabilities from twin analyses were similar for high intelligence and for the rest of the distribution, using concordances, liability–threshold analysis, and DF extremes analysis. As explained earlier, DF extremes analysis not only indicates substantial heritability of high intelligence and of individual differences in intelligence in the normal distribution but also suggests substantial genetic correlation between them. Importantly, our twin results are highly similar to the results of the only other large twin study of high intelligence (GHCA; Haworth, Wright, et al., 2009).

Table 5
DF extremes model-fitting results. Results are additive genetic (A), shared environmental (C) and residual (E; i.e., non-shared environment and error) components of variance. $N = 6235$ twin pairs.

Group ACE components (95% confidence intervals)		
A	C	E
0.40 (0.28–0.52)	0.37 (0.27–0.46)	0.23 (0.19–0.27)

For these reasons, we conclude that high intelligence is familial, heritable, and caused by the same genetic factors responsible for the normal distribution of intelligence. Stated more provocatively, high intelligence as we defined it appears to be nothing more than the quantitative extreme of the same genetic factors responsible for normal variation.

We found no support for the genetic Discontinuity Hypothesis that nonadditive genetic variance is greater for high intelligence, as suggested by the emergenesis hypothesis (Lykken, 1982, 2006). There was no evidence for nonadditive genetic variance for either high intelligence or for the entire sample, which is similar to GHCA results. One caveat concerns assortative mating. Assortative mating is much greater for intelligence (spouse correlations ~0.40) than for personality (spouse correlations ~0.10) or for physical characteristics such as height and weight (~0.20) (Plomin & Deary, 2014). In twin studies such as ours and GHCA that do not also include parental data, nonadditive genetic variance could be masked by assortative mating, and there is some evidence that this is the case for intelligence (Vinkhuyzen, van der Sluis, Maes, & Posthuma, 2012). If assortative mating were similar for high intelligence and the entire sample, it would not affect the interpretation of our results, which are based solely on the twin design. However, if assortative mating were greater for high intelligence, this could mask greater nonadditive genetic variance for high intelligence (Vinkhuyzen et al., 2012). We are not aware of any studies that have investigated whether assortative mating differs as a function of level of intelligence.

9.1. Environmental and genetic discontinuity

The GHCA study found a trend supporting the environmental Discontinuity Hypothesis, in that shared family environmental influence was somewhat greater for high intelligence. In the GHCA study, shared environment was estimated at 28% in the high intelligence group using liability–threshold modelling and 21% in the entire sample, although the difference was not nearly significant (95% confidence intervals were 0.19–0.37 and 0.17–0.25, respectively). In the present study, the results were 36% for high intelligence and 22% for the entire sample, with the difference again non-significant. The shared environmental estimate for high intelligence from DF extremes analysis was similar in the present study (37%), although DF extremes analysis is less comparable to the other analyses. The confidence intervals overlap substantially for all of these comparisons.

For these reasons, we conclude that high intelligence is caused by the same environmental factors responsible for the normal distribution of intelligence. However, it should be mentioned that the Continuity Hypothesis is essentially a null hypothesis of no difference between high intelligence and the normal distribution. Caution is warranted because insufficient power biases results in favour of the Continuity Hypothesis. Nonetheless, the similarity of results from the GHCA study with 11,000 twin pairs and the present study with 9000 twin pairs

affords strong, if not definitive, support for the Continuity Hypothesis.

As mentioned in Introduction, an important qualification for all of these conclusions is that more extreme cut-offs might yield different results. The GHCA study selected the top 15% of the distribution (although a case-control study with more extreme cut-offs is underway; Spain et al., in preparation), and the present study the top 5%. These cut-offs balance sample size and power. Twin studies are unlikely to reach adequate power using Galton's (1869) cut-offs of .025% for "eminent" and .0001% for "illustrious." If 1% of births are twins, a population of 80 million would be needed to obtain a mere 200 pairs of twins above the .025% cut-off. However, molecular genetic studies could be useful even for extreme cut-offs, in part because they do not require special populations such as twins.

9.2. Positive genetics

Nothing would advance genetic research on intelligence more than identifying some of the genes responsible for its substantial heritability. We now know that many genes of very small effect are responsible for the heritability of intelligence, as is the case for all common disorders and complex dimensions in the life sciences (Plomin & Deary, 2014). Nonetheless a polygenic score that adds up the effects of many genes of small effect size would provide a strong test of the prediction from the Continuity Hypothesis that genes associated with normal variation in intelligence will also be associated with high intelligence. It could also be used to test the Continuity Hypothesis for very high cut-offs.

If the Continuity Hypothesis is correct, high intelligence represents the positive end of a normal distribution. In contrast, most genome-wide association research has focused on the negative effects of genes on disorders, diseases, and disabilities (Visscher, Brown, McCarthy, & Yang, 2012). For intelligence, the problematic end of the distribution has also become a focus of research, as rare non-inherited mutations are emerging as a major source of severe intellectual disability (Ellison et al., 2013). Genetic exploration of the positive tail of normally distributed traits such as high intelligence is important conceptually because it moves away from the notion that we are all the same genetically except for rogue mutations that cause disorders, diseases and disabilities. The term positive genetics has been used to highlight genetic research on the positive end of distributions (Plomin et al., 2009).

The normal phenotypic distribution of intelligence makes it an obvious target for investigating the positive as well as negative extremes. Another possibly important feature of intelligence is that, like athletic ability, it is assessed as maximal performance, in contrast to other behavioural domains such as psychopathology and personality that involve typical behaviour. However, the larger significance of positive genetics is that these phenotypic considerations about the positive pole of the normal distribution have far-reaching implications for genomics. Polygenic scores created from genome-wide association studies are normally distributed, for disorders as well as for dimensions. In other words, polygenic scores have a positive pole with just as many people as the negative pole, even though the spotlight is typically on the negative end of the distribution of genetic "risk." This normal distribution of polygenic scores implies that at the level of DNA variation there are no common disorders, only normally distributed quantitative traits (Plomin et al., 2009).

Positive genetics and the Continuity Hypothesis have practical as well as conceptual implications for intelligence, for example, for identifying genes associated with intelligence. Rather than using the brute force strategy of getting ever-larger samples of unselected individuals to narrow the "missing heritability" gap (Plomin & Simpson, 2013), selecting individuals of high intelligence might increase power for gene-hunting based on the simple hypothesis that high-intelligence individuals are enriched for intelligence-enhancing alleles and harbour few intelligence-depleting alleles. In other words, intellectual development can be disrupted by any and many mutations, including non-inherited (*de novo*) mutations, but high intelligence requires that everything works correctly. This hypothesis provided the rationale for a genome-wide case-control association study for cases with extremely high intelligence (IQ > 150) compared to unselected control individuals (Spain et al., in preparation). However, in an initial report, this design does not appear to have found richer results either for identifying individual DNA variants, or for genomic approaches such as comparing the total number of rare variants (which generally have negative effects and might be expected to occur less frequently in the high-intelligence sample). Nonetheless, it is early days for the use of high-intelligence samples to increase power for gene-hunting.

Positive genetics raises the question: who are the people at the positive end of the polygenic distribution of "risk" for disorders? Are they merely individuals at low risk for problems, or do they have unusual positive traits? Thinking positively begins by thinking quantitatively — about "dimensions" rather than "disorders" and about genetic "variability" rather than genetic "risk." Intelligence makes it easy to think positively.

Acknowledgements

We thank the participants in the Twins Early Development Study (TEDS) for making the study possible. TEDS is supported by a programme grant to R.P. from the UK Medical Research Council (MRC) [G0901245, and previously G0500079], with additional support from the US National Institutes of Health [HD044454; HD059215]. N.G.S. and E.K. are supported by MRC studentships. R.P. is supported by a MRC Research Professorship award [G19/2] and also a European Research Council Advanced Investigator award [295366] which is specifically focused on the genetics of high cognitive abilities. The Swedish research and team were funded by the Swedish Research Council and Swedish Research Council for Health, Working Life and Welfare.

References

Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., et al. (2011). OpenMx: An open source extended structural equation modeling framework. *Psychometrika, 76*, 306–317. http://dx.doi.org/10.1007/s11336-010-9200-6.

Burks, B. (1928). *The relative influence of nature and nurture upon mental development: A comparative study on foster parent–foster child resemblance.* Yearbook of the National Society for the Study of Education, Part 127. (pp. 219–316), 219–316.

Carlstedt, B. (2000). *Cognitive abilities — aspects of structure, process and measurement.* Gothenburg, Sweden: Acta Universitatis Gothenburgensis.

Carlstedt, B., & Mårdberg, B. (1993). Construct validity of the Swedish Enlistment Battery. *Scandinavian Journal of Psychology, 34*, 353–362. http://dx.doi.org/10.1111/j.1467-9450.1993.tb01131.x.

Cherny, S.S., Cardon, L.R., Fulker, D.W., & DeFries, J.C. (1992). Differential heritability across levels of cognitive ability. *Behavior Genetics*, *22*, 153–162. http://dx.doi.org/10.1007/BF01066994.
Deary, I.J., Johnson, W., & Houlihan, L.M. (2009). Genetic foundations of human intelligence. *Human Genetics*, *126*, 215–232. http://dx.doi.org/10.1007/s00439-009-0655-4.
DeFries, J.C., & Fulker, D.W. (1985). Multiple regression analysis of twin data. *Behavior Genetics*, *15*, 467–473. http://dx.doi.org/10.1007/BF01066239.
DeFries, J.C., & Fulker, D.W. (1988). Multiple regression analysis of twin data: Etiology of deviant scores versus individual differences. *Acta Geneticae Medicae et Gemellologiae*, *37*, 205–216.
DeFries, J.C., Fulker, D.W., & LaBuda, M.C. (1987). Evidence for a genetic aetiology in reading disability of twins. *Nature*, *329*, 537–539. http://dx.doi.org/10.1038/329537a0.
Detterman, D.K. (2014). Introduction to the intelligence special issue on the development of expertise: Is ability necessary? *Intelligence*, *45*, 1–5. http://dx.doi.org/10.1016/j.intell.2014.02.004.
Ellison, J.W., Rosenfeld, J.A., & Shaffer, L.G. (2013). Genetic basis of intellectual disability. *Annual Review of Medicine*, *64*, 441–450. http://dx.doi.org/10.1146/annurev-med-042711-140053.
Epstein, D. (2013). *The sports gene: Inside the science of extraordinary athletic performance*. New York: Current.
Ericsson, K.A. (2014). Why expert performance is special and cannot be extrapolated from studies of performance in the general population: A response to criticisms. *Intelligence*, *45*, 81–103. http://dx.doi.org/10.1016/j.intell.2013.12.001.
Falconer, D.S. (1965). The inheritance of liability to certain diseases estimated from the incidence among relatives. *Annals of Human Genetics*, *29*, 51–76. http://dx.doi.org/10.1111/j.1469-1809.1965.tb00500.x.
Fisher, R.A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, *52*, 399–433. http://dx.doi.org/10.1017/S0080456800012163.
Freeman, F.N., Holzinger, K.J., & Mitchell, B. (1928). *The influence of environment on the intelligence, school achievement, and conduct of foster children. Yearbook of the National Society for the Study of Education, Part 127*. (pp. 103–217), 103–217.
Galton, F. (1869). *Hereditary genius: An enquiry into its laws and consequences*. Cleveland, OH: World.
Haworth, C.M.A., Dale, P.S., & Plomin, R. (2009). Generalist genes and high cognitive abilities. *Behavior Genetics*, *39*, 437–445. http://dx.doi.org/10.1007/s10519-009-9271-2.
Haworth, C.M.A., Wright, M.J., Luciano, M., Martin, N.G., de Geus, E.J.C., van Beijsterveldt, C.E.M., et al. (2010). The heritability of general cognitive ability increases linearly from childhood to young adulthood. *Molecular Psychiatry*, *15*, 1112–1120. http://dx.doi.org/10.1038/mp.2009.55.
Haworth, C.M.A., Wright, M.J., Martin, N.W., Martin, N.G., Boomsma, D.I., Bartels, M., et al. (2009). A twin study of the genetics of high cognitive ability selected from 11,000 twin pairs in six studies from four countries. *Behavior Genetics*, *39*, 359–370. http://dx.doi.org/10.1007/s10519-009-9262-3.
Howe, M.J.A., Davidson, J.W., & Sloboda, J.A. (1998). Innate talents: Reality or myth? *Behavioral and Brain Sciences*, *21*, 399–442. http://dx.doi.org/10.1017/S0140525X9800123X.
Jensen, A.R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger.
Kell, H.J., Lubinski, D., & Benbow, C.P. (2013). Who rises to the top? Early indicators. *Psychological Science*, *24*, 648–659. http://dx.doi.org/10.1177/0956797612457784.
Logan, J.A.R., Petrill, S.A., Hart, S.A., Schatschneider, C., Thompson, L.A., Deater-Deckard, K., et al. (2012). Heritability across the distribution: An application of quantile regression. *Behavior Genetics*, *42*, 256–267. http://dx.doi.org/10.1007/s10519-011-9497-7.
Lubinski, D., Benbow, C.P., Webb, R.M., & Bleske-Rechek, A. (2006). Tracking exceptional human capital over two decades. *Psychological Science*, *17*, 194–199. http://dx.doi.org/10.1111/j.1467-9280.2006.01685.x.
Lykken, D.T. (1982). Research with twins: The concept of emergenesis. *Psychophysiology*, *19*, 361–373. http://dx.doi.org/10.1111/j.1469-8986.1982.tb02489.x.
Lykken, D.T. (1998). The genetics of genius. In A. Steptoe (Ed.), *Genius and the mind: Studies of creativity and temperament in the historical record* (pp. 15–37). New York: Oxford University Press.
Lykken, D.T. (2006). The mechanism of emergenesis. *Genes, Brain and Behavior*, *5*, 306–310. http://dx.doi.org/10.1111/j.1601-183X.2006.00233.x.
Mårdberg, B., & Carlstedt, B. (1998). Swedish Enlistment Battery (SEB): Construct validity and latent variable estimation of cognitive abilities by the CAT-SEB. *International Journal of Selection and Assessment*, *6*, 107–114. http://dx.doi.org/10.1111/1468-2389.00079.
Merriman, C. (1924). The intellectual resemblance of twins. *Psychological Monographs*, *33*, 1–58. http://dx.doi.org/10.1037/h0093212.
Nichols, P.L. (1984). Familial mental retardation. *Behavior Genetics*, *14*, 161–170. http://dx.doi.org/10.1007/BF01065538.
Penrose, L.S. (1938). *A clinical and genetic study of 1280 cases of mental defect*. London, UK: HM Stationery Office.
Petrill, S.A., Kovas, Y., Hart, S.A., Thompson, L.A., & Plomin, R. (2009). The genetic and environmental etiology of high math performance in 10-year-old twins. *Behavior Genetics*, *39*, 371–379. http://dx.doi.org/10.1007/s10519-009-9258-z.
Petrill, S.A., Plomin, R., McClearn, G.E., Smith, D.L., Vignetti, S., Chorney, M.J., et al. (1997). No association between general cognitive ability and the A1 allele of the D2 dopamine receptor gene. *Behavior Genetics*, *27*, 29–31. http://dx.doi.org/10.1023/A:1025659124405.
Plomin, R. (1991). Genetic risk and psychosocial disorders: Links between the normal and abnormal. In M. Rutter, & P. Casaer (Eds.), *Biological risk factors for psychosocial disorders* (pp. 101–138). Cambridge, UK: Cambridge University Press.
Plomin, R., & Deary, I.J. (2014). Genetics and intelligence differences: Five special findings. *Molecular Psychiatry*. http://dx.doi.org/10.1038/mp.2014.105 (Advance online pub).
Plomin, R., DeFries, J.C., Knopik, V.S., & Neiderhiser, J.M. (2013). *Behavioral genetics* (6th ed.). New York: Worth Publishers.
Plomin, R., & Haworth, C. (2009). Genetics of high cognitive abilities. *Behavior Genetics*, *39*, 347–349. http://dx.doi.org/10.1007/s10519-009-9277-9.
Plomin, R., Haworth, C.M.A., & Davis, O.S.P. (2009). Common disorders are quantitative traits. *Nature Reviews Genetics*, *10*, 872–878. http://dx.doi.org/10.1038/nrg2670.
Plomin, R., & Kovas, Y. (2005). Generalist genes and learning disabilities. *Psychological Bulletin*, *131*, 592–617. http://dx.doi.org/10.1037/0033-2909.131.4.592.
Plomin, R., Shakeshaft, N.G., McMillan, A., & Trzaskowski, M. (2014). Nature, nurture, and expertise. *Intelligence*, *45*, 46–59. http://dx.doi.org/10.1016/j.intell.2013.06.008.
Plomin, R., & Simpson, M. (2013). The future of genomics for developmentalists. *Development and Psychopathology*, *25*, 1263–1278. http://dx.doi.org/10.1017/S0954579413000606.
Plomin, R., & Thompson, L.A. (1993). Genetics and high cognitive ability. In G.R. Bock, & K. Ackrill (Eds.), *The origins and development of high ability. CIBA Foundation Symposium*, *178*. (pp. 62–84). Chichester, UK: Wiley.
Rijsdijk, F.V., & Sham, P.C. (2002). Analytic approaches to twin data using structural equation models. *Briefings in Bioinformatics*, *3*, 119–133. http://dx.doi.org/10.1093/bib/3.2.119.
Rindermann, H., & Thompson, J. (2011). Cognitive capitalism: The effect of cognitive ability on wealth, as mediated through scientific achievement and economic freedom. *Psychological Science*, *22*, 754–763. http://dx.doi.org/10.1177/0956797611407207.
Ronald, A., Spinath, F., & Plomin, R. (2002). The aetiology of high cognitive ability in early childhood. *High Ability Studies*, *13*, 103–114. http://dx.doi.org/10.1080/1359813022000048761.
Saudino, K.J., Plomin, R., Pedersen, N.L., & McClearn, G.E. (1994). The etiology of high and low cognitive ability during the second half of the life span. *Intelligence*, *19*, 353–371. http://dx.doi.org/10.1016/0160-2896(94)90007-8.
Smith, C. (1974). Concordance in twins: Methods and interpretation. *American Journal of Human Genetics*, *26*, 454–466.
Spain, S.L., Pedroso, I., Kadeva, N., Miller, M.B., ALSPAC, TEDS, et al. (2014n). Genetic analysis of exonic variation in a sample with extreme high cognitive ability (in preparation).
Spearman, C. (1904). "General intelligence", objectively determined and measured. *American Journal of Psychology*, *15*, 201–292. http://dx.doi.org/10.2307/1412107.
Speliotes, E.K., Willer, C.J., Berndt, S.I., Monda, K.L., Thorleifsson, G., Jackson, A.U., et al. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics*, *42*, 937–948. http://dx.doi.org/10.1038/ng.686.
Sternberg, R.J. (1999). Intelligence as developing expertise. *Contemporary Educational Psychology*, *24*, 359–375. http://dx.doi.org/10.1006/ceps.1998.0998.
Theis, S.V.S. (1924). *How foster children turn out*. New York: State Charities Aid Association.
Thompson, L.A., Detterman, D.K., & Plomin, R. (1993). Differences in heritability across groups differing in ability, revisited. *Behavior Genetics*, *23*, 331–336. http://dx.doi.org/10.1007/BF01067433.
Vinkhuyzen, A.A.E., van der Sluis, S., Maes, H.H.M., & Posthuma, D. (2012). Reconsidering the heritability of intelligence in adulthood: Taking assortative mating and cultural transmission into account. *Behavior Genetics*, *42*, 187–198. http://dx.doi.org/10.1007/s10519-011-9507-9.
Visscher, P.M., Brown, M.A., McCarthy, M.I., & Yang, J. (2012). Five years of GWAS discovery. *American Journal of Human Genetics*, *90*, 7–24. http://dx.doi.org/10.1016/j.ajhg.2011.11.029.

Chapter 4:- Rotation is visualisation, 3D is 2D: using a novel measure to investigate the genetics of spatial ability

This chapter, introducing the “Bricks” battery and beginning the exploration of the aetiological architecture of spatial ability, is presented as a published paper. It is an exact copy of this publication:

Shakeshaft NG, Rimfeld K, Schofield KL, Selzam S, Malanchini M, Rodic M, Kovas Y, Plomin R (2016). Rotation is visualisation, 3D is 2D: using a novel measure to investigate the genetics of spatial ability. *Scientific Reports* 6: 30545. doi:10.1038/srep30545

Supplementary materials for this chapter, as detailed in the text, are attached as Appendix 2.

Rotation is visualisation, 3D is 2D: using a novel measure to investigate the genetics of spatial ability

Nicholas G. Shakeshaft[1], Kaili Rimfeld[1], Kerry L. Schofield[1], Saskia Selzam[1], Margherita Malanchini[2], Maja Rodic[3], Yulia Kovas[2,4] & Robert Plomin[1]

Spatial abilities–defined broadly as the capacity to manipulate mental representations of objects and the relations between them–have been studied widely, but with little agreement reached concerning their nature or structure. Two major putative spatial abilities are "mental rotation" (rotating mental models) and "visualisation" (complex manipulations, such as identifying objects from incomplete information), but inconsistent findings have been presented regarding their relationship to one another. Similarly inconsistent findings have been reported for the relationship between two- and three-dimensional stimuli. Behavioural genetic methods offer a largely untapped means to investigate such relationships. 1,265 twin pairs from the Twins Early Development Study completed the novel "Bricks" test battery, designed to tap these abilities in isolation. The results suggest substantial genetic influence unique to spatial ability as a whole, but indicate that dissociations between the more specific constructs (rotation and visualisation, in 2D and 3D) disappear when tested under identical conditions: they are highly correlated phenotypically, perfectly correlated genetically (indicating that the same genetic influences underpin performance), and are related similarly to other abilities. This has important implications for the structure of spatial ability, suggesting that the proliferation of apparent sub-domains may sometimes reflect idiosyncratic tasks rather than meaningful dissociations.

Spatial ability is one of the most widely-studied domains of cognitive ability, yet there is little consensus as to its nature or structure. It has been found to be a strong predictor of important outcomes, such as science, technology, engineering and maths (STEM) performance[1], but its usefulness in this regard is limited by the lack of understanding about its basic architecture. Broadly defined, the spatial domain comprises the processes involved in perceiving, memorising and manipulating mental representations of visual scenes[2], including two-dimensional (2D) and three-dimensional (3D) objects[1,3] and the relationships between them[4]. Putative processes, categories and sub-domains–such as visualisation[5], spatial orientation[6], mental rotation[7], spatial relations[6] and many others–have proliferated in the literature, often with overlapping definitions, to the extent that the term "spatial ability" itself is difficult even to define with precision[8,9].

A great many spatial tests have been developed and are commonly used, with varying intercorrelations among them, and several theories have been proposed to describe the multifactorial structure suggested by these relationships[4,9,10]. Two major putative sub-domains (among many others) are "mental rotation" and "visualisation". Definitions vary, but mental rotation involves rotating mental models of objects into different orientations, and visualisation describes various complex mental manipulations of spatial information, including identifying hidden or partially occluded objects from incomplete information[11]. Theories differ as to the nature of these abilities and the relationship between them, with some proposing that they represent distinct sub-domains of spatial ability[5], while others suggest that visualisation is a major sub-domain, of which mental rotation is merely a component or exemplar[9]. Similarly, investigating the effects of the dimensionality of stimuli has led to contradictory

[1]Medical Research Council Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, United Kingdom. [2]Goldsmiths, University of London, New Cross, London, United Kingdom. [3]University of Sussex, Sussex House, Falmer, Brighton, United Kingdom. [4]Tomsk State University, Tomsk, Russia. Correspondence and requests for materials should be addressed to N.G.S. (email: nicholas.shakeshaft@kcl.ac.uk)

results, with some studies[3,12] finding differences between the processing of 2D and 3D stimuli, and other results[9,13] suggesting otherwise. One possible explanation for some of the inconsistent findings in the literature is that the available tests may not be "pure", in the sense that their items may conflate multiple cognitive processes such that factor analyses cannot distinguish them[4]. Another possibility, primarily concerning the apparent distinction between 2D and 3D stimuli, is that test items may differ substantially in complexity[3].

Behavioural genetic methods may provide a different perspective, as yet largely unexplored, from which to clarify the nature of spatial abilities and the aetiology of their interrelationships. These methods concern individual differences, rather than the normative focus of much cognitive work. Several studies have observed substantial familiality (i.e., resemblance among related individuals) for spatial abilities[14–17]. Adoption[18] and twin[19–21] studies have found this familiality to be substantially genetic in origin, with average heritability estimates at around 50% for spatial ability in adulthood. However, for the purpose of elucidating the structure of individual differences within and between domains, multivariate genetic analyses–permitting calculation of the genetic and environmental influences shared between multiple observed traits[22]–are more informative: if two traits are meaningfully and fundamentally dissociable (in their neurobiological basis, for example), we might reasonably predict this to be reflected in their genetic aetiology. Such methods have been applied to investigate the degree to which spatial ability shares common genetic influences with other cognitive domains such as mathematical ability[23], finding a moderate overlap. However, to date no multivariate genetic studies have been published examining the genetic architecture *within* the spatial domain itself.

Thus the present study had two main aims. First, a novel battery of spatial tests was developed and validated with the express purpose of allowing i) mental rotation and ii) visualisation *without* rotation (e.g., picturing a whole object from incomplete information) to be tested in isolation from one another, using both 2D and 3D stimuli of approximately equivalent complexity. In this way, the relationship between mental rotation and visualisation, and between 2D and 3D stimuli, could be examined without confounds. Second, this new battery was administered to a large twin sample, together with other cognitive measures, in order to assess the extent to which any dissociation between these different types of stimuli may be attributed to genetic or environmental factors.

Results

Data. The Twins Early Development Study (TEDS) is a longitudinal cohort study of more than 10,000 pairs of British twins, born between 1994 and 1996. The sample is representative of the population of the United Kingdom, and has been described previously[24]. For the present study, a representative subsample was selected from among the older twins in the cohort, who had completed a battery of cognitive tests on a previous occasion (at age 16), assessing their verbal ability (with the Mill Hill Vocabulary Scale[25]) and non-verbal ability (Raven's Progressive Matrices[26]), from which a proxy of their general cognitive ability (*g*) could be derived as the mean of these two standardised scores.

This TEDS subsample was asked to complete a novel battery of spatial tests: the "Bricks" battery. This consisted of six subtests, assessing either mental rotation alone, spatial visualisation alone (without rotation), or both together, using either two- or three-dimensional stimuli. Three "functional" composites ("Rotation", "Visualisation", and "Rotation/Visualisation combined", each being the mean of the 2D and 3D subtests of that type), and two "dimensional" composites ("2D" and "3D", each being the mean of the three corresponding subtests) were derived from these subtest scores. As a marker of overall spatial ability (for reference), an "Overall Bricks" composite was also derived as the mean of all six subtest scores. Details are presented in Methods, with examples of the stimuli in Fig. 1. These stimuli were prepared using purpose-built software allowing computer-generated objects to be manipulated dynamically; this software is freely available here: https://www.forepsyte.com/resources/public

The data were cleaned and prepared as described in Methods. The final dataset comprised 2,913 participants: 1,250 twin pairs (528 monozygotic (MZ), 722 dizygotic (DZ)), and an additional 413 unpaired individuals (104 from MZ and 309 from DZ pairs). The participants were 63% female (the gender imbalance reflecting a disparity in response rates), with a mean age of 20.3 years ($\pm$0.47 SD) on completing the Bricks tests.

Sample sizes and descriptive statistics for the Bricks subtests and composites, and the other measures, are presented in Supplementary Table S1. The reliability of the Bricks battery was assessed with regard to Cronbach's alphas, and also test-retest correlations in an independent pilot sample. Bricks composite alphas ranged from 0.63 to 0.85, and test-retest correlations from 0.62 to 0.83; for details, see Supplementary Table S2.

For each measure, an analysis of variance assessed the mean effects of sex and zygosity (Supplementary Table S1). As is often observed for spatial abilities[27], a main effect of sex was found for all Bricks measures, representing a slight male advantage (average $R^2 = 0.03$ for the Bricks composites). Mean sex differences are irrelevant to twin analyses, which examine variances, but common practice for twin studies is to analyse sex- and age-corrected residuals (see Methods). For all subsequent analyses, the data were regressed on age and sex, normality-transformed and standardised.

Phenotypic analyses. For all phenotypic analyses, one twin was selected at random per pair to create an independent sample.

Preliminary analyses suggested immediately that the putative distinctions in some of the literature between mental rotation and spatial visualisation, and between 2D and 3D stimuli, were not supported. The modest intercorrelations among the six subtest scores (*r* ranging from 0.25 to 0.42; see Supplementary Table S3) revealed no apparent clusters of stronger or weaker associations. For example, the 2D subtests showed no consistently stronger correlations with one another than with the 3D subtests, nor were the Rotation subtests associated more substantially with each other than with the Visualisation subtests. To examine this more formally, the subtest scores were subjected to factor analysis, producing only a single factor on which all six subtests were strongly loaded (with factor loadings ranging from 0.57 to 0.70; see Supplementary Table S4).

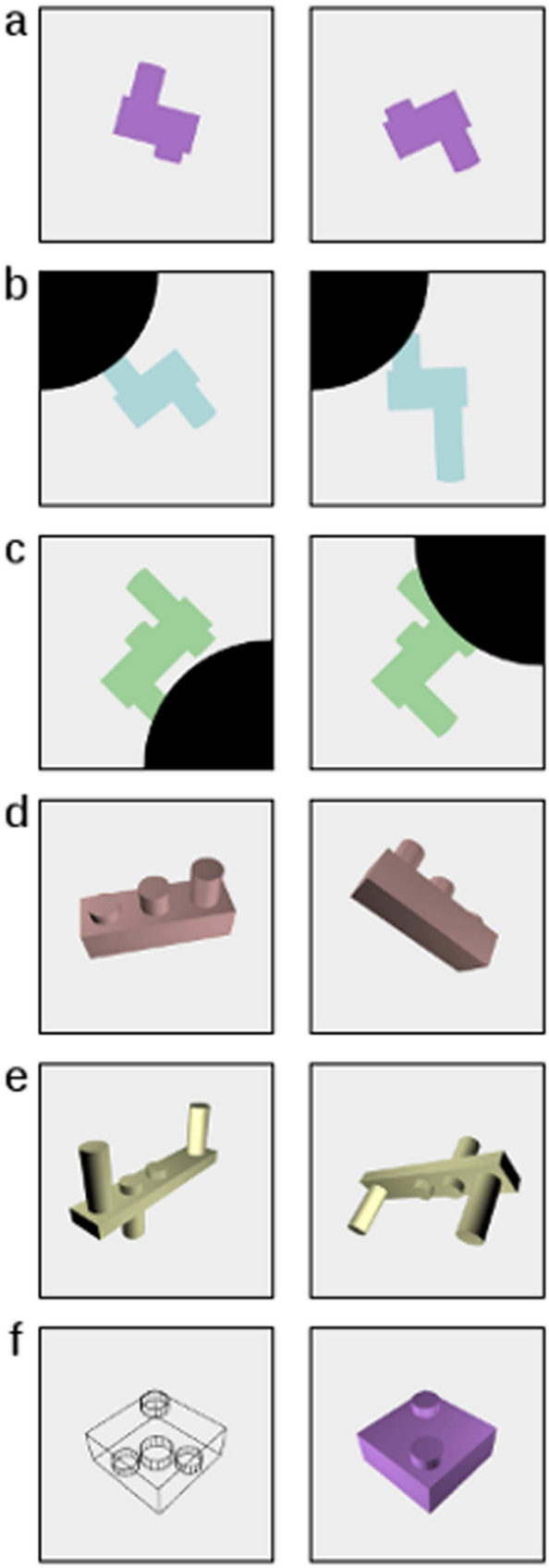


Figure 1. Sample stimuli. Sample target images (left) and correct responses (right) for the six Bricks subtests: (**a**) 2D Rotation, (**b**) 2D Rotation/Visualisation combined, (**c**) 2D Visualisation, (**d**) 3D Rotation/Visualisation combined, (**e**) 3D Rotation, and (**f**) 3D Visualisation.

	Intrapair twin correlations		Variance component estimates			Sample (numbers of pairs)	
	MZ	DZ	h^2	c^2	e^2	MZ	DZ
Rotation	0.33 (0.26–0.41)	0.21 (0.14–0.28)	0.25	0.09	0.67	520	714
Rotation/Visualisation	0.38 (0.31–0.46)	0.22 (0.14–0.28)	0.34	0.05	0.62	521	714
Visualisation	0.45 (0.38–0.51)	0.22 (0.15–0.29)	0.45	0.00	0.55	516	711
2D	0.47 (0.40–0.53)	0.25 (0.18–0.31)	0.44	0.02	0.53	526	724
3D	0.41 (0.33–0.48)	0.20 (0.13–0.27)	0.41	0.00	0.59	508	697
Overall Bricks	0.56 (0.49–0.61)	0.27 (0.20–0.33)	0.56	0.00	0.44	522	720

Table 1. Twin correlations and approximated variance components. Intraclass twin correlations (95% confidence intervals) for MZ and DZ twins, for the Bricks composites. Variance component estimates are heritability (h^2: double the difference between the MZ and DZ correlations, constrained not to exceed the former–MZ twins are genetically identical, so heritability cannot exceed their correlation), shared environment (c^2: the MZ correlation minus h^2), and unique environment + error of measurement (e^2: 1-h^2-c^2). Sample sizes shown are complete pairs, after exclusions and data cleaning.

However, it must be noted that the subtests were not intended for use in this way, being very short individually in comparison to most cognitive tests–and thus not very highly reliable–in order to keep the administration of the whole battery within a reasonable time limit. The results from the individual subtests should therefore be treated with caution, and the Bricks composites were created on the original theoretical grounds, to assess whether clearer distinctions might emerge from the more reliable constructs.

The resulting functional composites were moderately intercorrelated. If mental rotation and spatial visualisation are functionally distinct, we would predict the Rotation and Visualisation composites to be correlated more modestly with each other than either is with Rotation/Visualisation combined. In fact, the results showed that the association between Rotation and Visualisation ($r = 0.46$, $p < 0.0001$, $N = 1411$) was identical to that between Rotation and Rotation/Visualisation combined ($r = 0.46$, $p < 0.0001$, $N = 1423$), and the correlation between Visualisation and Rotation/Visualisation combined ($r = 0.54$, $p < 0.0001$, $N = 1426$; the slight variations in sample size result from losses during data cleaning, described in the Supplementary Methods online) did not differ substantially (although the small difference was significant in this large sample; $p < 0.001$). However, these correlations are far from unity, as is that between the 2D and 3D composites ($r = 0.56$, $p < 0.0001$, $N = 1413$), which suggests some specificity between the composites. The nature of this specificity is the subject of the multivariate genetic analyses below.

The Bricks composites correlated modestly with verbal ability (average $r = 0.20$), and moderately with non-verbal ability ($r = 0.43$) and *g* ($r = 0.38$); see Supplementary Table S5. It was considered that the associations among the Bricks scores could be driven in part by more domain-general abilities or processes captured by these other measures, which could potentially obscure the "true" relationships among the Bricks subtests and composites. Accordingly, the Bricks subtests and composites were regressed separately on verbal ability (a conservative under-correction for domain-general processes; see Methods), on non-verbal ability (perhaps an over-correction including some of the variance in spatial ability, reflected in its higher correlations with Bricks), and on *g* (their mean). The strength of the relationships among the resulting subtest and composite residuals was reduced slightly and uniformly, with no different patterns emerging among either the subtests (see Supplementary Tables S6–S8) or composites (Supplementary Tables S9–S12). The factor analysis results were similarly unaffected (Supplementary Table S13), implying that *g* does not mask differentiation among the spatial subtests.

Univariate genetic analyses. Intraclass twin correlations are presented in Table 1 for the Bricks composites, and in Supplementary Table S14 for the Bricks subtests and other cognitive measures. These intraclass correlations may be used to calculate initial estimates for the "heritability" (additive genetic influences), "shared environment" (environmental factors promoting similarity) and "non-shared" or "unique environment" (environmental factors not contributing to similarity between twins, and also any measurement error) influencing the trait–see Table 1 for details. The resulting estimates (Table 1) indicate substantial genetic influence on all measures, up to 56% for the Overall Bricks composite.

To establish these estimates more precisely, and to obtain model fit statistics and confidence intervals (CIs), the data for each measure were subjected to maximum-likelihood model-fitting to estimate the portions of variance attributable to additive genetic (A), shared environmental (C) and non-shared (unique) environmental components (E, also including measurement error). See Methods for details. The results confirm that all Bricks composites are moderately heritable (Table 2), with no significant differences in the magnitude of the genetic influences between the various functional composites, or between the two dimensional composites. There were substantial non-shared, but no significant shared environmental influences. Results for the individual Bricks subtests and other cognitive measures are presented for reference in Supplementary Table S15.

Multivariate genetic analyses. Bivariate correlated factors solutions (see Methods) were fitted to each pair of Bricks composites in turn, from which their phenotypic correlations could be decomposed into the proportions attributable to genetic, shared and non-shared environmental influences. The results (Fig. 2, with precise estimates and CIs in Supplementary Table S16) indicate that the phenotypic correlations are largely (70–80%) genetic in origin, with the remainder due to non-shared environmental influences. Similar patterns appear between the individual subtests (Supplementary Tables S17 and S18). The correlations between the Bricks composites and the

	A	C	E
Rotation	0.23 (0.03–0.40)	*0.10 (0.00–0.26)*	0.67 (0.60–0.75)
Rotation/Visualisation	0.34 (0.14–0.45)	*0.05 (0.00–0.20)*	0.62 (0.55–0.69)
Visualisation	0.43 (0.24–0.50)	*0.01 (0.00–0.16)*	0.56 (0.50–0.63)
2D	0.45 (0.27–0.52)	*0.02 (0.00–0.16)*	0.53 (0.48–0.60)
3D	0.41 (0.22–0.47)	*0.00 (0.00–0.15)*	0.59 (0.53–0.66)
Overall Bricks	0.55 (0.42–0.60)	*0.00 (0.00–0.11)*	0.45 (0.40–0.50)

Table 2. Univariate model-fitting results. Model-fitting estimates (95% confidence intervals) for additive genetic (A), shared environmental (C) and residual (E; i.e., non-shared environment and error) components of variance. Italicised estimates are non-significant (their confidence intervals include zero).

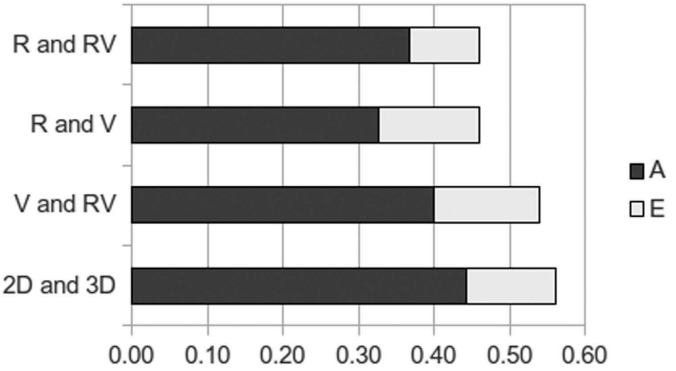


Figure 2. Decomposition of phenotypic correlations. Correlated factor solution analyses, indicating the proportion of the phenotypic correlations (line length) among the Bricks composites attributable to genetic (**A**) shared environmental (**C**) and non-shared environmental influences/error (**E**). R = Rotation, RV = Rotation/Visualisation combined, V = Visualisation.

other cognitive measures are also substantially genetically driven, with shared genetic influences accounting for approximately all of the relationships with verbal ability, and a majority (64% on average) of the stronger relationships with non-verbal ability (Supplementary Table S19).

As these results only decompose the phenotypic correlations, they do not directly estimate the portions of variance that are unique to each variable–that is, they do not reveal what proportions of the *total* influences on each composite are shared with others. This is the purpose of Cholesky decomposition (Methods). These results (Fig. 3 and Supplementary Tables S20–S23) suggest, for each bivariate relationship among the Bricks composites, that 100% of the substantial genetic influences on each composite measure is shared with all the others. This can be seen in Fig. 3: in each model, all of the genetic variance of the second variable (on the right) is shared with the first, resulting in a loading of 0 for the residual genetic path for the second variable.

This pattern is revealed even more starkly by the genetic correlations, which indicate the correlation between genetic influences on the two variables independent of their heritabilities (Methods). These are all at unity among the Bricks composites (Supplementary Tables S24 and S25). Even for the comparatively unreliable individual subtests, the genetic correlations are all either at unity or have CIs including unity (Supplementary Table S26).

As there are no significant shared environmental influences on any of the Bricks measures, there are no meaningful correlations between these components. However, the correlations between *non*-shared environmental influences (Supplementary Tables S24, S25 and S27) indicate that there are modest "unique" environmental effects in common between the measures (i.e., effects unique to each individual, but affecting multiple traits), up to a maximum rE = 0.23 between Bricks composites.

The genetic correlations between the Bricks composites and the other cognitive measures (Supplementary Table S28) indicate a substantial genetic overlap (average rA = 0.55) with verbal ability, higher still with non-verbal ability (average rA = 0.71), and the association with *g* (their mean) unsurprisingly in between (average rA = 0.65).

As with the phenotypic results, it was considered that the genetic associations among the Bricks measures could reflect domain-general influences shared with other cognitive abilities, too, rather than influences specific to spatial abilities. Multivariate Cholesky decompositions (see Methods) were performed for Rotation and Visualisation, and for 2D and 3D, first accounting for the genetic influences on verbal ability, non-verbal ability, or both, and then examining the residual relationships between the Bricks composites. In these trivariate models, verbal ability accounts for less than one third of the heritability of the Bricks composites, non-verbal ability for around half (but the difference is non-significant), and *g* (their mean) in between. In two quadrivariate models (entering verbal and non-verbal ability separately, then Rotation and Visualisation or 2D and 3D), the verbal and non-verbal cognitive measures accounted in total for around half of the heritability of the Bricks measures. In every model, substantial genetic influence remains that is unique to spatial ability as a whole, supporting it as a distinct cognitive domain from *g*. However, none of the genetic variance is unique to any specific Bricks composite–all genetic influences are shared between all Bricks measures.

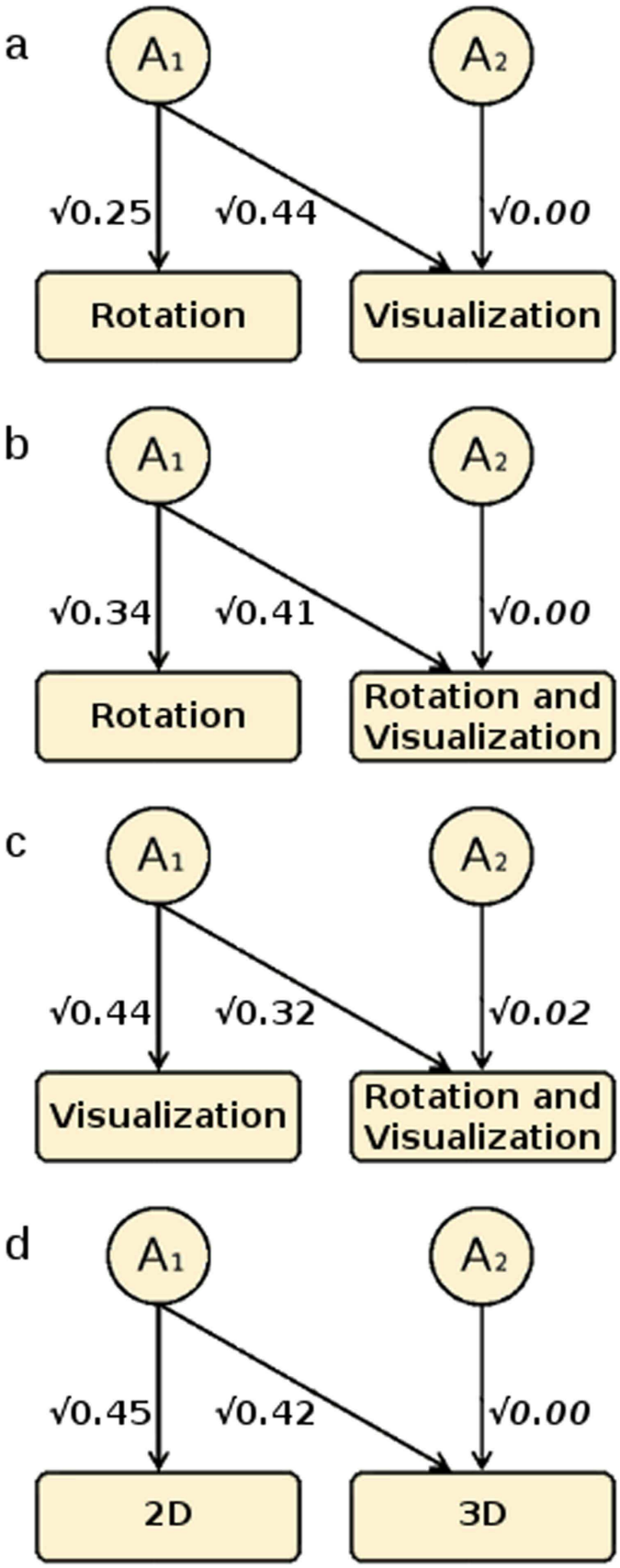


Figure 3. Decomposition of heritability. Four bivariate Cholesky decompositions indicating the genetic relationship between (**a**) Rotation and Visualisation, (**b**) Rotation and Rotation/Visualisation combined, (**c**) Visualisation and Rotation/Visualisation combined, and (**d**) 2D and 3D. Independent paths (italicised) are all non-significant.

Detailed results are presented in the Supplementary Materials online: Supplementary Figs S1 and S2 for illustration, and full details in Supplementary Tables S29–S36. Fit statistics for the Bricks composite models are presented in Supplementary Tables S37–S40.

Discussion

The Bricks battery was designed with the express purpose of differentiating between mental rotation and spatial visualisation, and to assess them equally in 2D and in 3D. The key multivariate genetic results all show a strong and consistent pattern, for the functional composites (Rotation, Visualisation, Rotation/Visualisation combined), dimensional composites (2D, 3D), and even for the individual subtests: it is impossible, genetically at least, to distinguish between any of these spatial constructs (Fig. 3). Once the genetic influences on any one of these measures are accounted for, nothing remains. As specific genes are identified that are associated with any of these spatial abilities, it is expected that these genes will be similarly associated with all of them.

Phenotypically, the results arguably present a more ambiguous picture, since the intercorrelations are modest among the Bricks subtests, and moderate (average $r = 0.51$) even among the more reliable composites. There are many reasons why phenotypic correlations might be imperfect, of course, without this reflecting theoretically meaningful dissociations–there can be unintended test-specific differences, for example. However, the most likely explanation in this instance is reliability: the test-retest correlations for the Bricks composites are respectable but far from unity (average $r = 0.69$), so the measures do share a large majority of their *reliable* phenotypic variance (i.e., 74% overall). In any case, the other phenotypic results show no evidence of any dissociations: factor analysis produces only a single factor with no substantial differences in loadings between the subtests, and the Bricks measures all present very similar patterns of correlations with the other cognitive measures assessed. Taken together, there is no more evidence of meaningful dissociations phenotypically than genetically.

While the genetic associations between the Bricks measures account for a majority of the phenotypic correlations between them (Fig. 2), a significant minority is driven by modest correlations between their non-shared environmental influences (Supplementary Tables S24, S25 and S27); i.e., E in the ACE models (Methods). These are environmental influences unique to each participant, making co-twins less similar to one another, but which influence multiple traits and increase their correlations–these could be personal traits affecting performance across multiple tests, or indeed situational factors such as the participant's testing environment. This non-shared component is the only source of environmental influence common to multiple Bricks measures, and the absence of any significant *shared* environmental influences (i.e., C in the ACE models) is striking. For the Bricks measures and everything they capture, genetic influences are the only source of familial similarity.

The tests were developed specifically to differentiate cleanly between mental rotation and spatial visualisation. The lack of any genetic (or even any unambiguous phenotypic) specificity between the Rotation and Visualisation composites would seem to provide strong support, therefore, for the previous literature[9] suggesting that they do not represent meaningfully dissociable tasks, and to refute the suggestions[5] to the contrary. While we cannot draw any conclusions about the specific mechanisms of action of any influences, it also suggests an absence of distinguishable cognitive processes underlying them. Stated more boldly, mental rotation is nothing more than visualisation, and likewise visualisation recruits no distinct processes even when rotation is not required. Where differentiation has been observed previously in this area, it seems plausible that this reflects task-specific effects or reliability issues, rather than theoretically meaningful differences.

Some of the previous reports of dissociation between 2D and 3D stimuli suggested that the difference might relate to 3D objects being more complex, and therefore more time being required to encode their mental representations[3]. While response times were not included directly in the Bricks scores reported here, the 2D and 3D Bricks composites were intended to be approximately equal in difficulty, and the inclusion of restrictive item time limits (see the Supplementary Methods online) would have been expected to affect scores if the 3D items had been substantially harder than the 2D items; there is no evidence of this (indeed the 3D mean score is marginally higher than 2D; Supplementary Table S1). This suggests that the 2D and 3D Bricks composites are indeed of broadly equivalent difficulty. Coupled with the clear lack of differentiation between these composites in the results, this supports the contention that differences in difficulty–rather than fundamental differences in the processes involved–are responsible for the dissociations sometimes observed.

It must be emphasised that there are a great many putative sub-domains of spatial ability not included in the present study. Likewise, even the definition of "visualisation" used here is quite narrow–definitions vary in the literature, but visualisation is sometimes taken to include more complex mental manipulations than those operationalised in the Bricks measures. The present results should not be over-interpreted beyond the abilities assessed, therefore, but it is hoped that they may indicate a fruitful approach. In subsequent work, we will apply these methods to more diverse abilities sampled from across the spatial domain.

The importance of spatial ability for outcomes such as STEM performance[1] is well documented, and it is to be hoped that clarifying the nature and structure of this domain will refine its measurement and increase its utility further. It should be noted that, while no differentiation *within* the spatial domain was supported by these results, the correlations between the Bricks measures and the other cognitive measures examined were only moderate, both phenotypically and genetically (Supplementary Tables S5 and S28), despite the probable inclusion of some spatial elements within the non-verbal cognitive measure itself (Methods). This certainly supports the existence of spatial ability as a distinct cognitive domain in its own right.

As noted above, the structure of this distinct spatial domain is hotly contested, and seemingly always growing in its apparent size and complexity. Where previous findings have suggested meaningful dissociations between visualisation and mental rotation, though, and between 2D and 3D stimuli, the present study suggests that it is possible to shrink it, too.

Methods

Measures. The Bricks battery comprises six subtests of nine items each (12 items of each type were actually administered, so that the nine psychometrically best-performing items could be selected to form the final battery). Each item consisted of a target stimulus image depicting a 2D or 3D object (a "brick"), and four multiple-choice response images, one of which (the correct answer) showed the same object as the target, following an appropriate manipulation. Correct answers were summed to create subtest scores, from which composite scores were derived as described in Results. Participants completed the subtests in the following sequence. i) 2D Rotation: the 2D target object is rotated in the picture plane. ii) 2D Rotation/Visualisation combined: the rotating target is partially obscured behind an (immobile) occluding shape. iii) 2D Visualisation: the target remains static while the occluding shape changes location. iv) 3D Rotation/Visualisation combined: the object rotates freely in three dimensions. v) 3D Rotation: the 3D object rotates only in the picture plane. vi) 3D Visualisation: the target is a wireframe diagram, and the correct response is the "solid" object depicted. Examples of stimuli (targets and correct responses) are presented in Fig. 1, and these measures are described in greater detail in the Supplementary Methods online.

Two other cognitive measures were also available for this sample. The Mill Hill Vocabulary Scale[25] was used as an index of verbal ability: across 33 trials, participants selected which of six multiple-choice options was closest in meaning to a target word. Non-verbal ability was assessed with Raven's Progressive Matrices[26], in which participants selected which of eight options completed a visual pattern, across 30 trials. Correct responses for each measure were summed and standardised, and the mean of these scores was used as a proxy of general cognitive ability (*g*). Participants completed these measures four years earlier than the Bricks battery, but since the genetic influences on *g* are highly stable over time[28,29], this is unlikely to have influenced results. Where these measures were used as a control for domain-general cognitive processes, it should be noted that the verbal ability measure is probably an under-correction (as verbal ability is only a portion of *g*[22]), and that the non-verbal ability measure is in all likelihood an *over*-correction, as Raven's Progressive Matrices have a substantial spatial component[30].

Participants were contacted by post, but participated online via the TEDS websites. The measures administered at age 16 were implemented using the Flash browser plugin. The Bricks items were developed with "Building Bricks", a web application developed for the purpose, and administered using the "psy.js" JavaScript library; both of these tools are open-source and freely available (see the Supplementary Methods online).

Twin data. DZ twins share 50% of their segregating genes on average, while MZ twins share 100%, but environments are shared to approximately the same extent for both MZ and DZ twins. Genetic influence on a trait is therefore indicated by the degree to which the intrapair MZ correlation exceeds the DZ correlation, and cross-twin cross-trait correlations (i.e., the correlation between twin 1 on the first trait and twin 2 on the second) allow the genetic influences common to multiple traits to be estimated.

MZs and same-sex DZs are perfectly correlated for sex, and all twins are for age; it is therefore common practice to regress twin data on sex and age, to avoid the artificially inflated estimates of shared environmental influences which would otherwise result[31]. In addition, for each measure in the present study, outliers beyond 3 SD from the mean were removed, along with any data for those participants suspected to have suffered technical errors or to have responded randomly or carelessly (see the Supplementary Methods online). Participants with severe physical or psychological disabilities, or whose mothers had experienced serious perinatal complications, were also excluded from analysis. All variables were standardised, and since the Bricks variables were slightly skewed, a van der Waerden rank transformation[32] was performed to ensure that all data were normally distributed, as required for the model-fitting procedures.

The study was approved by the appropriate King's College London ethics committee, and was conducted in accordance with the approved guidelines. Participants provided informed consent.

Model-fitting. The data were subjected to full-information maximum-likelihood (FIML) model-fitting procedures, accounting for missing data and combining both same- and opposite-sex DZ twins to maximise power. Univariate ACE models[33] were fitted to the data, which use the expected genetic and environmental correlations between the twins (additive genetic influences correlating 1.0 for MZs and 0.5 for DZs; shared environment 1.0 for both; non-shared environment 0 for both) to apportion the variance into components attributable to: i) additive genetic influences (A); ii) shared (or "common") environmental influences making people raised in the same family more similar to each other (C); and iii) non-shared (unique) environmental influences making them less similar (E, which also includes any measurement error). Individual components may be dropped in nested sub-models, but the full ACE models were used here despite C being non-significant for the Bricks measures, both because this tends to produce the most conservative heritability estimates, and for consistency with the other cognitive measures used (as C is significant for Raven's Progressive Matrices; see Supplementary Table S15). All model-fitting was conducted using OpenMx[34], an R package for structural equations.

Multivariate ACE model-fitting uses cross-twin cross-trait correlations[22] to estimate the genetic and environmental sources of covariance, revealing the architecture underpinning two or more traits[35]. This calculates the genetic correlations (rA) between each pair of variables, which are independent from the heritability estimates of either trait, and indicate the degree to which they share genetic influences–i.e., common genes. A "correlated factors" solution then estimates common A, C and E influences, and thus allows phenotypic correlations to be decomposed into these sources of covariance (as in Fig. 2 and Supplementary Tables S16–S19). Alternatively, the (algebraically equivalent) Cholesky decomposition focuses instead on the *total* influences on each trait in sequence, and determines at each step the proportion of its A, C and E components that are shared with, or independent from, each variable. This process is analogous to stepwise multiple regression, accounting for the influences on each variable in turn, in order to determine the residual portions at each stage. Thus in bivariate models (as in Fig. 3 and Supplementary Tables S20–S23), path estimates show the proportion of each component

that is common to both variables, and the proportion unique to the second variable. Similarly, trivariate and further extensions (as in Supplementary Figs S1 and S2 and Supplementary Tables S29–S36) indicate the influences in common to all variables, then those common to all but the first, and so on, and finally those influences unique to the last variable.

References

1. Wai, J., Lubinski, D. & Benbow, C. P. Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *J. Educ. Psychol.* **101,** 817–835 (2009).
2. Lohman, D. F. In *Encyclo*pedia *of* int*elligence* (ed. Sternberg, R. J.) 2, 1000–1007 (Macmillan, 1994).
3. Shepard, S. & Metzler, D. Mental rotation: effects of dimensionality of objects and type of task. *J. Exp. Psychol. Hum. Percept. Perform*. **14,** 3–11 (1988).
4. Hegarty, M. & Waller, D. A. In *T*he *Ca*mbridge Handb*ook of Visuospatial Thinking* (eds. Shah, P. & Miyake, A.) (Cambridge University Press, 2005).
5. Lohman, D. F. *Spatial Ability: A Review and Reanalysis of the Correlational Literature*. **8,** 226 (1979).
6. Lohman, D. F., Pellegrino, J. W., Alderton, D. L. & Regian, J. W. In *Intelligence and Cognition: Contemporary Frames of R*eferenc*e* (eds. Irvine, S. H. & Newstead, S. E.) 253–312 (Springer Netherlands, 1987).
7. Shepard, R. N. & Metzler, J. Mental rotation of three-dimensional objects. *Science* **171,** 701–703 (1971).
8. Eliot, J. & Smith, I. M. *An International Directory of Spatial Tests*. (NFER-Nelson, 1983).
9. Carroll, J. B. *Human cognitive abilities: a survey of factor-analytic studies*. (Cambridge University Press, 1993).
10. Kozhevnikov, M. & Hegarty, M. A dissociation between object manipulation spatial ability and spatial orientation ability. *Mem. Cognit*. **29,** 745–756 (2001).
11. Linn, M. C. & Petersen, A. C. Emergence and characterization of sex differences in spatial ability: a meta-analysis. *Child Dev.* **56,** 1479–1498 (1985).
12. Pellegrino, J. W. & Kail, R. In *Advances in the psychology of human intelligence* (ed. Sternberg, R. J.) 1, 311–366 (Lawrence Erlbaum Associates, 1982).
13. Ho, C.-H., Eastman, C. & Catrambone, R. An investigation of 2D and 3D spatial and mathematical abilities. *Des. Stud*. **27,** 505–524 (2006).
14. DeFries, J. C., Vandenberg, S. G. & McClearn, G. E. Genetics of specific cognitive abilities. *Annu. Rev. Genet.* **10,** 179–207 (1976).
15. Loehlin, J. C., Sharan, S. & Jacoby, R. In pursuit of the 'spatial gene': a family study. *Behav. Genet.* **8,** 27–41 (1978).
16. McGee, M. G. Human spatial abilities: psychometric studies and environmental, genetic, hormonal, and neurological influences. *Psychol. Bull.* **86,** 889–918 (1979).
17. Smalley, S. L., Thompson, A. L., Spence, M. A., Judd, W. J. & Sparkes, R. S. Genetic influences on spatial ability: transmission in an extended kindred. *Behav. Genet.* **19,** 229–240 (1989).
18. Alarcón, M., Plomin, R., Fulker, D. W., Corley, R. & DeFries, J. C. Multivariate path analysis of specific cognitive abilities data at 12 years of age in the Colorado Adoption Project. *Behav. Genet.* **28,** 255–264 (1998).
19. Pedersen, N. L., Plomin, R., Nesselroade, J. R. & McClearn, G. E. A quantitative genetic analysis of cognitive abilities during the second half of the life span. *Psychol. Sci.* **3,** 346–352 (1992).
20. McClearn, G. E. *et al.* Substantial genetic influence on cognitive abilities in twins 80 or more years old. *Science* **276,** 1560–1563 (1997).
21. Rietveld, M. J. H., Dolan, C. V., Baal, G. C. M. van & Boomsma, D. I. A Twin Study of Differentiation of Cognitive Abilities in Childhood. *Behav. Genet.* **33,** 367–381 (2003).
22. Plomin, R., DeFries, J. C., Knopik, V. S. & Neiderhiser, J. M. *Behavioral genetics*. (Worth Publishers, 2013).
23. Tosto, M. G. *et al.* Why do spatial abilities predict mathematical performance? *Dev. Sci.* **17,** 462–470 (2014).
24. Haworth, C. M. A., Davis, O. S. P. & Plomin, R. Twins Early Development Study (TEDS): A Genetically Sensitive Investigation of Cognitive and Behavioral Development From Childhood to Young Adulthood. *Twin Res. Hum. Genet.* **16,** 117–125 (2012).
25. Raven, J., Raven, J. C. & Court, J. H. In *Manual for Raven's Progressive Matrices and Vocabulary Scales* (Harcourt Assessment, 1998).
26. Raven, J. C., Court, J. H. & Raven, J. In *Manual for Raven's Progressive Matrices and Vocabulary Scales* (Oxford Psychologists Press, 1996).
27. Voyer, D., Voyer, S. & Bryden, M. P. Magnitude of sex differences in spatial abilities: a meta-analysis and consideration of critical variables. *Psychol. Bull.* **117,** 250–270 (1995).
28. Deary, I. J. *et al.* Genetic contributions to stability and change in intelligence from childhood to old age. *Nature* **482,** 212–215 (2012).
29. Lyons, M. J. *et al.* Genes determine stability and the environment determines change in cognitive ability during 35 years of adulthood. *Psychol. Sci.* **20,** 1146–1152 (2009).
30. Schweizer, K., Goldhammer, F., Rauch, W. & Moosbrugger, H. On the validity of Raven's matrices test: Does spatial ability contribute to performance? *Personal. Individ. Differ*. **43,** 1998–2010 (2007).
31. McGue, M. & Bouchard, T. Jr. Adjustment of twin data for the effects of age and sex. *Behav. Genet.* **14,** 325–343 (1984).
32. Lehmann, E. L. *Nonparametrics: Statistical Methods Based on Ranks*. (Springer, 2006).
33. Rijsdijk, F. V. & Sham, P. C. Analytic approaches to twin data using structural equation models. *Brief. Bioinform.* **3,** 119–133 (2002).
34. Boker, S. *et al.* OpenMx: An Open Source Extended Structural Equation Modeling Framework. *Psychometrika* **76,** 306–317 (2011).
35. Loehlin, J. C. The Cholesky approach: A cautionary note. *Behav. Genet.* **26,** 65–69 (1996).

Acknowledgements

We thank the twins in the Twins Early Development Study (TEDS) for making the study possible. TEDS is supported by a program grant to RP from the UK Medical Research Council (MRC) [MR/M021475/1; previously G0901245 and G0500079], with additional support from the US National Institutes of Health [HD044454; HD059215; NIA046938]. NGS and KR are supported by MRC studentships. SS is supported by an MRC studentship and EU Framework Programme 7 [602768]. RP is supported by a Medical Research Council Research Professorship award [G19/2] and a European Research Council Advanced Investigator award [295366].

Author Contributions

N.G.S., K.R., K.L.S., S.S., M.M., M.R., Y.K. and R.P. designed the study. N.G.S. conducted the analyses. N.G.S. and R.P. wrote the manuscript. All authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Shakeshaft, N. G. *et al.* Rotation is visualisation, 3D is 2D: using a novel measure to investigate the genetics of spatial ability. *Sci. Rep.* **6**, 30545; doi: 10.1038/srep30545 (2016).

Chapter 5:- Spatial ability or spatial abilities? Investigating the phenotypic and genetic structure of spatial ability

This chapter, introducing the “King's Challenge” battery and expanding the examination of spatial ability across the breadth of this cognitive domain, has been adapted from a manuscript currently under review at *PNAS*:

Rimfeld K*[1], **Shakeshaft NG***[1], Malanchini M[1,2], Rodic M[3,5], Selzam S[1], Schofield KL[1], Dale PS[4], Kovas Y[2,5], Plomin R[1] (2016). Spatial ability or spatial abilities? Investigating the phenotypic and genetic structure of spatial ability. *PNAS*.

* These authors contributed equally to this work.

1 King's College London, MRC Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology & Neuroscience, London, SE5 8AF, UK

2 Goldsmiths, University of London, Department of Psychology, London, SE14 6NW, UK

3 University of Sussex, Sussex House, Falmer, Brighton BN1 9RH, UK

4 University of New Mexico, Department of Speech and Hearing Sciences, Albuquerque, NM, 87131, USA

5 Tomsk State University, Tomsk, 634050, Russia

Supplementary materials for this chapter, as detailed in the text, are attached as Appendix 3.

Significance

Spatial ability is a strong predictor of several important outcomes, including success in science, technology, engineering and mathematics (STEM) subjects and careers. This ability is widely believed to be multifactorial, with numerous components and sub-domains, such as "mental rotation", "scanning", and "mechanical reasoning." For the first time, this large twin study allows the genetic and environmental aetiology of diverse putative spatial abilities to be explored. The results indicate that this domain is in fact unifactorial, albeit dissociable from general intelligence, suggesting that its structure is much simpler than the sprawling literature suggests. This will aid gene-hunting efforts, and allow this ability and its consequences to be examined with greater precision.

Abstract

Spatial abilities encompass several skills differentiable from general cognitive ability (*g*). Importantly, spatial abilities have been shown to be significant predictors of many life outcomes, even after controlling for *g*. To date, no studies have analysed the genetic architecture of diverse spatial abilities using a multivariate approach. We developed novel, "gamified" measures of diverse putative spatial abilities. The battery of 10 tests was administered online to 1,367 twin pairs (age 19-21) from the UK-representative Twins Early Development Study (TEDS).

We show that spatial abilities constitute a single factor, both phenotypically and genetically, even after controlling for *g*. This spatial ability factor is highly heritable (69%). We draw three conclusions: (1) the high heritability of spatial ability makes it a good target for gene-hunting research; (2) some genes will be specific to spatial ability, independent of *g*; and (3) these genes will be associated with all components of spatial ability.

Introduction

Spatial ability is a vital skill that we use daily to understand and operate within the physical world around us. Spatial ability can be defined as the ability to produce, recall, store and modify spatial relations among objects (1), and to visualise the transformation of these relations due to changes in perspective or other manipulations – although many competing definitions exist (1–4). Spatial ability has a unique role in predicting many life outcomes. It has been found to be a strong predictor of academic achievement and career success in STEM-related fields (Science, Technology, Engineering and Mathematics), even after controlling for *g* (3, 5–8). STEM-related abilities are likely to become ever more important in our rapidly developing technological world, so it is important to understand this cognitive domain better. Research to date suggests that spatial ability includes several factors that are differentiable from general cognitive ability (*g*, intelligence). However, the structure of spatial ability is not clear (2, 9) and little is known about the genetic and environmental aetiology of individual differences. The purpose of the present study is to investigate the structure and aetiology of spatial ability using a genetically sensitive design.

Many components of spatial ability have been proposed, including “spatial visualisation” (complex, multi-stage manipulations of spatial information); “mental rotation” (mentally rotating spatial forms); “spatial relations” (apprehending the relations between objects); “closure speed” (understanding spatial form in the presence of distracting content, such as combining visual stimuli into a meaningful whole); and “closure flexibility” (searching the visual field to find a particular spatial form); as well as other related abilities such as “spatial scanning”, “movement detection”, “mechanical reasoning”, “length estimation”, and “directional thinking”, among many others (9). However, these proposed components of spatial ability often overlap in their definitions and there is little consensus as to the structure of this domain. This could be partly due to the fact that most spatial tests are complex, involving multiple mental processes, such as apprehending and encoding spatial forms, mentally rotating them, using non-verbal reasoning, etc. (10). In addition, spatial manipulations can use 3D or 2D stimuli, and the tests may involve operations between multiple objects (such as combining pieces to make a whole) or within a single object (such as understanding and visualising its structure) (11). These manipulations can be done on a small scale (such as object rotation) or a large scale (such as understanding the map of a building) (12). These processes have been studied in a wide variety of permutations, producing

inconsistent results. It is unclear to what extent these processes are independent, rather than reflecting a single general spatial ability factor.

Even less is known about the genetic architecture of spatial ability than about its phenotypic structure. Family, twin and adoption studies have shown that spatial ability is moderately heritable (30 – 50%), with heritability estimates varying depending on the particular tests used (13–20). There is evidence for partial genetic overlap between spatial ability and general intelligence (with genetic correlations around 0.60, although the estimates vary greatly depending on the spatial measures used (21–23)). However, little is known about the genetic associations among different components of spatial ability. The present study is the first to use a multivariate genetic design to investigate the genetic, as well as phenotypic, architecture among the putative components of spatial ability, as well as the relationship between spatial ability and *g*.

We measured spatial ability using a novel, "gamified" battery of 10 spatial tests that cover a wide range of the major putative factors across this broad domain. Specifically, we investigated three questions: 1) To what extent do genetic factors account for individual differences in spatial ability (or spatial abilities)? 2) Is spatial ability unifactorial or multifactorial, both phenotypically and genetically? 3) To what extent is spatial ability (or the factors of spatial ability) genetically associated with *g*?

Results

Phenotypic analyses

Our battery comprised 10 measures of spatial ability; see Figure 1 for examples and Methods for a description, with full details in the Supplementary Materials (Supplementary Table S10). For our 10 measures of spatial ability, Table S1 presents the means and standard deviations for the whole sample, males and females separately, and for all five sex and zygosity groups: monozygotic (MZ) males, dizygotic (DZ) males, MZ females, DZ females and DZ opposite-sex twin pairs. Males outperformed females by an average of around half a standard deviation (there was no significant effect of zygosity); however, ANOVA results show that sex and zygosity together explain only around 6% of variance on average. For the subsequent analyses, the data were corrected for mean sex differences, as described in Methods.

Exploratory principal components analysis (PCA) was conducted using the 10 spatial measures. One member of each twin pair was randomly selected to maintain the independence of data (the results remained the same when the analysis was repeated after selecting the other member of the twin pair). As shown in Figure 2A, the PCA results indicated that the ten tests assess a single spatial ability factor, suggesting that spatial ability is unifactorial phenotypically. The first principal component accounted for 42% of the variance (See Supplementary Figure S1 for the scree plot and Supplementary Table S2 for the correlation matrix and reproduced correlation matrix). We repeated the analyses after regressing out general cognitive ability (*g)* from the spatial ability scores. Figure 2B shows that the factor structure of spatial ability remains unchanged after correcting the scores for *g*. The first principal component then accounted for 35% of the variance.

Confirmatory factor analysis (CFA) was conducted to test whether the one-factor model of spatial ability fit better than a two-factor solution. CFA, as shown in Table 1A, confirms that spatial ability is unifactorial phenotypically, as the unifactorial model fit significantly better than the two-factor model. All parameters such as AIC and BIC were worse for the two-factor model compared to the one-factor model of spatial ability. The root mean square error approximation was less than 0.05 for the one-factor model, but was 0.16 for the two-factor model, indicating that the one-factor model fits the data much better. The results remained the same when using residuals regressed on *g* scores, as presented in Table 1B.

Since these results clearly indicate a unifactorial structure, the first principal component of spatial ability was used in subsequent analyses as a composite measure of spatial ability.

As a simple check for the possibility that the gamified administration of the tests could inflate their correlations (i.e., by method-specific variance), the main phenotypic analyses were repeated for the non-gamified preliminary pilot data (see Methods). The samples were too small for adequate power, but these analyses nonetheless yielded very similar results to those presented here.

Twin analyses

The full sex-limitation model was used to investigate possible quantitative and qualitative sex

differences (see Methods) for the composite spatial ability score and for the 10 spatial ability tests. We found no evidence for qualitative sex differences for either the composite measure or the individual tests – in other words, the same genetic and environmental factors contributed to the variability in spatial performance for males and females. A few quantitative sex differences emerged for individual spatial ability tests; however, the differences were small when examining the ACE estimates for males and females separately. (Full model fit statistics with nested models are presented in Supplementary Table S3; ACE estimates with 95% confidence intervals for males and females separately are presented in Supplementary Table S4.) Even with over 1300 twin pairs, the sample size is not sufficiently large for sex-limitation models to reliably detect quantitative and qualitative sex differences of this small magnitude (24), so little confidence can be placed in these differences, as is evident from the large confidence intervals around the estimates when calculated for males and females separately. For the general spatial factor, no significant quantitative or qualitative sex differences emerged (see Supplementary Tables S3 and S4 for details). For these reasons and to increase power, the full sample was used in subsequent analyses, combining males and females, and same- and opposite-sex twin pairs.

Figure 3 presents the ACE estimates for the general spatial ability score and for the 10 spatial tests. General spatial ability was substantially heritable (69%), with a small proportion of variance explained by shared environmental factors (8%) and the rest of the variance explained by non-shared environmental factors (23%). Heritability was lower for the individual 10 tests, ranging from 18% to 59%. Twin intra-class correlations and full model fit statistics with confidence intervals are presented in Supplementary Table S5.

Common and independent pathway models were fitted to the data (see Methods). Comparison of the model fit between the common pathway model and the independent pathway model indicated that the independent pathway model was the best fit for the data (see Supplementary Table S6). Figure 4 presents the standardised squared path estimates for the independent pathway model. All spatial tests loaded substantially on the common A factor, with no significant specific genetic influence remaining after controlling for the common genetic factor (Figure 4A). On average, the common A factor accounted for 85% of the heritabilities of the 10 spatial tests (for example, the heritability of the Mazes test was 37% (the sum of the common path: 0.25; and the specific path: 0.12), therefore the proportion of heritability accounted for by the common factor is 0.25/0.37=68%). The spatial tests are differentiated by E factors, which indicate test-specific environmental influences and measurement error

specific to each test. The standardised squared path estimates with 95% confidence intervals are presented in Supplementary Table S7a. Figure 4B shows the results for the same analysis after correcting the spatial scores for *g*. A common genetic factor still explained most of the heritability across the 10 tests, although loadings on the common A factor were reduced by about one third. For these *g*-corrected scores, the common A factor accounts for 79% of the heritabilities of the 10 spatial tests on average. The standardised squared path estimates for the g-corrected model with 95% confidence intervals are presented in Supplementary Table S7b. The results of the common pathway model are presented in Supplementary Table S8 for completeness, but yield the same conclusions.

Cholesky analysis was conducted to assess the extent to which spatial ability is distinct from verbal and non-verbal abilities. As shown in Figure 5, the heritability of spatial ability is estimated at 0.70 (i.e., 0.17 + 0.23 + 0.30) (precise estimates vary between the models used). Of the 0.70 heritability of spatial ability, 24% (0.17/0.70) was shared with verbal ability, an additional 33% (0.23/0.70) was shared with non-verbal ability independent of verbal ability, and 43% (0.30/0.70) of the variance in spatial ability was specific to spatial ability independent of both verbal and non-verbal ability. The small amount of shared environmental influence in all cognitive measures was in common between verbal, non-verbal and spatial measures, while non-shared environmental factors were largely specific to each cognitive measure.

We repeated the Cholesky analysis using a broader measure of intelligence (a composite *g* measure from ages 7-16; see Methods). The results remained the same, as shown in Supplementary Figure S2. The heritability of spatial ability in this model was estimated at 0.66, of which 41% (0.27/0.66) was shared with *g* and 59% (0.39/0.66) was specific to spatial ability independent of *g*.

Discussion

A new “gamified” battery was developed to test the phenotypic and genetic structure of spatial abilities, covering a diverse range of the putative components of this cognitive domain. Our results indicate, for the first time, that spatial ability is unifactorial both phenotypically (Figure 2A; Table 1) and genetically (Figure 4A). We show that performance on different spatial tests was influenced by the same genetic factors. Non-shared environmental

influences, on the other hand, were largely specific to each spatial test (Figure 4A); this could be due to specific environmental influences, or more likely due to test-specific measurement error.

We show that all spatial tests are moderately to substantially influenced by genetic factors (Figure 3), with the highest heritability shown for the composite spatial factor (69%). The single spatial tests were less heritable than the composite spatial factor, suggesting that measuring spatial ability with multiple tests increases the reliability of the construct. This can also be seen from the relatively low MZ correlations for single tests compared to the composite spatial factor (Supplementary Table S5). Since the reliable portions of spatial ability are shared between all tests – i.e., it is unifactorial, with the reliable variance in common between them – this finding suggests that using multiple tests (or perhaps a single, long test composed of many items) will capture spatial ability more reliably.

It is important to emphasise that heritability refers to the extent to which inherited differences in the DNA sequence explain the observed individual differences in a particular population, at a particular time (13). It describes what is, but not what could be; in other words, it only reflects the proportion of variance attributable to genetic influences under present conditions. We found that only a modest proportion (8%) of individual differences can be accounted for by shared environmental factors, such as school and family influences (Figure 3). The rest of the individual differences were explained by non-shared environmental influences, which are environmental factors that do not contribute to similarities between twins; for example, different groups of friends or each individual's perceptions of his/her environment. The estimate of non-shared environmental factors also includes any measurement error; since the magnitude of the non-shared environment component is greatly reduced for the (more highly reliable) overall spatial ability factor, in comparison to the individual tests, it seems likely that measurement error explains much of this component.

It might be reasonable to assume that the unifactorial structure of spatial ability is explained by general cognitive ability (*g*). However, our results show that the factor structure was not explained by *g*, remaining unchanged both phenotypically (Figure 2B) and genetically (Figure 4B) after correcting for it. Further, the latent factor of spatial ability is a specific cognitive ability in its own right, genetically distinguishable from intelligence (Figure 5; Supplementary Figure S2), as indicated by its significant and substantial genetic specificity (at least 40%). The unifactorial genetic structure of this spatial domain (i.e., its pleiotropy; see Methods)

could indicate that the same general processes contribute to all aspects of spatial ability; alternatively, since these spatial tests were administered late in development, genetic factors influencing some specific aspects may in turn drive the development of others.

Research has shown that spatial ability contributes importantly to positive life outcomes, especially achievement in STEM fields (3). For this reason, we argue that it is important to clarify the phenotypic and genetic structure of this domain, in order to make its measurement both more precise and more useful. We included all the main putative domains of spatial ability in our test battery, with the aim of differentiating between possible spatial factors. However, the results indicate strongly that spatial ability is unifactorial, as we found no evidence of differentiation either phenotypically or genetically.

It would be a mistake to interpret weak shared environmental influence, as found in this study, to suggest that training spatial ability would not be possible. These analyses only decompose the observed variance under current conditions, and therefore the findings do not limit the possibility of successful training programs that do not currently contribute to the variance in these measures. Various interventions have been proposed, and research to date suggests that training spatial ability can be effective, with an average improvement of around 0.5 standard deviations (11, 25). However, our findings suggest that individuals differ widely in spatial ability, and that these differences result in part from genetic differences between them. It is possible that training will be more successful if it is tailored to these (partly genetically-driven) differences in spatial ability, and for example detecting any weaknesses early in development and tailoring intervention programs to individual needs.

The high heritability of spatial ability at ages 19-21 suggests that this phenotype is a good candidate for gene-hunting efforts attempting to identify specific genetic variants. As predicted, our results show partial genetic overlap between spatial ability and *g*, so it is likely that as genes associated with *g* are identified, some of these genes will also be associated with spatial ability; however, since there is substantial genetic variance independent of *g*, there are also likely to be DNA differences that explain spatial ability specifically, independent of general intelligence.

Nothing would advance the field more than identifying specific genetic factors associated with cognitive abilities. However, research has shown that the heritability of complex traits, such as intelligence, is influenced by many DNA differences, possibly thousands, with each

individual genetic variant having a very small effect size (26). The structure of spatial ability has important potential consequences for the success of these efforts: if this domain were multifactorial, the genetic influences on each intricate component would be considerably harder to isolate, not least because of the diverse spatial tests in common use. However, the results of the present study suggest that, as genetic variants associated with spatial ability are identified, they will be related to general spatial ability, rather than with individual subcomponents. Any study with genetic data available for any spatial test may therefore be used to identify associations with spatial ability in general. That said, a composite of diverse measures may still be preferable, in order to ensure that the breadth of genetic influences on spatial ability are captured reliably.

The limitations of the present study include the usual limitations of the twin method, described in detail elsewhere (13, 27). Another limitation is that our diverse battery of spatial testing did not include navigation abilities, such as way-finding or map-reading skills, which have been argued to be multifactorial in their own right (12). Tests of navigation abilities will be included in our ongoing research.

Given its associations with STEM outcomes, it seems likely that spatial ability will become ever more important in our increasingly technological society, but tests of this domain are of limited use if it is unclear exactly what they measure. Identifying the specific genetic and environmental influences driving this ability, and the interactions between them, may ultimately refine its measurement, but the first step is to clarify its structure. The present results offer some insight here, suggesting that spatial ability can be differentiated from *g* and has a much simpler phenotypic and genetic architecture than previously supposed. Clarifying the structure of this domain is an important step towards understanding its aetiology, correlates and consequences.

Methods

Participants

The sample was drawn from the Twins Early Development Study (TEDS). TEDS is a large longitudinal study in the UK that recruited over 16,000 twin pairs born in England and Wales between 1994 and 1996. Although there has been some attrition, more than 10,000 twin pairs

remain actively involved in the study. Importantly, TEDS is a representative sample of the UK population (28–30). Zygosity was assessed using a parent questionnaire of physical similarity, which has been shown to be over 95% accurate when compared to DNA testing (31). DNA testing was conducted when zygosity was not clear from the physical similarity questionnaire criteria.

A randomly selected subsample of the older participants from the TEDS study (aged 19-21) participated in the present study, excluding individuals with major medical or psychiatric problems. After exclusions, the total number of individuals with spatial data available was 2,734 (1367 twin pairs), of whom 543 pairs were monozygotic (MZ), 432 were same-sex dizygotic (DZss) and 392 pairs were opposite-sex dizygotic (DZos). When DZos data are available, the aetiology of sex differences can be explored (32). The results of the full sex-limitation model fitting are presented in the Results section. Since little evidence was found for aetiological sex differences for spatial ability, and to increase power, we used the full sample, including DZos pairs in the genetic analyses.

Measures

A novel, online, gamified test battery, called the “King’s Challenge”, was used to test diverse measures of spatial ability. Examples of the test are provided in Figure 1. A demonstration of the battery is available here: http://teds.ac.uk/research/collaborators-and-data/public-datasets. The King’s Challenge battery is available on request for other researchers to use.

The development of the King’s Challenge began with an extensive literature review of the various measures used to test spatial ability. We assembled all available measures of spatial abilities, including mental rotation, spatial visualisation, spatial scanning, spatial reasoning, perspective-taking and mechanical reasoning. After a series of feasibility and pilot studies, we modified the existing measures and developed new tests as appropriate, to create a preliminary battery of 27 measures, administered as paper-and-pencil tests in the first feasibility study and as a computer-based test in the second feasibility study. Based on psychometric analyses and test-retest reliability, we ultimately reduced the number of tests to 10; these represented the psychometrically best-performing tests while eliminating redundancy between tests and capturing a diverse range of proposed spatial abilities. We removed all tests that did not show normal (or close to normal) distributions, as we were

interested in spatial ability in the general population and did not want to have tests that were too easy or too difficult for the participants; we removed all tests with low test-retest reliability (test retest $r < 0.5$); additionally we removed redundant tests (those that correlated with each other $r < 0.65$).

Each of these tests with their psychometric properties, as well as the test-retest correlations between paper-pencil tests and computerised tests is presented in Supplementary Table S10. These 10 spatial tests captured the major putative dimensions of spatial ability, comprising: a mazes task (searching for a way through a 2D maze in a speeded task), 2D drawing (sketching a 2D layout of a 3D object from a specified viewpoint), Elithorn mazes (joining together as many dots as possible from an array), pattern assembly (visually combining pieces of objects together to make a whole), mechanical reasoning (multiple-choice naïve physics questions), paper folding (visualising where the holes are situated after a piece of paper is folded and a hole is punched through it), 3D drawing (sketching a 3D drawing from a 2D diagram), mental rotation (mentally rotating objects), perspective-taking (visualising objects from a different perspective) and cross-sections (visualising cross-sections of objects). The creation of the King's Challenge is summarised in the Supplementary material. Each test started with a practice item, for which feedback was given (unlike other items). To promote participation, the final battery was "gamified" with the help of IT developers, Helmes Ltd (http://www.helmes.ee), meaning that the tests were embedded in a game-like narrative.

We piloted the King's Challenge battery on 100 unrelated individuals; all measures produced good test-retest reliability (r=0.65 on average for the 10 spatial tests): Pattern assembly r=0.56; Shapes rotation r=0.56; Paper folding r=.58; Cross-section r=0.64; Perspective taking r=0.56; Mechanical reasoning r=0.65; Elithorn maze r=.0.69; 3D drawing r=0.63; 2D drawing r=0.68; Maze r=0.46). All tests were taken using laptop or desktop computers (not smartphones or tablets) in web browsers.

Verbal and non-verbal ability were assessed online as an index of *g* when the participants were 16 years old. The Mill Hill Vocabulary Scale (33) was used to assess verbal ability. This test consists of multiple-choice vocabulary items. For each item a single word is presented at the top of the screen, and participants choose the answer closest in meaning to the target word. Non-verbal ability was assessed using Raven's Progressive Matrices, which consists of a series of incomplete patterns (34). This is also a multiple-choice test, in which the participant identifies the missing part of the pattern. General cognitive ability ('*g*',

intelligence) was indexed as the mean of the standardised verbal and non-verbal scores.

We also created a more robust measure of *g*, combining the general cognitive ability measures collected in TEDS longitudinally. At age 7, *g* was calculated as the mean of conceptual grouping (35) , a WISC similarities test (36), a WISC vocabulary test (36), and a WISC picture completion test (36), all conducted via telephone with parents' or guardians' assistance. At age 9, *g* was calculated as the mean of a shapes test (CAT3 Figure Classification) (37), a WISC vocabulary test (37), a WISC general knowledge task (37), and a puzzle test (CAT3 Figure Analogies) (36), all collected with booklets sent to the twins by post. At age 10, the *g* measure was calculated as the mean of the Ravens Standard Progressive Matrices (34), a WISC vocabulary test (37), WISC picture completion (38), and a WISC general knowledge test (37), all collected via web-based testing. At age 12, *g* was calculated exactly as at age 10. At age 14, *g* was computed as the mean of Raven's Progressive Matrices (34) and a WISC vocabulary test (36). Finally at age 16, *g* was measured as described above. The cross-age intelligence score was calculated as the mean of the *g* scores across the five ages (or however many time points for which each individual had data).

Prior to genetic analyses, all measures were corrected for age and sex differences using the regression method (39) by creating standardised residual scores. This procedure was used to avoid inflation of estimates of shared environment, as both members of twin pairs are identical for age and MZ twin pairs are also identical for sex. Finally all scores were transformed using the rank-based van der Waerden transformation (40, 41) to correct for a slight positive skew in some tests.

Analyses

Descriptive statistics. We compared means and variances for male and female participants and identical (MZ) and fraternal (DZ) twins for the whole sample (after exclusions). Since the present study used a twin sample, we maintained the independence of data for all phenotypic analyses by randomly selecting one twin per pair. The mean differences for sex and zygosity across all the measures and the interaction between sex and zygosity were tested using univariate analyses of variance (ANOVA).

Factor analyses. Exploratory principal component analyses were conducted to assess the

factor structure of spatial abilities. The factor structure was also tested by using the other half of the data (we randomly assigned members of twin pairs to two sub-samples). The statistical software SPSS was used for the analyses. The factor structure was also assessed by confirmatory factor analyses, using the statistical software package MPlus (42). Both exploratory and confirmatory factor analyses were conducted again after correcting the spatial scores for general intelligence.

Twin analyses. The twin method was used to estimate the relative contribution of the additive genetic (A), shared environmental (C) and non-shared environmental (E) components of variance of the spatial factor, and the covariance between the spatial tests (13). The twin method offers a powerful natural experiment by comparing the similarity of scores within MZ and DZ twin pairs, as MZ twins share 100% of their DNA, while DZ twins share on average 50% of their segregating genes, like any other siblings. Shared environmental influences are assumed to be 1.0 and the same for MZ and DZ twin pairs growing up in the same family. The rest of the variance is attributed to non-shared environmental influence, which includes error of measurement.

ACE parameters can be estimated by comparing cross-twin correlations for MZ and DZ twins. The A component may be approximated by doubling the difference between the MZ and DZ correlations; C is indexed by deducting the heritability from the MZ correlation; and E can be assessed by deducting the MZ correlation from unity. These ACE parameters and their 95% confidence intervals were estimated more precisely using structural equation modelling. In the present study, we used the structural equation program, OpenMx (43).

Univariate twin analysis of the variance of a single trait can be extended to multivariate analysis to estimate ACE parameters for the covariance between traits. Multivariate analysis also estimates additional statistics: the genetic correlation (rG), shared environmental correlation (rC) and non-shared environmental correlation (rE). Genetic correlation is an index of pleiotropy, the extent to which the same genetic variants influence multiple traits. Importantly, the genetic correlation is estimated independently of the heritabilities of the traits; that is, the heritabilities of two traits could be low, but the genetic correlation between the traits could be high. A shared environmental correlation of 1.0 indicates that the same environmental factors that make twins similar on one trait also make twins similar on another trait. Likewise, for non-shared environment (which is not shared between individuals, but may influence multiple traits for each individual), a correlation of zero indicates that completely

different non-shared environmental influences affect the two traits (13).

The independent pathway model is a multivariate genetic model that allows for estimation of the extent to which the genetic and environmental factors influencing the traits can be attributed to common latent ACE factors (44). The common factors have specific paths (standardised partial regressions) to each trait. In addition, residual paths index the extent to which the variance of the traits is not shared with other traits in the model (44–46).

The common pathway model is a multivariate genetic model in which the aetiology of all the variables in the analysis can be reduced to a common latent factor. That is, all genetic, shared environmental and non-shared environmental influences on all observed variables in the analyses will load onto a single latent factor. The common pathway model is considered to be more stringent than the independent pathway model, as it assumes that a single latent factor mediates the genetic, shared and non-shared environmental effects; compared to the independent pathway model that specifies both common and specific genetic and environmental causes (45–47).

Cholesky decomposition is a multivariate genetic analysis that is conceptually similar to hierarchical regression. This method estimates the extent to which the heritability of one trait is explained by the heritability of another trait. When entering a third variable in the model, it estimates the extent to which the heritability of trait three is explained by the heritability of trait one, and by the heritability of trait two when controlling for the heritability of trait one. Importantly this method also allows for estimation of the genetic correlation between pairs of variables, which is an index of pleiotropy, indicating the extent to which the same genetic variants influence two traits. Likewise, shared environmental and non-shared environmental correlations can be estimated (45, 46, 48). In the present study this method allows for the estimation of how much heritability in spatial ability is explained by the heritability of verbal ability, and how much heritability in spatial ability is explained by non-verbal ability when controlling for verbal ability.

Sex-limitation model. When data are available for DZos as well as DZss twins, the standard univariate model can be extended to a sex-limitation model to test for differences in the ACE aetiologies of sex differences, by comparing all five sex and zygosity groups: MZ males, MZ females, DZ males, DZ females and DZ opposite-sex twin pairs (13). Differences in the magnitude of ACE estimates for males and females are called quantitative sex differences.

Qualitative sex differences indicate whether different genetic or environmental factors affect males and females. Sex-limitation model-fitting was conducted by fitting a series of nested models and then testing the relative drop of the fit in the models (43). The sex-limitation model is described in detail elsewhere (32).

Acknowledgements

We gratefully acknowledge the ongoing contribution of the participants in the Twins Early Development Study (TEDS) and their families. TEDS is supported by a program grant to RP from the UK Medical Research Council [MR/M021475/1 and previously G0901245], with additional support from the US National Institutes of Health [HD044454; HD059215]. KR, NGS and SS are supported by a Medical Research Council studentship. MM is supported by the Economic and Social Research Council. RP is supported by a Medical Research Council Research Professorship award [G19/2] and a European Research Council Advanced Investigator award [295366]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

Conceived and designed the experiments: KR, NGS, MM, MR, SS, KS, PSD, YK, RP. Analysed the data: KR, NGS. Wrote the paper: KR, NGS, MM, RP. All authors approved the final draft of the paper.

Competing financial interests

The authors declare no competing financial interests.

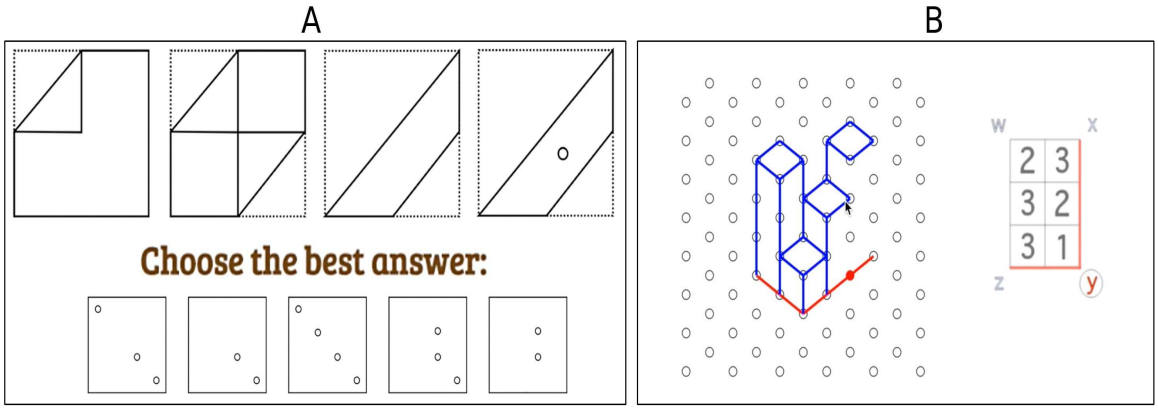
Table and Figures

Table 1. Confirmatory factor analyses.

	Model	AIC	BIC	X^2	RMSEA	CFI	TLI	SRMR
A	1-factor model	39450.91	39595.71	92.47**	0.04	0.98	0.98	0.02
	2-factor model	40235.14	40379.94	876.70**	0.16	0.71	0.63	0.24
B	1-factor model	21575.94	21717.44	77.15**	0.04	0.97	0.96	0.03
	2-factor model	22002.73	22144.23	503.94**	0.13	0.68	0.59	0.17

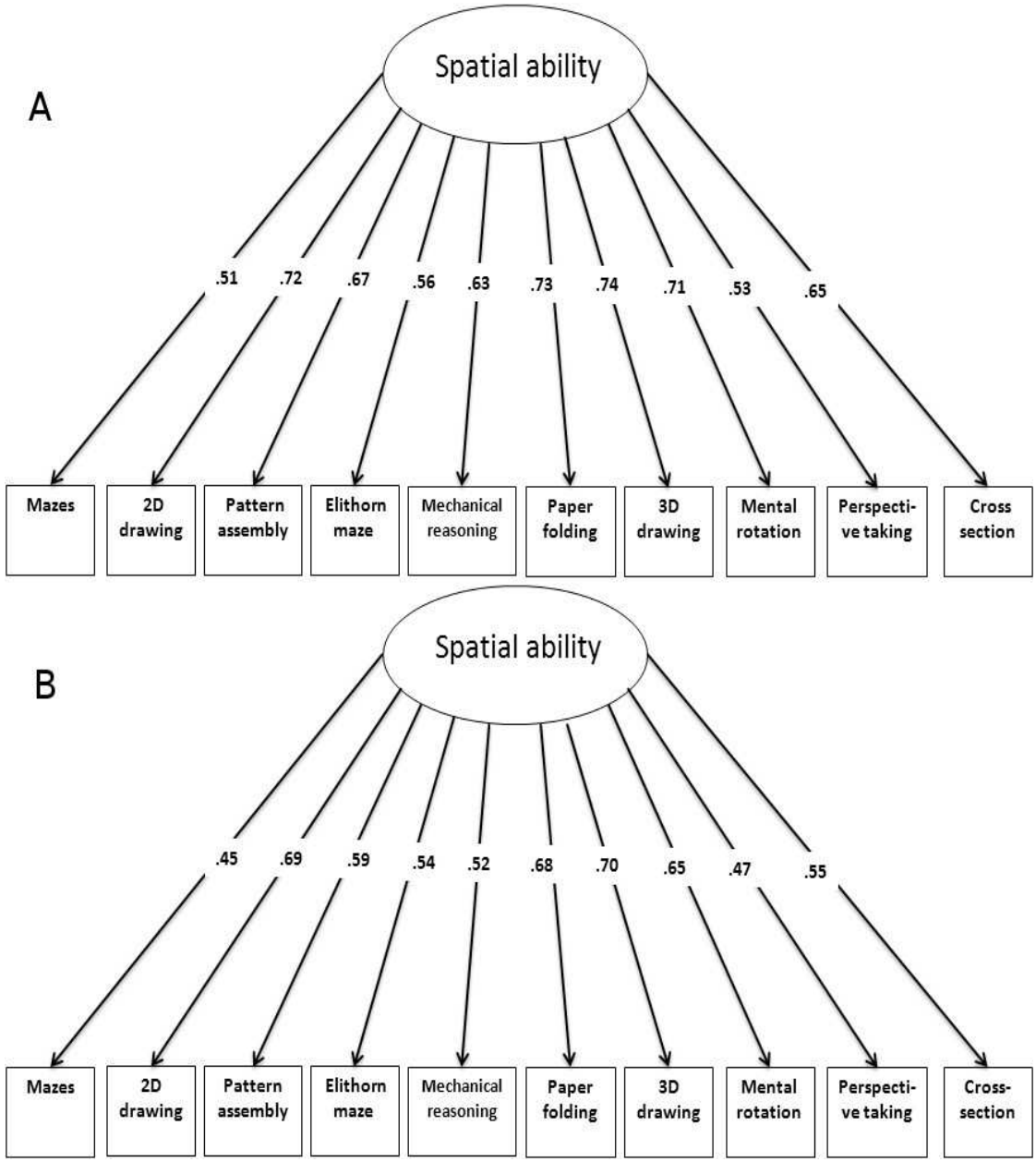
Model fit statistics between a 1-factor model and 2-factor model of spatial ability. (A) 10 spatial tests; (B) 10 spatial tests after correction for general intelligence scores using the regression method. AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; RMSEA = Root Mean Square Error of Approximation; CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; SRMR = Standardised Root Mean Square Residual; ** = p< 0.01. Note: In the 2-factor model, allowing the factors to correlate results in a correlation of 0.99; thus this correlation was constrained to zero to force orthogonality.

Figure 1. Sample stimuli.



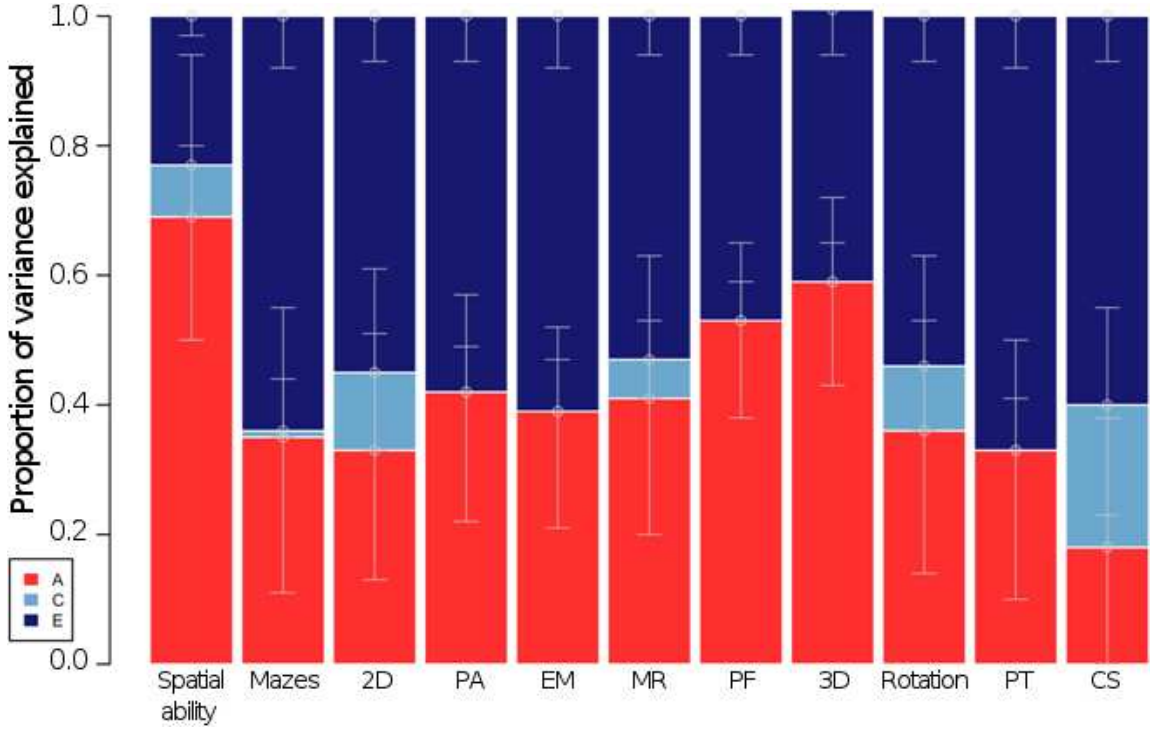
Example of the “King’s Challenge’ spatial battery with sample stimuli for the paper-folding (A) and 3D drawing (B) subtests. Examples of all 10 subtests, together with others included in pilot work, are presented in Supplementary Table S10.

Figure 2. Exploratory principal components analyses.



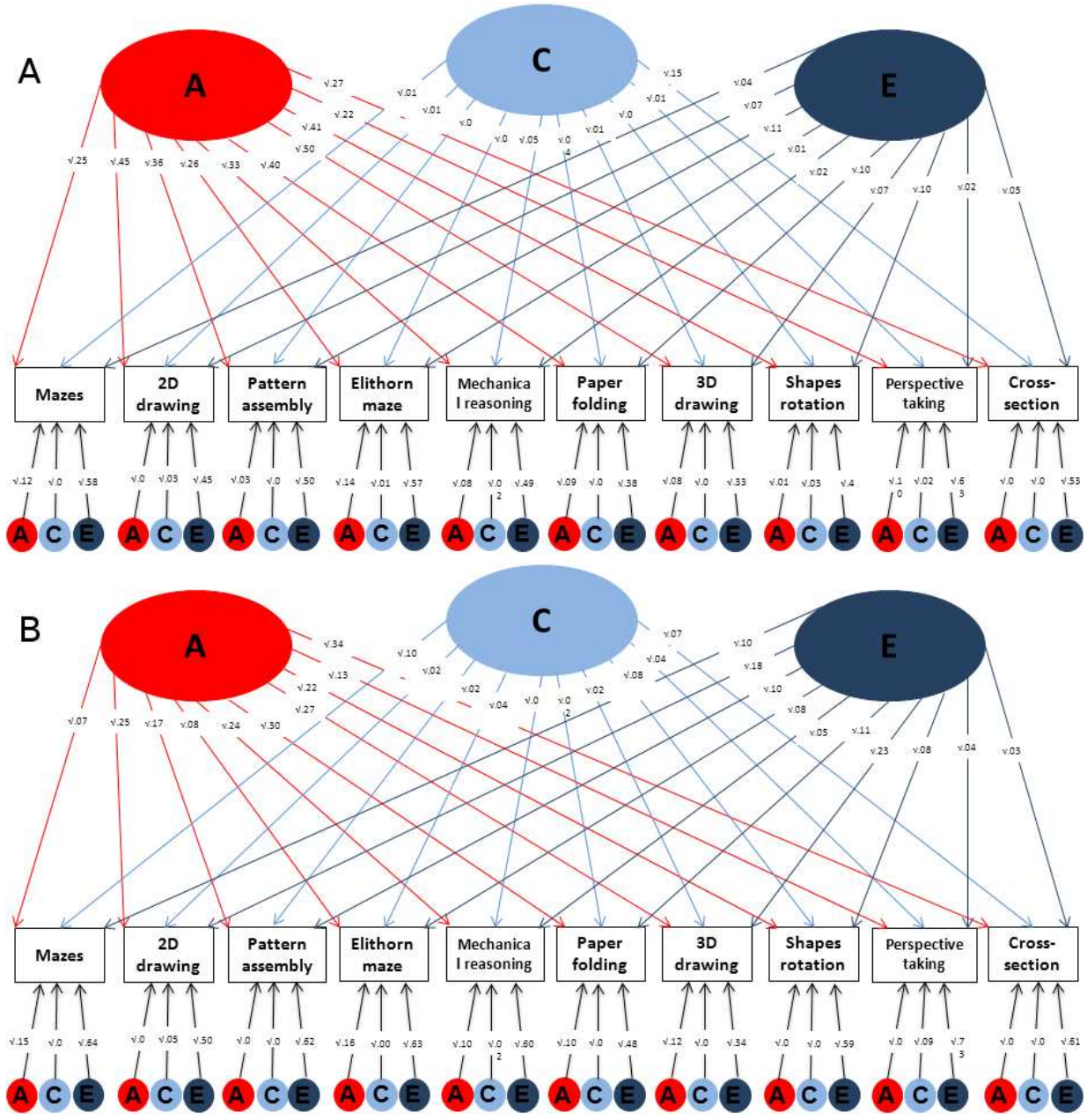
Exploratory principal components analyses. (A) Factor loadings for the 10 spatial tests. (B) Factor loadings after correction for general intelligence using the regression method.

Figure 3. Univariate model-fitting results.



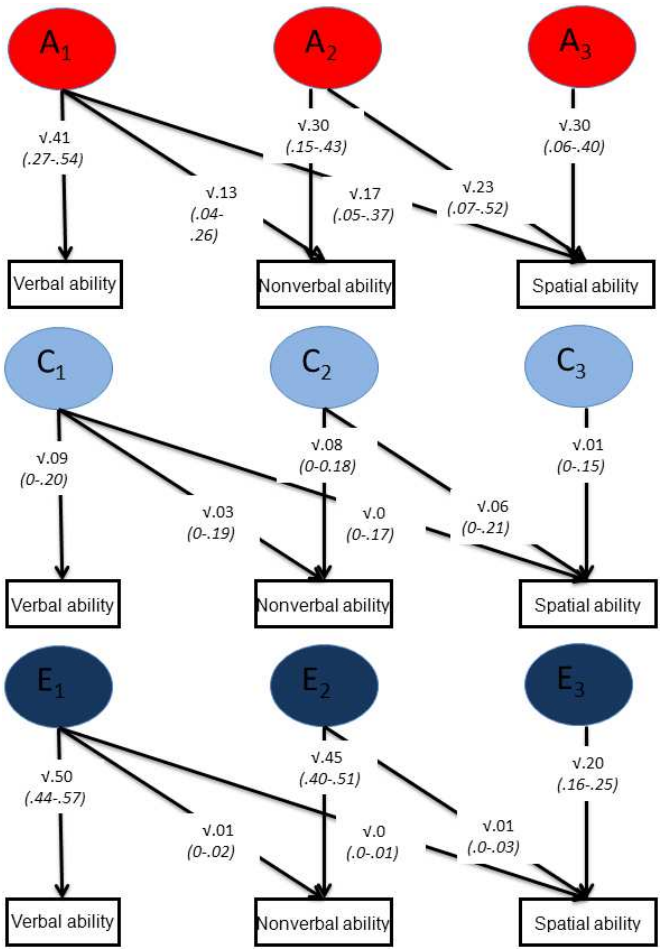
Genetic and environmental estimates for spatial tests: univariate model-fitting results (error bars representing 95% confidence intervals). A=additive genetic, C=shared environmental and E=non-shared environmental components of variance. 2D=2D drawing; PA=pattern assembly; EM=Elithorn maze, MR=mechanical reasoning, PF=paper folding; 3D=3D drawing, Rotation=mental rotation; PT=perspective-taking; CS=cross-sections.

Figure 4. Independent pathway model results.



Independent pathway model presenting the standardised squared path estimates. A=additive genetic, C=shared environmental and E=non-shared environmental components of variance. (A) Path estimates for the 10 spatial tests. (B) Path estimates for 10 tests after correction for general intelligence using the regression method.

Figure 5. Trivariate Cholesky decomposition.



Trivariate genetic Cholesky decomposition for verbal ability, non-verbal ability and spatial ability (with 95% confidence intervals in parentheses). A=additive genetic, C=shared environmental and E=non-shared environmental components of variance.

References

1. Lohman D (1996) in *Human abilities: Their nature and measurement*, eds Dennis I, Tapsfield P (Lawrence Erlbaum Associates Inc. New Juersey, USA), pp 97–116.
2. Mackintosh N, Mackintosh NJ (2011) *IQ and human intelligence* (Oxford University Press., Oxford, UK).
3. Wai J, Lubinski D, Benbow CP (2009) Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *J Educ Psychol* 101:817–835.
4. Newcombe NS, Shipley TF (2009) in *Studying visual and spatial reasoning for design creativity*, ed Gero J. (Springer Netherlands), pp 179–192.
5. Rhodes SM, Riby DM, Fraser E, Campbell LE (2011) The extent of working memory deficits associated with Williams syndrome: exploration of verbal and spatial domains and executively controlled processes. *Brain Cogn* 77:208–14. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21889249 [Accessed February 24, 2014].
6. Kell HJ, Lubinski D, Benbow CP, Steiger JH (2013) Creativity and Technical Innovation: Spatial Ability's Unique Role. *Psychol Sci*.
7. Gohm CL, Humphreys LG, Yao G (1998) Underachievement among spatially gifted students. *Am Educ Res J* 35:515–531.
8. Benbow CP, Shea DL, Lubinski D (2001) Importance of assessing spatial ability in intellectually talented young adolescents: A 20-year longitudinal study. *J Educ Psychol* 93:604–614.
9. Carroll JB (1993) *Human cognitive abilities: A survey of factor-analytic studies* (Cambridge University Press, Cambridge, UK).
10. Colom R, Contreras MJ, Shih PC, Santacreu J (2003) The assessment of spatial ability with a single computerized test. *Eur J Psychol Assess* 19:92–100.
11. Uttal DH, Miller DI, Newcombe NS (2013) Exploring and Enhancing Spatial Thinking: Links to Achievement in Science, Technology, Engineering, and Mathematics? *Curr Dir Psychol Sci* 22:367–373.
12. Weisberg SM, Schinazi VR, Newcombe NS, Shipley TF, Epstein R a (2014) Variations in cognitive maps: understanding individual differences in navigation. *J Exp Psychol Learn Mem Cogn* 40:669–82.
13. Plomin R, DeFries JC, Knopik VS, Neiderhiser JM (2013) *Behavioral Genetics. 6th ed* (Worth Publishers, New York).
14. Bratko D (1996) Twin study of verbal and spatial abilities. *Pers Individ Dif* 21:621–624.
15. Rietveld MJH, Dolan C V, van Baal GCM, Boomsma DI (2003) A Twin Study of Differentiation of Cognitive Abilities in Childhood. *Behav Genet* 33:367–381.
16. Pedersen NL, Plomin R, Nesselroade JR, McLearn GE (1992) A quantitative genetic analysis of cognitive abilities during the second half of the life span. *Psychol Sci* 3:346–353.
17. Kan K-J, Wicherts JM, Dolan C V., van der Maas HLJ (2013) On the Nature and Nurture of Intelligence and Specific Cognitive Abilities: The More Heritable, the More Culture Dependent. *Psychol Sci*.
18. Tosto MG et al. (2014) Why do spatial abilities predict mathematical performance? *Dev Sci* 17:462–470.
19. DeFries JC et al. (1979) Familial resemblance for specific cognitive abilities. *Behav Genet* 9:23–43.

20. DeFries JC, Vandenberg SG, McClearn GE (1976) Genetics of specific cognitive abilities. *Annu Rev Genet* 10:179–207.

21. Robinson EB et al. (2015) The genetic architecture of pediatric cognitive abilities in the Philadelphia Neurodevelopmental Cohort. *Mol Psychiatry* 20:454–8.

22. Deary IJ, Spinath FM, Bates TC (2006) Genetics of intelligence. *Eur J Hum Genet* 14:690–700.

23. Petrill SA et al. (1998) The Genetic and Environmental Relationship Between General and Specific Cognitive Abilities in Twins Age 80 and Older. *Psychol Sci* 9:183–189.

24. Neale MC, Cardon LR (2004) *Methodology for Genetic Studies of Twins and Families.* (Kluwer Academic Publishers B.V., Dordrecht).

25. Meadow NG et al. (2012) The Malleability of Spatial Skills: A Meta-Analysis of Training Studies. *Psychol Bull*.

26. Plomin R, Deary IJ (2014) Genetics and intelligence differences: five special findings. *Mol Psychiatry*:1–11.

27. Rijsdijk F V, Sham PC (2002) Analytic approaches to twin data using structural equation models. *Brief Bioinform* 3:119–133.

28. Kovas Y, Haworth CMA, Dale PS, Plomin R (2007) The genetic and environmental origins of learning abilities and disabilities in the early school years. *Monogr Soc Res Child Dev* 72:vii–160.

29. Oliver BR, Plomin R (2007) Twins' Early Development Study (TEDS): a multivariate, longitudinal genetic investigation of language, cognition and behavior problems from childhood through adolescence. *Twin Res Hum Genet* 10:96–105.

30. Haworth CMA, Davis OSP, Plomin R (2013) Twins Early Development Study (TEDS): A genetically sensitive investigation of cognitive and behavioral development from childhood to young adulthood. *Twin Res Hum Genet* 16:117–25.

31. Price TS et al. (2000) Infant zygosity can be assigned by parental report questionnaire data. *Twin Res* 3:129–133.

32. Medland SE (2004) Alternate parameterization for scalar and non-scalar sex-limitation models in Mx. *Twin Res* 7:299–305.

33. Raven JC, Raven J, Court JH (1998) *The Mill Hill Vocabulary Scale* (Oxford:OPP).

34. Raven J, Raven JC, Court J (1996) *Manual for Raven's Progressive Matrices and Vocabulary Scales* (Oxford: Oxford University Press).

35. McCarthy D (1972) *McCarthy Scales of Children's Abilities* (New York: The Psychological Corporation.).

36. Smith P, Fernandes C, Strand S (2001) *Cognitive Abilities Test 3 (CAT3)* (Windsor: nferNELSON.).

37. Kaplan E, Fein D, Kramer J, Delis D, Morris R (1999) *WISC-III As a Process Instrument (WISC-III-PI)* (New York: The Psychological Corporation.).

38. Wechsler D (1992) *Wechsler Intelligence Scale for Children (3rd Ed. UK)* (The Psychological Corporation).

39. McGue M, Bouchard TJ (1984) Adjustment of twin data for the effects of age and sex. *Behav Genet* 14:325–343.

40. Lehmann E (1975) *Nonparametric Statistical Methods Based on Ranks* (Holden-Day, San Francisco, CA).

41. Van Der Waerden BL (1975) On the sources of my book Moderne Algebra. *Hist Math* 2:31–

40.

42. Muthén L, Muthén B (2010) *Mplus (6th ed.)*.

43. Boker S et al. (2011) OpenMx: an open source extended structural equation modeling framework. *Psychometrika* 76:306–317.

44. Rijsdijk F V. (2005) in *Encyclopedia of Statistics in Behavioral Science*, eds Everitt BS, Howell DC (John Wiley & Sons Ltd., Chichester, UK), pp 913–914.

45. Neale MC, Maes H. (2001) *Methodology for genetic studies of twins and families* (Dordrecht, The Netherlands: Kluwer Academic Publishers B.V).

46. Neale MC, Cardon L. (1992) *Methodology for genetic studies of twins and families* (Dordrecht, The Netherlands: Kluwer Academic Publications).

47. Rijsdijk F V. (2005) in *Encycl. Stat. Behav. Sci.*, eds Everitt BS, Howell DC (John Wiley & Sons Ltd.), pp 330–331.

48. Loehlin JC (1996) The Cholesky approach: A cautionary note. *Behav Genet* 26:65–69.

Chapter 6:- Genetic specificity of face recognition

This chapter, investigating the aetiology of face recognition and its relationship with other cognitive abilities, is presented as a published paper. It is an exact copy of this publication:

Shakeshaft NG, Plomin R (2015). Genetic specificity of face recognition. *PNAS* 112(41): 12887–12892. doi:10.1073/pnas.1421881112

Supplementary materials for this chapter, as detailed in the text, are attached as Appendix 4.

Genetic specificity of face recognition

Nicholas G. Shakeshaft[1] and Robert Plomin

Medical Research Council Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London SE5 8AF, United Kingdom

Edited by Nancy Kanwisher, Massachusetts Institute of Technology, Cambridge, MA, and approved August 25, 2015 (received for review December 24, 2014)

Specific cognitive abilities in diverse domains are typically found to be highly heritable and substantially correlated with general cognitive ability (*g*), both phenotypically and genetically. Recent twin studies have found the ability to memorize and recognize faces to be an exception, being similarly heritable but phenotypically substantially uncorrelated both with *g* and with general object recognition. However, the genetic relationships between face recognition and other abilities (the extent to which they share a common genetic etiology) cannot be determined from phenotypic associations. In this, to our knowledge, first study of the genetic associations between face recognition and other domains, 2,000 18- and 19-year-old United Kingdom twins completed tests assessing their face recognition, object recognition, and general cognitive abilities. Results confirmed the substantial heritability of face recognition (61%), and multivariate genetic analyses found that most of this genetic influence is unique and not shared with other cognitive abilities.

face perception | behavioral genetics | cognitive psychology | twin study

Specific cognitive abilities correlate substantially with general cognitive ability (*g*). This finding holds true for domains as diverse as literacy (1), spatial reasoning (2), mathematical ability (3), and visual and verbal memory (4). In addition, these diverse specific abilities, and *g* itself, are typically found to be substantially heritable (5). Genetic correlations between abilities (i.e., the degree to which genetic influences are correlated between them, indicating pleiotropy: common genes influencing multiple traits) tend to be at least as strong as their phenotypic associations (the correlations between task scores or other behavioral measures) (6), and *g* typically accounts for almost all of the genetic variance in each domain (7). Even though the nature of *g* itself remains unclear, these phenotypic and genetic intercorrelations among diverse abilities suggest that cognitive domains form a single hierarchy (8). At the apex of this hierarchy is *g*, explaining on average 40% of the total phenotypic variance in each domain (9) and—via pleiotropic "generalist genes" (10)—almost all of their genetic variance.

Two recent twin studies have suggested that face recognition, the ability to memorize and recognize human faces, may represent an exception to this model. Faces have long been argued to be "special" as a category of visual stimulus, showing both cortical specificity (11) and a wide range of face-specific perceptual effects (12). Whether such effects suggest true domain specificity or merely reflect a highly specialized form of learned expertise (acquired almost universally among typically developing children) has long been the subject of debate (13), with proponents of the former suggesting evolutionary specificity for face recognition (14). In this context, the findings of two recent twin studies (15, 16) are informative. Individual differences in face recognition were found to be substantially heritable: 68% in one study (15) and 39% in the other (16)—the difference perhaps reflecting the different tasks used, or perhaps insufficient power to establish precise point estimates due to the modest sample sizes of these studies (289 and 173 twin pairs, respectively). The ability was also found to be phenotypically largely unrelated either to visual or verbal memory (15) or to *g* (16).

These findings seem consistent with the argument for evolutionary—and thus genetic—specificity (13), although it should be noted that the etiology of within-species variation may be unrelated to the evolutionary origins of a trait. However, a low phenotypic correlation between two traits does not inevitably indicate the absence of common genetic influences. Their genetic correlation may be high (even at unity, in principle) when their phenotypic correlation is low, if the heritability of either trait is relatively low (9). Even two highly heritable but phenotypically largely uncorrelated traits could still have a substantial genetic correlation if, for example, a negative environmental correlation counterbalanced a positive genetic correlation. For example, if environmental factors positively influencing the ability to recognize nonface objects (e.g., by promoting interest in activities that provide relevant practice) also tended to have a negative influence on face recognition ability (e.g., by reducing social interaction or attention), then this negative environmental correlation would offset the positive genetic correlation between these traits and confound the interpretation of any study unable to examine their genetic relationship directly.

Unambiguously establishing the architecture of genetic influences on multiple traits is the purpose of multivariate genetic analyses, which have not been reported by any study conducted in this field to date, presumably due to the large samples required for adequate power. The present study administered tests assessing face recognition, general (nonface) object recognition, and *g* to a large sample of twins to examine directly the degree to which face recognition is genetically distinct from other perceptual and cognitive abilities.

Results

Data. The Twins Early Development Study (TEDS) is a longitudinal cohort study of twins born in England and Wales between 1994 and 1996, with more than 10,000 pairs still enrolled. The recruitment and characteristics of this sample have been described previously (17, 18). Zygosity was assessed at enrollment using a

Significance

Diverse cognitive abilities have typically been found to intercorrelate highly and to be strongly influenced by genetics. Recent twin studies have suggested that the ability to recognize human faces is an exception: it is similarly highly heritable, but largely uncorrelated with other abilities. However, assessing genetic relationships—the degree to which traits are influenced by the same genes—requires very large samples, which have not previously been available. This study, using data from more than 2,000 twins, shows for the first time, to our knowledge, that the genetic influences on face recognition are almost entirely unique. This finding provides strong support for the view that face recognition is "special" and may ultimately illuminate the nature of cognitive abilities in general.

Author contributions: N.G.S. and R.P. designed research; N.G.S. performed research; N.G.S. analyzed data; and N.G.S. and R.P. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. Email: nicholas.shakeshaft@kcl.ac.uk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1421881112/-/DCSupplemental.

parental questionnaire shown to be more than 95% accurate compared with direct genetic testing (19), with DNA testing conducted where results were unclear. For the present study, a representative subsample was selected from the oldest twins in this cohort (who had passed the age of majority, 18 years of age), and data were obtained from 2,149 participants—924 complete pairs [375 monozygotic (MZ), 549 dizygotic (DZ)]—plus an additional 301 unpaired individuals. Individuals with severe physical or psychological disabilities, or whose mothers had experienced serious medical complications during pregnancy, were excluded. The resulting dataset was 58% female, with a mean age of 19.5 years of age (±0.3 SD) on completion of the face and object recognition tests.

Face recognition ability was assessed with the widely used Cambridge Face Memory Test (CFMT) (20), requiring participants to memorize a series of unfamiliar faces, from images cropped to exclude cues such as hair and clothing, and then to identify them among distractors in a variety of viewpoints and lighting conditions. General (nonface) object recognition ability was measured using the Cambridge Car Memory Test (CCMT) (21), designed to be matched precisely to the CFMT but using computer-generated 3D models of cars instead of faces. See Fig. 1 for sample stimuli for both tests. General cognitive ability (*g*) was assessed during an earlier testing phase for this cohort at age 16, as a verbal/nonverbal composite: the mean of standardized scores from the Mill Hill Vocabulary Scale (22) and Raven's Progressive Matrices (23). See *Methods* for more details on these measures.

Sample sizes and descriptive statistics for these measures are presented in Table 1. The distributions demonstrate a large amount of variability in the sample for these abilities, with face and nonface recognition scores ranging from chance to (in very rare cases) perfect scores, and do not differ significantly from those obtained with the original reference samples (20, 21) for these tests. The face and object recognition tasks were newly administered to the TEDS sample, so care was taken to ensure their reliability. Cronbach's alpha was high for both measures: 0.893 for the CFMT, 0.875 for the CCMT (see *SI Appendix*, Table S1 for more details).

An analysis of variance was performed for each measure to assess the mean effects of sex and zygosity. The only significant mean difference found was a main effect of sex on object

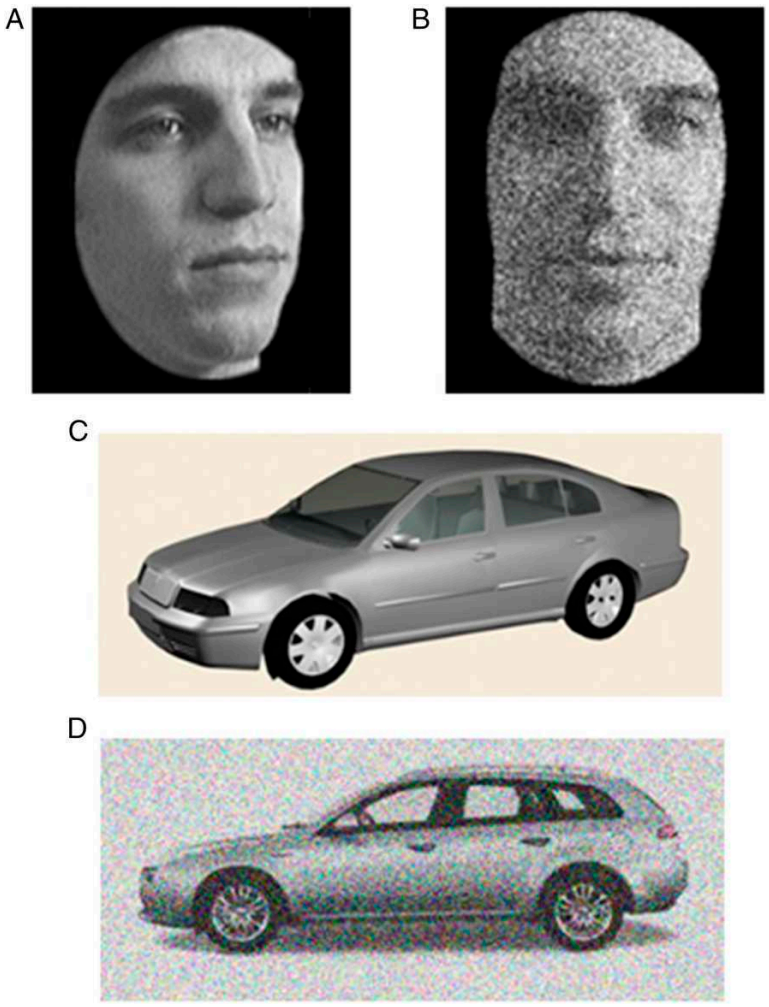


Fig. 1. Sample stimuli. Sample images for the Cambridge Face Memory Test (20), for both the clean (*A*) and degraded (*B*) conditions (see *Methods*), and for the Cambridge Car Memory Test (21), both clean (*C*) and degraded (*D*).

Table 1. Descriptive statistics

	n	Whole sample	Males	Females	MZ	DZss	DZos	Sex	Zyg	Sex × zyg	R^2
Face recognition	1,068	54.10 (9.44)	53.45 (9.62)	54.54 (9.30)	54.04 (9.28)	54.14 (9.55)	54.30 (9.52)	2.78	0.20	0.07	0.00
Object recognition	1,042	50.54 (9.75)	53.84 (10.05)	48.33 (8.90)	50.10 (9.95)	50.81 (9.62)	51.68 (9.52)	85.61**	0.09	1.64	0.07
g	758	0.05 (0.97)	0.08 (1.02)	0.03 (0.93)	−0.03 (0.98)	0.10 (0.95)	0.24 (0.94)	0.39	3.42	0.02	0.01

Mean scores (SDs) for the whole sample, separately by sex, and for monozygotic (MZ) and same-sex (ss) and opposite-sex (os) dizygotic (DZ) twins. *n* = sample size (sample shown is fully independent, randomly selecting one individual per twin pair). ANOVA was performed on cleaned, normality-transformed data to test effects of sex and zygosity. Results = F statistic. $^{**}P < 0.001$. R^2 = proportion of variance explained by sex, zygosity (Zyg), and their interaction.

recognition (Table 1), explaining 7% of the variance, perhaps relating (as argued by the test's authors) (21) to differential average interest in or experience with cars. Twin analyses are concerned with variances, so a mean sex difference is irrelevant provided (as in our results) the distribution is not restricted. In any case, per standard practice for twin studies (*Methods*), the mean effects of sex were regressed out. All subsequent analyses were conducted using sex- and age-regressed, normality-transformed, standardized data.

Phenotypic Analyses. Phenotypic analyses were conducted using a fully independent sample, randomly selecting one twin per pair. Face recognition ability was moderately correlated with nonface object recognition [$r = 0.29$, 95% confidence interval (CI) 0.23–0.34, $P < 0.001$, $n = 1{,}042$], and modestly with *g* ($r = 0.16$, CI 0.09–0.23, $P < 0.001$, $n = 718$). Nonface object recognition and *g* were similarly modestly correlated ($r = 0.15$, 95% CI 0.08–0.22, $P < 0.001$, $n = 706$). The phenotypic relationship between face recognition and *g* largely survived controlling for general object recognition (partial correlation, $r = 0.12$, $P < 0.001$, $n = 706$). Similarly, much of the association between face recognition and general object recognition was independent of *g* (partial correlation, $r = 0.25$, $P < 0.001$, $n = 706$). The smaller samples for those analyses involving *g* reflect the intersection between the datasets produced at the two testing phases.

Taken together, these results indicate that face recognition is largely, but not wholly, phenotypically independent both from general cognitive ability and from general object recognition. The significant partial correlations suggest that the associations between face recognition and each of these other two measures are largely independent from one another.

Univariate Genetic Analyses. Intraclass twin correlations for monozygotic (MZ) and same- and opposite-sex dizygotic (DZ) twins are presented in Table 2. MZ correlations are consistently significantly higher than those for DZ twins, suggesting genetic influence. From these intraclass correlations, initial estimates may be obtained for heritability (additive genetic influences on the trait), shared environmental influences (environmental factors making twins more similar), and unique (nonshared) environmental influences (the remaining variance, including influences making twins dissimilar, and also any error of measurement)—see Table 2 for calculation details. These estimates (Table 2) suggest that genetic influence is substantial for all measures.

These initial estimates were tested formally with full-information maximum-likelihood model fitting (accounting for missing data, and using the full dataset including both same-sex and opposite-sex DZ twins) to estimate the variance attributable to additive genetic (A), shared environmental (C), and unique environmental/error (E) components (*Methods*). The results (Fig. 2) confirm substantial genetic influence for all three measures, with heritability estimated at 61% for face recognition, 56% for object recognition, and 48% for *g*, very similar to the rough estimates (Table 2). Also (similar to the estimates in Table 2), almost no shared environmental influences were detected (i.e., environmental influence was apportioned to E, representing nonshared influences and error of measurement, rather than C). Precise estimates and confidence intervals are presented in *SI Appendix*, Table S2, and fit statistics (*Methods*) in *SI Appendix*, Table S3.

Multivariate Genetic Analyses. The main focus of this study was to examine the genetic relationships between face recognition and other abilities, as indexed by *g* and general object recognition. This aim may be achieved with twin data using bi- and multivariate model-fitting analyses (*Methods*). Two bivariate correlated factors solution models indicate the genetic, shared, and unique environmental correlations between the traits and (derived from these results) the proportions of the phenotypic correlations (between face recognition and each other variable) attributable to each component (Fig. 3 and *SI Appendix*, Table S4). These phenotypic correlations are substantially genetic in origin: 66% of the correlation with general object recognition and 88% of the correlation with *g*, the latter being the only component of the correlation with *g* whose estimate is significant.

However, as the phenotypic correlations are modest, the proportion of the total variance of face recognition ability included in these results is low. However, the genetic correlations with face recognition, which are independent of the phenotypic correlations and heritabilities, are also low (0.31 with object recognition, 0.32 with *g*; see *SI Appendix*, Table S4), indicating substantial genetic independence. Bivariate Cholesky decomposition analyses (*Methods*) provide another way to quantify these relationships: These analyses indicate the proportion of the heritability of a trait that is due to genetic effects shared with another trait. These analyses

Table 2. Twin correlations and approximated variance components

	Intrapair twin correlations			Variance component estimates			Sample (nos. of pairs)		
	MZ	DZss	DZos	h^2	c^2	e^2	MZ	DZss	DZos
Face recognition	0.60 (0.54–0.66)	0.30 (0.19–0.40)	0.17 (0.05–0.28)	0.60	0.00	0.40	374	289	256
Object recognition	0.58 (0.50–0.64)	0.15 (0.03–0.26)	0.30 (0.18–0.41)	0.58	0.00	0.42	358	276	244
g	0.58 (0.49–0.65)	0.37 (0.25–0.48)	0.28 (0.14–0.42)	0.42	0.16	0.42	285	226	170

Intraclass twin correlations (95% confidence intervals) for monozygotic (MZ) and same-sex (ss) and opposite-sex (os) dizygotic (DZ) twins. Variance component estimates are heritability (h^2, double the difference between the MZ and DZss correlations, constrained not to exceed the former—MZ twins are genetically identical, so heritability cannot exceed their correlation), shared environment (c^2, the MZ correlation minus h^2), and unique environment/error of measurement ($e^2 = 1 - h^2 - c^2$). Sample sizes shown are complete pairs, after exclusions and data cleaning.

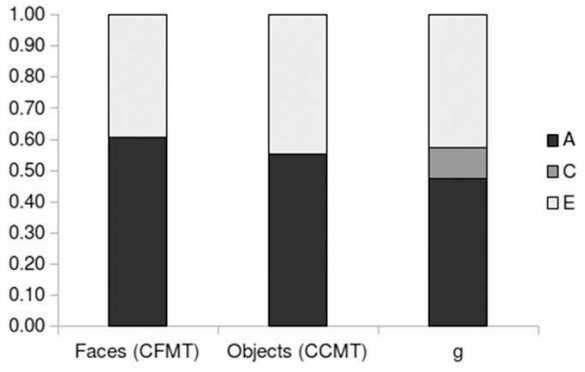


Fig. 2. Model-fitting estimates. Variance due to additive genetic (A), shared environmental (C), and nonshared environmental influences/error (E).

(Fig. 4*A* and *SI Appendix*, Table S5) show that the genetic effects constituting the heritability of face recognition are largely specific to this trait (~90%), rather than shared either with general object recognition or with *g*. That is, only 10% of the heritability of face recognition, representing 6% of its total variance, is due to genetic effects shared with object recognition. Similarly, 10% of the heritability of face recognition (6% of total variance) is due to genetic effects shared with *g*. Path estimates for these model-fitting analyses are presented in *SI Appendix*, Fig. S1.

However, subjecting the object recognition measure to the same analysis (bivariate Cholesky decomposition, predicted by *g*) reveals a similar pattern to that observed with face recognition. Shared genetic influences between *g* and object recognition account for only 10% of the heritability of the latter (6% of total variance), perhaps suggesting that this *g* composite undercorrects for domain-general processes involved in the face and object recognition tasks (*Discussion*). Details are presented in *SI Appendix*, Table S6 (with fit statistics for all bivariate models in *SI Appendix*, Table S7).

Separate bivariate analyses cannot determine the proportion of influences that might be common to multiple predictor variables. Multivariate extension of the Cholesky decomposition allows the shared and independent components of variance to be estimated sequentially for multiple predictors. Details, fit statistics, and path estimates are presented in *SI Appendix*, Tables S8 and S9 and Fig. S2, respectively, but the main finding (Fig. 4*B*) is that only 11% of the heritability of face recognition (representing 6% of the total variance in this trait) is accounted for by genetic influences shared both with *g* and with general object recognition. Although the point estimate suggests that an additional 5% of its heritability (3% of total variance) is explained by genetic influences shared only with object recognition, independently from *g*, this estimate is nonsignificant—indicated both by the confidence interval of this estimate intersecting zero (*SI Appendix*, Table S8) and a submodel with this path constrained to zero resulting in no significant deterioration in fit (*Methods* and *SI Appendix*, Table S9). This result suggests that all of the genetic influences shared between face and object recognition are also shared with *g*. However, the large majority of the heritability of face recognition (85% in this model, representing 51% of its total variance) is due to genetic effects that are not shared with either of these other measures.

Since the *g* composite used here is the mean of two standardized test scores (see *Methods*), a more complete multivariate model would incorporate the two scores individually, ensuring that all of the shared variance between these measures is included. An additional model therefore entered the Mill Hill, Raven's, and object recognition scores independently. In this model, the first entered variable (Mill Hill) accounted for 8% of the heritability of face recognition (5% of its total variance). Raven's and object recognition accounted for no significant additional genetic variance, and again the large majority is unique. Details, fit statistics, and path estimates are provided in *SI Appendix*, Tables S10 and S11 and Fig. S3, respectively.

Race. The "other race" effect, meaning significantly decreased recognition accuracy for faces of less familiar races, has been demonstrated with the CFMT (24). Because the stimuli used here were Caucasian faces, the key analyses were repeated with the sample restricted to Caucasian participants (93% of the sample). The results were virtually identical to those obtained with the full sample, both in test performance (*SI Appendix*, Table S12) and the genetic independence of face perception from other measures (*SI Appendix*, Tables S13 and S14).

Discussion

We show for the first time, to our knowledge, that the substantial heritability of face recognition is due to genetic influences that are mostly specific to this ability, rather than shared either with general object recognition or general intelligence. The phenotypic and univariate genetic analyses broadly supported the findings of the two previous twin studies in this area (15, 16): Face recognition was phenotypically correlated only quite modestly with general object recognition (0.29) and very modestly with *g* (0.16), as indexed by these measures. The substantial heritability of face recognition (61%) is also in line with previous literature. However, one of the main strengths of quantitative genetic methods is their ability to establish the genetic architecture surrounding multiple traits (25). Thus, the main purpose of data collection on this scale from a twin sample was to provide the opportunity to conduct multivariate genetic analyses.

The results of these analyses indicate, first, that face recognition is not wholly distinct genetically either from general object recognition or from *g* and that its phenotypic correlations with each are largely due to shared genetic influences (Fig. 3). It seems that face recognition, as measured here, does after all fall within the traditional cognitive hierarchy to some degree—this result is perhaps unsurprising because the CFMT measure necessarily involves memory, attention, and other cognitive capacities. Most of the substantial heritability of face recognition, however, is due to genetic influences that are not shared either with general object recognition or with *g* (Fig. 4). Since *g* usually accounts for a large proportion of the genetic variance within any specific cognitive domain (7, 10), this finding offers support for the special nature of faces.

However, the CCMT object recognition measure shows similar results: It is genetically largely independent from *g* (*SI Appendix*, Table S6). At first glance, this observation would seem to undermine the argument that face recognition is special, but another

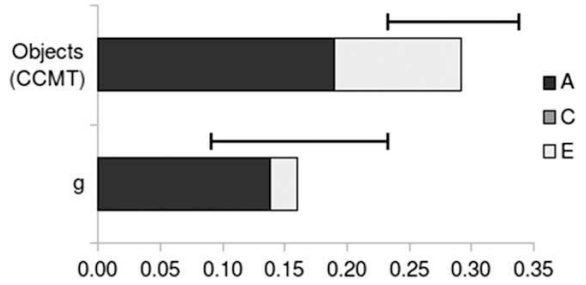


Fig. 3. Decomposition of phenotypic correlations with face recognition. Correlated factor solution analyses, indicating the proportion of the phenotypic correlation between face recognition and each other variable (line length, with 95% confidence intervals) attributable to genetic (A), shared environmental (C), and nonshared environmental influences/error (E).

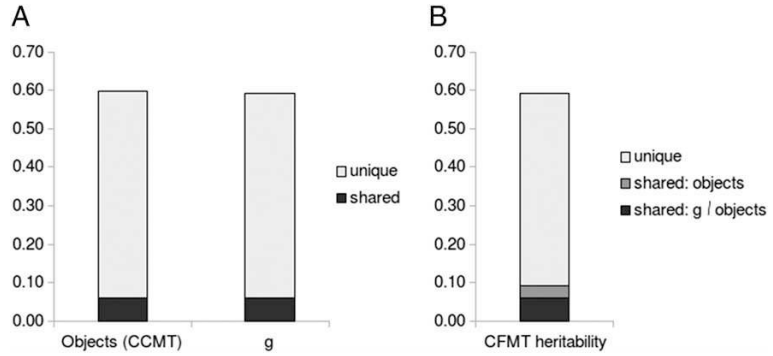


Fig. 4. Decomposition of heritability of face recognition. Cholesky bivariate (*A*) and trivariate (*B*) decomposition analyses, indicating that genetic influences on face recognition ability are largely independent from the genetic influences on general object recognition and *g*.

possible explanation is that the *g* composite used substantially undercorrects for domain-general processes: For example, neither of the component scores in this composite recruit memory. To evaluate this argument, the CCMT measure is the perfect control: The attentional, memory, and other requirements of these tasks are identical, and yet the genetic influences on face recognition, accounting for the majority of its total variance, are almost entirely independent from this measure, too. The multivariate models including both the *g* measure (whether as a composite or its components) and the CCMT show the same pattern: Face perception is almost entirely genetically distinct from both.

An alternative explanation for the CCMT results, of course, is that general object recognition is genuinely genetically dissociable from *g* to the same extent as face recognition. Although there is no reason from the literature to suspect that general object recognition may be special in this way (unlike the case for face recognition), the possibility merits further study, using both additional object recognition measures and also a broader *g* composite including memory performance. Even if it were true, however, the present findings would still be striking: Face recognition is genetically independent both from general object recognition and from all of the general cognitive abilities and processes captured both by this perfectly-matched task and by the *g* composite itself.

Quantitative genetic methods can estimate pleiotropic influences among traits—that is, the degree to which common genes influence multiple traits and drive the observable associations between them—even where the specific genes involved have not yet been identified (25). Among other things, calculating the genetic correlations between traits thus supports both the search for the genes themselves (by predicting the patterns of traits with which they are likely to be associated) and also theorizing about their possible mechanisms and modes of action. Caution is warranted, however: The shared and unique genetic influences on face recognition could reflect the genetic etiology of domain-general and domain-specific cortical development, for example, but equally they could represent influences detectable only as higher level aspects of cognition, behavior, or personality. In any case, our results indicating that face recognition is largely genetically distinct from other cognitive and perceptual domains suggest that identifying genes associated with domain-general cognitive processes (or indeed general cortical development) will be of limited use in understanding face recognition.

It should be emphasized that these results do not rule out the "expertise" hypothesis of face recognition: In principle, the unique genetic influences identified could be unique to the skilled recognition of very highly familiar (i.e., learned from an early age) categories of objects, rather than to faces specifically. Equally, nor do they confirm that those apparently unique influences truly affect faces alone because it remains possible that they may be shared with abilities or traits not captured adequately by the object recognition or *g* measures used. One possibility in the latter direction is suggested by the various attempts to conceive a broad domain of "social intelligence" (26, 27): Perhaps socially relevant abilities other than face recognition, such as emotion recognition or theory of mind, may be found to share some or all of its genetic etiology.

The very fact that *g* correlates with diverse specific cognitive abilities by only 0.4 on average (9) means that much, and usually most, of the variance in each domain is specific rather than general: All domains are special in this sense. The genetic near-independence of face recognition both from *g* (in contrast to the usual rule) and from a perfectly matched task of general object recognition is striking, however, and lends weight to the view that this ability is more special than most.

Methods

Measures. The Cambridge Face Memory Test (CFMT) (20) instructs participants to memorize six male Caucasian faces, each from three images showing the face in different orientations. Images show the faces with neutral expressions, edited to remove any distinguishing facial blemishes and cropped to remove hair and clothing. Test stimuli present a target face alongside two distractors, and participants make keyboard responses to identify the target. Trials fall into three distinct phases, the first following immediately after the memorization of each face (three trials for each, identifying that face among distractors), the second being a series of 30 trials in which the target can be any of the six memorized faces, and the third a series of 24 trials (again with any target) using impoverished images degraded with Gaussian noise. Correct responses are summed to give a total score out of 72.

The Cambridge Car Memory Test (CCMT) (21), a nonsocial object recognition task, was developed as a matched companion task to the CFMT. The stimuli are computer-generated 3D images of cars, viewed in various orientations. The cars depicted are loosely modeled on contemporary real-world designs, but altered so as not to be identifiable as real cars with which participants may be familiar. The procedure for memorization, testing and scoring is identical to that for the CFMT.

The general cognitive ability (*g*) composite used in this study is the mean of participants' standardized scores on two measures. The first, assessing verbal cognitive ability, is the Mill Hill Vocabulary Scale (22), a multiple-choice test of vocabulary: In each of 33 trials, a target word is presented, and participants select (by clicking on screen) which of six options is closest to it in meaning; correct responses are summed. Nonverbal cognitive ability was assessed using Raven's Progressive Matrices (23): Participants are shown an incomplete pattern and asked each time to select which of eight options completes it; correct responses are summed across 30 trials. Although these measures were administered 3 years earlier than the CFMT and CCMT, this gap is highly unlikely to have influenced the results because the genetic influences on *g* have been found to be highly stable over considerably longer periods (28, 29).

Ethical approval was granted by the relevant ethics committee (Psychiatry, Nursing & Midwifery, at King's College London), and informed consent was obtained. TEDS participants were contacted by post but completed all

measures online via websites developed for the purpose. The *g* battery (administered at age 16) was developed using the Flash browser plugin, and the CFMT and CCMT (administered separately; see *Results, Data*) were developed in Javascript, using the open-source "psy.js" library (https://www.forepsyte.com/resources/public/psy-1.56.js).

Twin Data. Twin studies analyze the intrapair concordances or correlations between monozygotic (MZ) and dizygotic (DZ) twins (9). MZ twins share all their genes whereas DZ twins share (on average) only half of their segregating genes; both share their environments to approximately the same extent. The degree to which the MZ is higher than the DZ correlation thus indicates the degree of genetic influence on a trait, and cross-twin cross-trait correlations allow genetic covariance between traits to be quantified. The twin method relies upon certain assumptions (such as the assumption that MZ and DZ twins share their environments to approximately the same degree), but these assumptions have been widely tested, and other study designs with different assumptions typically replicate results.

Twins are perfectly correlated for age, and MZ and half of DZ twins also for sex. Any effects of age or sex on a trait would thus distort the "true" intrapair correlations and inflate the apparent role of shared environmental influences (30). For this reason, standard practice for twin data is to analyze residuals corrected for any mean effects of age and sex. (This practice does not preclude sex differences being analyzed where appropriate because twin analyses are concerned with variances, which are unaffected by correcting for mean differences.) In addition, for the present data, outliers were removed for each measure beyond three SDs from the mean, random responders were removed (defined as participants with infeasibly low median item response times, under 1.5 SDs below the sample mean), and the dataset was normalized with a van der Waerden transformation (accounting for a slight negative skew present in the raw CFMT and CCMT data).

Model Fitting. Twin analyses were conducted using model-fitting procedures, allowing point estimates and confidence intervals to be established for the variance component estimates, and the goodness of fit of the model to the data to be tested (31). This test may be achieved by comparing the fit statistics of the model to a fully saturated model in which all parameters are allowed to vary, and no particular structure is imposed on the data—if the fit of the constrained model is not significantly worse than that of the saturated model, it may be considered a good fit. A series of maximum-likelihood nested models were applied and fitted to the data (32), based upon the expected genetic and environmental correlations (additive genetic influences correlating 1.0 for MZ twins and 0.5 for DZ twins, and shared environmental influences 1.0 for both). Additive genetic (A), shared environmental (C), and unique environmental (E) influences were estimated, and nested submodels tested which components were required. Any error of measurement was included in the E estimate, deflating A and C equally. All model-fitting was conducted in R, using the structural equation program OpenMx (33).

Multivariate model fitting, based upon cross-twin cross-trait correlations, decomposed phenotypic covariance between traits into genetic and environmental components of covariance. Two algebraically equivalent models with different analytic foci (34) were analyzed in this study. First, a "correlated factors" solution permitted the common A, C, and E influences underpinning multiple traits to be estimated, and thus the phenotypic correlation between them to be decomposed into these components of covariance (as in *SI Appendix*, Table S4). Second, in a manner analogous to a phenotypic stepwise multiple regression, Cholesky decomposition permitted the ACE influences shared between two or more traits to be determined sequentially, estimating at each step the proportion of the A, C, and E components shared with, and independent from, each variable. Thus, in *SI Appendix*, Fig. S1, showing the structure of additive genetic influences (A), path estimates indicate the proportion of genetic influences common to both the predictor variable and to face recognition, and the proportion unique to the latter. *SI Appendix*, Fig. S2 illustrates a trivariate extension, showing the genetic influences common to all three variables (*g*, object recognition, and face recognition), then the influences common to object recognition and face recognition (but not to *g*), then finally the residual influences unique to face recognition. Further extensions may likewise include additional variables (as in *SI Appendix*, Fig. S3).

ACKNOWLEDGMENTS. We thank the twins in the Twins Early Development Study (TEDS) for making the study possible. TEDS is supported by a program grant to R.P. from the United Kingdom Medical Research Council (MRC) (Grant G0901245 and, previously, Grant G0500079), with additional support from US National Institutes of Health Grants HD044454 and HD059215. N.G.S. is supported by an MRC studentship. R.P. is supported by Medical Research Council Research Professorship Award G19/2 and European Research Council Advanced Investigator Award 295366.

1. Alloway TP, Gregory D (2013) The predictive ability of IQ and working memory scores in literacy in an adult population. *Int J Educ Res* 57:51–56.
2. Ashton MC, Vernon PA (1995) Verbal and spatial abilities are uncorrelated when g is controlled. *Pers Individ Dif* 19(3):399–401.
3. Alloway TP, Passolunghi MC (2011) The relationship between working memory, IQ, and mathematical skills in children. *Learn Individ Differ* 21(1):133–137.
4. Duff K, Schoenberg MR, Scott JG, Adams RL (2005) The relationship between executive functioning and verbal and visual learning and memory. *Arch Clin Neuropsychol* 20(1):111–122.
5. Plomin R (1988) The nature and nurture of cognitive abilities. *Advances in the Psychology of Human Intelligence*, ed Sternberg R (Lawrence Erlbaum Associates, Hillsdale, NJ), Vol 4, pp 1–33.
6. Petrill SA (1997) Molarity versus modularity of cognitive functioning? A behavioral genetic perspective. *Curr Dir Psychol Sci* 6(4):96–99.
7. Plomin R, Spinath FM (2002) Genetics and general cognitive ability (g). *Trends Cogn Sci* 6(4):169–176.
8. Carroll JB (1993) *Human Cognitive Abilities* (Cambridge Univ Press, New York).
9. Plomin R, DeFries JC, Knopik VS, Neiderhiser JM (2013) *Behavioral Genetics* (Worth, New York), 6th Ed.
10. Kovas Y, Plomin R (2006) Generalist genes: Implications for the cognitive sciences. *Trends Cogn Sci* 10(5):198–203.
11. Ishai A (2008) Let's face it: It's a cortical network. *Neuroimage* 40(2):415–419.
12. Lee K, Anzures G, Quinn PC, Pascalis O, Slater A (2011) Development of face processing expertise. *The Oxford Handbook of Face Perception*, eds Calder AJ, Rhodes G, Johnson MH, Haxby JV (Oxford Univ Press, New York), pp 753–778.
13. McKone E, Palermo R (2010) A strong role for nature in face recognition. *Proc Natl Acad Sci USA* 107(11):4795–4796.
14. McKone E, Kanwisher N, Duchaine BC (2007) Can generic expertise explain special processing for faces? *Trends Cogn Sci* 11(1):8–15.
15. Wilmer JB, et al. (2010) Human face recognition ability is specific and highly heritable. *Proc Natl Acad Sci USA* 107(11):5238–5241.
16. Zhu Q, et al. (2010) Heritability of the specific cognitive ability of face perception. *Curr Biol* 20(2):137–142.
17. Dale PS, et al. (1998) Genetic influence on language delay in two-year-old children. *Nat Neurosci* 1(4):324–328.
18. Haworth CMA, Davis OSP, Plomin R (2013) Twins Early Development Study (TEDS): A genetically sensitive investigation of cognitive and behavioral development from childhood to young adulthood. *Twin Res Hum Genet* 16(1):117–125.
19. Price TS, et al. (2000) Infant zygosity can be assigned by parental report questionnaire data. *Twin Res* 3(3):129–133.
20. Duchaine B, Nakayama K (2006) The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia* 44(4):576–585.
21. Dennett HW, et al. (2012) The Cambridge Car Memory Test: A task matched in format to the Cambridge Face Memory Test, with norms, reliability, sex differences, dissociations from face memory, and expertise effects. *Behav Res Methods* 44(2):587–605.
22. Raven JC, Court JH, Raven J (1998) *Manual for Raven's Progressive Matrices and Vocabulary Scales* (H. K. Lewis, London).
23. Raven JC, Court JH, Raven J (1996) *Manual for Raven's Progressive Matrices and Vocabulary Scales* (Oxford Univ Press, Oxford).
24. McKone E, et al. (2012) A robust method of measuring other-race and other-ethnicity effects: The Cambridge Face Memory Test format. *PLoS One* 7(10):e47956.
25. Haworth CMA, Plomin R (2010) Quantitative genetics in the era of molecular genetics: Learning abilities and disabilities as an example. *J Am Acad Child Adolesc Psychiatry* 49(8):783–793.
26. Goleman D (2007) *Social Intelligence: The New Science of Human Relationships* (Bantam, New York).
27. Petrides KV (2011) Social intelligence. *Encyclopedia of Adolescence*, eds Brown BB, Prinstein MJ (Academic, San Diego), pp 342–352.
28. Deary IJ, et al. (2012) Genetic contributions to stability and change in intelligence from childhood to old age. *Nature* 482(7384):212–215.
29. Lyons MJ, et al. (2009) Genes determine stability and the environment determines change in cognitive ability during 35 years of adulthood. *Psychol Sci* 20(9):1146–1152.
30. McGue M, Bouchard TJ, Jr (1984) Adjustment of twin data for the effects of age and sex. *Behav Genet* 14(4):325–343.
31. Rijsdijk FV, Sham PC (2002) Analytic approaches to twin data using structural equation models. *Brief Bioinform* 3(2):119–133.
32. Neale MC, Boker SM, Xie G, Maes HH (2006) *Mx: Statistical Modeling* (Virginia Commonwealth University, Richmond, VA).
33. Boker S, et al. (2011) OpenMx: An open source extended structural equation modeling framework. *Psychometrika* 76(2):306–317.
34. Loehlin JC (1996) The Cholesky approach: A cautionary note. *Behav Genet* 26(1):65–69.

Chapter 7:- STEM is spatial, English is social: genetic dissociations in the prediction of educational achievement

This chapter, using spatial ability and face recognition to predict educational outcomes, has been adapted and expanded from a manuscript currently being prepared for submission as a paper to *Scientific Reports*:

Shakeshaft NG[1], Rimfield K[1], Malanchini M[1,2], Schofield KL[1], Rodic M[3], Selzam S[1], Plomin R[1] (in preparation). STEM is spatial, English is social: genetic dissociations in the prediction of educational achievement.

[1]Medical Research Council Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, United Kingdom

[2]Goldsmiths, University of London, New Cross, London, United Kingdom

[3]University of Sussex, Sussex House, Falmer, Brighton, United Kingdom

Supplementary materials for this chapter, as detailed in the text, are attached as Appendix 5.

Abstract

Understanding and predicting differences in educational achievement has considerable societal value, with great potential to improve outcomes if their origins can be identified. Spatial ability has been found to be a strong predictor, particularly for science, technology, engineering and mathematics (STEM) subjects. Social and emotional skills have also been suggested to play an important role, contributing to students' motivation and engagement, perhaps disproportionately for non-STEM subjects. In both cases, the nature of the relationships are not well understood.

4000 twins from the Twins Early Development Study provided their examination grades at the end of compulsory education. Together with measures of spatial ability and face recognition (an important social skill), the associations and their aetiology are explored in detail. Results confirm that the strong relationship between spatial ability and STEM subjects is largely genetic in nature, and show that it rests on complex manipulations rather than simple perceptual processes. Face recognition is shown to be phenotypically and genetically predictive of English grades, but not of STEM subjects, suggesting that social skills have differential relationships with different subjects. Taken together, the results indicate that the prediction of educational achievement should be considered separately for different academic subjects.

Introduction

The differences between children in their academic achievements, particularly in their final grades at the end of compulsory education, have profound consequences for their lives. In contrast to common assumptions (Asbury and Plomin, 2013), previous research has shown these differences to be substantially genetic in origin, with common genes associated to some extent with grades across diverse academic subjects (Rimfeld *et al.*, 2015), and environmental influences accounting for a little under half of the variation in scores (Nicoletti and Rabe, 2013; Shakeshaft *et al.*, 2013). The factors found to be predictive of academic outcomes are extremely complex, with diverse traits investigated as possible correlates or causes, such as general cognitive ability (intelligence, or *g*; Deary *et al.*, 2007; Mackintosh, 2011), working memory (Alloway and Alloway, 2010), as well as non-cognitive traits such as self-efficacy (Zuffianò, 2013), personality (Stankov, 2013 ; Briley *et al.*, 2014), behaviour problems (Pingault *et al.*, 2011) and physical health (De Ridder *et al.*, 2013). The genetic component of academic achievement has been found to be related strongly to many non-cognitive traits as well as to *g*, but the latter appears to explain as much of the heritability of educational outcomes as dozens of non-cognitive traits and influences combined (Krapohl *et al.*, 2014).

Spatial ability

While *g* has been shown to be highly predictive of educational achievement in general (Kaufman *et al.*, 2012), abilities in specific domains are thought to have varying degrees of association with achievement in different academic subjects. A substantial body of research in this area has focused on spatial ability, which is widely considered to be multifactorial, and may be defined loosely as the recognition and mental manipulation of visual stimuli and the relations between them (Lohman, 1994; Hegarty and Waller, 2005). This specific cognitive ability has been found to be associated with both verbal and mathematical achievement (Shea *et al.*, 2001), but is typically found to be especially predictive for the so-called “STEM” subjects – Science, Technology, Engineering and Mathematics – for which it explains a substantial portion of variance even with *g* controlled (Rohde and Thomson, 2007; Wai *et al.*, 2009).

Although this is a common finding, little is known about the aetiology of the association. It

has been observed that spatial and numerical representations appear to invoke overlapping parietal networks (Hubbard *et al.*, 2005), and the single multivariate behavioural genetic study to date found the substantial correlation between spatial ability and mathematical performance to be 60% genetic in origin (Tosto *et al.*, 2014) – but the exact nature of the relationship remains unclear. In science subjects, it has been noted that spatial ability is associated most strongly with questions requiring the restructuring of information rather than mere rote learning (Carter *et al.*, 1987); and similarly in mathematics, that more gifted students tend to solve problems using visuospatial, schematic representations of the required information, whereas less able students use simpler pictorial strategies (Van Garderen and Montague, 2003). Such observations have led to suggestions that certain aspects of spatial ability – such as those involving visualisation and mental manipulation – may be more relevant to STEM success than others (such as perceptual speed; Rohde and Thomson, 2007). However, this is yet to be tested, and would be at odds with recent findings suggesting (very much in contrast to the wider literature; Carroll, 1993; Hegarty and Waller, 2005) that spatial ability is in fact unifactorial, both phenotypically and genetically – this was the subject of two recent behavioural genetic studies (Shakeshaft *et al.*, 2016; Rimfeld, Shakeshaft *et al.*, under review), but the measures used in these studies have yet to be directly compared.

Social and emotional skills

Another set of specific abilities researched widely for their potential to predict educational achievement is that of social and emotional intelligence (Petrides, 2011; Mayer *et al.*, 2008). These domains, encompassing the regulation and recognition of emotions and other socially-relevant skills, have been argued to influence academic outcomes significantly in several ways: it may promote the development of intrinsic motivation to study (Baumeister *et al.*, 1994), be reflected in different (and more or less adaptive) strategies for managing stress (MacCann *et al.*, 2011), or influence relationships with teachers (Halberstadt and Hall, 1980). Regardless of the mechanisms, a number of studies have reported modest relationships between social or emotional competence and educational achievement (Nowicki and Duke, 1992; Teo *et al.*, 1996; Brackett *et al.*, 2004; Parker *et al.*, 2004; Graziano *et al.*, 2007; Costa and Faria, 2015) (although significant associations are not always found; Newsome *et al.*, 2000; Kashani *et al.*, 2012). As with spatial ability, differential relationships have been observed for different academic subjects, but with less consistency: emotion recognition, for example, has been reported to be equally predictive for mathematics as for reading

comprehension and spelling (Nowicki and Duke, 1994), but other studies find it to have lower correlations (or none) with mathematics (Petrides *et al.*, 2004; Downey *et al.*, 2008). It is possible that different components of social and emotional intelligence have different educational associations (Downey *et al.*, 2008), but the somewhat underspecified nature of some of these components makes it difficult to investigate these relationships with precision.

One highly specific socially-important skill is the ability to recognise faces (Wilmer *et al.*, 2010). This ability is (at least partially) dissociable from other social abilities such as recognising emotions, even from faces (Fitousi and Wenger, 2013; Chen, 2014). Face recognition is highly heritable (Wilmer *et al.*, 2010; Shakeshaft and Plomin, 2015) and almost entirely distinct from general object recognition and from *g*, both phenotypically and genetically (Ishai, 2008; Shakeshaft and Plomin, 2015). This does not preclude it correlating genetically with educational achievement for reasons unrelated to *g*, however, given that non-cognitive traits form a substantial portion of the heritability of the latter (Krapohl *et al.*, 2014). In light of the literature suggesting that social competence plays an important role in educational outcomes (Teo *et al.*, 1996), therefore, face identity recognition should be predicted to reveal these relationships, notwithstanding its independence from other abilities, provided that emotional intelligence (rather than social intelligence more widely) does not mediate this relationship entirely. To date, no studies have investigated this.

Present study

There are many gaps in our understanding of the relationships between specific abilities and educational achievement. For spatial ability, it is unknown whether the substantial genetic association with mathematics revealed by the only behaviour genetic study to date (Tosto *et al.*, 2014) extends to other STEM subjects or to non-STEM subjects; nor is it known whether some putative aspects of spatial ability might be better predictors than others. With social abilities, it is unclear whether they are more strongly associated with some academic subjects than with others, whether these associations are genetic or environmental in origin, and whether they rely on emotional intelligence (such as emotion regulation promoting the development of motivation; Baumeister *et al.*, 1994) rather than “social competence” (Teo *et al.*, 1996) more generally.

In the present study, educational achievement data from a large sample of twins were analysed

(General Certificate of Secondary Education grades; GCSEs – see Methods), together with scores on two spatial ability measures and one face recognition measure. The spatial scores reflect performance on a "narrow" measure, focusing on visualisation alone, and a "broad" measure assessing more diverse spatial skills. Face recognition makes an ideal comparison to spatial ability, we contend, as both involve the discrimination and identification of visual stimuli in different orientations. If simple, low-level perceptual processes shared between these tasks were sufficient to drive the association between spatial ability and STEM achievement, these same relationships should be found for face recognition, too; but if more complex mental manipulations crucially underpin this relationship as some of the literature suggests (Rohde and Thomson, 2007), spatial ability will be unmatched. Thus there were three main aims:

i) To clarify the strength and nature of the relationships between spatial ability and educational outcomes, by investigating the phenotypic and aetiological associations with STEM and non-STEM academic subjects, as well as with overall achievement.

ii) To establish which aspects of spatial ability are the key predictors of STEM success. If the previous literature were correct and spatial ability is fractionated, the "broad" and "narrow" measures should be discriminable, and the broad measure (accounting for more spatial variance) should be the better predictor of educational achievement. Further, if the relationship is driven by complex mental manipulations rather than simpler perceptual processes, spatial ability should be a much better predictor than face recognition.

iii) To clarify the nature and aetiology of the relationships between social skills and educational outcomes (overall and for specific academic subjects), by assessing the predictive potential of face identity recognition. If emotional intelligence, specifically, underlies the relationships previously observed, then face recognition – despite its social value – should have no associations with educational achievement at all.

Results

Data

This study was conducted as part of the Twins Early Development Study (TEDS), a

longitudinal study of more than 10,000 pairs of British twins (Haworth *et al.*, 2012). From this sample, every participant was included for whom the necessary data were available: their GCSE grades, and scores for at least one of the three predictors of interest. The GCSE variables analysed were composites representing examination grades in three specific subject areas – Mathematics, Science and English – and four broader composites representing aggregate scores across multiple subjects. The key predictor variables were: i) the "Bricks" test (Shakeshaft *et al.*, 2016), a narrow spatial ability measure of visualisation and mental rotation; ii) the "King's Challenge" (KC) battery (Rimfield, Shakeshaft *et al.*, under review), a much broader spatial ability battery of ten diverse tests, from which a single composite score was derived; and iii) a "pure" measure of face recognition, created by regressing a widely-used face recognition test (Duchaine and Nakayama, 2006) on a matched non-face object recognition test (Dennett *et al.*, 2012). A measure of verbal ability was also used for some analyses, as assessed with the Mill Hill Vocabulary Scale (Raven *et al.*, 1998). These measures are described in more detail in Methods.

Data preparation procedures are described in Methods. The resulting dataset contained 3916 participants: 1729 complete twin pairs – 698 monozygotic (MZ), 1031 dizygotic (DZ) – and 458 unpaired individuals with no data available for their co-twins (118 from MZ pairs, 340 DZ). This sample was 61.0% female, and completed the verbal ability measure at a mean age of 16.5 years (±0.26 SD), shortly after sitting their GCSE exams. The participants completed the face and object recognition tests at a mean age of 19.5 years (±0.32 SD), the Bricks battery at 20.3 years (±0.48 SD), and the KC measures at 20.6 years (±0.48 SD).

Descriptive statistics for all measures are shown in Supplementary Table S1. Analyses of variance (ANOVAs) were performed to assess the mean effects of sex and zygosity for each measure. Significant effects of sex were found for several variables, most notably representing modest advantages for males on the spatial ability measures, and for females on face recognition. Mean differences are not relevant to twin analyses, which focus on variance. Nonetheless, the data for all subsequent analyses were regressed on sex as well as age (see Methods), normality-transformed and standardised.

Phenotypic analyses

A fully independent sample was used for all phenotypic analyses, by selecting one twin

randomly per pair.

The phenotypic correlations among the predictors (i.e., the face/object recognition and spatial ability measures) are presented in Supplementary Table S2. The correlation between the Bricks and KC spatial ability measures (r = 0.65) is close to the previously-reported test-retest reliabilities of these measures – r = 0.83 for Bricks (Shakeshaft *et al.*, 2016), r = 0.75 for KC (Rimfield, Shakeshaft *et al.*, under review) – meaning that they share a large majority of their reliable variance. However, this leaves a modest proportion of reliable variance not shared between them, which is consistent with the observation that they have slightly different phenotypic relationships with other measures: Bricks is modestly correlated with face recognition (r = 0.16) whereas KC is not, and Bricks also has a slightly stronger relationship with object recognition (r = 0.31) than KC does (r = 0.21). This difference does not persist for the "pure" face recognition measure used for subsequent analyses – i.e., face recognition regressed on non-face object recognition – which has no significant correlation with either Bricks or KC. This indicates that general perceptual or other domain-general processes (and perhaps a very modest portion of spatial ability captured by both the face- and non-face recognition measures) are all that face recognition has in common with the spatial ability variables.

Controlling verbal ability does not significantly alter any of the predictors' intercorrelations (Supplementary Table S3), indicating that vocabulary size does not mediate any of these relationships. This suggests that these associations are not significantly driven by verbal ability itself, by domain-general considerations such as understanding the instructions, or indeed by any other general test-taking factors (such as attention or motivation) which may be captured by the vocabulary measure.

The GCSE measures are all highly intercorrelated, as shown in Supplementary Table S4. This is unremarkable for the cross-subject composites (which share part-whole overlaps with all other GCSE measures), but the strength of the associations is not substantially lower even between the four completely independent subject variables: English, Humanities, Maths and Science. Among these, the strongest relationship is between Maths and Science (r = 0.79), and the lowest between Maths and Humanities (r = 0.63). This difference is significant ($p < 0.001$), as are most of the comparisons in this very large and highly-powered sample, but the majority of these differences are not substantial – for example, the relationship between English and Humanities (r = 0.72) is little different from that between English and Maths (r =

0.69). Controlling verbal ability reduces these GCSE intercorrelations only very modestly (0.05 on average – see Supplementary Table S5), and this reduction is quite uniform, suggesting that vocabulary size (and other general test-taking factors) are not significantly more responsible for the relationship between, for example, English and Humanities than it is for that between English and Maths.

The correlations between the predictors and the GCSE measures are presented in Table 1. Bricks is moderately correlated with all GCSE measures (r = 0.35 on average), and its associations with English (r = 0.31) and Humanities (r = 0.28) are modestly but significantly (all $p < 0.001$) weaker than those with Maths (r = 0.43) and Science (r = 0.36). The King's Challenge measure has slightly stronger associations with the GCSE grades overall (r = 0.40 on average), but this difference is derived entirely from its stronger association with the STEM subjects: its relationships with English (r = 0.30) and Humanities (r = 0.30) do not differ from those of Bricks, but its correlations with Maths (r = 0.51) and Science (r = 0.43) are significantly stronger (both $p < 0.001$). “Pure” face recognition is a modest predictor of all the GCSE measures (r = 0.13 on average); the differences among its relationships with different subjects are not very substantial, but its correlation with English (r = 0.17) is almost double that with Maths (r = 0.09), and this difference is significant ($p < 0.01$).

Non-face object recognition is weakly associated with Maths (r = 0.13) and Science (r = 0.07), but not at all with English or Humanities. Presumably owing to the influence of the latter, it is not significantly associated with the mean GCSE grade. With the sole exception of its non-significantly ($p = 0.40$) stronger relationship with Maths (r = 0.13, compared to 0.09 for “pure” face recognition), its associations with the GCSE variables are substantially weaker overall than those of face recognition.

Controlling verbal ability (Supplementary Table S6) reduces these associations only very modestly (0.04 on average). This reduction is highest for the spatial ability measures: 0.08 on average, representing around one fifth of their uncorrected correlations with the GCSE composites, thus indicating that verbal ability (and any other domain-general factors captured by the vocabulary measure) does account for a substantial portion of the spatial ability measures' prediction of educational outcomes. For “pure” face recognition, however, there is barely any reduction at all: 0.01 on average, none of these differences significant (all $p > 0.05$). For each predictor, the reduction in the strength of association is fairly uniform across all GCSE measures, although the spatial ability measures' correlations with Humanities and

English are perhaps reduced slightly more substantially (by 0.10 on average) than those with Science and Maths (0.07).

Univariate genetic analyses

Univariate genetic results for these data – estimating the proportions of variance in the measures attributable to additive genetic (A), shared environmental (C) and non-shared environmental (E) components – have been presented previously, for the GCSE data (Shakeshaft *et al.*, 2013), for the face and object recognition measures (Shakeshaft and Plomin, 2015), and for the Bricks (Shakeshaft *et al.*, 2016) and King's Challenge (Rimfield, Shakeshaft *et al.*, under review) spatial ability measures. The sample for the present analyses does not overlap perfectly with those used in the previous analyses, however – and the "pure" face recognition measure has not been presented previously – so Supplementary Table S7 presents the univariate model-fitting results for all measures with the present sample, together with an indication of the sample size available for each variable. As shown, genetic influence is substantial for all measures, with heritabilities mostly above 50%. The key predictors all have heritabilities around 60%.

Shared environment accounts for 30% of the variance on average for the GCSE variables, but none of the predictor measures show any significant shared environmental influence at all. The latter observation is borne out by the fit statistics for the univariate models (Supplementary Table S8), indicating that AE submodels (dropping C) show no deterioration in fit for the predictors, but fit much more poorly for the GCSE variables. This prompted the design of the multivariate models used for all of the subsequent genetic analyses below, in which all shared environmental (C) paths were constrained to zero, except for those influencing the GCSE scores uniquely (see Methods).

The present study focuses on the multivariate relationships between the GCSE scores and the three key predictors: Bricks, KC and "pure" face recognition. The phenotypic correlations between these predictors and the GCSE variables (Table 1), and among the GCSE variables themselves (Supplementary Table S4), suggested that several of the latter were redundant: the three summary GCSE composites (the overall mean grade, the number of A*-C grades, and the mean grade for "core" subjects) are highly intercorrelated and show similar associations with the predictors, as do the Humanities and English variables. Thus to simplify the analyses,

only four of the GCSE variables were retained for further analysis: Maths, Science, English, and the "core subjects" composite representing the mean of these three subjects. The raw face and object recognition measures were also dropped from subsequent analyses, leaving only the three predictors of interest: Bricks, KC and "pure" face recognition.

Bivariate genetic analyses: predictor interrelationships

The relationships among the three predictor variables themselves were examined. Bivariate correlated factors solutions (Methods) were fitted to each pair of variables, allowing their phenotypic correlations (Supplementary Table S2) to be decomposed into genetic, shared and non-shared environmental components. These results, presented in Supplementary Table S9, indicate that the substantial phenotypic correlation between Bricks and KC ($r = 0.65$) is overwhelmingly genetic in origin: 89%. This is similarly reflected in the genetic and non-shared environmental correlations between Bricks and KC (Supplementary Tables S10 and S11): the non-shared environmental correlation is significant but modest (rE = 0.23), but their genetic correlation (the degree to which they are driven by common genes) is virtually at unity: rA = 0.98. Genetically at least, there is no distinction between these spatial measures.

For completeness, results are also included in these tables (Supplementary Tables S9-S11) similarly examining the relationships between the spatial measures and face recognition. However, as their phenotypic correlations are non-significant in each case (Table S2), these are not meaningful. The only significant result is a modest genetic correlation (rA = 0.20) between face recognition and Bricks, but since this does not differ significantly from the non-significant genetic correlation between face recognition and KC (i.e., the confidence intervals overlap), and since Bricks and KC themselves correlate genetically at unity, this seems most likely to be a chance result. Taken together, it appears that spatial ability and "pure" face recognition are entirely independent.

Finally, bivariate Cholesky decompositions (Methods) revealed, for each predictor, what proportions of their variance components were shared with, and unique from, each other predictor. These results (Supplementary Tables S12-S14) confirm the expectations arising from the analyses above: the substantial genetic influence on the spatial measures is shared entirely between both, whereas the equally substantial genetic component of face recognition is completely unique to it. Non-shared environment is substantial for each measure

individually, but no significant common paths emerge.

Fit statistics for these models are presented in Supplementary Table S15.

Bivariate genetic analyses: predictors and GCSEs

Twelve further bivariate genetic analyses were then conducted, representing each pairing of the three predictors and four GCSE measures, in order to examine the aetiology underpinning their phenotypic relationships.

Bivariate factor solutions decomposed the phenotypic correlations into portions attributable to genetic and non-shared environmental influences. (Common shared environmental influences were constrained to zero, as described in Methods, in view of the fact that none of the predictors show any shared environmental influence). The results, presented in full in Supplementary Table S16, show that an overwhelming majority of the relationships between each predictor and each GCSE measure are mediated by common genetic influences – these account for 90% of the phenotypic correlations on average, with minimal (sometimes not even significant) non-shared environmental influences accounting for the remainder. It is striking how similar all of these results are: the magnitude of the genetic portion of shared variance does not differ significantly between any of the twelve analyses, with the point estimates ranging only from 0.85 to 0.93. The phenotypic correlations in question (Table 1) vary considerably in magnitude, of course – spatial ability correlates two or three times more highly with each GCSE measure than “pure” face recognition, for example – but the aetiology of these relationships is remarkably uniform, regardless of size.

The magnitude of the aetiological overlap between the measures is revealed by the genetic and non-shared environmental correlations, presented in Tables 2 and 3, respectively. The genetic correlation is significant for every pairing of predictors and GCSE measures. For the two spatial ability measures (Bricks and KC), they are also very substantial: 60% on average for the four models with Bricks as the predictor, and 58% for KC. (These averages are slightly artificial, as they include the “core subjects” composite which is itself a mean of the other three measures, but are useful for illustration).

The genetic correlations with each GCSE measure do not differ significantly between Bricks

and KC. This is unsurprising considering that Bricks and KC correlate genetically almost at unity, as noted above, but offers no explanation for the KC measure's significantly higher phenotypic correlations with Maths and Science – but not English – compared with those for Bricks (Table 1). Instead, the non-shared environment correlations provide the probable answer (Table 3): these are 14% on average for Bricks, but 26% for KC. The difference is significant only for Science, but the point estimates are higher for KC with all of the GCSE measures – with English as the sole exception. To the modest extent that KC offers stronger prediction of STEM GCSE subjects, it does so due to greater non-shared environmental effects in common with them.

For face recognition, the genetic correlations with the GCSE measures are a third of those of the spatial measures, at 19% on average (Table 2), and it has no significant non-shared environmental correlations at all (Table 3). The genetic correlations are informative, however: there is little difference between them for Maths (12%) and Science (16%), but the magnitude is roughly double for English (28%). This difference is not significant (the confidence intervals overlap, albeit only barely for Maths and English) – but is perhaps instructive nonetheless, as it mirrors the phenotypic correlations, in which “pure” face recognition correlates twice as highly with English (r = 0.17) as it does with Maths (r = 0.09). The inverse is clear for the two spatial measures, which correlate genetically at 68% on average with Maths, 62% with Science, but only 43% with English, again mirroring the pattern of the phenotypic correlations (Table 1).

As with the predictor interrelationships above, Cholesky decompositions provide another way to examine these associations, directly estimating the portions of variance shared between the variables, and those unique to each. The results for all twelve bivariate decompositions are presented in Supplementary Tables S17-S20. Bricks and KC show no significant differences from one another in their relationships with any of the GCSE measures – indeed even the modest differences between their non-shared environmental correlations with Maths and Science (Table 3) are partially obscured in these models (probably for lack of power, the estimates being very small), as only the shared path for KC and Science reaches significance.

For Maths (Supplementary Table S17), genetic influences shared with the spatial predictors account for almost half of its genetic variance (e.g., with KC, 0.30 / (0.30 + 0.34) = 47%), and 30% of the total variance in Maths GCSE scores. With face recognition, the results are starkly different: the shared genetic path does not reach significance at all, leaving all of the

substantial heritability of Maths in the residual (unique) path. For Science (Supplementary Table S18), the results are similar: around one third of the genetic influence on this GCSE subject is shared with the spatial measures (e.g., for KC, the shared genetic path accounts for 0.21 / (0.21 + 0.41) = 34% of the heritability of Science, and 21% of its total variance), while face recognition again has no significant shared aetiology with Science at all.

English (Supplementary Table S19) shows an entirely different pattern. The spatial predictors do share a significant genetic relationship with this GCSE subject, but it is considerably weaker – KC, for example, accounts for only 16% of the heritability of English (0.10 / (0.10 + 0.51)) and 10% of its total variance. This is still double the genetic variance in common between English and face recognition, but – unlike the other GCSEs – the latter relationship is significant, with face recognition accounting for 8% of the heritability of English, and 5% of its total variance.

For the "core subjects" composite (Supplementary Table S20) representing the mean of these three individual GCSE subjects, all three predictors show significant genetic (and no environmental) overlap: KC accounts for 40% of its heritability and 25% of its total variance (Bricks being virtually identical), while face recognition explains 3% of its heritability and 2% of total variance.

See Supplementary Tables S17-S20 for full results and confidence intervals. Fit statistics for these models are presented in Supplementary Table S21.

Multivariate genetic analyses: verbal ability, predictors and GCSEs

The bivariate Cholesky decompositions above indicate the strength of the aetiological relationships between each pair of predictors and educational outcome measures, but cannot provide any insight into what exactly their overlap represents: for example, the extent to which the predictor might be capturing domain-general rather than more specific variance. To examine this, the bivariate models were expanded to include verbal ability as the first-entered variable. Using verbal ability as a conservative proxy for domain-general abilities and effects (see Methods), these trivariate analyses reveal the residual relationships between the three predictors of interest and the four GCSE measures, once general factors are accounted for. Path estimates for these twelve additional models are presented in Supplementary Tables S22-

S25. There are no significant common non-shared environmental paths (with the sole exception of KC and Science again, as with the bivariate models), so the focus is on the genetic relationships revealed. Bricks and KC again show essentially identical patterns.

As in the bivariate analyses, Maths and Science (Supplementary Tables S22 and S23) show very similar patterns of relationships. For Maths, verbal ability accounts for almost half of its genetic variance (e.g., 0.30 / (0.30 + 0.13 + 0.22) = 46% for the model including KC), and around 30% of its total variance. A similar result emerges for Science, with verbal ability representing 37% of its total variance, and more than half of its heritability (e.g., 58% in the model with KC). With this domain-general shared variance accounted for, spatial ability explains an additional ~20% of the heritability of Maths, and ~10% for Science (e.g., for KC and Maths, the residual genetic path represents 0.13 / (0.30 + 0.13 + 0.22) = 20% of heritability). With face recognition, however, these models show no residual genetic relationship at all with these GCSE subjects: once the variance captured by the vocabulary measure is accounted for, face recognition has no genetic relationship whatsoever with Maths or Science.

With English (Supplementary Table S24), the results are completely different. Verbal ability explains 39% of the total variance in this GCSE subject, representing over 60% of its heritability. With this accounted for, none of the predictors – Bricks, KC or face recognition – explain *any* additional significant genetic variance at all: at first glance, then, it appears that vocabulary (and any other relevant variance reflected in the verbal measure) is all that any of these predictors has in common with English GCSE grades. However, for the residual path with face recognition, the point estimate and the upper bound of the 95% confidence interval are both marginally higher than they are in the equivalent models for Maths and Science, raising the possibility that meaningful residual relationships could exist which these models are underpowered to detect. This is explored further below.

The “core subjects” composite (Supplementary Table S25) shows a similar pattern to that of Maths and Science – unsurprisingly, since these are two of the three subjects included. Genetic influences shared with verbal ability account for 66% of the heritability of this composite (43% of total variance), with spatial ability accounting for an additional ~11% (7% of total variance). Face recognition has no residual relationship with this composite once verbal ability is accounted for.

To establish how much of the heritability of each GCSE measure can be accounted for in total by verbal ability, face recognition and spatial ability together, four quadrivariate models were fitted to the data: see Supplementary Tables S26 (for Maths), S27 (Science), S28 (English) and S29 (the core subjects composite). Only one spatial measure was used, since KC and Bricks correlate genetically at unity: KC was chosen, since it is theoretically the "broader" measure (Methods). Unsurprisingly, since face recognition had no significant residual relationships in any of the trivariate analyses above, these expanded models did not advance on the proportion of GCSE heritability explained: face recognition had a loading at/near zero in each case, and the other variables' relationships were virtually identical to those in the corresponding trivariate models.

Fit statistics for these trivariate and quadrivariate models are presented in Supplementary Tables S30 and S31, respectively.

Multivariate genetic analyses: verbal-regressed predictors and GCSEs

The phenotypic and genetic analyses above could be taken to indicate strongly that face recognition has little if any meaningful relationship with educational achievement at all, and more specifically: i) that spatial ability is a substantially stronger predictor of the GCSE measures than face recognition for *all* subjects, not just for Maths and Science; and ii) that accounting for the influences shared with verbal ability abolishes all genetic relationships between face recognition and GCSE scores completely. However, two observations suggest otherwise. First, for the residual genetic path between face recognition and English with verbal ability accounted for (Supplementary Table S24), the point estimate and confidence interval upper bound is slightly (non-significantly) higher than the equivalent paths in the models predicting other GCSE subjects. More substantially, the phenotypic partial correlations in Supplementary Table S6 suggest that, with verbal ability controlled, the residual relationship between spatial ability and English is not substantially stronger than that between "pure" face recognition and English.

This raises the prospect of meaningful residual genetic relationships with face recognition that the models above were underpowered to detect. To assess this possibility, the spatial (KC) and "pure" face recognition variables were phenotypically regressed on verbal ability, and the resulting residuals were subjected to a final set of trivariate Cholesky decompositions. In this

way, analyses analogous to the four quadrivariate models (Supplementary Tables S26-S29) could be conducted while reducing their complexity and thus increasing power.

The genetic results are presented in Fig. 1, with full results and fit statistics in Supplementary Tables S32 and S33. For Maths, Science and the core subjects composite, these are little different from those above: face recognition has no significant genetic association with the GCSE measures, despite now being the first-entered variable in these models. Spatial ability explains 40% of the heritability of Maths, 24% that of Science, and 26% that of the "core" subjects composite – the decreases in comparison to the corresponding bivariate models (Supplementary Tables S17-S20) presumably reflecting the fact that the GCSE measures in the present models retain the variance shared with verbal ability, while the predictors do not.

For English, however, the results are illuminating: face recognition and spatial ability each explains a modest portion of genetic variance, and the difference in magnitude between these paths is not significant. With the domain-general factors captured by the vocabulary measure removed from each variable, face recognition and spatial ability are equally, and independently, predictive of English GCSE scores.

Discussion

Spatial ability and face recognition both predict educational outcomes significantly, but these relationships are very different and entirely independent. The one commonality between all of the associations between predictors and outcomes, without exception, is that they are overwhelmingly genetic in origin (Supplementary Table S16). To whatever extent each predictor explains the variation in GCSE grades, they do so largely as a result of genetics.

Spatial ability and STEM subjects

The associations between spatial ability and educational outcomes are substantial. Phenotypically (Table 1), these are strongest with Maths, slightly weaker for Science, and somewhat weaker again for English and the Humanities. Even with the non-STEM subjects, though, these are still substantial relationships ($r \sim 0.30$ for all comparisons), and only a relatively small proportion is accounted for by controlling verbal ability (Supplementary

Table S6). The genetic results are more revealing: the spatial ability measures are highly genetically correlated with Maths (Table 2), explaining half of its heritability (Supplementary Table S17), and not significantly less so with Science, explaining a third of its heritability (Supplementary Table S18). For English, however, the genetic correlations are considerably and significantly lower, explaining only a sixth of its heritability. These results offer strong support for the common finding that spatial ability is much more strongly predictive of STEM than of non-STEM subjects, and indicate that this is largely due to differences in the degree of genetic overlap between them.

The expanded models including verbal ability (Supplementary Tables S22-S24) are more revealing still. Verbal ability, representing a conservative proxy for domain-general abilities and influences (see Methods), was found to account for a substantial portion of the genetic variance previously attributed to the spatial ability measure, but – for the STEM subjects, at least – by no means all of it. With this removed, spatial ability still accounts for an additional 20% of the heritability of Maths and 10% for Science, but explains no significant additional variance for English. The final analyses presented, accounting for verbal ability by regression prior to the model-fitting in order to improve power (Fig. 1), suggest that a small amount of significant shared genetic variance with English actually does survive, but nonetheless the contrast with the STEM subjects is marked. Once domain-general factors are taken into account, spatial ability still explains a unique and substantial portion of the heritability of Maths and Science; whereas with English, very little of its previous relationship remains.

The “narrow” Bricks and “broad” KC spatial measures correlate phenotypically at close to the ceiling of their reliability, and genetically at unity. Their genetic relationships with the GCSE measures are essentially identical. This would seem to confirm that spatial ability is not fractionated, as the previous studies with these measures indicated (Shakeshaft *et al.*, 2016; Rimfield, Shakeshaft *et al.*, under review), and therefore to suggest that there are no specific subcomponents of spatial ability which could be considered better predictors of educational achievement than others, in contrast to previous predictions (Rohde and Thomson, 2007). Two possible notes of caution are warranted, however. First, the portion of the correlation with Maths explained genetically in the present study (~90% for both measures) is considerably higher than that reported in the only previous behaviour genetic study in this area (60%; Tosto *et al.*, 2014), suggesting that the specific spatial measures used may make a difference; however, that study was conducted with younger participants (age 12) and used different achievement variables to those in the present study, so is perhaps not really

comparable. Second, while the Bricks and KC measures performed identically in their *genetic* relationships, the same was not true environmentally: KC had higher non-shared environmental correlations with the STEM subjects than Bricks (Table 3), presumably accounting for the modestly higher phenotypic correlations observed (Table 1). It is difficult to speculate about the nature of the environmental influences driving this difference between the measures, but further investigation is warranted. On a practical note meanwhile, though, it suggests that broader, more diverse measures of spatial ability may perhaps be (slightly) better predictors of educational achievement, capturing environmental sources of variation that narrower measures may miss.

In the prediction of STEM subjects, any minor distinction between the spatial measures is dwarfed by the comparison with face recognition, which has very weak relationships with STEM subjects even with domain-general factors uncontrolled (Tables 1 and 2), and none at all once these are accounted for (Supplementary Tables S22 and S23). Face recognition is discussed in detail below, but the implication for spatial ability is clear: the general, low-level visual perceptual processes shared with face recognition do not account for the association between STEM subjects and spatial ability. This is consistent with the observations suggesting that more complex manipulations are involved (Carter *et al.*, 1987; Van Garderen and Montague, 2003), and indicates that spatial ability has a unique relationship with science and mathematics.

Face recognition and English

Previous research found face recognition to be largely distinct from other cognitive abilities (Ishai, 2008; Shakeshaft and Plomin, 2015). In the "pure" form that was the focus of most of the present analyses, though, this dissociation is total: face recognition is completely unrelated to spatial ability, either phenotypically or genetically. Despite this uniqueness in comparison to other cognitive domains – particularly once verbal ability is controlled, too (Supplementary Table S6) – it is perhaps surprising that any associations with the GCSE measures persist at all, yet they do. No significant partial correlation remains with Maths, only a marginal one survives (albeit highly significant with this large sample) for Science, but quite a respectable (if modest) relationship remains with English and several of the cross-subject composites. In fact, its associations are mostly much stronger than those of the object recognition task, which correlates weakly (or not at all) with any of the GCSE measures – in other words, face

recognition is a significantly better predictor of educational achievement than general object recognition, despite these tests being precisely matched to one another in form, and despite object recognition being correlated with the (highly predictive) spatial ability measures (Supplementary Table S3) whereas “pure” face recognition is not. The portion of GCSE variance predicted by this measure really is specific to face recognition.

To some extent, the genetic results are the inverse of those with spatial ability. The genetic correlations between face recognition and educational achievement (Table 2) reveal modest but significant genetic overlaps with every GCSE measure, but more than twice as strong for English as for Maths. Face recognition does not explain any significant portion of the heritability of Maths and Science, but does explain a small portion for English (Supplementary Tables S17-S19). This does not survive in the trivariate model accounting for verbal ability (Supplementary Table S24), but the subsequent analysis accounting for this by regression instead to improve power (Fig. 1) indicates that around half of its initial genetic relationship with English remains. The point estimate for the residual genetic relationship with spatial ability is slightly higher, but does not differ significantly from the contribution of face recognition. Genetically, then, once domain-general influences are accounted for, face recognition and spatial ability are approximately equal predictors of English. This mirrors the phenotypic results: with verbal ability controlled (Supplementary Table S6), face recognition and the spatial measures do not differ substantially in their relationships with this subject. The modest residual relationship with spatial ability may perhaps just represent inadequately controlled domain-general variance, but the same cannot be true for face recognition, as it shows no such relationship with the STEM subjects. Social intelligence – or even the small portion of it captured by face recognition – appears to predict non-STEM performance selectively.

The strength of this association should not be overstated: the phenotypic and genetic correlations are small. However, it is noteworthy that this genetic overlap between face recognition and English (Supplementary Table S19) is no weaker than its relationship with either object recognition or *g* (Shakeshaft and Plomin, 2015) – even in this “pure” form with object recognition regressed out. Indeed it is not significantly weaker (although the point estimate is lower) even with verbal ability controlled, too (Fig. 1).

Face recognition is only one specific social ability, and may capture only a fraction of the possible relationships with more diverse social/emotional intelligence measures. Nonetheless,

these results offer support to the suggestion that social skills play a significant role in educational outcomes, and also perhaps offer some clarity to the mixed findings reported previously: social skills may have significant relationships with *some* academic subjects, but not with others. Since face recognition is not a facet of emotional intelligence, these results also suggest that broader social competence (Teo *et al.*, 1996) is implicated.

Summary

This study has shown a dissociation between spatial and social test performance on achievement in different academic subjects. These results lend considerable weight to the predictive power of spatial ability for STEM success, and conversely suggest that non-STEM subjects may be selectively influenced by social skills. In both cases, the relationships are overwhelmingly genetic in origin.

Some study limitations should be noted. First, some of the analyses were underpowered, as not every participant completed every measure, so the available sample sizes varied considerably. Second, the predictor measures were administered around four years later than the GCSE outcome measures, making this a prediction after the fact (although the genetic influences on cognitive ability are highly stable (Deary *et al.*, 2012), so this is unlikely to have influenced the results). In any case, the present results were highly substantial and significant.

The “core subjects” composite, representing a proxy for overall educational achievement, unsurprisingly showed relationships more similar to the STEM subjects (which form two thirds of it) than to English. However, while overall achievement is an important measure, the present results suggest that it may be of limited utility in understanding the aetiology of academic outcomes. Different subjects may be highly intercorrelated (Supplementary Table S4), but they are not the same. In order to understand the factors driving and limiting educational success, it may be necessary to consider very different sets of predictors for different domains.

In the context of education, any discussion of genetics can provoke considerable resistance. Although this point has been made often before, it is therefore worth emphasising that genetic influence does not imply determinism: the results of behavioural genetic studies are

population statistics, describing the origins of variation in the current population under current conditions, and certainly do not imply that outcomes are not amenable to intervention. On the contrary, as studies such as this reveal the aetiology of achievement, it is hoped that hidden barriers may be revealed, and better interventions developed to raise them.

Methods

Measures

Participants provided their GCSE scores by post, and completed all other measures online via purpose-built websites. The educational achievement measures analysed were General Certificate of Secondary Education (GCSE) grades, administered at the end of compulsory education in the UK, usually at age 16. These may be taken in a wide variety of subjects, but English, Mathematics and Science (among others) are compulsory. Syllabi vary, and the measures used here for these three subjects represent the mean of grades for the specific examinations taken. Also analysed here were four cross-subject composites: “core subjects” (the mean of scores for these three subjects), a “Humanities” composite (the mean of all the most commonly taken art and humanities subjects, excluding English and other languages), an “overall” score (the mean of grades for every subject attempted), and a score representing the number of passes at grades A*-C (a metric commonly used for university admissions). These measures are described in more detail elsewhere (Shakeshaft *et al.*, 2013).

The “Bricks” spatial measures, described in detail previously (Shakeshaft *et al.*, 2016), are six short subtests of mental rotation and visualisation, using 2D and 3D stimuli. The mean of these six subtest scores is used here as a “narrow” measure of spatial ability, in the sense that it includes only those two putative subdomains of spatial ability (mental rotation and visualisation), out of the many discussed in the literature (Hegarty and Waller, 2005). The previously-published (Shakeshaft *et al.*, 2016) results found no evidence that even these two were phenotypically or genetically dissociable. The Bricks score may thus be regarded as a measure of spatial visualisation.

The “King's Challenge” (KC) battery, described previously (Rimfield, Shakeshaft *et al.*, under review), comprises ten diverse spatial tests, collectively representing ability across the entire spatial domain. The prior results found no evidence for phenotypic or genetic dissociations

among these tests, suggesting that they form a single factor; however, this battery was administered separately from the Bricks measures, and they have not previously been directly compared. The derived score used here is the first principal component of the ten subtest scores, thus representing a “broad” measure of overall spatial ability.

The face and object recognition measures, described elsewhere (Duchaine and Nakayama, 2006; Dennett *et al.*, 2012; Shakeshaft and Plomin, 2015), require participants to memorise stimuli (faces and cars, respectively) then recognise them in different orientations and conditions. Both tests are reliable (Wilmer *et al.*, 2010; Dennett *et al.*, 2012). They are exactly matched in form, so the objects test provides an ideal control for domain-general factors (memory, attention, etc.) reflected in the faces test (Shakeshaft and Plomin, 2015). A “pure” face recognition measure was thus created by regressing the latter on the former.

Verbal ability was assessed with the Mill Hill Vocabulary Scale (Raven *et al.*, 1998), a 33-item measure of vocabulary size. Although verbal ability represents only a portion of *g* (Plomin *et al.*, 2013), it was considered to be the only portion that could be used as a control for domain-general abilities without risking removing some of the spatial ability variance of interest. Raven's Progressive Matrices, for example, a common measure of non-verbal ability, contains a substantial spatial component (Schweizer *et al.*, 2007). For this reason, verbal ability was used as a conservative proxy for domain-general factors (including *g*), while acknowledging that it is probably an under-correction.

The data were cleaned and prepared prior to analysis. The specific procedures were described previously for each of the key variables, but in brief: participants suspected to have experienced technical errors or who had severe relevant disabilities were excluded, the data were regressed on age and sex (per standard practice for twin data; McGue and Bouchard, 1984), outliers beyond ±3 SD from the mean were removed, and the measures were rank-transformed to account for minor skew (Lehmann, 2006) and standardised. These procedures were applied individually to each Bricks and KC subtest, and to the face and object recognition measures, before creating the three composite predictors as above.

Twin analyses

Identical (monozygotic; MZ) twin pairs share all of their segregating genes, while fraternal

(dizygotic; DZ) twin pairs share on average only half, but both share their environments to approximately the same extent. The heritability of a trait (the degree to which the phenotypic variance is attributable to genetic variance) may thus be derived from the degree to which MZ exceed DZ intrapair correlations. The environmental portion of variance may be further decomposed into “shared” influences (effects promoting intrapair similarity) and “non-shared” influences unique to each individual. Multivariate genetic analyses similarly compare the cross-twin cross-trait correlations for MZ and DZ twins (between trait 1 for twin 1, and trait 2 for twin 2) to estimate the aetiological structure of multiple traits: the degree to which they are driven by the same genes or environmental influences as one another.

To derive the best estimates and confidence intervals, model-fitting procedures were used that account for missing data, and same- and opposite-sex DZ twins were both included to maximise power. Model-fitting was conducted using OpenMx (Boker *et al.*, 2011). Univariate ACE models (Rijsdijk and Sham, 2002) decomposed the variance in each measure into additive genetic (A), shared environmental (C), and non-shared environmental (E) portions, the latter term also including any error of measurement.

Multivariate ACE twin model-fitting calculates the genetic correlations (rA) between traits – these are independent of their heritability, and indicate the degree to which genetic influences are shared between them. Similar correlations are derived for shared environment (rC) and non-shared environment (rE), the latter indicating environmental effects unique to the individual, but shared between traits. A “correlated factors” solution estimates the common influences between the traits, allowing their phenotypic correlation to be attributed to specific components of covariance – these estimates can be expressed as proportions (as in Supplementary Tables S9 and S16), reflecting the correlation between the traits for that component, weighted by the univariate components (e.g., the proportion due to A is the genetic correlation weighted by the product of the square roots of the two univariate heritability estimates). A reorganised but algebraically equivalent presentation of the data, the Cholesky decomposition (Loehlin, 1996), focuses on the total influences on each trait in the model in sequence, estimating the ACE components shared with each previous variable, or independent from them; see Fig. 1 for an example. Precise estimates vary between models.

Submodels may be created by constraining certain paths to zero, and then testing for a significant drop in the goodness of fit. For the univariate models, for example, the ACE model was compared to a fully saturated model imposing no expectations on the data

(Supplementary Table S8). In the larger multivariate models, the same comparisons are provided for reference (e.g., see Supplementary Table S30), but are not very meaningful: the model assumes equality of intrapair covariances for each pair of traits (i.e., trait 1 for twin 1 with trait 2 for twin 2, and vice versa), and small, random differences accumulate exponentially. A more meaningful comparison (Supplementary Table S8) was that between the ACE model and the AE submodel (dropping C), revealing that none of the predictors have any significant shared environmental influences, whereas the GCSE measures do. The multivariate models used to derive the present results were designed accordingly: all common C paths were removed (i.e., constrained to zero) except those influencing only the GCSE variables. In almost every case, the resulting submodel showed no significant deterioration in fit from the full ACE model (although see the footnote to Supplementary Table S30 for discussion of the exceptions).

Acknowledgements

We thank the twins in the Twins Early Development Study (TEDS) for their ongoing participation in the study. TEDS is supported by a programme grant to RP from the UK Medical Research Council (MRC) [MR/M021475/1; previously G0901245 and G0500079], with additional support from the US National Institutes of Health [HD044454; HD059215; NIA046938]. NGS and KR are supported by MRC studentships. SS is supported by an MRC studentship and EU Framework Programme 7 [602768]. RP is supported by a Medical Research Council Research Professorship award [G19/2] and a European Research Council Advanced Investigator award [295366].

Author contributions

NGS, KR, MM, KLS, MR, SS and RP created the spatial measures and designed the study. NGS conducted the analyses. NGS and RP wrote the manuscript. All authors reviewed the manuscript.

Competing interests

The authors declare no competing financial interests.

Tables and Figure

Table 1. Correlations between predictors and GCSEs.

		GCSE mean	No. A*-C	"Core" subjects	Humanities	English	Science	Maths
Bricks	*r*	0.37 **	0.34 **	0.41 **	0.28 **	0.31 **	0.36 **	0.43 **
	N	*1326*	*1336*	*1327*	*1165*	*1328*	*1274*	*1317*
King's Challenge	*r*	0.43 **	0.34 **	0.47 **	0.30 **	0.30 **	0.43 **	0.51 **
	N	*808*	*809*	*805*	*716*	*805*	*786*	*802*
Face recognition	*r*	0.17 **	0.15 **	0.16 **	0.15 **	0.17 **	0.13 **	0.12 **
	N	*899*	*906*	*895*	*774*	*895*	*860*	*888*
Object recognition	*r*	*0.05*	0.11 **	0.09 **	*0.04*	*0.02*	0.07 *	0.13 **
	N	*889*	*896*	*885*	*765*	*885*	*850*	*878*
"Pure" face recognition	*r*	0.16 **	0.12 **	0.14 **	0.14 **	0.17 **	0.11 **	0.09 *
	N	*889*	*896*	*885*	*765*	*885*	*850*	*878*

Correlations (Pearson's *r*) between the predictors and GCSEs. The sample is fully independent, with one individual selected randomly from each twin pair. N = sample size, ** = p < 0.01, * = p < 0.05.

Table 2. Genetic correlations between predictors and GCSEs.

	Maths	Science	English	"Core" subjects
Bricks	0.67 (0.60 – 0.74)	0.64 (0.57 – 0.71)	0.45 (0.38 – 0.53)	0.65 (0.58 – 0.72)
King's Challenge	0.69 (0.62 – 0.75)	0.59 (0.51 – 0.66)	0.41 (0.33 – 0.49)	0.63 (0.56 – 0.70)
"Pure" face recognition	0.12 (0.03 – 0.22)	0.16 (0.06 – 0.26)	0.28 (0.18 – 0.37)	0.19 (0.10 – 0.29)

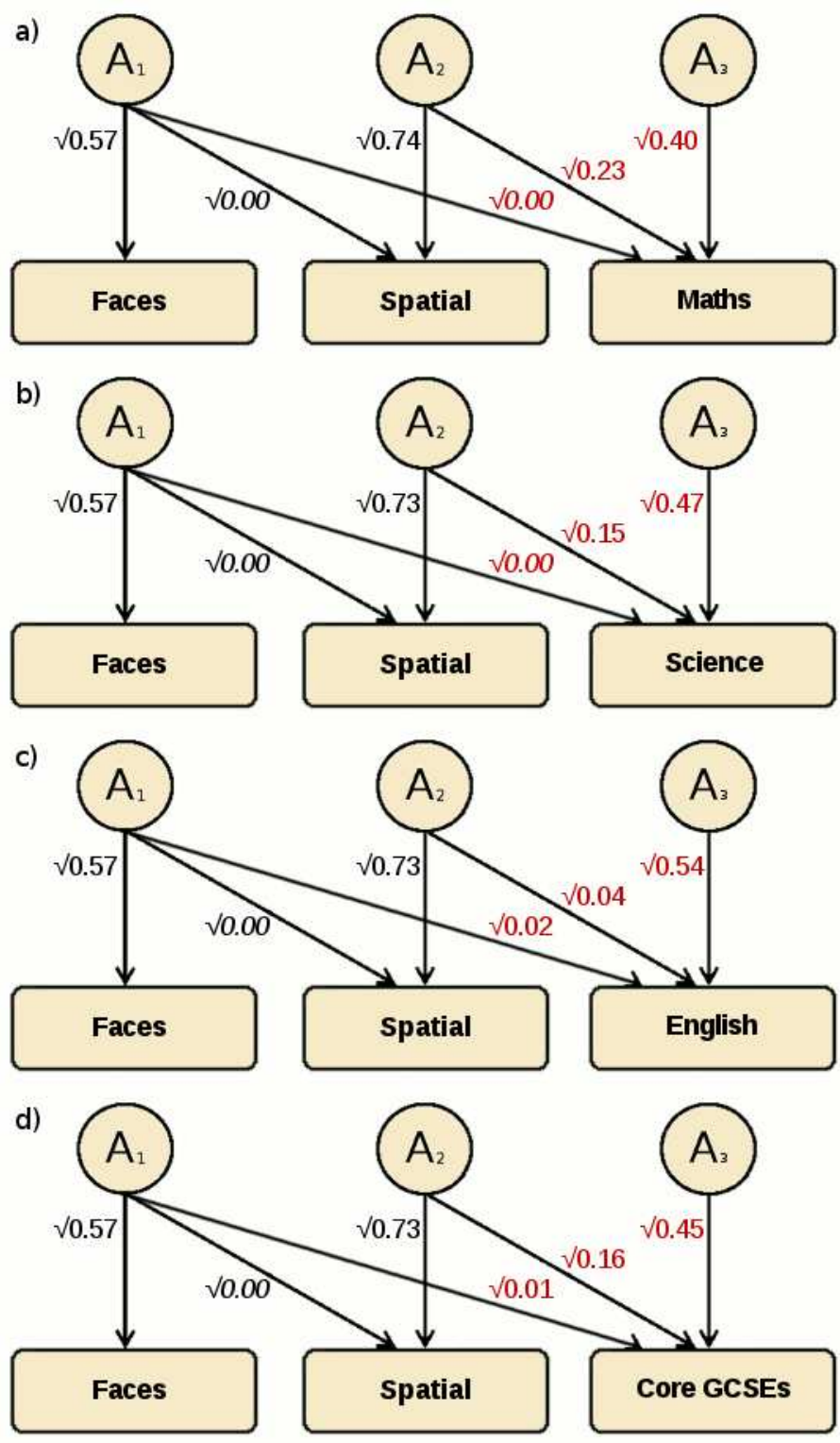
Genetic correlations (95% confidence intervals) between predictor variables and GCSE grades.

Table 3. Non-shared environmental correlations between predictors and GCSEs.

	Maths	Science	English	"Core" subjects
Bricks	0.15 (0.07 – 0.24)	0.11 (0.02 – 0.19)	0.13 (0.05 – 0.21)	0.17 (0.08 – 0.25)
King's Challenge	0.24 (0.12 – 0.35)	0.35 (0.22 – 0.46)	*0.12 (-0.01 – 0.24)*	0.31 (0.18 – 0.42)
"Pure" face recognition	*0.02 (-0.09 – 0.13)*	*0.06 (-0.06 – 0.17)*	*0.05 (-0.06 – 0.16)*	*0.06 (-0.06 – 0.17)*

Non-shared environmental correlations (95% confidence intervals) between predictor variables and GCSE grades. Italicised estimates are non-significant (their CIs include zero).

Figure 1. Trivariate Cholesky decomposition genetic path estimates.



Path estimates (standardised) for four trivariate ACE Cholesky decompositions, showing the structure of additive genetic influences. The variables are "pure" face regression and spatial ability (King's Challenge), both regressed phenotypically on verbal ability, then finally the GCSE measure: a) Maths; b) Science, c) English; d) the "core" GCSE subjects composite. See Table S32 for more details. The paths in red indicate the genetic influences on the GCSE measure i) common to all three variables; ii) shared only between spatial ability and the GCSE measure but not with face recognition; and iii) those unique to the GCSE measure. Italicised paths are non-significant.

References

Alloway TP, Alloway RG (2010). Investigating the predictive roles of working memory and IQ in academic attainment. *J. Exp. Child Psychol.* 106(1): 20–29.

Asbury K, Plomin R (2013). *G is for genes: The impact of genetics on education and achievement*. London: Wiley-Blackwell.

Baumeister RF, Heatherton TF, Tice DM (1994). *How and Why People Fail at Self-Regulation.* San Diego, CA: Academic Press.

Boker S, Neale M, Maes H, Wilde M, Spiegel M, Brick T, Spies J, Estabrook R, Kenny S, Bates T, Mehta P, Fox J (2011). OpenMx: An Open Source Extended Structural Equation Modeling Framework. *Psychometrika* 76: 306–317.

Brackett MA, Mayer JD, Warner RM (2004). Emotional intelligence and its expression in everyday behavior. *Pers. Individ. Dif.* 36: 1387–1402.

Briley DA, Domiteaux M, Tucker-Drob EM (2014). Achievement-Relevant Personality: Relations with the Big Five and Validation of an Efficient Instrument. *Learn. Individ. Differ.* 32: 26–39.

Carroll JB (1993). *Human cognitive abilities: A survey of factor-analytic studies.* Cambridge: Cambridge University Press.

Carter CS, Larussa MA, Bodner GM (1987). A study of two measures of spatial ability as predictors of success in different levels of general chemistry. *J. Res. Sci. Teach.* 24(7): 645–657.

Chen J (2014). Face recognition as a predictor of social cognitive ability: Effects of emotion and race on face processing. *Asian J. Soc. Psychol.* 17(1): 61–69.

Costa A, Faria L (2015). The impact of Emotional Intelligence on academic achievement: A longitudinal study in Portuguese secondary school. *Learn. Individ. Differ.* 37: 38–47.

De Ridder KAA, Pape K, Johnsen R, Holmen TL, Westin S, Bjørngaard JH (2013). Adolescent health and high school dropout: A prospective cohort study of 9000 Norwegian adolescents (the Young-HUNT). *PLoS ONE* 8(9): e74954.

Deary IJ, Strand S, Smith P, Fernandes C (2007). Intelligence and educational achievement. *Intelligence* 35(1): 13–21.

Deary IJ, Yang J, Davies G, Harris SE, Tenesa A, Liewald D, Luciano M, Lopez LM, Gow AJ, Corley J, Redmond P, Fox HC, Rowe SJ, Haggarty P, McNeill G, Goddard ME, Porteous DJ, Whalley LJ, Starr JM, Visscher PM. (2012). Genetic contributions to stability and change in intelligence from childhood to old age. *Nature* 482: 212–215.

Dennett HW, McKone E, Tavashmi R, Hall A, Pidcock M, Edwards M, Duchaine B (2012). The Cambridge Car Memory Test: A task matched in format to the Cambridge Face Memory Test, with norms, reliability, sex differences, dissociations from face memory, and expertise effects. *Behav. Res. Methods* 44(2): 587–605.

Downey LA, Mountstephen J, Lloyd J, Hansen K, Stough C (2008). Emotional intelligence and scholastic achievement in Australian adolescents. *Aust. J. Psychol.* 60(1): 10–17.

Duchaine B, Nakayama K (2006). The Cambridge Face Memory Test: Results for neurologically

intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia* 44(4): 576–585.

Fitousi D, Wenger MJ (2013). Variants in independence in the perception of facial identity and expression. *J. Exp. Psychol. Hum. Percept. Perform.* 39: 133–155.

Graziano PA, Reavis RD, Keane SP, Calkins SD (2007). The role of emotion regulation in children's early academic success. *J. Sch. Psychol.* 45(1): 3–19.

Halberstadt AG, Hall JA (1980). Who's getting the message? Children's nonverbal skill and their evaluation by teachers. *Dev. Psychol.* 16: 564–573.

Haworth CMA, Davis OSP, Plomin R (2012). Twins Early Development Study (TEDS): A Genetically Sensitive Investigation of Cognitive and Behavioral Development From Childhood to Young Adulthood. *Twin Res. Hum. Genet.* 16: 117–125.

Hegarty M, Waller DA (2005). Individual Differences in Spatial Abilities. *The Cambridge Handbook of Visuospatial Thinking* (eds. Shah P, Miyake A), pp 121–169. Cambridge: Cambridge University Press.

Hubbard EM, Piazza M, Pinel P, Dehaene S (2005). Interactions between number and space in parietal cortex. *Nat. Rev. Neurosci.* 6(6): 435–448.

Ishai A (2008). Let's face it: It's a cortical network. *Neuroimage* 40(2): 415–419.

Kashani FL, Azimi AL, Vaziri S (2012). Relationship between Emotional Intelligence and Educational Achievement. *ICEEPSY* 69: 1270–1275.

Kaufman SB, Reynolds MR, Liu X, Kaufman AS, McGrew KS (2012). Are cognitive *g* and academic achievement *g* one and the same *g*? An exploration on the Woodcock–Johnson and Kaufman tests. *Intelligence* 40(2): 123–138.

Krapohl E, Rimfeld K, Shakeshaft NG, Trzaskowski M, McMillan A, Pingault JB, Asbury K, Harlaar N, Kovas Y, Dale PS, Plomin R (2014). The high heritability of educational achievement reflects many genetically influenced traits, not just intelligence. *PNAS* 111(42): 15273–15278.

Lehmann EL (2006). *Nonparametrics: Statistical Methods Based on Ranks*. New York: Springer.

Loehlin JC (1996). The Cholesky approach: A cautionary note. *Behav. Genet.* 26: 65–69.

Lohman DF (1994). Spatial ability. *Encyclopedia of intelligence* (ed. Sternberg, RJ), pp 1000–1007. Michigan: Macmillan.

MacCann C, Fogarty GJ, Zeidner M, Roberts RD (2011). Coping mediates the relationship between emotional intelligence (EI) and academic achievement. *Contemp. Educ. Psychol.* 36(1): 60– 70.

Mackintosh N (2011). *IQ and Human Intelligence*. Oxford: Oxford University Press.

Mayer JD, Salovey P, Caruso DR (2008b). Emotional intelligence: new ability or eclectic traits? *Am. Psychol.* 63(6): 503-517.

McGue M, Bouchard TJ (1984). Adjustment of twin data for the effects of age and sex. *Behav. Genet.* 14: 325–343.

Newsome S, Day AL, Catano VM (2000). Assessing the predictive validity of emotional intelligence. *Pers. Individ. Dif.* 29(6): 1005–1016.

Nicoletti C, Rabe B (2013). Inequality in pupils' test scores: How much do family, sibling type and neighbourhood matter? *Economica* 80: 197–218.

Nowicki S, Duke MP (1992). The association of children's nonverbal decoding abilities with their popularity, locus of control, and academic achievement. *J. Genet. Psychol.* 153(4): 385–393.

Nowicki S, Duke MP (1994). Individual differences in the nonverbal communication of affect: The diagnostic analysis of nonverbal accuracy scale. *J. Nonverbal Behav.* 19: 9–35.

Parker JD, Creque RE, Barnhart DL, Harris JI, Majeski SA, Wood LM, Bond BJ, Hogan MJ (2004). Academic achievement in high school: does emotional intelligence matter? *Pers. Individ. Dif.* 37(7): 1321–1330.

Petrides KV (2011). Social intelligence. *Encyclopedia of Adolescence* (eds. Brown BB, Prinstein MJ), pp 342–352. San Diego, CA: Academic Press.

Petrides KV, Frederickson N, Furnham A (2004). The role of trait emotional intelligence in academic performance and deviant behavior at school. *Pers. Individ. Dif.* 36(2): 277–293.

Pingault JB, Tremblay RE, Vitaro F, Carbonneau R, Genolini C, Falissard B, Côté SM (2011). Childhood trajectories of inattention and hyperactivity and prediction of educational attainment in early adulthood: A 16-year longitudinal population-based study. *Am. J. Psychiatry* 168(11): 1164–1170.

Plomin R, DeFries JC, Knopik VS, Neiderhiser JM (2013). *Behavioral genetics*. New York: Worth Publishers.

Raven J, Raven JC, Court JH (1998). Section 5: The Mill Hill Vocabulary Scale. *Manual for Raven's Progressive Matrices and Vocabulary Scales*. San Antonio, TX: Harcourt Assessment.

Rijsdijk FV, Sham PC (2002). Analytic approaches to twin data using structural equation models. *Brief. Bioinform.* 3: 119–133.

Rimfeld K, Kovas Y, Dale PS, Plomin R (2015). Pleiotropy across academic subjects at the end of compulsory education. *Sci. Rep.* 5: 11713.

Rimfeld K, Shakeshaft NG, Malanchini M, Rodic M, Selzam S, Schofield KL, Dale PS, Kovas Y, Plomin R (under review). Spatial ability or spatial abilities? Investigating the phenotypic and genetic structure of spatial ability. *PNAS*. See Chapter 5.

Rohde TE, Thompson LA (2007). Predicting academic achievement with cognitive ability. *Intelligence* 35(1): 83–92.

Schweizer K, Goldhammer F, Rauch W, Moosbrugger H (2007). On the validity of Raven's matrices test: Does spatial ability contribute to performance? *Pers. Individ. Dif.* 43: 1998–2010.

Shakeshaft NG, Plomin R (2015). Genetic specificity of face recognition. *PNAS* 112(41): 12887–12892. See Chapter 6.

Shakeshaft NG, Rimfeld K, Schofield KL, Selzam S, Malanchini M, Rodic M, Kovas Y, Plomin R (2016). Rotation is visualisation, 3D is 2D: using a novel measure to investigate the genetics of spatial ability. *Sci. Rep.* 6: 30545. See Chapter 4.

Shakeshaft NG, Trzaskowski M, McMillan A, Rimfeld K, Krapohl E, Haworth CMA, Dale PS, 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): e80341. See Chapter 2.

Shea DL, Lubinski D, Benbow CP (2001). Importance of assessing spatial ability in intellectually talented young adolescents: A 20-year longitudinal study. *J. Educ. Psychol.* 93(3): 604–614.

Stankov L (2013). Noncognitive predictors of intelligence and academic achievement: An important role of confidence. *Pers. Individ. Dif.* 55(7): 727–732.

Teo A, Carlson E, Mathieu PJ, Egeland B, Sroufe LA (1996). A prospective longitudinal study of psychosocial predictors of achievement. *J. Sch. Psychol.* 34(3): 285–306.

Tosto MG, Hanscombe KB, Haworth CMA, Davis OSP, Petrill SA, Dale PS, Malykh S, Plomin R, Kovas Y (2014). Why do spatial abilities predict mathematical performance? *Dev. Sci.* 17(3): 462–470.

Van Garderen D, Montague M (2003). Visual-spatial representation, mathematical problem solving, and students of varying abilities. *Learn. Disabil. Res. Pract.* 18(4): 246–254.

Wai J, Lubinski D, Benbow CP (2009). Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *J. Educ. Psychol.* 101(4): 817–835.

Wilmer JB, Germine L, Chabris CF, Chatterjee G, Williams M, Loken E, Nakayama K, Duchaine B (2010). Human face recognition ability is specific and highly heritable. *Proc. Natl. Acad. Sci. USA* 107(11): 5238–5241.

Zuffianò A, Alessandri G, Gerbino M, Kanacri BPL, Di Giunta L, Milioni M, Caprara GV (2013). Academic achievement: The unique contribution of self-efficacy beliefs in self-regulated learning beyond intelligence, personality traits, and self-esteem. *Learn. Individ. Differ.* 23: 158–162.

Chapter 8:- Discussion

The work presented in this thesis investigated the relationships among spatial ability, face recognition and general cognitive ability (*g*), and used measures of these abilities to predict some of the variability in educational achievement. Implications specific to each study are discussed within the individual chapters concerned and are not repeated here. This chapter briefly summarises the key findings, notes the limitations of the studies conducted, and then draws them together to consider broader implications and potential future directions.

Summary of results

Previous behavioural genetic research has found educational achievement to be highly heritable throughout early and middle childhood. In Chapter 2, this was shown to extend to the end of compulsory education in the UK. Individual differences in performance on the General Certificate of Secondary Education (GCSE) examinations, which heavily influence access to higher education and to job opportunities, were shown to be substantially heritable: on average across all of the academic subjects and composites analysed, 53% of the variance was attributable to genetic differences. This was almost double the portion of variance accounted for by shared environmental factors (30%), representing all of the environmental influences shared within families, schools and wider neighbourhoods (although even this degree of shared environmental influence is unusually high for behavioural traits, as discussed below). Familial resemblance in educational outcomes is therefore around two thirds genetic in origin. Non-shared environmental influences unique to each individual (including any error of measurement in the GCSE grades and composites) explained only 17% of the variance overall.

For analyses such as those conducted throughout these studies, the genetic and environmental influences on each variable are assumed to operate in a linear fashion throughout the distribution, with no discontinuities. In order to test the validity of this assumption for *g*, Chapter 3 subjected twin and sibling data to a series of analyses designed to determine whether the aetiology of the upper (highly intelligent) extreme of the distribution of *g* was different from that of the rest of the distribution. No evidence was found for any such discontinuities, either genetically or environmentally. This supports the use of *g* for linear

analyses across whole samples, although some important caveats are discussed as limitations below.

The present work focused primarily on the role of two specific cognitive abilities in predicting educational outcomes. The first of these, spatial ability, has previously been shown to be associated with achievement in science, technology, engineering and mathematics (STEM) subjects, but the usefulness of spatial ability for this purpose is greatly reduced by a substantial lack of clarity as to its structure, and therefore its proper measurement. Chapters 4 and 5 present the results of two large twin studies, in which two novel batteries of spatial tests (one "narrow", including only two of the putative spatial subdomains, and the other "broad") were administered in an effort to clarify the phenotypic and genetic structure of this ability. In both cases, spatial ability was found to be highly heritable, and significant and substantial dissociations were observed between this ability and *g*, confirming the former as a distinct cognitive domain. However, no dissociations were supported *within* spatial ability, either phenotypically or genetically, between any of its suggested subdomains. This indicates that a single, unifactorical measure of spatial ability is likely to capture it best, and that the dissociations often reported previously in the literature may (in large part, if not entirely) reflect unreliability in the measures rather than a meaningfully multifaceted structure.

For face recognition, the other specific ability of interest, Chapter 6 similarly examined its genetic relationship to other abilities. Face recognition has long been considered "special" in a variety of ways, including its phenotypic dissociation from general object recognition and *g*, but the genetic relationships had not been examined previously. Consistent with previous findings, face recognition was found to be highly heritable and phenotypically only very modestly correlated with general object recognition and *g*. Genetically, the dissociation was shown to be even more striking: the substantial heritability of the face recognition measure was almost entirely unique to it, not shared either with *g* or with the object recognition measure, despite the latter being exactly matched to the face recognition test in its administration and cognitive demands.

Finally, Chapter 7 combined the measures from the preceding chapters, using spatial ability and face recognition to predict the variance in GCSE results in Mathematics, Science and English. Since non-verbal cognitive ability has been shown to be substantially spatial in nature (Schweizer *et al.*, 2007), verbal ability was used as a conservative proxy for *g* when controlling for domain-general abilities, in order to retain all of the variance of interest from

the spatial measures. For spatial ability, the results confirmed previous findings of a substantial association with the STEM subjects. The “narrow” and “broad” measures were virtually identical in their genetic associations – as predicted, in light of the findings suggesting that spatial ability is unifactorial. The weaker association with English disappeared once verbal ability (the proxy for domain-general factors) was accounted for, strongly suggesting that spatial ability plays a specific role in STEM subjects; this was found to be largely due to common genetic influences. Face recognition, conversely, had no relationship whatsoever with the STEM subjects once domain-general factors were accounted for.

Face recognition was assessed in Chapter 7 in a “pure” form, regressed phenotypically on an exactly-matched general object recognition task in order to remove the influence of all domain-general factors shared between these measures, such as memory and attention. Despite all such domain-general factors being controlled, face recognition remained significantly associated with English, both phenotypically and genetically, even once verbal ability was accounted for, too. Face recognition (and perhaps social ability more generally, by implication) appears to predict non-STEM subjects selectively, albeit only weakly.

Limitations

The general limitations of the twin method apply to every study throughout this work; these are primarily the issue of representativeness of twin samples, and the “equal environments assumption” (see Plomin *et al.*, 2013 and the discussion in Chapter 1). More specific limitations applicable to each analysis are discussed in the chapters concerned.

Two further issues should be noted regarding the measures used. The first concerns the timing of administration: the measures were collected several years apart from one another. The GCSE examinations were taken when the participants were 16 years old, and the measures forming the *g* composite were administered at the same age. However, the face and object recognition measures were administered at age 19 on average, and the two spatial batteries at 20. As observed in Chapter 7, this makes the prediction of educational achievement retrospective. *g* is very stable genetically across time (Deary et al, 2012), so the difference in age between the collection of *g* data and the specific cognitive ability measures is unlikely to have had a substantial effect on the results of the latter. However, age-to-age genetic stability has not been assessed for face recognition or spatial ability, so the effect of the time of

administration on the strength of their associations with the GCSE measures is unknown. In principle, if the genetic stability of these traits were substantially lower than that of *g* (if different genetic influences on these abilities come online at different ages, for example), then it is possible that the associations reported between these abilities and educational achievement (Chapter 7) could be underestimates, in comparison to the results that might be obtained with contemporaneous measures.

Another potential limitation is the issue of discontinuity, outlined above and in Chapter 1: analyses such as those presented here assume a continuous distribution of aetiological influences on the measures. As the results and discussion in Chapter 3 indicate, this assumption certainly appears to be valid with respect to *g* – no evidence was found for discontinuity between the top of the distribution and the rest. One caveat, noted in that chapter, is that a more stringent cut-off (if this had been possible with the data available) could perhaps have produced different results. While this possibility is interesting in its own right, however, it is not really relevant to the present issue: even if there *were* a discontinuity affecting only a very small minority of individuals at the upper extreme of the distribution (just as there is at the lower extreme; Reichenberg *et al.*, 2016), the numbers affected would be negligible in a general population sample. The analyses in Chapter 3 are presented as a proxy for considering this issue for cognitive abilities in general, but in principle it is possible that specific abilities could show different aetiological patterns. There is no evidence in the literature for such discontinuities in spatial ability or face recognition (although none specifically against them, either), but the existence of specific impairments for face recognition (see Chapter 6) does perhaps raise the prospect of a genetic discontinuity for those most profoundly affected. In general, testing *all* variables for potential aetiological discontinuities, using analyses such as those used for *g* in Chapter 3, might be useful as a standard step in future research.

Even if no substantial aetiological discontinuities do exist within any of the measures themselves, however, it does not follow that their associations with each other must be linear. Such non-linear relationships have in fact been reported in the relationship between certain social abilities and educational outcomes – for example, emotional intelligence has been found to be predictive of academic achievement in participants with low or average *g*, but not in those with higher *g* (Petrides *et al.*, 2004; Agnoli *et al.*, 2012). It is unknown whether similar patterns could be found with face recognition (or indeed with non-social domains such as spatial ability), so assessing the possible role of *g* as a *moderating* as well as mediating

variable would be a useful future extension to the present work.

Implications and future directions

Implications specific to each study and their individual foci are discussed in detail within their own chapters. This section attempts to draw out more overarching issues. It considers the aetiology of educational achievement, the importance of reliability, and the potential of internet-based research. Proposals are made for future work. The chapter concludes with a discussion of the nature of "genetic influence" and the value of understanding it.

The aetiology of education

At the end of compulsory education, the variation in academic grades is substantially attributable to genetic differences (Chapter 2). This substantial heritability, representing over half of the total variance and around two thirds of familial resemblance, represents a wide array of cognitive and non-cognitive traits (Krapohl *et al.*, 2014). This diversity of genetic factors may explain why the heritability is so high: the GCSE scores are effectively composites, reflecting the cumulative influences of a large number of relevant traits, many of which are themselves heritable. Domain-general cognitive ability accounts for a large proportion of this genetic component of variance, and spatial ability independently explains another substantial portion for STEM subjects, but not (or at most very little) for English (Chapter 7). Face recognition, conversely, accounts for a significant (although very small) portion of the genetic and phenotypic variance in English, independently from domain-general factors, but not in STEM subjects. This suggests that the low-level perceptual processes in common between the spatial and face recognition measures cannot account for the specific association found between spatial ability and STEM fields – the more complex manipulations intrinsic to spatial tasks seem to be crucially implicated in this relationship.

Face recognition is almost entirely distinct from other abilities, both phenotypically and genetically (Chapter 6), and its genetic association with English seems to be reflected entirely in the portion of the heritability of achievement that is *not* related to *g*. To the extent that face recognition indexes social abilities more generally, it appears that social skills may differentially influence non-STEM subjects such as English. The genetic influences on

educational achievement appear to be diverse and complex.

This may be contrasted with the shared environmental portion of variance, accounting for around one third of familial resemblance. As noted in Chapter 2, even this modest shared environmental influence is unusually high for behavioural traits (Plomin, 2011). Consistent with this, all of the cognitive abilities used in the present work (i.e., spatial ability in Chapters 4 and 5, face recognition in Chapter 6, and *g* throughout) show no significant shared environmental component at all: familial resemblance in these abilities is almost entirely, if not completely, genetic. This implies that the shared environmental portion of variance in educational achievement – representing all of the familial resemblance attributable to family environments, schools, neighbourhoods and shared experiences – probably represents (mostly, if not exclusively) non-cognitive traits. To the modest extent that shared environmental factors promote similarity in achievement, they do so for reasons unrelated to ability.

As demonstrated by the dissociation between the educational correlates of spatial ability and face recognition, different predictors are relevant for different academic subjects, genetically as well as phenotypically. Even though the genetic correlations between all subjects are very high (Rimfeld *et al.*, 2015), therefore, this substantial pleiotropy should not be mistaken for a complete lack of heterogeneity between academic domains. Measures of "overall" educational achievement (such as that presented for reference in Chapter 7) may mask considerable aetiological differences between subjects.

In considering such potential dissociations, it is interesting to note the significant differences found between the (broadly defined) "sciences" and "humanities" composites presented in Chapter 2. The humanities composite was found to be significantly less heritable than science – arguably a counterintuitive result, as the sciences are often thought of as "taught", in comparison to "gifts" in the humanities and arts. It is possible that humanities subjects are simply less reliably measured than the sciences at GCSE level (the difference in heritability was offset by non-shared environment, which includes the error term). However, this would be difficult to reconcile with the observation that, in the phenotypic analyses presented in Chapter 7, English and Humanities showed very similar relationships to the various predictor variables – this makes the apparent difference in heritability between "humanities" and "sciences" difficult to account for, because English is *not* less heritable than Science (Chapter 2). A more detailed investigation is warranted: if this difference can be explained, it would provide a useful guide to ongoing research into the aetiological disassociations between

academic subjects, and could even suggest which variables of interest are plausible predictors for each domain.

Although not a substantial focus of this research, differences between males and females are another potential source of heterogeneity. No qualitative sex differences (different genes influencing males and females) were found for the GCSE measures analysed in Chapter 2, but some modest quantitative differences were observed (i.e., the same influences explaining different proportions of variance): heritability was around 10% greater for males, approximately counterbalanced by greater shared environmental influence for females. As noted in Chapter 2, this should be interpreted with caution until replicated, as the differences were small, with overlapping confidence intervals, and the same pattern was not observed at earlier ages (Kovas et al, 2007). However, it will be important to consider this issue further in future work, to ascertain (for example) whether the predictor variables show differential associations with outcomes between males and females.

Reliability and behavioural genetics

A crucial lesson to draw from the present work, particularly with regard to the investigations of spatial ability in Chapters 4 and 5, is the importance of reliable measures. It is in this context that one of the greatest strengths of behavioural genetic methods emerges.

It is almost invariably found that the genetic associations between traits closely mirror the phenotypic pattern, and it has been argued that genetic results therefore tend to provide little by way of additional information about these relationships (Turkheimer, 2016). The results in Chapters 4 and 5 provide a useful counterexample, illustrating one of the ways in which genetic analyses can help to clarify a murky phenotypic structure. While it is certainly true that the phenotypic and genetic results in these chapters are consistent with each other – only a single factor emerges from the phenotypic principal components analyses, for example, echoing the genetic findings – the picture is very much clearer in the genetic results, and this clarity matters: with the “Bricks” analyses (Chapter 4), for example, the spatial composites are only moderately intercorrelated phenotypically ($r \sim 0.50$), but correlate genetically at unity.

Chapter 4 notes that the likely explanation for the difference is reliability, but the significance

of this deserves emphasis. The “raw” data for twin analyses, in effect, are not the actual phenotypic scores themselves, but the intrapair correlations between twins (see Plomin *et al.*, 2013, and Chapter 1). This means not only that the reliability of a measure is the ceiling of its heritability estimate, but also, conversely, that the heritable variance is reliable. The importance of this reliability is demonstrated clearly by the inconsistent findings and dissociations that have marked the literature on spatial ability to date. If the conclusions in Chapters 4 and 5 are correct (and without independent replication, this is of course a substantial “if”), then many – perhaps even all – of the putative subdomains of spatial ability, proliferating over several decades of phenotypic research, have been phantoms created by unreliable measures and the illusory dissociations between them. In contrast, by focusing only on the *reliable* variance to explore the aetiological architecture, behavioural genetic methods strip out the chance dissociations, and the phantoms fade away.

Online testing

Most of the data presented in these studies were collected remotely, and none in traditional lab settings. Genetic analyses require large samples for adequate statistical power, and with the exception of data made available from routine testing – such as standardised school examinations (as in Chapters 2 and 7) or tests conducted during national military service (Chapter 3) – administering measures remotely is often unavoidable: there is simply no other practical way to collect the data at scale. This introduces difficulties especially for cognitive testing, which often requires rigorous enforcement of time limits or other test rules. Some success has been shown with testing via telephone (see Haworth *et al.*, 2012, for example), but a more versatile modern method is to administer measures using the internet, on participants' own computers or mobile devices. This was how most of the data in this work were collected, as described in the relevant chapters, and the implications of this increasingly common method deserve attention.

Online testing is a practical necessity for studies such as those described here, but the potential benefits extend much further. The limitations of much of the research in the psychological sciences are well documented, with the available samples often being highly unrepresentative of the general population demographically (Henrich *et al.*, 2010), and also substantially statistically underpowered. The latter especially is thought to be a major contributor to the widespread failure to replicate published results in the psychological

sciences (Open Science Collaboration, 2015), as evidenced by the observation that behavioural genetic findings – for which larger samples have long been the norm – have tended to buck the trend (Plomin *et al.*, 2016). Recruiting participants and administering measures online, where much larger and more representative samples are (in principle) readily available seems an attractive prospect, therefore, but there are substantial obstacles.

The first concern is whether data obtained online are as useful as they seem. In comparison with traditional lab-based research, it is commonly assumed that conducting studies via the internet – administering measures online and remotely – must inevitably involve a trade-off between increased sample size and decreased data quality (Germine *et al.*, 2012). Evidence is mounting against this concern, however: studies have been reported directly comparing online and lab samples, finding no significant or systematic differences in results, either for questionnaires (Casler *et al.*, 2013) or for cognitive measures (Germine *et al.*, 2012). There is even some evidence that participants pay greater attention to instructions when unsupervised (Ramsey *et al.*, 2016). One potential caveat concerns the manner of recruitment, with paid, "professional" test-takers sometimes being found to be less diligent than more traditional opportunity samples (Smith *et al.*, 2016), although even this is not consistent (e.g., Casler *et al.*, 2013) – and in any case this is not a consideration for the present research, since the web-administered measures were used only with a specific, longitudinal cohort sample. While online data collection is still relatively in its infancy, the indications so far suggest its validity and quality to be comparable to traditional lab-based research.

A more substantial obstacle simply concerns the practicalities involved: implementing measures in a format suitable for use online can be extremely difficult. Many commercial services exist for administering surveys or other questionnaires on the web, such as Mechanical Turk (https://www.mturk.com) or Qualtrics (https://www.qualtrics.com), but they are very limited in their capabilities, particularly for cognitive testing. For many types of measures, developing bespoke websites or other software remains the only option for conducting testing online. In the present research, the tests were integrated into purpose-built websites, hosted on in-house servers (see Chapter 1, and details for each measure in the chapters concerned). The development and maintenance of the necessary software and hardware represented a considerable investment of both time and budget – and at present, the expensive infrastructure and technical expertise required for such projects undoubtedly makes this impractical for many studies.

Despite the difficulties inherent in these relatively novel methods, their considerable potential seems clear. As technologies advance and the internet becomes ever more pervasive, it seems likely that these practical difficulties will ultimately be overcome, bringing the benefits of remote testing within reach of an ever-expanding range of research.

Future work

Several important directions for ongoing work have already been suggested by the considerations of heterogeneity and discontinuity above. All measures should be tested (as in Chapter 3) for possible aetiological discontinuities; *g* should be considered as a potential moderating (as well as mediating) variable affecting associations with educational achievement; the apparent aetiological differences between the sciences and humanities should be explored in greater depth; and the potential for substantial sex differences should be explored (even if only to rule them out).

A major question arising from the exploration of spatial ability in Chapters 4 and 5 is whether the batteries really do cover the full spatial domain. The work presented indicates strongly that spatial ability is unifactorial, but this is very much in opposition to the previous literature, and opponents could contend that the breadth of the spatial domain was not adequately captured. While a substantial effort was made to do so (particularly in the development of the "broad" measure presented in Chapter 5), ongoing work should consider potential omissions. One possibility in this direction is navigational ability, which has been suggested to represent a diverse array of skills in its own right (Wolbers and Hegarty, 2010). A battery of navigational tests has recently been administered to the same participants who completed the spatial measures presented here, so the relationships involved can now be tested.

In evaluating the appropriate measurement of spatial skills, it is important to consider the nature of the relationships with academic achievement. As discussed in Chapter 7, spatial ability predicts STEM subjects independently from *g*, while face recognition does not – this suggests that the more complex manipulations required in tests of spatial ability are intrinsic to their unique relationship with STEM outcomes. A potentially fruitful line of future research might therefore be to devise measures designed to alter progressively the nature and complexity of the manipulations required to solve each item. This would permit us to ask: what exactly constitutes a "spatial" test, as opposed to a visual, non-verbal test (such as face

recognition) which is *not* spatial? How complex do the manipulations involved need to be before the measure begins to predict STEM achievement, phenotypically and genetically, independently from *g*?

Given the association between face recognition and English, also independently from *g* (Chapter 7), it is reasonable to hypothesise that this ability has indeed indexed social skills, as intended; however, the relationship is very weak. Face recognition is only one, highly specific social skill among many (Petrides, 2011; Fitousi and Wenger, 2013), so the modest association identified here should prompt an expansion, considering potential aetiological relationships between educational achievement and social abilities more broadly: what other social or emotional measures might be predictive of outcomes, and will they show the same dissociation observed for face recognition between STEM and non-STEM subjects? Given the observation that the shared environmental influences on educational achievement appear to be (at least largely) non-cognitive in nature, might social abilities (albeit not face recognition, which was shown in Chapter 6 to have no shared environmental influences) form part of this component, too, rather than just the heritable variance?

Ultimately, finding specific genetic associations with educational outcomes will help to clarify the mechanisms involved. In the shorter term, polygenic scores derived from genome-wide association data (Wray *et al.*, 2014) have recently shown promise to explain a meaningful portion of the variance in overall educational achievement (Selzam *et al.*, 2016). Extending these analyses to consider individual academic subjects, and to test for differential associations with spatial and social predictors, would be a useful next step, which could be readily accomplished with available data. In light of the present results, we might hypothesise (for example) that a substantial relationship between spatial ability and educational achievement would be observed independently from *g*, showing a marked dissociation between STEM and non-STEM subjects, but that the association between face recognition and English may be too weak to be detectable using the polygenic scores currently available.

The meaning of genetic influence

As Turkheimer (2016) notes, “genetic influence” itself can be difficult to define: in itself, it does not automatically or intrinsically indicate anything more than an observed correlation between genetic and phenotypic similarity, for which a causal explanation is plausible but

unproven (i.e., that the proteins transcribed or regulated by the relevant genes partially drive differences in the phenotype, directly or indirectly). The significance of this theoretical argument is debatable given the implausibility of reverse causation – since no ordinary environmental events can change an individual's DNA – but the broader point is certainly valid: quantitative genetic methods such as twin studies may estimate the portion of variance and covariance attributable to "genetic influence", but cannot determine the actual mechanisms this influence represents. Despite recent advances, replicable associations between specific genetic variants and behavioural traits have proven extremely difficult to find; in large part, this is because no single variant in isolation ever accounts for a substantial portion of the variance in a trait, so the signal-to-noise ratio is very low (see Plomin *et al.*, 2013).

Even without yet understanding the biological mechanisms concerned, though, findings such as those presented here are useful. They allow better prediction of outcomes of interest and clearer interpretations of the relationships in question; they allow the associations between measures to be examined without potential confounding by genetic relatedness (Johnson *et al.*, 2009); they clarify the relationships between traits (as with spatial ability, above); and they aid the search for specific genetic mechanisms by identifying the expected associations to look for. As and when specific genes are discovered that are suggested to be related to spatial ability, for example, we can already predict their expected pattern of associations with different academic subjects, long before we know what they are or what they do.

In the eyes of the general public, at least, the most striking finding of this research may be the simplest: in examination grades at the end of compulsory education, the majority of the variation in achievement is genetic in origin. While no specific implications follow for educational policy, some possible interpretations are discussed in Chapters 2 and 7. Most notably: research such as this, identifying the genetic underpinnings of individual differences in educational outcomes, perhaps suggests that a personalised curriculum (rather than "one-size-fits-all") may be likely to achieve better results. As individual genetic testing for specific strengths and weaknesses ultimately becomes practical, this has the potential to inform the development and administration of such curricula – for example, by allowing potential difficulties to be detected before they manifest phenotypically, so that appropriate compensatory interventions can be administered early to prevent problems before they develop.

As noted in Chapters 2 and 7, however, there is often considerable resistance to any suggestion of genetic testing for educational purposes (see Asbury and Plomin, 2013, for a discussion) – and indeed sometimes even to research such as the present work being conducted at all. This resistance is often felt when any research is publicised finding genetic associations with abilities and outcomes, and typically reflects a fear of such findings being abused, or of genetic inequality being used to rationalise or excuse social inequality. While such concerns are understandable, they rely on a common and fundamental misunderstanding of behavioural genetics: that heritability implies determinism, such that interventions or social changes cannot be effective. Height provides a good illustration as to why this if not the case: human height is around 80-90% heritable, yet mean heights differ greatly between generations, and even between closely-related contemporaneous populations such as those of North and South Korea (Johnson *et al.*, 2009); these differences are generally attributed to environmental factors such as diet (e.g., Shams and Williams, 1997). As this example demonstrates, heritability has no bearing on the question of whether environmental interventions can be effective, or on whether the mean differences between groups are genetically driven. Heritability estimates describe the present population under present conditions, and nothing more – in fact, as argued in Chapter 2, it is even possible (perhaps counterintuitively) to interpret the heritability of educational outcomes as an index of equality of opportunity: if environmental sources of variation were reduced, genetic differences would be all that remained.

In closing, I would argue that resistance to genetic research, whether for educational outcomes or any other traits, is also misplaced for another, more fundamental reason: information is a tool. The risks of harm are undoubtedly real, if only by virtue of the results being misunderstood or misrepresented as above. Any tool can be used as a weapon, and vigilance against foreseeable risks is sensible – but the potential benefits are equally real. The appropriate response to risk is rarely to abandon our tools.

References

Agnoli S, Mancini G, Pozzoli T, Baldaro B, Russo PM, Surcinelli P (2012). The interaction between emotional intelligence and cognitive ability in predicting scholastic performance in school-aged children. *Pers. Individ. Dif.* 53(5): 660–665.

Asbury K, Plomin R (2013). *G is for genes: The impact of genetics on education and achievement*. London: Wiley-Blackwell.

Casler K, Bickel L, Hackett E (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Comput. Human Behav.* 29: 2156–2160.

Deary IJ, Yang J, Davies G, Harris SE, Tenesa A, Liewald D, Luciano M, Lopez LM, Gow AJ, Corley J, Redmond P, Fox HC, Rowe SJ, Haggarty P, McNeill G, Goddard ME, Porteous DJ, Whalley LJ, Starr JM, Visscher PM (2012). Genetic contributions to stability and change in intelligence from childhood to old age. *Nature* 482: 212–215.

Fitousi D, Wenger MJ (2013). Variants in independence in the perception of facial identity and expression. *J. Exp. Psychol. Hum. Percept. Perform.* 39: 133–155.

Germine L, Nakayama K, Duchaine BC, Chabris CF, Chatterjee G, Wilmer JB (2012). Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychon. Bull. Rev.* 19(5): 847–857.

Haworth CMA, Davis OSP, Plomin R (2012). Twins Early Development Study (TEDS): A Genetically Sensitive Investigation of Cognitive and Behavioral Development From Childhood to Young Adulthood. *Twin Res. Hum. Genet.* 16: 117–125.

Henrich J, Heine SJ, Norenzayan A (2010). The weirdest people in the world? *Behav. Brain. Sci.* 33(2-3): 61–83.

Johnson W, Turkheimer E, Gottesman II, Bouchard TJ (2009). Beyond Heritability: Twin Studies in Behavioral Research. *Curr. Dir. Psychol.* 18(4): 217–20.

Kovas Y, Haworth CMA, Dale PS, Plomin R (2007). The genetic and environmental origins of learning abilities and disabilities in the early school years. *Monogr. Soc. Res. Child Dev.* 72: 1–144.

Krapohl E, Rimfeld K, Shakeshaft NG, Trzaskowski M, McMillan A, Pingault JB, Asbury K, Harlaar N, Kovas Y, Dale PS, Plomin R (2014). The high heritability of educational achievement reflects many genetically influenced traits, not just intelligence. *Proc. Natl. Acad. Sci. USA* 111(42): 15273–15278.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science* 349: aac4716.

Petrides KV (2011). Social intelligence. *Encyclopedia of Adolescence* (eds. Brown BB, Prinstein MJ), pp 342–352. San Diego: Academic.

Petrides KV, Frederickson N, Furnham A (2004). The role of trait emotional intelligence in academic performance and deviant behavior at school. *Pers. Individ. Dif.* 36(2): 277–293.

Plomin R (2011). Commentary: Why are children in the same family so different? Non-shared environment three decades later. *Int. J. Epidemiol.* 40: 582–592.

Plomin R, DeFries JC, Knopik VS, Neiderhiser JM (2013). *Behavioral genetics*. New York: Worth Publishers.

Plomin R, DeFries JC, Knopik VS, Neiderhiser JM (2016). Top 10 Replicated Findings From Behavioral Genetics. *Perspect. Psychol. Sci.* 11(1): 3–23.

Ramsey SR, Thompson KL, McKenzie M, Rosenbaum A (2016). Psychological research in the internet age: The quality of web-based data. *Comput. Human Behav.* 58: 354–360.

Reichenberg A, Cederlöf M, McMillan A, Trzaskowski M, Kapara O, Fruchter E, Ginat K, Davidson M, Weiser M, Larsson H, Plomin R, Lichtenstein P (2016). Discontinuity in the genetic and environmental causes of the intellectual disability spectrum. *Proc. Natl. Acad. Sci. USA* 113(4): 1098–1103.

Rimfeld K, Kovas Y, Dale PS, Plomin R (2015). Pleiotropy across academic subjects at the end of compulsory education. *Sci. Rep.* 5: 11713.

Schweizer K, Goldhammer F, Rauch W, Moosbrugger H (2007). On the validity of Raven's matrices test: Does spatial ability contribute to performance? *Pers. Individ. Dif.* 43: 1998–2010.

Selzam S, Krapohl E, von Stumm S, O'Reilly PF, Rimfeld K, Kovas Y, Dale PS, Lee JJ, Plomin R (2016). Predicting educational achievement from DNA. *Mol. Psychiatry* Advance online publication.

Shams M, Williams R (1997). Generational changes in height and body mass differences between British Asians and the general population in Glasgow. *J. Biosoc. Sci.* 29(1): 101–9.

Smith SM, Roster CA, Golden LL, Albaum GS (2016). A multi-group analysis of online survey respondent data quality: Comparing a regular USA consumer panel to MTurk samples. *J. Bus. Res.* 69(8): 3139–3148.

Turkheimer E (2016). Weak Genetic Explanation 20 Years Later: Reply to Plomin et al. (2016). *Perspect. Psychol. Sci.* 11(1): 24–28.

Wolbers T, Hegarty M (2010). What determine our navigational abilities? *Trends Cogn. Sci.* 14: 138–146.

Wray NR, Lee SH, Mehta D, Vinkhuyzen AAE, Dudbridge F, Middeldorp CM (2014). Research review: polygenic methods and their application to psychiatric traits. *J. Child Psychol. Psychiatry* 55: 1068–1087.

Appendices

The supplementary materials for several chapters, as referenced in the text, are attached as appendices:

Appendix 1: Supplementary tables for Chapter 2

Appendix 2: Supplementary methods, figures and tables for Chapter 4

Appendix 3: Supplementary figures and tables for Chapter 5

Appendix 4: Supplementary figures and tables for Chapter 6

Appendix 5: Supplementary tables for Chapter 7

Appendix 1

Supplementary Online Material

Contents:

- Table S1: Details of construction for each subject / composite assessed.

- Table S2: Correlation matrix (Pearson's *r*) for all GCSE subjects in our dataset, excluding short-course GCSEs and subjects with sample sizes too small to analyse individually. Significance (2-tailed) and sample size (N) is shown for each.

- Table S3: Correlation matrix (Pearson's *r*) for GCSE subjects included in composites.

- Table S4: Sex limitation model fitting results (with 95% confidence intervals), showing A, C and E estimates separately for males and females.

- Tables S5-S11: Sex limitation sub-model comparisons.

Table S1: Construction of composites.

Mean grade for GCSE passes	All GCSE subjects in dataset, including those with sample sizes too small to analyse individually.
Number of GCSE passes at grade A*-C	All GCSE subjects in dataset, including those with sample sizes too small to analyse individually.
GCSE English mean grade	Mean of: English language, English literature (whichever taken)
GCSE science mean grade	Mean of: science core, science additional, biology, chemistry, physics (whichever taken)
Mathematics	Single GCSE; raw grade used
GCSE core subjects mean grade	Mean of: English composite, mathematics grade, science composite (requiring all three)
GCSE humanities mean grade	Mean of: media studies, history, religious education (RE), art, drama and music (whichever taken)

Table S2: Correlation matrix for all GCSE subjects (continues on next page).

		English Language	English Literature	Media Studies	Mathematics	Statistics	Science Core	Science Additional	Biology	Chemistry	Physics	History	Geography	RE	French	German	Spanish	DT	ICT	Business Studies	Art and Design	PE	Drama	Music
English Language	Correlation	1	.800	.609	.691	.646	.659	.631	.646	.633	.631	.729	.733	.683	.666	.634	.673	.611	.496	.609	.551	.546	.563	.526
	Sig. (2-tailed)		.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	N	5466	4791	461	5398	627	2890	2349	2159	2147	2142	2369	2082	2566	1884	780	731	2078	1166	775	1607	1325	692	562
English Literature	Correlation	.800	1	.549	.618	.597	.601	.586	.601	.599	.595	.696	.697	.638	.633	.619	.621	.599	.478	.601	.520	.500	.549	.535
	Sig. (2-tailed)	.000		.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	N	4791	4832	366	4775	592	2513	2109	2065	2053	2047	2239	1933	2335	1818	754	689	1826	1034	711	1436	1162	631	541
Media Studies	Correlation	.609	.549	1	.508	.489	.498	.520	.539	.429	.420	.591	.649	.434	.562	.559	.382	.532	.390	.537	.439	.338	.621	.674
	Sig. (2-tailed)	.000	.000		.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.011	.000	.000	.000	.000	.001	.000	.000
	N	461	366	467	455	57	298	239	125	124	124	145	120	188	109	57	44	150	104	71	128	96	73	33
Mathematics	Correlation	.691	.618	.508	1	.785	.752	.735	.734	.763	.775	.686	.721	.595	.643	.611	.670	.606	.545	.602	.483	.577	.423	.536
	Sig. (2-tailed)	.000	.000	.000		.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	N	5398	4775	455	5461	627	2889	2347	2157	2145	2139	2363	2080	2559	1878	783	727	2081	1162	773	1606	1320	691	567
Statistics	Correlation	.646	.597	.489	.785	1	.675	.661	.700	.669	.688	.625	.653	.614	.611	.597	.630	.605	.580	.600	.447	.425	.133	.548
	Sig. (2-tailed)	.000	.000	.000	.000		.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.335	.000
	N	627	592	57	627	629	276	247	333	334	333	269	257	290	236	115	90	232	168	104	171	157	55	75
Science Core	Correlation	.659	.601	.498	.752	.675	1	.831				.664	.720	.594	.612	.540	.613	.529	.460	.629	.422	.534	.439	.468
	Sig. (2-tailed)	.000	.000	.000	.000	.000		.000				.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	N	2890	2513	298	2889	276	2931	2372				1199	1025	1369	855	337	345	1165	708	444	908	789	421	280
Science Additional	Correlation	.631	.586	.520	.735	.661	.831	1				.650	.699	.568	.651	.600	.622	.550	.444	.550	.449	.523	.428	.505
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000					.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	N	2349	2109	239	2347	247	2372	2372				1018	883	1145	769	296	305	960	557	368	733	620	335	243
Biology	Correlation	.646	.601	.539	.734	.700			1	.828	.821	.678	.707	.620	.610	.574	.633	.535	.491	.646	.463	.530	.368	.399
	Sig. (2-tailed)	.000	.000	.000	.000	.000				.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	N	2159	2065	125	2157	333			2174	2145	2137	1093	987	1046	989	427	368	773	426	300	598	470	228	270
Chemistry	Correlation	.633	.599	.429	.763	.669			.828	1	.834	.681	.690	.609	.637	.544	.616	.548	.479	.612	.471	.541	.321	.402
	Sig. (2-tailed)	.000	.000	.000	.000	.000			.000		.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	N	2147	2053	124	2145	334			2145	2162	2141	1091	985	1032	983	429	365	766	425	298	590	467	221	269
Physics	Correlation	.631	.595	.420	.775	.688			.821	.834	1	.673	.651	.584	.615	.572	.583	.519	.469	.625	.448	.558	.354	.443
	Sig. (2-tailed)	.000	.000	.000	.000	.000			.000	.000		.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	N	2142	2047	124	2139	333			2137	2141	2157	1081	981	1035	977	428	366	766	423	294	585	467	226	269
History	Correlation	.729	.696	.591	.686	.625	.664	.650	.678	.681	.673	1	.745	.689	.643	.590	.620	.610	.504	.670	.505	.577	.410	.483
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000		.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	N	2369	2239	145	2363	269	1199	1018	1093	1091	1081	2388	828	1179	967	406	350	766	467	278	614	504	289	243
Geography	Correlation	.733	.697	.649	.721	.653	.720	.699	.707	.690	.651	.745	1	.678	.653	.606	.653	.673	.515	.658	.575	.639	.558	.517
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000		.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	N	2082	1933	120	2080	257	1025	883	987	985	981	828	2100	953	838	325	299	734	412	254	551	517	195	190

		English Language	English Literature	Media Studies	Mathematics	Statistics	Science Core	Science Additional	Biology	Chemistry	Physics	History	Geography	RE	French	German	Spanish	DT	ICT	Business Studies	Art and Design	PE	Drama	Music
RE	Correlation	.683	.638	.434	.595	.614	.594	.568	.620	.609	.584	.689	.678	1	.584	.520	.591	.591	.480	.558	.491	.543	.531	.555
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000		.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	N	2566	2335	188	2559	290	1369	1145	1046	1032	1035	1179	953	2587	930	377	352	992	599	369	751	619	345	273
French	Correlation	.666	.633	.562	.643	.611	.612	.651	.610	.637	.615	.643	.653	.584	1	.799	.776	.539	.479	.607	.524	.568	.401	.489
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000		.000	.000	.000	.000	.000	.000	.000	.000	.000
	N	1884	1818	109	1878	236	855	769	989	983	977	967	838	930	1896	143	202	637	360	240	535	433	241	241
German	Correlation	.634	.619	.559	.611	.597	.540	.600	.574	.544	.572	.590	.606	.520	.799	1	.867	.504	.450	.579	.405	.448	.400	.440
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000		.000	.000	.000	.000	.000	.000	.000	.000
	N	780	754	57	783	115	337	296	427	429	428	406	325	377	143	787	45	286	162	104	238	164	85	93
Spanish	Correlation	.673	.621	.382	.670	.630	.613	.622	.633	.616	.583	.620	.653	.591	.776	.867	1	.479	.343	.500	.473	.474	.538	.556
	Sig. (2-tailed)	.000	.000	.011	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000		.000	.000	.000	.000	.000	.000	.000
	N	731	689	44	727	90	345	305	368	365	366	350	299	352	202	45	735	224	145	102	191	152	88	78
DT	Correlation	.611	.599	.532	.606	.605	.529	.550	.535	.548	.519	.610	.673	.591	.539	.504	.479	1	.469	.566	.525	.487	.336	.335
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000		.000	.000	.000	.000	.000	.000
	N	2078	1826	150	2081	232	1165	960	773	766	766	766	734	992	637	286	224	2102	482	293	603	494	194	166
ICT	Correlation	.496	.478	.390	.545	.580	.460	.444	.491	.479	.469	.504	.515	.480	.479	.450	.343	.469	1	.452	.363	.316	.428	.415
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000		.000	.000	.000	.000	.000
	N	1166	1034	104	1162	168	708	557	426	425	423	467	412	599	360	162	145	482	1179	203	322	342	163	99
Business Studies	Correlation	.609	.601	.537	.602	.600	.629	.550	.646	.612	.625	.670	.658	.558	.607	.579	.500	.566	.452	1	.491	.567	.260	.465
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000		.000	.000	.026	.001
	N	775	711	71	773	104	444	368	300	298	294	278	254	369	240	104	102	293	203	784	170	214	73	45
Art and Design	Correlation	.551	.520	.439	.483	.447	.422	.449	.463	.471	.448	.505	.575	.491	.524	.405	.473	.525	.363	.491	1	.387	.327	.357
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000		.000	.000	.000
	N	1607	1436	128	1606	171	908	733	598	590	585	614	551	751	535	238	191	603	322	170	1629	280	180	141
PE	Correlation	.546	.500	.338	.577	.425	.534	.523	.530	.541	.558	.577	.639	.543	.568	.448	.474	.487	.316	.567	.387	1	.383	.514
	Sig. (2-tailed)	.000	.000	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000		.000	.000
	N	1325	1162	96	1320	157	789	620	470	467	467	504	517	619	433	164	152	494	342	214	280	1338	129	94
Drama	Correlation	.563	.549	.621	.423	.133	.439	.428	.368	.321	.354	.410	.558	.531	.401	.400	.538	.336	.428	.260	.327	.383	1	.613
	Sig. (2-tailed)	.000	.000	.000	.000	.335	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.026	.000	.000		.000
	N	692	631	73	691	55	421	335	228	221	226	289	195	345	241	85	88	194	163	73	180	129	700	86
Music	Correlation	.526	.535	.674	.536	.548	.468	.505	.399	.402	.443	.483	.517	.555	.489	.440	.556	.335	.415	.465	.357	.514	.613	1
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.001	.000	.000	.000	
	N	562	541	33	567	75	280	243	270	269	269	243	190	273	241	93	78	166	99	45	141	94	86	568

Note: Blank cells denote results which are incomputable due to lack of data (i.e., between subjects which are mutually exclusive in the GCSE syllabus, or those with very small samples).

Table S3: Correlation matrix for subjects included in composites.

		English Language	English Literature	Media Studies	Mathematics	Science Core	Science Additional	Biology	Chemistry	Physics	History	RE	Art and Design	Drama	Music
English Language	Correlation	1	.800	.609	.691	.659	.631	.646	.633	.631	.729	.683	.551	.563	.526
	N	5466	4791	461	5398	2890	2349	2159	2147	2142	2369	2566	1607	692	562
English Literature	Correlation	.800	1	.549	.618	.601	.586	.601	.599	.595	.696	.638	.520	.549	.535
	N	4791	4832	366	4775	2513	2109	2065	2053	2047	2239	2335	1436	631	541
Media Studies	Correlation	.609	.549	1	.508	.498	.520	.539	.429	.420	.591	.434	.439	.621	.674
	N	461	366	467	455	298	239	125	124	124	145	188	128	73	33
Mathematics	Correlation	.691	.618	.508	1	.752	.735	.734	.763	.775	.686	.595	.483	.423	.536
	N	5398	4775	455	5461	2889	2347	2157	2145	2139	2363	2559	1606	691	567
Science Core	Correlation	.659	.601	.498	.752	1	.831				.664	.594	.422	.439	.468
	N	2890	2513	298	2889	2931	2372				1199	1369	908	421	280
Science Additional	Correlation	.631	.586	.520	.735	.831	1				.650	.568	.449	.428	.505
	N	2349	2109	239	2347	2372	2372				1018	1145	733	335	243
Biology	Correlation	.646	.601	.539	.734			1	.828	.821	.678	.620	.463	.368	.399
	N	2159	2065	125	2157			2174	2145	2137	1093	1046	598	228	270
Chemistry	Correlation	.633	.599	.429	.763			.828	1	.834	.681	.609	.471	.321	.402
	N	2147	2053	124	2145			2145	2162	2141	1091	1032	590	221	269
Physics	Correlation	.631	.595	.420	.775			.821	.834	1	.673	.584	.448	.354	.443
	N	2142	2047	124	2139			2137	2141	2157	1081	1035	585	226	269
History	Correlation	.729	.696	.591	.686	.664	.650	.678	.681	.673	1	.689	.505	.410	.483
	N	2369	2239	145	2363	1199	1018	1093	1091	1081	2388	1179	614	289	243
RE	Correlation	.683	.638	.434	.595	.594	.568	.620	.609	.584	.689	1	.491	.531	.555
	N	2566	2335	188	2559	1369	1145	1046	1032	1035	1179	2587	751	345	273
Art and Design	Correlation	.551	.520	.439	.483	.422	.449	.463	.471	.448	.505	.491	1	.327	.357
	N	1607	1436	128	1606	908	733	598	590	585	614	751	1629	180	141
Drama	Correlation	.563	.549	.621	.423	.439	.428	.368	.321	.354	.410	.531	.327	1	.613
	N	692	631	73	691	421	335	228	221	226	289	345	180	700	86
Music	Correlation	.526	.535	.674	.536	.468	.505	.399	.402	.443	.483	.555	.357	.613	1
	N	562	541	33	567	280	243	270	269	269	243	273	141	86	568

Note: This table duplicates relevant correlations from SOM Table 2, for convenience. Mean correlations between the individual subjects included in composites (see SOM Table 1) are as follows:- English: .80, Science: .83, Humanities: .51, Core Subjects: .7. The overall mean correlation (.56) given in the text includes all subjects in the dataset with sufficient sample sizes to analyse (i.e., all subjects listed in SOM Table 2).

Table S4: Sex limitation A, C and E estimates (95% confidence intervals).

	Male			Female		
	A	C	E	A	C	E
Mean grade for GCSE passes	0.57 (0.49 - 0.66)	0.29 (0.20 - 0.37)	0.13 (0.12 - 0.15)	0.47 (0.41 - 0.54)	0.43 (0.36 - 0.49)	0.1 (0.09 - 0.11)
Number of GCSE passes at grade A*-C	0.56 (0.47 - 0.66)	0.26 (0.16 - 0.34)	0.19 (0.17 - 0.21)	0.46 (0.39 - 0.54)	0.38 (0.3 - 0.45)	0.16 (0.15 - 0.18)
GCSE English mean grade	0.54 (0.45 - 0.65)	0.26 (0.16 - 0.35)	0.20 (0.18 - 0.22)	0.48 (0.41 - 0.56)	0.36 (0.28 - 0.43)	0.16 (0.15 - 0.18)
GCSE science mean grade	0.68 (0.58 - 0.80)	0.13 (0.01 - 0.23)	0.19 (0.17 - 0.21)	0.50 (0.42 - 0.59)	0.33 (0.24 - 0.41)	0.17 (0.15 - 0.19)
Mathematics	0.57 (0.47 - 0.68)	0.22 (0.12 - 0.32)	0.21 (0.19 - 0.23)	0.54 (0.46 - 0.63)	0.29 (0.21 - 0.37)	0.17 (0.15 - 0.18)
GCSE core subjects mean grade	0.63 (0.53 - 0.73)	0.23 (0.12 - 0.32)	0.15 (0.13 - 0.16)	0.53 (0.46 - 0.62)	0.35 (0.27 - 0.43)	0.11 (0.10 - 0.12)
GCSE humanities mean grade	0.37 (0.25 - 0.49)	0.35 (0.23 - 0.45)	0.28 (0.25 - 0.32)	0.47 (0.37 - 0.58)	0.29 (0.19 - 0.39)	0.23 (0.21 - 0.26)

Note: All subjects are composites, except for mathematics (the single GCSE). Heritability is consistently higher for males, and shared environment higher for females – except for the humanities composite, for which this pattern is reversed. For females, heritability is similar (around 50%) for all the subjects and composites assessed. For males, it is more variable, reaching 68% for science (the highest heritability found for any subject), and falling almost in half to 37% for humanities (the lowest found for any subject). However, caution is warranted in interpreting these results, for reasons discussed in the text. In addition, the confidence intervals are quite wide, and often overlapping between males and females.

Tables S5-S11

Note: These tables present fit indices for sex-limitation sub-model comparisons. All comparisons are made with the full sex-limitation model. The same pattern of results is shown for all variables: significant quantitative but no qualitative sex differences (however, the null model is the most informative in this instance, as discussed in the text). ep = estimated parameters, Δχ2 = change in chi-square (-2 log-likelihood) between the models, Δdf = change in degrees of freedom.

Table S5: Sex limitation sub-model comparisons: Mean grade for GCSE passes

Model	ep	χ2	df	AIC	Δχ2	Δdf	p
Full sex-limited	9	17292.3	10965	-4637.702	-	-	-
Quantitative differences	6	17352.23	10968	-4583.77	59.93	3	< .01
Qualitative differences (fixed rG)	8	17292.3	10966	-4639.702	0	1	1.00
Qualitative differences (fixed rC)	9	17292.3	10965	-4637.702	0	0	1.00
Null model	5	17352.23	10969	-4585.77	59.93	4	< .01

Table S6: Sex limitation sub-model comparisons: Number of GCSE passes at grade A*-C

Model	ep	χ2	df	AIC	Δχ2	Δdf	p
Full sex-limited	9	18211.53	11050	-3888.469	-	-	-
Quantitative differences	6	18250.71	11053	-3855.294	39.17	3	< .01
Qualitative differences (fixed rG)	8	18211.53	11051	-3890.469	0	1	1.00
Qualitative differences (fixed rC)	9	18211.53	11050	-3888.469	0	0	1.00
Null model	5	18250.71	11054	-3857.294	39.17	4	< .01

Table S7: Sex limitation sub-model comparisons: GCSE English mean grade

Model	ep	χ2	df	AIC	Δχ2	Δdf	p
Full sex-limited	9	18162.12	10882	-3601.883	-	-	-
Quantitative differences	6	18221.53	10885	-3548.472	59.41	3	< .01
Qualitative differences (fixed rG)	8	18162.12	10883	-3603.883	0	1	1.00
Qualitative differences (fixed rC)	9	18162.12	10882	-3601.883	0	0	1.00
Null model	5	18221.53	10886	-3550.472	59.41	4	< .01

Table S8: Sex limitation sub-model comparisons: GCSE science mean grade

Model	ep	χ2	df	AIC	Δχ2	Δdf	p
Full sex-limited	9	17123.15	10124	-3124.847	-	-	-
Quantitative differences	6	17133.86	10127	-3120.136	10.71	3	0.01
Qualitative differences (fixed rG)	8	17123.15	10125	-3126.847	0	1	1.00
Qualitative differences (fixed rC)	9	17123.15	10124	-3124.847	0	0	1.00
Null model	5	17133.86	10128	-3122.136	10.71	4	0.03

Table S9: Sex limitation sub-model comparisons: Mathematics

Model	ep	χ2	df	AIC	Δχ2	Δdf	p
Full sex-limited	9	18118.01	10806	-3493.987	-	-	-
Quantitative differences	6	18130.59	10809	-3487.407	12.58	3	< .01
Qualitative differences (fixed rG)	8	18118.01	10807	-3495.987	0	1	1.00
Qualitative differences (fixed rC)	9	18118.01	10806	-3493.987	0	0	1.00
Null model	5	18130.59	10810	-3489.407	12.58	4	0.01

Table S10: Sex limitation sub-model comparisons: GCSE core subjects mean grade

Model	ep	χ2	df	AIC	Δχ2	Δdf	p
Full sex-limited	9	16224.42	9995	-3765.582	-	-	-
Quantitative differences	6	16252.67	9998	-3743.33	28.25	3	< .01
Qualitative differences (fixed rG)	8	16224.42	9996	-3767.582	0	1	1.00
Qualitative differences (fixed rC)	9	16224.42	9995	-3765.582	0	0	1.00
Null model	5	16252.67	9999	-3745.33	28.25	4	< .01

Table S11: Sex limitation sub-model comparisons: GCSE humanities mean grade

Model	ep	χ2	df	AIC	Δχ2	Δdf	p
Full sex-limited	9	16295.03	9314	-2332.971	-	-	-
Quantitative differences	6	16358.53	9317	-2275.469	63.50	3	< .01
Qualitative differences (fixed rG)	8	16295.03	9315	-2334.971	0	1	1.00
Qualitative differences (fixed rC)	9	16295.03	9314	-2332.971	0	0	1.00
Null model	5	16358.53	9318	-2277.469	63.50	4	< .01

Appendix 2

Rotation is visualisation, 3D is 2D: using a novel measure to investigate the genetics of spatial ability

Nicholas G. Shakeshaft, Kaili Rimfield, Kerry L. Schofield, Saskia Selzam, Margherita Malanchini, Maja Rodic, Yulia Kovas & Robert Plomin

Supplementary Information

Supplementary methods – the Bricks battery

Rationale, development and description of the Bricks measures.

Figures

Tables

Supplementary methods – the Bricks battery

Rationale

As discussed in the main text, the literature on spatial abilities is inconsistent regarding the relationship between mental rotation and spatial visualisation, and between 2D and 3D stimuli. If rotation and visualisation were dissociable processes, it was reasoned that traditional 2D and 3D mental rotation stimuli may engage them differently. With 3D mental rotation stimuli, target objects commonly rotate freely in three dimensions, such that key identifiable features are out of view or disguised by foreshortening. However, with 2D stimuli, in which the object rotates only in the picture plane (i.e., as though rotating the whole image itself, rather than the object), full information about the object is always available, and there is no need to visualise missing or disguised features.

The Bricks battery was therefore developed to isolate rotation and visualisation cleanly, and to include stimuli depicting 2D objects with concealed features (as a closer match to common 3D stimuli), and 3D objects which do *not* obscure features (as with common 2D stimuli). In this way, the putative rotation and visualisation processes could be assessed both separately and together, equally in 2D and in 3D.

Design

Six subtests were conceived. Each consists of a series of items with a stimulus image containing a “target” object, and four multiple-choice response images, only one of which (the correct response) depicts the same object as the target, following a suitable transformation. Participants completed the subtests in the following order:

i) 2D Rotation: the most “natural” form of 2D rotation, in which the target (a two-dimensional object) is rotated only in the picture plane, and the target stimulus and correct response contain exactly the same information.

ii) 2D Rotation / Visualisation combined: to add the element of incomplete information commonly found in 3D stimuli, the target object is partially obscured behind an “occluder” - a square or circle quadrant partially obscuring the target. In the correct response, the target has rotated (in the picture plane) but the occluder is immobile.

iii) 2D Visualisation: the target remains entirely motionless and unchanged, but the occluder is in a different location in the correct response, thereby revealing a different portion of the target.

iv) 3D Rotation / Visualisation combined: the most “natural” form of 3D rotation, in which the target (a three-dimensional object, computer-generated and rendered with simple overhead “lighting”) has been rotated freely in three dimensions in the correct response.

v) 3D Rotation: corresponding to 2D rotation but with an image of an apparently three-dimensional object – in the correct response, the target is rotated only in the picture plane (i.e., as though the whole image had rotated, or the “camera” showing the scene had rotated on the spot). As with 2D rotation, the target stimulus and correct response therefore contain invariant information, with even the lighting and shadows remaining unchanged.

vi) 3D Visualisation: to assess visualisation without rotation, the target stimulus depicts a wireframe drawing of an object, and the correct response shows the “solid” version, otherwise unchanged. The participant must therefore use the available information to determine how the solid will appear (e.g., which features are in view from the current perspective and which are obscured by others).

Development

A JavaScript web application, “Building Bricks”, was developed to enable appropriate stimuli to be created for each subtest. This allows the creation of images of “bricks” (rectangular blocks, either 2D or 3D) of variable size, including one or more “studs” – protrusions of arbitrary length emerging from the main body of the brick, from the “top”, “bottom” or both. 2D bricks may be rotated in the picture plane, 3D bricks in any direction, and the camera may be rotated to simulate picture-plane rotation for 3D objects. Occluders

(squares or circle quadrants) of arbitrary size may be added to any corner of the image. Various other options such as colours or camera distances may be altered as required, and bricks may be presented in wireframe or solid form.

This software is freely available online under the open-source MIT license, and researchers are welcome to experiment with it to see how the constructs were operationalised, or to create their own items. It is accessible via this page: https://www.forepsyte.com/resources/public

For each subtest, 12 items of varying difficulty were created and administered, but with a view to reducing this to 9 items post hoc before the calculation of scores. This allowed the final selection to be approximately equated for difficulty between subtests, and for 'experimental' items (e.g., those with potentially counterintuitive responses) to be included in the initial battery before being discarded on the basis of their psychometric properties. Examples of stimulus images and the corresponding correct responses are shown in Fig. 1.

Participants completed the Bricks battery online, via a website created for the purpose using the open-source “psy.js” JavaScript library, which was developed specifically for the administration of psychometric measures such as questionnaires and cognitive tests. This library is also freely available at the link above.

Procedure

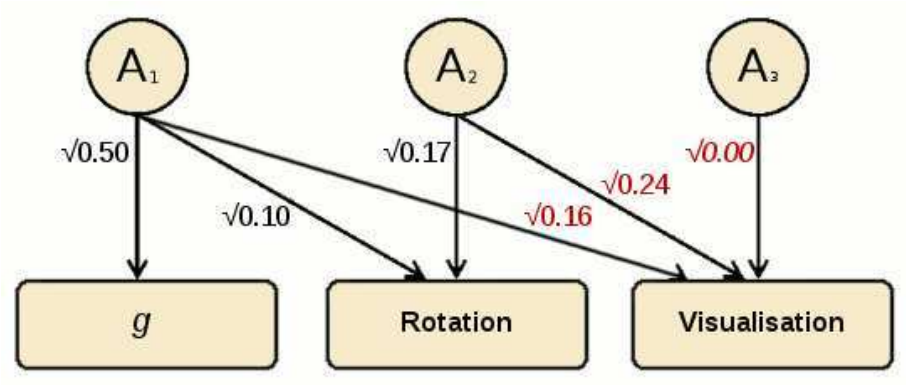
For each subtest, participants read appropriate instructions, completed two simple practice items (which provided feedback and clarification of the subtest rules), and then completed the test items in a fixed sequence of approximately increasing difficulty (selected based on pilot work). A time limit of 20 seconds was allowed for each item – the time remaining was displayed to participants via a timer at the top right of the screen. If participants made four consecutive incorrect responses, they were discontinued from the current subtest and began the next. Including the time spent reading instructions and reviewing practice items, the battery typically took 20-25 minutes to complete.

Data cleaning and scoring

After the participant exclusions described in the main text (e.g., excluding those with relevant severe disabilities), and prior to the data preparation procedures described (outlier removal, etc.), additional exclusions were made on the basis of suspected random or thoughtless responding. Conservative cut-offs were used to identify participants with very low variability in their responses – 3SD below the mean, indicating that they had clicked on the same response option repeatedly for most or all items – or with mean reaction times of less than one second per item. Participants falling below these cut-offs were excluded from analysis.

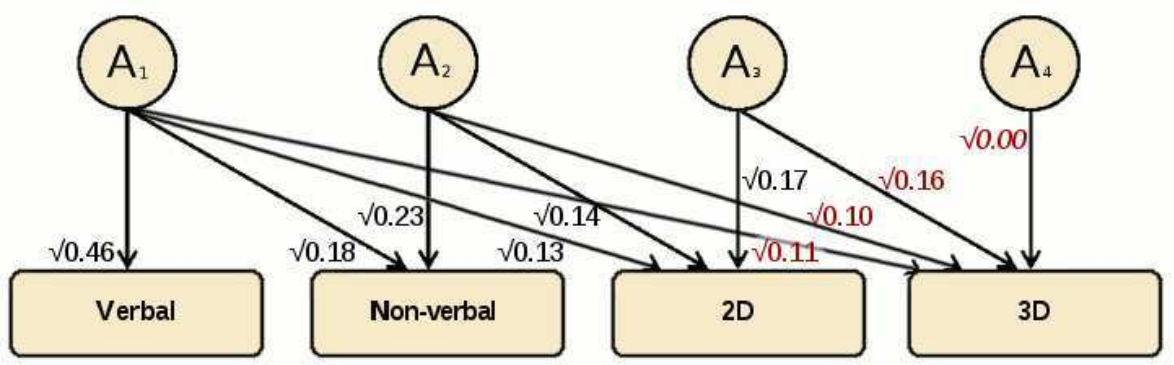
For each item, a score of 1 was awarded for a correct response, or 0 for incorrect responses, no response or the item being skipped due to discontinuation. Scores from the nine items in the final battery were summed to yield subtest scores. These individual subtest scores were then cleaned and combined into “functional”, “dimensional” and “overall Bricks” composites, as described in the main text.

Fig. S1. Trivariate Cholesky decomposition path estimates: *g*, Rotation, Visualisation.



Path estimates (standardised) for the structure of additive genetic influences on *g*, Rotation and Visualisation (see Table S31 for more details). The paths in red indicate the genetic influences on Visualisation (the last variable in the model): i) those common to all three variables; ii) those shared only between Rotation and Visualisation but not with *g* (suggesting influences specific to spatial ability); and iii) those unique to Visualisation alone. The latter (italicised) is non-significant – i.e., all genetic influences on Visualisation are shared with Rotation.

Fig. S2. Quadrivariate Cholesky decomposition path estimates: Verbal, non-verbal, 2D, 3D.



Path estimates (standardised) for the structure of additive genetic influences on verbal ability, non-verbal ability, and the 2D and 3D Bricks composites (see Table S36 for more details). The paths in red indicate the genetic influences on 3D (the last variable in the model): i) those common to all four variables; ii) those shared between non-verbal ability, and the 2D and 3D Bricks composites, but not with verbal ability; iii) those shared only between 2D and 3D but not with verbal or non-verbal ability (suggesting influences specific to spatial ability); and iv) those unique to 3D alone. The latter (italicised) is non-significant – i.e., all genetic influences on 3D are shared with 2D.

Table S1. Descriptive statistics.

	N	Whole sample	Males	Females	MZs	DZs	Sex	Zyg	Sex x zyg	R^2
2D Rotation	1451	5.58 (1.68)	5.77 (1.75)	5.47 (1.64)	5.47. (1.72)	5.65 (1.65)	12.73 **	2.80	1.19	0.01
2D Rotation / Visualisation	1443	4.87 (1.94)	5.16 (2.01)	4.70 (1.89)	4.76 (1.96)	4.95 (1.93)	19.48 **	2.35	1.92	0.02
2D Visualisation	1434	5.05 (2.02)	5.32 (1.99)	4.90 (2.02)	5.08 (2.00)	5.04 (2.03)	15.92 **	0.37	0.24	0.01
3D Rotation	1403	6.41 (1.69)	6.53 (1.71)	6.34 (1.67)	6.44 (1.67)	6.39 (1.70)	4.91 *	0.56	0.16	0.00
3D Rotation / Visualisation	1426	5.58 (1.34)	5.85 (1.42)	5.42 (1.26)	5.56 (1.33)	5.59 (1.34)	33.70 **	0.00	1.88	0.02
3D Visualisation	1427	5.61 (2.07)	6.01 (2.01)	5.37 (2.07)	5.56 (2.09)	5.64 (2.06)	36.40 **	0.14	1.48	0.03
Rotation	1435	5.99 (1.34)	6.17 (1.36)	5.88 (1.32)	5.97 (1.36)	6.00 (1.33)	17.40 **	0.02	1.92	0.01
Rotation / Visualisation	1440	5.23 (1.36)	5.52 (1.43)	5.06 (1.29)	5.17 (1.34)	5.27 (1.37)	39.00 **	0.78	2.63	0.03
Visualisation	1429	5.35 (1.69)	5.69 (1.61)	5.15 (1.70)	5.35 (1.65)	5.35 (1.71)	35.83 **	0.07	0.00	0.03
2D	1451	5.17 (1.42)	5.44 (1.43)	5.02 (1.39)	5.12 (1.41)	5.21 (1.43)	31.41 **	0.70	1.04	0.02
3D	1414	5.88 (1.28)	6.15 (1.30)	5.72 (1.25)	5.89 (1.24)	5.87 (1.31)	45.88 **	0.31	1.81	0.03
Overall Bricks	1443	5.52 (1.21)	5.79 (1.23)	5.37 (1.18)	5.49 (1.19)	5.54 (1.23)	46.90 **	0.08	1.28	0.03
Verbal	1442	15.61 (4.00)	15.72 (4.05)	15.55 (3.97)	15.30 (3.98)	15.82 (4.00)	0.01	3.40	0.02	0.00
Non-verbal	1437	14.06 (3.63)	14.33 (3.76)	13.91 (3.54)	13.95 (3.61)	14.14 (3.64)	3.40	0.01	0.82	0.00
g	1439	0.06 (0.97)	0.12 (0.99)	0.03 (0.96)	0.00 (0.97)	0.10 (0.98)	1.26	2.12	0.84	0.00

Mean scores (standard deviations) for the whole sample, separately by sex, and for MZ and DZ twins, for the six Bricks subtests, the three functional and two dimensional composites, the single overall Bricks mean, and the other cognitive measures. N = sample size (the sample shown is fully independent, selecting one individual randomly per twin pair). ANOVA performed on cleaned, normality-transformed data to test effects of sex and zygosity. Results = F statistic; ** = $p < 0.01$; * = $p < 0.05$; R^2 = proportion of variance explained by sex, zygosity and their interaction.

Table S2. Internal consistency and test-retest reliability of Bricks measures.

	Consistency		**Test-retest reliability**
	Alpha	N	Pearson's *r* (N = 45, p < 0.01)
2D Rotation	0.45	1453	0.42
2D Rotation / Visualisation	0.60	1443	0.63
2D Visualisation	0.63	1434	0.52
3D Rotation	0.62	1429	0.59
3D Rotation / Visualisation	0.47	1431	0.51
3D Visualisation	0.71	1427	0.57
Rotation	0.63	1429	0.62
Rotation / Visualisation	0.66	1431	0.75
Visualisation	0.75	1427	0.62
2D	0.74	1434	0.77
3D	0.78	1427	0.67
Overall Bricks	0.85	1427	0.83

Consistency (Cronbach's alpha) and test-retest reliability (Pearson's *r*) for the six Bricks subtests, the three functional and two dimensional composites, and the single overall mean. The consistency sample is fully independent, with one individual selected randomly from each twin pair. Test-retest reliability was assessed with a separate pilot sample.

Table S3. Subtest intercorrelations.

		Overall Bricks	2D R	2D R / V	2D V	3D R	3D R / V	3D V
2D Rotation	*r*	**0.57**	1					
	N	***1441***	*1451*					
2D Rotation / Visualisation	*r*	**0.71**	0.31	1				
	N	***1433***	*1441*	*1443*				
2D Visualisation	*r*	**0.69**	0.27	0.42	1			
	N	***1424***	*1432*	*1434*	*1434*			
3D Rotation	*r*	**0.62**	0.25	0.34	0.35	1		
	N	***1401***	*1401*	*1403*	*1403*	*1403*		
3D Rotation / Visualisation	*r*	**0.59**	0.27	0.32	0.31	0.26	1	
	N	***1420***	*1424*	*1426*	*1426*	*1403*	*1426*	
3D Visualisation	*r*	**0.68**	0.27	0.37	0.34	0.34	0.38	1
	N	***1417***	*1425*	*1427*	*1427*	*1401*	*1422*	*1427*

Correlations (Pearson's *r*) between the six subtests, and between each subtest and the overall Bricks mean. The sample is fully independent, with one individual selected randomly from each twin pair. R = Rotation; R / V = Rotation and Visualisation; V = Visualisation. All correlations significant at p < 0.0001.

Table S4. Subtest factor analysis.

	Factor loading
2D Rotation	0.57
2D Rotation / Visualisation	0.70
2D Visualisation	0.68
3D Rotation	0.65
3D Rotation / Visualisation	0.63
3D Visualisation	0.68

Factor loadings of Bricks subtests on the first (and only) principal component produced by factor analysis of the six subtest scores. The factor accounts for 2.56 eigenvalues, 42.6% of total variance.

Table S5. Bricks correlations with other measures.

		Mill Hill (verbal)	Raven's Matrices (non-verbal)	*g*
Rotation	*r*	0.13	0.35	0.30
	N	*1414*	*1410*	*1412*
Rotation / Visualisation	*r*	0.24	0.45	0.41
	N	*1419*	*1414*	*1416*
Visualisation	*r*	0.21	0.44	0.39
	N	*1410*	*1405*	*1408*
2D	*r*	0.19	0.44	0.39
	N	*1430*	*1425*	*1427*
3D	*r*	0.22	0.46	0.41
	N	*1393*	*1389*	*1391*
Overall Bricks	*r*	0.22	0.50	0.44
	N	*1422*	*1417*	*1419*

Correlations (Pearson's *r*) with other cognitive measures for the three functional and two dimensional Bricks composites, and the single overall mean. Mill Hill and Raven's Matrices correlate $r = 0.31$ with each other in this sample (N = 1420). All correlations significant at $p < 0.0001$.

N.B. The Rotation correlations with each other measure are significantly lower than those of Visualisation (all $p < 0.01$); however, since the 'Rotation / Visualisation combined' correlations do *not* differ significantly from those of Visualisation (despite the 'Rotation / Visualisation combined' conditions including both elements), this seems most likely to be related to the slightly lower reliability of one of the Rotation subtests (2D rotation) compared to the others, coupled with the highly-powered sample size, rather than representing a theoretically meaningful difference.

Table S6. Subtest intercorrelations, regressed on verbal ability.

		Overall Bricks	2D R	2D R / V	2D V	3D R	3D R / V	3D V
2D Rotation	*r*	**0.57**	1					
	N	***1420***	*1430*					
2D Rotation / Visualisation	*r*	**0.70**	0.29	1				
	N	***1412***	*1420*	*1422*				
2D Visualisation	*r*	**0.68**	0.26	0.39	1			
	N	***1403***	*1411*	*1413*	*1413*			
3D Rotation	*r*	**0.62**	0.24	0.33	0.33	1		
	N	***1380***	*1380*	*1382*	*1382*	*1382*		
3D Rotation / Visualisation	*r*	**0.57**	0.26	0.29	0.29	0.25	1	
	N	***1399***	*1403*	*1405*	*1405*	*1382*	*1405*	
3D Visualisation	*r*	**0.66**	0.26	0.34	0.32	0.32	0.36	1
	N	***1396***	*1404*	*1406*	*1406*	*1380*	*1401*	*1406*

Correlations (Pearson's *r*) between the six subtest residuals after regression on verbal ability (Mill Hill scores), and between each subtest and the overall Bricks mean. R = Rotation; R / V = Rotation and Visualisation; V = Visualisation. All correlations significant at $p < 0.0001$.

Table S7. Subtest intercorrelations, regressed on non-verbal ability.

		Overall Bricks	2D R	2D R / V	2D V	3D R	3D R / V	3D V
2D Rotation	*r*	**0.52**	1					
	N	***1415***	*1425*					
2D Rotation / Visualisation	*r*	**0.65**	0.22	1				
	N	***1407***	*1415*	*1417*				
2D Visualisation	*r*	**0.64**	0.19	0.33	1			
	N	***1398***	*1406*	*1408*	*1408*			
3D Rotation	*r*	**0.58**	0.18	0.25	0.27	1		
	N	***1376***	*1376*	*1378*	*1378*	*1378*		
3D Rotation / Visualisation	*r*	**0.50**	0.19	0.21	0.20	0.17	1	
	N	***1394***	*1398*	*1400*	*1400*	*1378*	*1400*	
3D Visualisation	*r*	**0.60**	0.18	0.25	0.24	0.25	0.28	1
	N	***1391***	*1399*	*1401*	*1401*	*1376*	*1396*	*1401*

Correlations (Pearson's *r*) between the six subtest residuals after regression on non-verbal ability (Raven's Matrices scores), and between each subtest and the overall Bricks mean. R = Rotation; R / V = Rotation and Visualisation; V = Visualisation. All correlations significant at $p < 0.0001$.

Table S8. Subtest intercorrelations, regressed on *g*.

		Overall Bricks	2D R	2D R / V	2D V	3D R	3D R / V	3D V
2D Rotation	*r*	**0.55**	1					
	N	***1417***	*1427*					
2D Rotation / Visualisation	*r*	**0.66**	0.24	1				
	N	***1409***	*1417*	*1419*				
2D Visualisation	*r*	**0.65**	0.21	0.35	1			
	N	***1400***	*1408*	*1410*	*1410*			
3D Rotation	*r*	**0.59**	0.21	0.28	0.29	1		
	N	***1378***	*1378*	*1380*	*1380*	*1380*		
3D Rotation / Visualisation	*r*	**0.52**	0.21	0.23	0.23	0.19	1	
	N	***1396***	*1400*	*1402*	*1402*	*1380*	*1402*	
3D Visualisation	*r*	**0.62**	0.20	0.27	0.26	0.27	0.30	1
	N	***1393***	*1401*	*1403*	*1403*	*1378*	*1398*	*1403*

Correlations (Pearson's *r*) between the six subtest residuals after regression on *g* (the mean of verbal and non-verbal ability scores), and between each subtest and the overall Bricks mean. R = Rotation; R / V = Rotation and Visualisation; V = Visualisation. All correlations significant at $p < 0.0001$.

Table S9. Functional composite intercorrelations, regressed on verbal ability.

		Rotation	Rotation / Visualisation	Visualisation
Rotation	*r*	1		
	N	*1414*		
Rotation / Visualisation	*r*	0.44	1	
	N	*1402*	*1419*	
Visualisation	*r*	0.44	0.51	1
	N	*1392*	*1407*	*1410*

Correlations (Pearson's *r*) between the three functional Bricks composite residuals after regression on verbal ability (Mill Hill scores). All correlations significant at $p < 0.0001$.

Table S10. Functional composite intercorrelations, regressed on non-verbal ability.

		Rotation	Rotation / Visualisation	Visualisation
Rotation	*r*	1		
	N	*1410*		
Rotation / Visualisation	*r*	0.35	1	
	N	*1398*	*1414*	
Visualisation	*r*	0.36	0.42	1
	N	*1388*	*1402*	*1405*

Correlations (Pearson's *r*) between the three functional Bricks composite residuals after regression on non-verbal ability (Raven's Matrices scores). All correlations significant at $p < 0.0001$.

Table S11. Functional composite intercorrelations, regressed on *g*.

		Rotation	Rotation / Visualisation	Visualisation
Rotation	*r*	1		
	N	*1412*		
Rotation / Visualisation	*r*	0.38	1	
	N	*1400*	*1416*	
Visualisation	*r*	0.38	0.44	1
	N	*1391*	*1405*	*1408*

Correlations (Pearson's *r*) between the three functional Bricks composite residuals after regression on *g* (the mean of verbal and non-verbal ability scores). All correlations significant at $p < 0.0001$.

Table S12. Dimensional composite correlation, regressed on other measures.

		Regressed variable		
		Mill Hill (verbal)	Raven's Matrices (non-verbal)	*g*
2D and 3D	*r*	0.54	0.44	0.47
	N	*1392*	*1388*	*1390*

Correlation between 2D and 3D dimensional Bricks composites, after regression on verbal ability, non-verbal ability or *g* (their mean). All correlations significant at $p < 0.0001$.

Table S13. Subtest factor analysis, regressed on other measures.

	Regressed variable		
	Mill Hill (verbal)	Raven's Matrices (non-verbal)	*g*
2D Rotation	0.57	0.51	0.54
2D Rotation / Visualisation	0.69	0.64	0.65
2D Visualisation	0.67	0.62	0.64
3D Rotation	0.64	0.61	0.62
3D Rotation / Visualisation	0.62	0.54	0.57
3D Visualisation	0.67	0.62	0.63
Variance explained	*41.3% (2.48 eigenvalues)*	*35.1% (2.10 eigenvalues)*	*37.0% (2.22 eigenvalues)*

Factor loadings of Bricks subtests on the first (and only) principal component produced by factor analysis of the six subtest scores, after regression on verbal ability, non-verbal ability or *g*.

Table S14. Twin correlations and approximated variance components.

	Intrapair twin correlations		Variance component estimates			Sample (numbers of pairs)	
	MZ	DZ	h^2	c^2	e^2	MZ	DZ
2D Rotation	0.21 (0.12 – 0.29)	0.11 (0.03 – 0.18)	0.20	0.01	0.79	528	722
2D Rotation / Visualisation	0.28 (0.20 – 0.36)	0.16 (0.09 – 0.23)	0.24	0.04	0.72	525	718
2D Visualisation	0.24 (0.16 – 0.32)	0.17 (0.10 – 0.24)	0.14	0.10	0.76	521	713
3D Rotation	0.21 (0.13 – 0.29)	0.13 (0.06 – 0.21)	0.16	0.05	0.79	502	684
3D Rotation / Visualisation	0.27 (0.19 – 0.35)	0.12 (0.05 – 0.20)	0.27	0.00	0.73	516	704
3D Visualisation	0.34 (0.26 – 0.41)	0.08 (0.01 – 0.15)	0.34	0.00	0.66	517	710
Verbal	0.48 (0.43 – 0.54)	0.27 (0.21 – 0.32)	0.43	0.05	0.52	729	1173
Non-verbal	0.51 (0.45 – 0.56)	0.32 (0.26 – 0.37)	0.39	0.12	0.49	700	1086
g	0.58 (0.52 – 0.62)	0.34 (0.29 – 0.39)	0.47	0.10	0.42	697	1084

Intraclass twin correlations (95% confidence intervals) for MZ and DZ twins, for the six Bricks subtests and for verbal ability (Mill Hill), non-verbal ability (Raven's Matrices) and *g* (their mean). For Bricks composites, see Table 1. Variance component estimates are heritability (h^2: double the difference between the MZ and DZ correlations, constrained not to exceed the former – MZ twins are genetically identical, so heritability cannot exceed their correlation), shared environment (c^2: the MZ correlation minus h^2), and unique environment + error of measurement (e^2: 1 - h^2 - c^2). Sample sizes shown are complete pairs, after exclusions and data cleaning.

Table S15. Univariate model-fitting results.

	A	C	E
2D Rotation	*0.19 (0.00 – 0.28)*	*0.01 (0.00 – 0.17)*	0.80 (0.72 – 0.88)
2D Rotation / Visualisation	0.23 (0.03 – 0.34)	*0.04 (0.00 – 0.20)*	0.73 (0.66 – 0.81)
2D Visualisation	*0.13 (0.00 – 0.31)*	*0.11 (0.00 – 0.24)*	0.76 (0.69 – 0.84)
3D Rotation	*0.15 (0.00 – 0.29)*	*0.06 (0.00 – 0.21)*	0.79 (0.71 – 0.87)
3D Rotation / Visualisation	0.26 (0.07 – 0.33)	*0.00 (0.00 – 0.14)*	0.74 (0.67 – 0.81)
3D Visualisation	0.30 (0.20 – 0.36)	*0.00 (0.00 – 0.07)*	0.70 (0.64 – 0.77)
Verbal	0.46 (0.32 – 0.55)	*0.04 (0.00 – 0.15)*	0.50 (0.45 – 0.55)
Non-verbal	0.40 (0.28 – 0.54)	0.12 (0.01 – 0.23)	0.48 (0.43 – 0.53)
g	0.49 (0.36 – 0.62)	*0.10 (0.00 – 0.21)*	0.41 (0.37 – 0.46)

Model-fitting estimates (95% confidence intervals) for additive genetic (A), shared environmental (C) and residual (E; i.e., non-shared environment and error) components of variance, for the six Bricks subtests and for verbal ability (Mill Hill), non-verbal ability (Raven's Matrices) and *g* (their mean). For Bricks composites, see Table 2. Italicised estimates are non-significant (their confidence intervals include zero).

Table S16. Decomposition of phenotypic correlations.

Variables in model	Variance component estimates		
	A	C	E
Rotation Rotation / Visualisation	0.80 (0.54 – 0.89)	*-0.02* *(-0.06 – 0.19)*	0.22 (0.13 – 0.32)
Rotation Visualisation	0.71 (0.43 – 0.85)	*-0.00* *(-0.09 – 0.22)*	0.29 (0.20 – 0.39)
Visualisation Rotation / Visualisation	0.74 (0.47 – 0.85)	*0.01* *(-0.06 – 0.22)*	0.26 (0.18 – 0.35)
2D 3D	0.79 (0.60 – 0.86)	*-0.01* *(-0.04 – 0.15)*	0.22 (0.15 – 0.30)

Bivariate correlated factors solutions of four models: three between the functional Bricks composites, and one between the dimensional composites. Results indicate the phenotypic correlations between the two composites in each model, decomposed into proportions attributable to additive genetic (A), shared environmental (C) or non-shared environmental/error (E) components (with 95% confidence intervals). The proportions explained reflect the correlation between the traits for that component, weighted by the two univariate component estimates – for example, the proportion of the phenotypic correlation due to A equals the genetic correlation weighted by the product of the square roots of the two univariate heritabilities estimated by the model. Italicised estimates are non-significant (their CIs include zero). Totals may exceed 1.00 due to rounding.

Table S17. Proportions of Bricks subtest correlations due to common genetic influences.

	2D R	2D R / V	2D V	3D R	3D R / V	3D V
2D Rotation	1					
2D Rotation / Visualisation	0.82 (0.48 – 0.97)	1				
2D Visualisation	0.73 (0.28 – 1.01)	0.47 (0.12 – 0.80)	1			
3D Rotation	0.60 (0.02 – 1.05)	0.69 (0.29 – 0.95)	*0.21* *(-0.13 – 0.66)*	1		
3D Rotation / Visualisation	0.91 (0.50 – 1.09)	0.79 (0.36 – 1.04)	0.75 (0.37 – 0.94)	*0.50* *(-0.02 – 0.89)*	1	
3D Visualisation	0.86 (0.52 – 1.09)	0.76 (0.50 – 0.90)	0.79 (0.46 – 0.96)	0.62 (0.29 – 0.82)	0.65 (0.30 – 0.82)	1

Bivariate correlated factors solutions, indicating the proportions of the phenotypic correlations between subtests due to common genetic influences (with 95% confidence intervals). Italicised estimates are non-significant (their CIs include zero). R = Rotation; R / V = Rotation and Visualisation; V = Visualisation.

N.B. The figures shown are proportions of the total covariance, so the two lower and non-significant estimates in this table reflect the correspondingly higher non-shared environment components (Table S17) for those associations (and the wide CIs), rather than a meaningful distinction from the other correlations.

Table S18. Proportions of Bricks subtest correlations due to common non-shared environmental influences.

	2D R	2D R / V	2D V	3D R	3D R / V	3D V
2D Rotation	1					
2D Rotation / Visualisation	0.18 (0.03 – 0.34)	1				
2D Visualisation	*0.18 (-0.00 – 0.36)*	0.31 (0.18 – 0.45)	1			
3D Rotation	*0.23 (-0.02 – 0.47)*	0.20 (0.05 – 0.36)	0.45 (0.28 – 0.63)	1		
3D Rotation / Visualisation	*0.10 (-0.09 – 0.29)*	*0.14 (-0.03 – 0.33)*	0.28 (0.12 – 0.45)	0.44 (0.23 – 0.66)	1	
3D Visualisation	*0.16 (-0.04 – 0.37)*	0.23 (0.10 – 0.36)	0.17 (0.04 – 0.30)	0.34 (0.19 – 0.49)	0.32 (0.18 – 0.48)	1

Bivariate correlated factors solutions, indicating the proportions of the phenotypic correlations between subtests due to common non-shared environmental influences (with 95% confidence intervals). Italicised estimates are non-significant (their CIs include zero). R = Rotation; R / V = Rotation and Visualisation; V = Visualisation.

Table S19. Proportions of correlations with other measures due to common genetic influences.

	Mill Hill (verbal)	Raven's Matrices (non-verbal)	*g*
Rotation	1.24 (0.50 – 1.90	0.56 (0.26 – 0.89)	0.69 (0.34 – 1.05)
Rotation / Visualisation	0.79 (0.35 – 1.23)	0.75 (0.50 – 1.00)	0.77 (0.51 – 1.03)
Visualisation	1.00 (0.49 – 1.40)	0.58 (0.33 – 0.82)	0.66 (0.40 – 0.92)
2D	1.08 (0.56 – 1.47)	0.74 (0.50 – 0.98)	0.80 (0.54 – 1.05)
3D	0.97 (0.54 – 1.32)	0.59 (0.36 – 0.81)	0.69 (0.45 – 0.92)
Overall Bricks	0.99 (0.64 – 1.29)	0.72 (0.52 – 0.90)	0.77 (0.57 – 0.95)

Bivariate correlated factors solutions, indicating the proportions of the phenotypic correlations (with 95% confidence intervals) between each Bricks composite and other cognitive measures which are attributable to common genetic influences.

N.B. the proportions above unity (with verbal ability) are offset by negative environmental contributions, but the wide CIs preclude any meaningful interpretations.

Table S20. Bivariate Cholesky decomposition: Rotation, Visualisation.

	Genetic paths		**Shared environment paths**		**Non-shared environment paths**	
	Rotation	Visualisation	Rotation	Visualisation	Rotation	Visualisation
1. Rotation	0.25 (0.11 – 0.38)		*0.09 (0.00 – 0.21)*		0.65 (0.57 – 0.73)	
2. Visualisation	0.44 (0.20 – 0.50)	*0.00 (0.00 – 0.18)*	*0.00 (0.00 – 0.09)*	*0.00 (0.00 – 0.11)*	0.03 (0.01 – 0.05)	0.53 (0.46 – 0.60)

Path estimates (standardised and squared, with 95% confidence intervals) for bivariate ACE Cholesky decomposition. The influences on the first entered variable (Rotation) are as in the univariate model for that variable (precise estimates vary between models), but those on the second (Visualisation) are decomposed into influences shared with the first variable, and those unique to the second. Italicised estimates are non-significant (their CIs include zero).

Table S21. Bivariate Cholesky decomposition: Rotation, Rotation/Visualisation.

	Genetic paths		**Shared environment paths**		**Non-shared environment paths**	
	Rotation	Rotation / Visualisation	Rotation	Rotation / Visualisation	Rotation	Rotation / Visualisation
1. Rotation	0.34 (0.19 – 0.42)		*0.03 (0.00 – 0.15)*		0.62 (0.55 – 0.70)	
2. Rotation / Visualisation	0.41 (0.25 – 0.47)	*0.00 (0.00 – 0.09)*	*0.00 (0.00 – 0.12)*	*0.00 (0.00 – 0.09)*	0.02 (0.01 – 0.03)	0.57 (0.50 – 0.64)

Path estimates (standardised and squared, with 95% confidence intervals) for bivariate ACE Cholesky decomposition. The influences on the first entered variable (Rotation) are as in the univariate model for that variable (precise estimates vary between models), but those on the second (Rotation / Visualisation combined) are decomposed into influences shared with the first variable, and those unique to the second. Italicised estimates are non-significant (their CIs include zero).

Table S22. Bivariate Cholesky decomposition: Visualisation, Rotation/Visualisation.

	Genetic paths		**Shared environment paths**		**Non-shared environment paths**	
	Visualisation	Rotation / Visualisation	Visualisation	Rotation / Visualisation	Visualisation	Rotation / Visualisation
1. Visualisation	0.44 (0.26 – 0.49)		*0.00 (0.00 – 0.14)*		0.56 (0.49 – 0.63)	
2. Rotation / Visualisation	0.32 (0.17 – 0.45)	*0.02 (0.00 – 0.13)*	*0.03 (0.00 – 0.18)*	*0.01 (0.00 – 0.10)*	0.03 (0.01 – 0.06)	0.58 (0.51 – 0.65)

Path estimates (standardised and squared, with 95% confidence intervals) for bivariate ACE Cholesky decomposition. The influences on the first entered variable (Visualisation) are as in the univariate model for that variable (precise estimates vary between models), but those on the second (Rotation / Visualisation combined) are decomposed into influences shared with the first variable, and those unique to the second. Italicised estimates are non-significant (their CIs include zero).

Table S23. Bivariate Cholesky decomposition: 2D, 3D.

	Genetic paths		**Shared environment paths**		**Non-shared environment paths**	
	2D	3D	2D	3D	2D	3D
1. 2D	0.45 (0.31 – 0.52)		*0.02 (0.00 – 0.14)*		0.53 (0.46 – 0.60)	
2. 3D	0.42 (0.28 – 0.48)	*0.00 (0.00 – 0.08)*	*0.00 (0.00 – 0.12)*	*0.00 (0.00 – 0.07)*	0.03 (0.01 – 0.05)	0.54 (0.47 – 0.61)

Path estimates (standardised and squared, with 95% confidence intervals) for bivariate ACE Cholesky decomposition. The influences on the first entered variable (2D) are as in the univariate model for that variable (precise estimates vary between models), but those on the second (3D) are decomposed into influences shared with the first variable, and those unique to the second. Italicised estimates are non-significant (their CIs include zero).

Table S24. Correlations between influences on functional composites.

	Genetic correlations (rA)		**Shared environment correlations (rC)**		**Non-shared environment correlations (rE)**	
	Rotation	Rotation / Visualisation	Rotation	Rotation / Visualisation	Rotation	Rotation / Visualisation
Rotation / Visualisation	1.00 (0.88 – 1.00)		*-1.00 (-1.00 – 1.00)*		0.17 (0.11 – 0.24)	
Visualisation	1.00 (0.74 – 1.00)	0.97 (0.95 – 1.00)	*-1.00 (-1.00 – 1.00)*	*0.86 (-1.00 – 1.00)*	0.23 (0.16 – 0.30)	0.23 (0.16 – 0.30)

Genetic, shared and non-shared environmental correlations (95% confidence intervals) between the functional Bricks composites. Italicised estimates are non-significant.

Table S25. Correlations between influences on dimensional composites.

	Genetic correlation (rA)	**Shared environment correlation (rC)**	**Non-shared environment correlation (rE)**
2D and 3D	1.00 (0.90 – 1.00)	*-1.00 (-1.00 – 1.00)*	0.22 (0.16 – 0.29)

Genetic, shared and non-shared environmental correlations (95% confidence intervals) between the dimensional Bricks composites. Italicised estimates are non-significant.

Table S26. Genetic correlations among Bricks subtests.

	2D R	2D R / V	2D V	3D R	3D R / V	3D V
2D Rotation	1					
2D Rotation / Visualisation	1.00 (0.83 – 1.00)	1				
2D Visualisation	1.00 (0.74 – 1.00)	1.00 (0.75 – 1.00)	1			
3D Rotation	0.86 (0.04 – 1.00)	1.00 (0.79 – 1.00)	0.61 (-0.97 – 1.00)	1		
3D Rotation / Visualisation	1.00 (0.77 – 1.00)	0.95 (0.94 – 1.00)	1.00 (0.69 – 1.00)	0.63 (-0.09 – 1.00)	1	
3D Visualisation	0.96 (0.95 – 1.00)	1.00 (1.00 – 1.00)	1.00 (0.91 – 1.00)	0.95 (0.76 – 1.00)	0.86 (0.65 – 1.00)	1

Genetic correlations (95% confidence intervals) among the individual Bricks subtests. R = Rotation; R / V = Rotation and Visualisation; V = Visualisation.

N.B. Two of these (3D Rotation's correlations with 2D Visualisation and with 3D Rotation/Visualisation) are technically non-significant, with CIs including zero; but given the high point estimates, and since this subtest's genetic correlations with other subtests have generally wider CIs than others, it seems likely that this reflects differences in the reliability of the subtests (or indeed chance differences) rather than a meaningful distinction from the other associations.

Table S27. Non-shared environmental correlations among Bricks subtests.

	2D R	2D R / V	2D V	3D R	3D R / V	3D V
2D Rotation	1					
2D Rotation / Visualisation	0.07 (0.01 – 0.14)	1				
2D Visualisation	*0.06* *(-0.00 – 0.13)*	0.15 (0.09 – 0.22)	1			
3D Rotation	*0.07* *(-0.01 – 0.14)*	0.09 (0.02 – 0.16)	0.19 (0.12 – 0.26)	1		
3D Rotation / Visualisation	*0.03* *(-0.03 – 0.10)*	*0.06* *(-0.01 – 0.13)*	0.11 (0.05 – 0.18)	0.15 (0.08 – 0.23)	1	
3D Visualisation	*0.06* *(-0.01 – 0.13)*	0.12 (0.05 – 0.19)	0.08 (0.02 – 0.15)	0.16 (0.09 – 0.23)	0.16 (0.09 – 0.23)	1

Genetic correlations (95% confidence intervals) among the individual Bricks subtests. R = Rotation; R / V = Rotation and Visualisation; V = Visualisation.

N.B. Most subtests have modest non-shared environmental influences in common. Some of these correlations are non- significant, but only barely (their 95% CIs are just below zero) and all the CIs overlap, so this is unlikely to reflect meaningful distinctions.

(The corresponding matrix for shared environment correlations is omitted, as there are no significant shared environmental influences on the Bricks measures).

Table S28. Genetic correlations with other measures.

	Mill Hill (verbal)	Raven's Matrices (non-verbal)	*g*
Rotation	0.61 (0.24 – 1.00)	0.62 (0.41 – 1.00)	0.60 (0.36 – 1.00)
Rotation / Visualisation	0.51 (0.27 – 0.87)	0.81 (0.67 – 1.00)	0.72 (0.58 – 1.00)
Visualisation	0.54 (0.29 – 0.87)	0.64 (0.46 – 0.84)	0.60 (0.44 – 0.81)
2D	0.53 (0.29 – 0.82)	0.74 (0.60 – 0.93)	0.66 (0.51 – 0.88)
3D	0.55 (0.35 – 0.87)	0.74 (0.60 – 0.91)	0.67 (0.54 – 0.85)
Overall Bricks	0.51 (0.35 – 0.74)	0.77 (0.66 – 0.92)	0.69 (0.57 – 0.80)

Genetic correlations (95% confidence intervals) with verbal ability, non-verbal ability and *g* (their mean).

Table S29. Trivariate Cholesky decomposition: verbal ability, Rotation, Visualisation.

	Genetic paths		
	Verbal	Rotation	Visualisation
1. Verbal	0.46 (0.32 – 0.55)		
2. Rotation	0.09 (0.01 – 0.21)	0.18 (0.03 – 0.35)	
3. Visualisation	0.13 (0.04 – 0.27)	0.30 (0.06 – 0.41)	*0.00 (0.00 – 0.00)*

Genetic path estimates (standardised and squared, with 95% confidence intervals) for trivariate ACE Cholesky decomposition. The last row indicates the genetic influences on Visualisation i) shared both with verbal ability (Mill Hill) and with Rotation, ii) shared only with Rotation, and iii) unique to Visualisation. The italicised estimate is non-significant (its CI includes zero).

Table S30. Trivariate Cholesky decomposition: non-verbal ability, Rotation, Visualisation.

	Genetic paths		
	Non-verbal	Rotation	Visualisation
1. Non-verbal	0.41 (0.27 – 0.54)		
2. Rotation	0.12 (0.03 – 0.26)	0.12 (0.02 – 0.23)	
3. Visualisation	0.18 (0.07 – 0.32)	0.21 (0.05 – 0.28)	*0.00 (0.00 – 0.17)*

Genetic path estimates (standardised and squared, with 95% confidence intervals) for trivariate ACE Cholesky decomposition. The last row indicates the genetic influences on Visualisation i) shared both with non-verbal ability (Raven's Matrices) and with Rotation, ii) shared only with Rotation, and iii) unique to Visualisation. The italicised estimate is non-significant (its CI includes zero).

Table S31. Trivariate Cholesky decomposition: *g*, Rotation, Visualisation.

	Genetic paths		
	g	Rotation	Visualisation
1. *g*	0.50 (0.37 – 0.62)		
2. Rotation	0.10 (0.03 – 0.23)	0.17 (0.04 – 0.28)	
3. Visualisation	0.16 (0.07 – 0.29)	0.24 (0.07 – 0.31)	*0.00 (0.00 – 0.16)*

Genetic path estimates (standardised and squared, with 95% confidence intervals) for trivariate ACE Cholesky decomposition. The last row indicates the genetic influences on Visualisation i) shared both with *g* (the mean of verbal and non-verbal ability) and with Rotation, ii) shared only with Rotation, and iii) unique to Visualisation. The italicised estimate is non-significant (its CI includes zero).

Table S32. Trivariate Cholesky decomposition: verbal ability, 2D, 3D.

	Genetic paths		
	Verbal	2D	3D
1. Verbal	0.47 (0.33 – 0.55)		
2. 2D	0.13 (0.04 – 0.26)	0.32 (0.14 – 0.44)	
3. 3D	0.12 (0.05 – 0.25)	0.31 (0.13 – 0.39)	*0.00 (0.00 – 0.08)*

Genetic path estimates (standardised and squared, with 95% confidence intervals) for trivariate ACE Cholesky decomposition. The last row indicates the genetic influences on 3D i) shared both with verbal ability (Mill Hill) and with 2D, ii) shared only with 2D, and iii) unique to 3D. The italicised estimate is non-significant (its CI includes zero).

Table S33. Trivariate Cholesky decomposition: non-verbal ability, 2D, 3D.

	Genetic paths		
	Non-verbal	2D	3D
1. Non-verbal	0.41 (0.28 – 0.54)		
2. 2D	0.27 (0.14 – 0.42)	0.17 (0.06 – 0.25)	
3. 3D	0.20 (0.11 – 0.33)	0.16 (0.08 – 0.21)	*0.00 (0.00 – 0.07)*

Genetic path estimates (standardised and squared, with 95% confidence intervals) for trivariate ACE Cholesky decomposition. The last row indicates the genetic influences on 3D i) shared both with non-verbal ability (Raven's Matrices) and with 2D, ii) shared only with 2D, and iii) unique to 3D. The italicised estimate is non-significant (its CI includes zero).

Table S34. Trivariate Cholesky decomposition: *g*, 2D, 3D.

	Genetic paths		
	g	2D	3D
1. *g*	0.50 (0.37 – 0.62)		
2. 2D	0.21 (0.11 – 0.34)	0.24 (0.11 – 0.32)	
3. 3D	0.19 (0.10 – 0.31)	0.21 (0.11 – 0.27)	*0.00 (0.00 – 0.08)*

Genetic path estimates (standardised and squared, with 95% confidence intervals) for trivariate ACE Cholesky decomposition. The last row indicates the genetic influences on 3D i) shared both with *g* (the mean of verbal and non-verbal ability) and with 2D, ii) shared only with 2D, and iii) unique to 3D. The italicised estimate is non-significant (its CI includes zero).

Table S35. Quadrivariate Cholesky decomposition: verbal, non-verbal, Rotation, Visualisation.

	Genetic paths			
	Verbal	Non-verbal	Rotation	Visualisation
1. Verbal	0.45 (0.31 – 0.55)			
2. Non-verbal	0.18 (0.08 – 0.32)	0.22 (0.06 – 0.34)		
3. Rotation	0.08 (0.02 – 0.20)	*0.05 (0.00 – 0.18)*	0.11 (0.02 – 0.22)	
4. Visualisation	0.11 (0.04 – 0.24)	*0.07 (0.00 – 0.20)*	0.20 (0.03 – 0.26)	*0.00 (0.00 – 0.16)*

Genetic path estimates (standardised and squared, with 95% confidence intervals) for quadrivariate ACE Cholesky decomposition. The last row indicates the genetic influences on Visualisation i) shared with verbal ability (Mill Hill), non-verbal ability (Raven's Matrices) and Rotation, ii) shared only with non-verbal ability and Rotation (but not verbal ability), iii) shared only with Rotation, and iv) unique to Visualisation. The italicised estimates are non-significant (their CIs include zero).

Table S36. Quadrivariate Cholesky decomposition: verbal, non-verbal, 2D, 3D.

	Genetic paths			
	Verbal	Non-verbal	2D	3D
1. Verbal	0.46 (0.32 – 0.55)			
2. Non-verbal	0.18 (0.09 – 0.30)	0.23 (0.07 – 0.36)		
3. 2D	0.13 (0.06 – 0.25)	0.14 (0.04 – 0.32)	0.17 (0.05 – 0.25)	
4. 3D	0.11 (0.05 – 0.16)	0.10 (0.02 – 0.25)	0.16 (0.05 – 0.21)	*0.00 (0.00 – 0.07)*

Genetic path estimates (standardised and squared, with 95% confidence intervals) for quadrivariate ACE Cholesky decomposition. The last row indicates the genetic influences on 3D i) shared with verbal ability (Mill Hill), non-verbal ability (Raven's Matrices) and 2D, ii) shared only with non-verbal ability and 2D (but not verbal ability), iii) shared only with 2D, and iv) unique to 3D. The italicised estimate is non-significant (its CI includes zero).

Table S37. Fit statistics: univariate Bricks composite models.

	Model	ep	χ2	df	AIC	Δχ2	Δdf	p
Rotation	Saturated	10	8106.82	2881	2344.82	-	-	-
	ACE	4	8108.17	2887	2334.17	1.35	6	0.97
Rotation / Visualisation	Saturated	10	8079.54	2880	2319.54	-	-	-
	ACE	4	8081.90	2886	2309.90	2.37	6	0.88
Visualisation	Saturated	10	7988.85	2860	2268.85	-	-	-
	ACE	4	7992.22	2866	2260.22	3.36	6	0.76
2D	Saturated	10	8084.56	2902	2280.56	-	-	-
	ACE	4	8086.42	2908	2270.42	1.86	6	0.93
3D	Saturated	10	7936.63	2831	2274.63	-	-	-
	ACE	4	7939.27	2837	2265.27	2.64	6	0.85
Overall Bricks	Saturated	10	7976.85	2889	2198.85	-	-	-
	ACE	4	7979.72	2895	2189.72	2.87	6	0.82

Comparison of univariate ACE models to fully saturated models. ep = estimated parameters; χ2 = -2 log-likelihood; df = degrees of freedom, AIC = Akaike information criterion. The p-values indicate no significant deterioration in fit between the saturated and constrained models (i.e., the ACE models fit well).

Table S38. Fit statistics: bivariate Bricks composite models.

	Model	Ep	χ2	df	AIC	Δχ2	Δdf	p
R, V	Saturated	28	15433.66	5733	3967.66	-	-	-
	ACE	11	15442.00	5750	3942.01	8.34	17	0.96
R, R / V	Saturated	28	15483.91	5753	3977.91	-	-	-
	ACE	11	15499.11	5770	3959.11	15.20	17	0.58
V, R / V	Saturated	28	15233.62	5732	3769.62	-	-	-
	ACE	11	15246.77	5749	3748.77	13.14	17	0.73
2D, 3D	Saturated	28	15041.06	5725	3591.06	-	-	-
	ACE	11	15049.10	5742	3565.10	8.04	17	0.97

Comparison of bivariate ACE models to fully saturated models. Variables were entered in the order specified. R = Rotation; R / V = Rotation and Visualisation; V = Visualisation; ep = estimated parameters; χ2 = -2 log-likelihood; df = degrees of freedom, AIC = Akaike information criterion. The p-values indicate no significant deterioration in fit between the saturated and constrained models (i.e., the ACE models fit well).

Table S39. Fit statistics: trivariate Bricks composite models.

	Model	ep	χ2	df	AIC	Δχ2	Δdf	p
Verbal, R, V	Saturated	54	26544.39	9792	6960.39	-	-	-
	ACE	21	26583.78	9825	6933.78	39.40	33	0.21
Non-verbal, R, V	Saturated	54	25357.27	9549	6259.27	-	-	-
	ACE	21	25392.17	9582	6228.17	34.91	33	0.38
g, R, V	Saturated	54	25353.00	9542	6269.00	-	-	-
	ACE	21	25394.66	9575	6244.67	41.67	33	0.14
Verbal, 2D, 3D	Saturated	54	26118.06	9784	6550.06	-	-	-
	ACE	21	26151.77	9817	6517.77	33.71	33	0.43
Non-verbal, 2D, 3D	Saturated	54	24852.31	9541	5770.31	-	-	-
	ACE	21	24879.46	9574	5731.46	27.15	33	0.75
g, 2D, 3D	Saturated	54	24869.37	9534	5801.37	-	-	-
	ACE	21	24897.55	9567	5763.55	28.18	33	0.71

Comparison of trivariate ACE models to fully saturated models. Variables were entered in the order specified. R = Rotation; R / V = Rotation and Visualisation; V = Visualisation; ep = estimated parameters; χ2 = -2 log-likelihood; df = degrees of freedom, AIC = Akaike information criterion. The p-values indicate no significant deterioration in fit between the saturated and constrained models (i.e., the ACE models fit well).

Table S40. Fit statistics: quadrivariate Bricks composite models.

	Model	ep	χ2	df	AIC	Δχ2	Δdf	p
Verbal, non-verbal, R, V	Saturated	88	36165.57	13600	8965.57	-	-	-
	ACE	34	36237.93	13654	8929.93	72.37	54	0.048
Verbal, non-verbal, 2D, 3D	Saturated	88	35650.66	13592	8466.66	-	-	-
	ACE	34	35714.20	13646	8422.21	63.54	54	0.18

Comparison of quadrivariate ACE models to fully saturated models. Variables were entered in the order specified. R = Rotation; R / V = Rotation and Visualisation; V = Visualisation; ep = estimated parameters; χ2 = -2 log-likelihood; df = degrees of freedom, AIC = Akaike information criterion.

N.B. The p-value for the first of these models indicates a significant deterioration in fit between the saturated and constrained model (albeit barely). This may be a chance effect, given the large number of models tested, or the sample size may be underpowered for these larger, more complex models.

Appendix 3

Supplementary Information

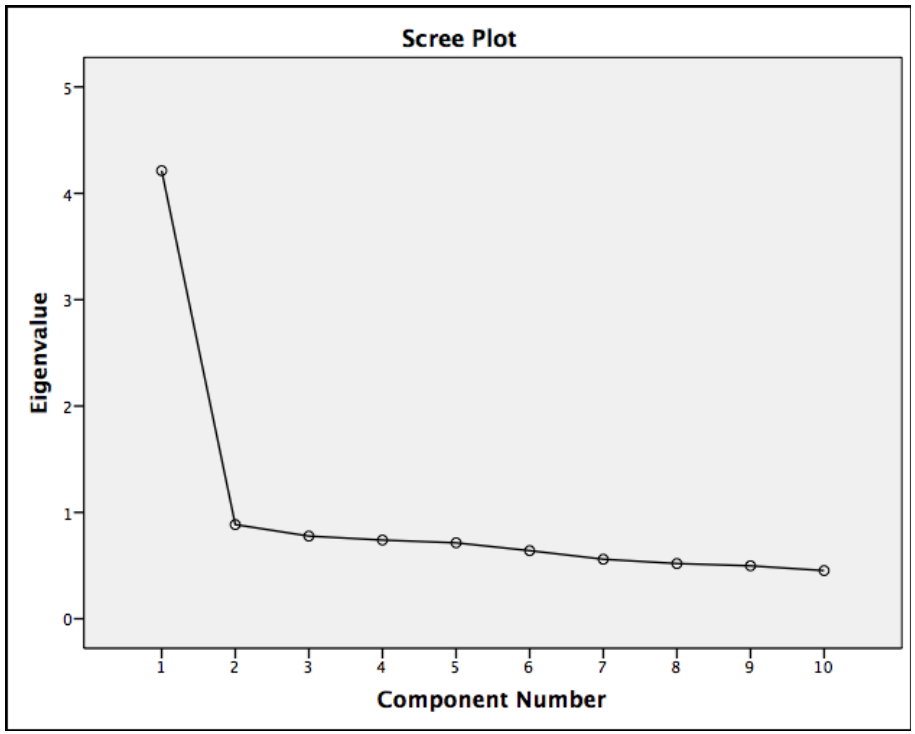
Spatial ability or spatial abilities? Investigating the phenotypic and genetic structure of spatial ability

Kaili Rimfeld, Nicholas G. Shakeshaft, Margherita Malanchini, Maja Rodic, Saskia Selzam, Kerry Schofield, Philip S. Dale, Yulia Kovas & Robert Plomin

Figures

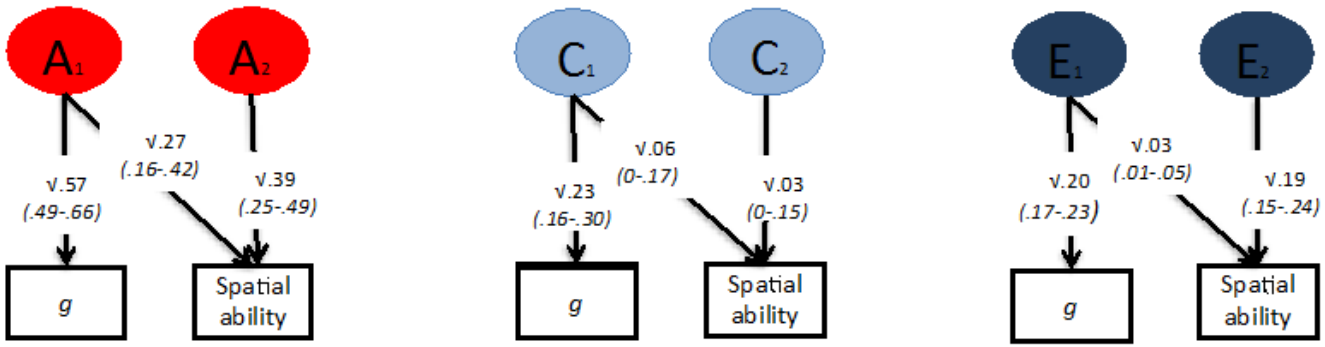
Tables

Figure S1. Scree plot illustrating the proportion of variance explained by the extracted factors from the ten spatial tests.



The scree plot taken from the principal components analysis (PCA) illustrating the factor structure of spatial ability. Only a single factor emerges with an eigenvalue above 1, indicating that spatial ability is unifactorial.

Figure S2. Bivariate Cholesky decomposition for intelligence (ages 7-16) and spatial ability (95% confidence intervals in parentheses).



Standardized path estimates for genetic (A), shared environmental (C) and non-shared environmental influences (E) on spatial ability, shared with and independent from *g* (intelligence).

Table S1. Mean scores (standard deviations) for ten spatial tests. N=sample size after exclusions (one randomly selected twin per pair); MZ=monozygotic; DZ=dizygotic; m=male; f=female; os=opposite sex. ANOVA analyses tested the effect of sex and zygosity: results = F statistic; ** = p<.01; R^2 = proportion of variance explained by sex, zygosity and their interaction.

Subject	N	Whole Sample	Male	Female	MZm	DZm	MZf	DZf	Dzos	Sex	Zyg	Sex x Zyg	R^2
Mazes	1213	5.77 (1.85)	6.23 (1.80)	5.49 (1.82)	6.39 (1.79)	6.19 (1.88)	5.42 (1.86)	5.50 (1.78)	5.83 (1.80)	40.10**	0.10	0.28	0.04
2D drawing	1345	3.58 (1.07)	3.90 (0.92)	3.40 (1.10)	3.98 (0.91)	3.95 (0.90)	3.33 (1.12)	3.47 (1.14)	3.60 (1.0)	56.49**	0.09	1.91	0.06
Pattern assembly	1300	6.62 (3.37)	7.43 (3.36)	6.15 (3.28)	7.33 (3.18)	7.33 (3.57)	5.91 (3.19)	6.14 (3.40)	7.03 (3.33)	25.12**	2.45	0.01	0.03
Elithorn maze	1160	7.68 (1.45)	8.26 (1.19)	7.32 (1.48)	8.34 (0.95)	8.21 (1.43)	7.19 (1.46)	7.30 (1.59)	7.87 (1.31)	108.07**	0.76	1.99	0.11
Mechanical reasoning	1314	9.28 (2.53)	10.28 (2.46)	8.67 (2.37)	10.38 (2.55)	10.43 (2.51)	8.66 (2.33)	8.57 (2.44)	9.38 (2.41)	123.71**	0.62	0.32	0.13
Paper folding	1262	8.02 (3.81)	8.70 (3.77)	7.63 (3.77)	8.64 (3.98)	8.86 (3.70)	7.37 (3.65)	7.69 (3.83)	8.25 (3.78)	20.99**	2.26	0.78	0.02
3D drawing	1211	2.95 (1.80)	3.52 (1.79)	2.62 (1.71)	3.70 (1.73)	3.51 (1.82)	2.57 (1.68)	2.64 (1.81)	2.97 (1.76)	66.49**	0.74	2.40	0.07
Mental rotation	1202	8.20 (4.03)	9.19 (3.77)	7.62 (4.06)	9.24 (3.54)	9.40 (4.10)	7.29 (4.06)	7.60 (4.41)	8.54 (3.84)	30.36**	2.10	0.01	0.04
Perspective taking	1222	4.41 (3.84)	5.87 (4.22)	3.54 (3.31)	6.11 (4.15)	5.86 (4.38)	3.56 (3.23)	3.30 (3.28)	4.61 (3.90)	81.66**	0.09	0.48	0.09
Cross section	1367	6.49 (3.58)	7.47 (3.61)	5.92 (3.44)	7.61 (3.68)	7.68 (3.46)	5.50 (3.59)	6.27 (3.23)	6.60 (3.50)	50.45**	1.80	2.94	0.05

Mean scores (standard deviations) from five sex and zygosity groups. ANOVA results indicate that sex and zygosity together explain between 2% and 13% of the variance in each spatial test.

Table S2. a) Correlation matrix and b) residual correlation matrix for ten spatial tests.

a) Correlation matrix

Correlations	Cross sections	2D drawing	Pattern assembly	Elithorn maze	Mechanical reasoning	Paper folding	3D drawing	Mental rotation	Perspective taking	Mazes
Cross sections	1									
2D drawing	.449**	1								
Pattern assembly	.396**	.494**	1							
Elithorn maze	.264**	.364**	.318**	1						
Mechanical reasoning	.445**	.428**	.379**	.280**	1					
Paper folding	.462**	.529**	.499**	.321**	.468**	1				
3D drawing	.448**	.571**	.445**	.360**	.417**	.526**	1			
Mental rotation	.398**	.482**	.514**	.395**	.410**	.510**	.483**	1		
Perspective taking	.335**	.347**	.289**	.210**	.315**	.328**	.367**	.343**	1	
Mazes	.288**	.334**	.345**	.266**	.321**	.371**	.365**	.377**	.232**	1

b) Reproduced and residual correlation matrices

Reproduced Correlations	Cross sections	2D drawing	Pattern assembly	Elithorn maze	Mechanical reasoning	Paper folding	3D drawing	Mental rotation	Perspective taking	Mazes
Cross sections	.428a									
2D drawing	0.472	.521a								
Pattern assembly	0.436	0.481	.444a							
Elithorn maze	0.366	0.404	0.373	.313a						
Mechanical reasoning	0.41	0.453	0.418	0.351	.393a					
Paper folding	0.477	0.527	0.486	0.408	0.457	.532a				
3D drawing	0.484	0.535	0.494	0.414	0.464	0.54	.548a			
Mental rotation	0.461	0.509	0.47	0.394	0.442	0.514	0.522	.498a		
Perspective taking	0.347	0.383	0.353	0.296	0.332	0.387	0.392	0.374	.281a	
Mazes	0.33	0.364	0.337	0.282	0.316	0.368	0.374	0.356	0.268	.255a
Residuals										
Cross sections										
2D drawing	-0.061									
Pattern assembly	-0.085	-0.05								
Elithorn maze	-0.117	-0.06	-0.072							
Mechanical reasoning	0.007	-0.107	-0.094	-0.067						
Paper folding	-0.037	-0.051	-0.044	-0.096	-0.031					
3D drawing	-0.068	0.001	-0.084	-0.051	-0.1	-0.047				
Mental rotation	-0.094	-0.085	-0.014	-0.001	-0.086	-0.083	-0.074			
Perspective taking	-0.03	-0.062	-0.098	-0.101	-0.031	-0.091	-0.053	-0.053		
Mazes	-0.106	-0.085	-0.046	-0.024	-0.065	-0.07	-0.058	-0.047	-0.063	

Note: ** Correlation is significant at the 0.01 level; a- reproduced communalities. Reproduced and residual correlation matrices extracted from principal components analyses. Reproduced correlations are based on the extracted factors; they are similar to the original correlations, indicating that the factor extracted accounts for a large proportion of covariance in these tests. The residual correlations are calculated as the difference between the original and reproduced correlations. The residuals are small in magnitude, confirming that the first principal component accounts for almost all covariance. There are many non-redundant residuals with values greater than 0.05 (presumably reflecting *g*, which was not controlled in this analysis.

Table S3. Sex-limitation model-fitting sub-model comparisons. FullHetACE=full genetic heterogeneity model, rG=Free; HetACE= quantitative heterogeneity model; cFullHetACE=full environmental heterogeneity model, rC=Free; HomACE= homogeneity model (no sex differences at all); ep=estimated parameters; minus2LL= minus 2 log-likelihood; df= degrees of freedom; AIC= Akaike information criterion; diffLL= change in log-likelihood; diffdf= change in degrees of freedom (significant differences are marked in bold).

Spatial ability							
Qualitative genetic differences:	ep	-2LL	df	AIC	diffLL	diffdf	p
FullHetACE	9	4798.23	1785	1228.23	-	-	-
HetACE	8	4798.37	1786	1226.37	0.13	1	0.71
Qualitative environmental differences:	ep	-2LL	df	AIC	diffLL	diffdf	p
cFullHetACE	9	4798.23	1785	1228.23	-	-	-
HetACE	8	4798.37	1786	1226.37	0.13	1	0.71
Quantitative differences:	ep	-2LL	df	AIC	diffLL	diffdf	p
HetACE	8	4798.37	1786	1226.37	-	-	-
HomACE	5	4800.75	1789	1222.75	2.38	3	0.50

Mazes							
Qualitative genetic differences:	ep	-2LL	df	AIC	diffLL	diffdf	p
FullHetACE	9	6739.17	2402	1935.17	-	-	-
HetACE	8	6739.3	2403	1933.3	0.13	1	0.72
Qualitative environmental differences:	ep	-2LL	df	AIC	diffLL	diffdf	p
cFullHetACE	9	6739.19	2402	1935.19	-	-	-
HetACE	8	6739.3	2403	1933.3	0.11	1	0.74
Quantitative differences:	ep	-2LL	df	AIC	diffLL	diffdf	p
HetACE	8	6739.3	2403	1933.3	-	-	-
HomACE	5	6744.91	2406	1932.91	5.61	3	0.13

2D Drawing							
Qualitative genetic differences:	ep	-2LL	df	AIC	diffLL	diffdf	p
FullHetACE	9	7396.63	2674	2048.63	-	-	-
HetACE	8	7396.88	2675	2046.88	0.25	1	0.62
Qualitative environmental differences:	ep	-2LL	df	AIC	diffLL	diffdf	p
cFullHetACE	9	7396.63	2674	2048.63	-	-	-
HetACE	8	7396.88	2675	2046.88	0.25	1	0.62
Quantitative differences:	ep	-2LL	df	AIC	diffLL	diffdf	p
HetACE	8	7396.88	2675	2046.88	-	-	-
HomACE	5	7444.77	2678	2088.77	47.89	3	**0**

Pattern assembly							
Qualitative genetic differences:	ep	-2LL	df	AIC	diffLL	diffdf	p
FullHetACE	9	7199.89	2573	2053.89	-	-	-
HetACE	8	7199.89	2574	2051.89	0	1	1
Qualitative environmental differences:	ep	-2LL	df	AIC	diffLL	diffdf	p
cFullHetACE	9	7199.89	2573	2053.89	-	-	-
HetACE	8	7199.89	2574	2051.89	0	1	1
Quantitative differences:	ep	-2LL	df	AIC	diffLL	diffdf	p
HetACE	8	7199.89	2574	2051.89	-	-	-
HomACE	5	7205.97	2577	2051.97	6.08	3	0.11

Elithorn maze							
Qualitative genetic differences:	ep	-2LL	df	AIC	diffLL	diffdf	p
FullHetACE	9	6427.18	2293	1841.18	-	-	-
HetACE	8	6427.36	2294	1839.36	0.19	1	0.67
Qualitative environmental differences:	ep	-2LL	df	AIC	diffLL	diffdf	p
cFullHetACE	9	6428.67	2293	1842.67	-	-	-
HetACE	8	6427.36	2294	1839.36	-1.31	1	1
Quantitative differences:	ep	-2LL	df	AIC	diffLL	diffdf	p
HetACE	8	6427.36	2294	1839.36	-	-	-
HomACE	5	6440.54	2297	1846.54	13.17	3	**0**

Mechanical reasoning							
Qualitative genetic differences:	ep	-2LL	df	AIC	diffLL	diffdf	p
FullHetACE	9	7229.19	2602	2025.19	-	-	-
HetACE	8	7229.19	2603	2023.19	0	1	1
Qualitative environmental differences:	ep	-2LL	df	AIC	diffLL	diffdf	p
cFullHetACE	9	7229.19	2602	2025.19	-	-	-
HetACE	8	7229.19	2603	2023.19	0	1	1
Quantitative differences:	ep	-2LL	df	AIC	diffLL	diffdf	p
HetACE	8	7229.19	2603	2023.19	-	-	-
HomACE	5	7237.08	2606	2025.08	7.9	3	**0.05**

Paper folding							
Qualitative genetic differences:	ep	-2LL	df	AIC	diffLL	diffdf	p
FullHetACE	9	6949.91	2511	1927.91	-	-	-
HetACE	8	6949.91	2512	1925.91	0	1	1
Qualitative environmental differences:	ep	-2LL	df	AIC	diffLL	diffdf	p
cFullHetACE	9	6949.91	2511	1927.91	-	-	-
HetACE	8	6949.91	2512	1925.91	0	1	1
Quantitative differences:	ep	-2LL	df	AIC	diffLL	diffdf	p
HetACE	8	6949.91	2512	1925.91	-	-	-
HomACE	5	6958.43	2515	1928.43	8.52	3	**0.04**

3D drawing							
Qualitative genetic differences:	ep	-2LL	df	AIC	diffLL	diffdf	p
FullHetACE	9	6583.55	2389	1805.55	-	-	-
HetACE	8	6583.55	2390	1803.55	0	1	1
Qualitative environmental differences:	ep	-2LL	df	AIC	diffLL	diffdf	p
cFullHetACE	9	6583.55	2389	1805.55	-	-	-
HetACE	8	6583.55	2390	1803.55	0	1	1
Quantitative differences:	ep	-2LL	df	AIC	diffLL	diffdf	p
HetACE	8	6583.55	2390	1803.55	-	-	-
HomACE	5	6586.89	2393	1800.89	3.34	3	0.34

Mental rotation							
Qualitative genetic differences:	ep	-2LL	df	AIC	diffLL	diffdf	p
FullHetACE	9	6715.01	2409	1897.01	-	-	-
HetACE	8	6715.01	2410	1895.01	0	1	1
Qualitative environmental differences:	ep	-2LL	df	AIC	diffLL	diffdf	p
cFullHetACE	9	6715.01	2409	1897.01	-	-	-
HetACE	8	6715.01	2410	1895.01	0	1	1
Quantitative differences:	ep	-2LL	df	AIC	diffLL	diffdf	p
HetACE	8	6715.01	2410	1895.01	-	-	-
HomACE	5	6716.3	2413	1890.3	1.29	3	0.73

Perspective taking							
Qualitative genetic differences:	ep	-2LL	df	AIC	diffLL	diffdf	p
FullHetACE	9	6726.35	2426	1874.35	-	-	-
HetACE	8	6726.35	2427	1872.35	0	1	0.99
Qualitative environmental differences:	ep	-2LL	df	AIC	diffLL	diffdf	p
cFullHetACE	9	6726.35	2426	1874.35	-	-	-
HetACE	8	6726.35	2427	1872.35	0	1	1
Quantitative differences:	ep	-2LL	df	AIC	diffLL	diffdf	p
HetACE	8	6726.35	2427	1872.35	-	-	-
HomACE	5	6826.25	2430	1966.25	99.9	3	**0**

Cross-sections							
Qualitative genetic differences:	ep	-2LL	df	AIC	diffLL	diffdf	p
FullHetACE	9	7524.12	2699	2126.12	-	-	-
HetACE	8	7524.12	2700	2124.12	0	1	1
Qualitative environmental differences:	ep	-2LL	df	AIC	diffLL	diffdf	p
cFullHetACE	9	7524.12	2699	2126.12	-	-	-
HetACE	8	7524.12	2700	2124.12	0	1	1
Quantitative differences:	ep	-2LL	df	AIC	diffLL	diffdf	p
HetACE	8	7524.12	2700	2124.12	-	-	-
HomACE	5	7528.59	2703	2122.59	4.47	3	0.21

Full sex limitation model results show that there were no significant qualitative sex differences in any of the spatial tests (i.e., no different genetic or environmental factors affecting males and females), but there were some significant quantitative sex differences (differences in the magnitude of ACE estimates for males and females). Significant results are indicated in bold. As noted in the text, little confidence can be placed in these differences, as the sex-limitation models are underpowered to detect differences of this small magnitude; nonetheless, separate ACE estimates for males and females are presented for reference in Table S4.

Table S4. Sex-limitation model-fitting results, showing A, C, E, estimates separately for males and females. A=additive genetic; C=shared environmental; E=non-shared environmental proportions of the variance (95% confidence intervals).

Overall spatial ability	**A**	**C**	**E**
Males	0.68 (0.32; 0.85)	0.12 (0; 0.45)	0.20 (0.15; 0.28)
Females	0.65 (0.39; 0.80)	0.09 (0; 0.33)	0.25 (0.20; 0.32)

Mazes	**A**	**C**	**E**
Males	0.39 (0.17; 0.58)	0.08 (0; 0.23)	0.53 (0.42; 0.67)
Females	0.16 (0; 0.41)	0.15 (0; 0.35)	0.69 (0.59; 0.80)

2D drawing	**A**	**C**	**E**
Males	0.40 (0.09; 0.65)	0.17 (0.00; 0.42)	0.43 (0.34; 0.55)
Females	0.29 (0.02; 0.50)	0.12 (0.00; 0.35)	0.58 (0.50; 0.68)

Pattern assembly	**A**	**C**	**E**
Males	0.35 (0.08; 0.57)	0.10 (0; 0.28)	0.55 (0.43; 0.70)
Females	0.40 (0.15; 0.49)	0.00 (0; 0.20)	0.60 (0.51; 0.70)

Elithorn maze	**A**	**C**	**E**
Males	0.38 (0.12; 0.54)	0.03 (0.00; 0.19)	0.59 (0.46; 0.77)
Females	0.34 (0.03; 0.50)	0.06 (0.00; 0.32)	0.59 (0.50; 0.71)

Mechanical reasoning	**A**	**C**	**E**
Males	0.03 (0.00; 0.35)	0.45 (015; 0.56)	0.52 (0.43; 0.61)
Females	0.41 (0.20; 0.52)	0.05 (0.00; 0.22)	0.54 (0.46; 0.64)

Paper folding	**A**	**C**	**E**
Males	0.04 (0; 0.41)	0.48 (0.13; 0.60)	0.48 (0.39; 0.59)
Females	0.53 (0.35; 0.62)	0.02 (0.00; 0.17)	0.45 (0.38; 0.53)

3D drawing	**A**	**C**	**E**
Males	0.40 (0.06; 0.68)	0.20 (0; 0.50)	0.40 (0.31; 0.51)
Females	0.58 (0.41; 0.65)	0.00 (0; 0.14)	0.42 (0.35; 0.50)

Mental rotation	**A**	**C**	**E**
Males	0.25 (0; 0.57)	0.21 (0; 0.22)	0.54 (0.42; 0.69)
Females	0.39 (0.09; 0.53)	0.07 (0; 0.32)	0.54 (0.46; 0.64)

Perspective taking	**A**	**C**	**E**
Males	0.28 (0; 0.42)	0.00 (0.00; 0.31)	0.72 (0.58; 0.88)
Females	0.32 (0; 0.46)	0.04 (0.00; 0.33)	0.64 (0.54; 0.76)

Cross-sections	**A**	**C**	**E**
Males	0.01 (0.00; 0.38)	0.40 (0.08; 0.50)	0.48 (0.48; 0.69)
Females	0.21 (0.00; 0.41)	0.17 (0.03; 0.37)	0.61 (0.53; 0.71)

A few quantitative sex differences emerged for individual spatial ability tests (Table S3); however, the differences were small when examining the ACE estimates for males and females separately. Even with over 1300 twin pairs, the sample size is not sufficiently large for sex-limitation models to reliably detect quantitative and qualitative sex differences of this small magnitude, so little confidence can be placed in these differences, as is evident from the large confidence intervals around the estimates when calculated for males and females separately.

Table S5. Model-fitting results for univariate analyses of spatial ability tests, with twin intraclass correlations (N=complete twin pairs). A=additive genetic; C=shared environmental; E=non-shared environmental proportions of the variance (95% confidence intervals).

				Twin intraclass correlations	
	A	**C**	**E**	**MZ**	**DZ**
Spatial ability	0.69 (0.50; 0.80)	0.08 (0.00; 0.25)	0.23 (0.20; 0.29)	0.77 (0.71; 0.82) (N=229)	0.41 (0.31; 0.50) (N=305)
Mazes	0.35 (0.11; 0.44)	0.01 (0.00; 0.20)	0.64 (0.56; 0.73)	0.36 (0.27; 0.44) (N=384)	0.17 (0.08; 0.26) (N=494)
2D drawing	0.33 (0.13; 0.51)	0.12 (0.00; 0.28)	0.55 (0.48; 0.62)	0.45 (0.38; 0.53) (N=432)	0.26 (0.18; 0.34) (N=574)
Pattern assembly	0.42 (0.22; 0.49)	0.00 (0.00; 0.15)	0.58 (0.51; 0.66)	0.40 (0.32; 0.48) N=412	0.20 (0.12; 0.28) (N=540)
Elithorn maze	0.39 (0.21; 0.47)	0.00 (0.00; 0.13)	0.61 (0.53; 0.70)	0.39 (0.29; 0.47) (N=342)	0.16 (0.07; 0.25) (N=456)
Mechanical reasoning	0.41 (0.20; 0.53)	0.06 (0.00; 0.22)	0.53 (0.47; 0.61)	0.48 (0.40; 0.55) (N=427)	0.26 (0.18; 0.33) (N=557)
Paper folding	0.53 (0.38; 0.59)	0.00 (0.00; 0.12)	0.47 (0.41; 0.53)	0.54 (0.46; 0.60) (N=396)	0.24 (0.16; 0.32) (N=525)
3D drawing	0.59 (0.43; ;0.65)	0.00 (0.00; 0.13)	0.42 (0.35; 0.47)	0.57 (0.50; 0.64) (N=385)	0.28 (0.19; 0.36) (N=479)
Mental rotation	0.36 (0.14; 0.53)	0.10 (0.00; 0.27)	0.54 (0.47; 0.63)	0.44 (0.36; 0.52) (N=376)	0.27 (0.18; 0.35) (N=492)
Perspective taking	0.33 (0.10; 0.41)	0.00 (0.00; 017)	0.67 (0.59; 0.76)	0.31 (0.21; 0.39) (N=391)	0.18 (0.09; 0.26) (N=501)
Cross sections	0.18 (0.00; 0.38)	0.22 (0.05; 0.37)	0.60 (0.53; 0.68)	0.40 (0.32; 0.48) (N=428)	0.30 (0.22; 0.38) (N=574)

General spatial ability was substantially heritable (69%), with a small proportion of variance explained by shared environmental factors (8%) and the rest of the variance explained by non-shared environmental factors (23%). Heritability was lower for the individual 10 tests, ranging from 18% to 59%.

Table S6. Model fit statistics a) comparing Cholesky decomposition to Independent pathway model and Common pathway model; b) comparing Independent pathway model to Common pathway model. CholACE= Cholesky model; IPACE= Independent pathway model; CPACE= Common pathway model; ep=estimated parameters; minus2LL= minus 2 log-likelihood; df= degrees of freedom; AIC= Akaike information criterion; diffLL= change in log-likelihood; diffdf= change in degrees of freedom.

a)

base	comparison	ep	minus2LL	df	AIC	diffLL	diffdf	p
CholACE	<NA>	175	62189.93	24893	12403.93	NA	NA	NA
CholACE	IPACE	70	62306.29	24998	12310.29	116.3653	105	0.21
CholACE	CPACE	53	62405.58	25016	12373.58	215.6527	123	0.00

b)

base	comparison	ep	minus2LL	df	AIC	diffLL	diffdf	p
IPACE	CPACE	53	62405.58	25016	12373.58	99.28742	18	0.00

a) Comparing the Cholesky ACE model and the independent pathway model shows that there is no significant deterioration in fit (indicated by the p-value). Comparing the Cholesky ACE model and the common pathway model shows a significant deterioration in fit.
b) Comparing the independent pathway model and common pathway model indicates that the former fits the data better than the latter (a significant deterioration of fit is indicated by the p-value).

Table S7. Independent pathway model presenting the standardized squared path estimates (95% CI). Cp=common path; SP= specific path; A=additive genetic; C=common environmental; E=non-shared environmental; 1=mazes; 2=2D drawing, 3=Pattern assembly, 4=Elithorn maze, 5=Mechanical reasoning, 6=Paper folding, 7=3D drawing, 8=Mental rotation, 9=Perspective taking, 10=Cross-sections. a) 10 spatial tests; b) 10 spatial tests after correction for general intelligence using the regression method.

a) 10 spatial tests

CpA2[1,1]	0.25 (0.15; 0.30)
CpA2[2,2]	0.45 (0.31; 0.52)
CpA2[3,3]	0.36 (0.30; 0.50)
CpA2[4,4]	0.26 (0.16; 0.32)
CpA2[5,5]	0.33 (0.20; 0.43)
CpA2[6,6]	0.40 (0.30; 0.53)
CpA2[7,7]	0.50 (0.32; 0.58)
CpA2[8,8]	0.41 (0.34; 0.55)
CpA2[9,9]	0.22 (0.12; 0.27)
CpA2[10,10]	0.27 (0.12; 0.40)

CpC2[1,1]	0.01 (0.00; 0.05)
CpC2[2,2]	0.01 (0.00; 0.11)
CpC2[3,3]	0.00 (0.00; 0.08)
CpC2[4,4]	0.00 (0.00; 0.05)
CpC2[5,5]	0.05 (0.00; 0.17)
CpC2[6,6]	0.04 (0.00; 0.16)
CpC2[7,7]	0.01 (0.00; 0.10)
CpC2[8,8]	0.00 (0.00; 0.08)
CpC2[9,9]	0.01 (0.00; 0.08)
CpC2[10,10]	0.15 (0.05; 0.31)

CpE2[1,1]	0.04 (0.01; 0.14)
CpE2[2,2]	0.07 (0.03; 0.18)
CpE2[3,3]	0.11 (0.01; 0.19)
CpE2[4,4]	0.01 (0.00; 0.09)
CpE2[5,5]	0.02 (0.00; 0.08)
CpE2[6,6]	0.10 (0.02; 0.16)
CpE2[7,7]	0.07 (0.03; 0.27)
CpE2[8,8]	0.10 (0.01; 0.17)
CpE2[9,9]	0.02 (0.00; 0.12)
CpE2[10,10]	0.05 (0.01; 0.12)

SpA2[1,1]	0.12 (0.00; 0.22)
SpA2[2,2]	0.00 (0.00; 0.09)
SpA2[3,3]	0.03 (0.00; 0.08)
SpA2[4,4]	0.14 (0.00; 0.23)
SpA2[5,5]	0.08 (0.00; 0.16)
SpA2[6,6]	0.09 (0.00; 0.13)
SpA2[7,7]	0.08 (0.00; 0.17)
SpA2[8,8]	0.01 (0.00; 0.10)
SpA2[9,9]	0.10 (0.00; 0.22)
SpA2[10,10]	0.00 (0.00; 0.04)

SpC2[1,1]	0.00 (0.00; 0.15)
SpC2[2,2]	0.02 (0.00; 0.08)
SpC2[3,3]	0.00 (0.00; 0.05)
SpC2[4,4]	0.01 (0.00; 0.15)
SpC2[5,5]	0.02 (0.00; 0.11)
SpC2[6,6]	0.00 (0.00; 0.09)
SpC2[7,7]	0.00 (0.00; 0.11)
SpC2[8,8]	0.03 (0.00; 0.08)
SpC2[9,9]	0.02 (0.00; 0.13)
SpC2[10,10]	0.00 (0.00; 0.03)

SpE2[1,1]	0.59 (0.49; 0.67)
SpE2[2,2]	0.45 (0.36; 0.49)
SpE2[3,3]	0.50 (0.43; 0.59)
SpE2[4,4]	0.57 (0.50; 0.66)
SpE2[5,5]	0.49 (0.44; 0.56)
SpE2[6,6]	0.38 (0.33; 0.48)
SpE2[7,7]	0.33 (0.19; 0.40)
SpE2[8,8]	0.45 (0.38; 0.54)
SpE2[9,9]	0.63 (0.53; 0.71)
SpE2[10,10]	0.53 (0.46; 0.57)

b) 10 spatial tests after correction for *g*

CpA2[1,1]	0.07 (0.02; 0.15)
CpA2[2,2]	0.25 (0.16; 0.32)
CpA2[3,3]	0.17 (0.06; 0.31)
CpA2[4,4]	0.08 (0.03; 0.16)
CpA2[5,5]	0.24 (0.18; 0.29)
CpA2[6,6]	0.30 (0.19; 0.37)
CpA2[7,7]	0.27 (0.17; 0.35)
CpA2[8,8]	0.22 (0.09; 0.36)
CpA2[9,9]	0.13 (0.08; 0.19)
CpA2[10,10]	0.34 (0.17; 0.41)

CpC2[1,1]	0.05 (0.00; 0.11)
CpC2[2,2]	0.02 (0.00; 0.10)
CpC2[3,3]	0.10 (0.00; 0.21)
CpC2[4,4]	0.04 (0.00; 0.10)
CpC2[5,5]	0.00 (0.00; 0.04)
CpC2[6,6]	0.02 (0.00; 0.10)
CpC2[7,7]	0.02 (0.00; 0.10)
CpC2[8,8]	0.10 (0.00; 0.23)
CpC2[9,9]	0.00 (0.00; 0.05)
CpC2[10,10]	0.02 (0.00; 0.16)

CpE2[1,1]	0.10 (0.05; 0.15)
CpE2[2,2]	0.18 (0.12; 0.25)
CpE2[3,3]	0.10 (0.05; 0.17)
CpE2[4,4]	0.08 (0.04; 0.14)
CpE2[5,5]	0.05 (0.02; 0.08)
CpE2[6,6]	0.11 (0.07; 0.17)
CpE2[7,7]	0.23 (0.16; 0.31)
CpE2[8,8]	0.08 (0.04; 0.14)
CpE2[9,9]	0.04 (0.01; 0.08)
CpE2[10,10]	0.03 (0.01; 0.07)

SpA2[1,1]	0.15 (0.00; 0.22)
SpA2[2,2]	0.00 (0.00; 0.10)
SpA2[3,3]	0.00 (0.00; 0.08)
SpA2[4,4]	0.16 (0.00; 0.25)
SpA2[5,5]	0.10 (0.00; 0.18)
SpA2[6,6]	0.10 (0.00; 0.15)
SpA2[7,7]	0.12 (0.00; 0.20)
SpA2[8,8]	0.00 (0.00; 0.06)
SpA2[9,9]	0.00 (0.00; 0.19)
SpA2[10,10]	0.00 (0.00; 0.05)

SpC2[1,1]	0.00 (0.00; 0.13)
SpC2[2,2]	0.05 (0.00; 0.09)
SpC2[3,3]	0.00 (0.00; 0.05)
SpC2[4,4]	0.00 (0.00; 0.16)
SpC2[5,5]	0.02 (0.00; 0.13)
SpC2[6,6]	0.00 (0.00; 0.09)
SpC2[7,7]	0.02 (0.00; 0.13)
SpC2[8,8]	0.00 (0.00; 0.06)
SpC2[9,9]	0.09 (0.00; 0.15)
SpC2[10,10]	0.00 (0.00; 0.05)

SpE2[1,1]	0.64 (0.56; 0.72)
SpE2[2,2]	0.50 (0.44; 0.56)
SpE2[3,3]	0.62 (0.55; 0.66)
SpE2[4,4]	0.63 (0.55; 0.73)
SpE2[5,5]	0.60 (0.53; 0.67)
SpE2[6,6]	0.48 (0.42; 0.55)
SpE2[7,7]	0.34 (0.27; 0.42)
SpE2[8,8]	0.59 (0.53; 0.63)
SpE2[9,9]	0.73 (0.63; 0.79)
SpE2[10,10]	0.61 (0.56; 0.67)

Standardized path estimates (following from Figure 4), with 95% confidence intervals, for the independent pathway model.

a) All spatial tests loaded substantially on the common A factor, with no significant specific genetic influence remaining after controlling for the common genetic factor. On average, the common A factor accounted for 85% of the heritabilities of the 10 spatial tests (for example the heritability of the Mazes task was 37% (the sum of common path, .25, and the specific path, .12), so the proportion of heritability accounted for by the common factor is .25/.37=68%). The spatial tests are differentiated by E factors, which indicate test-specific environmental influences and measurement error specific to each test.
b) These results show the same analysis after correcting the spatial scores for *g*. A common genetic factor still explained most of the heritability across the 10 tests, although loadings on the common A factor were reduced by about one third.

Table S8. Common pathway model presenting the standardized path estimates. A- additive genetic, C- shared environmental and E- non-shared environmental components of variance. a) 10 spatial tests; b) 10 spatial tests when corrected for intelligence using the regression method.

a) 10 spatial tests

Spatial ability (Latent Factor)	
A	0.80 (0.64-0.89)
C	0.06 (0-0.21)
E	0.14 (0.11-0.18)

Loadings to Spatial ability factor	
Cross sections	0.61 (0.68-0.64)
2D drawing	0.73 (0.71-0.75)
Pattern assembly	0.66 (0.64-0.69)
Elithorn maze	0.50 (0.46-0.69)
Mechanical reasoning	0.62 (0.59-0.63)
Paper folding	0.72 (0.70-0.75)
3D drawing	0.77 (0.75-0.79)
Mental rotation	0.70 (0.68 -0.72)
Perspective taking	0.50 (0.47-0.54)
Mazes	0.51 (0.47-0.54)

Residual variance			
	A	**C**	**E**
Cross sections	0.00 (0-0.15)	0.09 (0-0.13)	0.54 (0.47-0.59)
2D drawing	0.00 (0-0.06)	0.02 (0-0.05)	0.44 (0.40 -0.49)
Pattern assembly	0.04 (0-0.09)	0.00 (0-0.06)	0.52 (0.46-0.58)
Elithorn maze	0.16 (0-0.25)	0.02 (0-0.16)	0.57 (0.50-0.66)
Mechanical reasoning	0.09 (0-0.18)	0.04 (0-0.14)	0.48 (0.43-0.55)
Paper folding	0.09 (0-0.13)	0.00 (0-0.07)	0.39 (0.34-0.44)
3D drawing	0.09 (0-0.13)	0.04 (0-0.08)	0.32 (0.28-0.38)
Mental rotation	0.01 (0-0.10)	0.04 (0-0.08)	0.46 (0.40-0.51)
Perspective taking	0.10 (0-0.19)	0.02 (0-0.13)	0.63 (0.55-0.71)
Mazes	0.15 (0-0.22)	0.00 (0-0.14)	0.59 (0.52-0.67)

b) 10 spatial tests after correction for *g*

Spatial ability (Latent factor)	
A	0.57 (0.35-0.75)
C	0.12 (0-0.31)
E	0.31 (0.25-0.38)

Loadings to Spatial ability factor	
Cross sections	0.52 (0.52-0.55)
2D drawing	0.67 (0.64-0.70)
Pattern assembly	0.58 (0.55-0.62)
Elithorn maze	0.44 (0.40-0.48)
Mechanical reasoning	0.51 (0.48 -0.55)
Paper folding	0.65 (0.62-0.68)
3D drawing	0.71 (0.68-0.74)
Mental rotation	0.61 (0.58-0.64)
Perspective taking	0.42 (0.38-0.46)
Mazes	0.44 (0.38-0.47)

Residual variance			
	A	**C**	**E**
Cross sections	0.02 (0-0.18)	0.09 (0-0.15)	0.63 (0.55-0.69)
2D drawing	0.00 (0-0.08)	0.03 (0-0.08)	0.51 (0.46-0.51)
Pattern assembly	0.06 (0-0.13)	0.00 (0-0.08)	0.60 (0.53-0.67)
Elithorn maze	0.16 (0-0.25)	0.01 (0-0.17)	0.64 (0.55-0.74)
Mechanical reasoning	0.12 (0-0.22)	0.03 (0-0.08)	0.59 (0.52-0.67)
Paper folding	0.11 (0-0.16)	0.00 (0-0.10)	0.47 (0.41-0.54)
3D drawing	0.10 (0-0.18)	0.02 (0-0.13)	0.37 (0.31-0.44)
Mental rotation	0.00 (0-0.10)	0.05 (0-0.10)	0.57 (0.51-0.63)
Perspective taking	0.01 (0-0.19)	0.09 (0-0.15)	0.72 (0.63-0.79)
Mazes	0.16 (0-0.24)	0.00 (0-0.15)	0.65 (0.57-0.73)

A, C and E influences on the common latent factor show that the spatial factor is highly heritable. The factor loadings on the latent factor are very substantial. There is some residual variance left after accounting for the latent factor, but this is very small in magnitude, with the A estimates for the residual variance not significant. It should be noted that the independent pathway model fitted the data better than the common pathway model (Supplementary Table S6), but the common pathway model results are presented here for completeness.

Table S9. Genetic, shared environmental and non-shared environmental correlations between 10 spatial tests.

Genetic correlations	Cross sections	2D drawing	Pattern assembly	Elithorn maze	Mechanical reasoning	Paper folding	3D drawing	Mental rotation	Perspective taking	Mazes
Cross sections	1.000									
2D drawing	0.883	1.000								
Pattern assembly	0.770	0.962	1.000							
Elithorn maze	0.860	0.805	0.732	1.000						
Mechanical reasoning	0.876	0.875	0.744	0.766	1.000					
Paper folding	0.966	0.933	0.885	0.862	0.824	1.000				
3D drawing	0.894	0.952	0.912	0.791	0.833	0.920	1.000			
Mental rotation	0.813	0.940	0.950	0.832	0.842	0.897	0.892	1.000		
Perspective taking	0.760	0.913	0.878	0.756	0.834	0.786	0.946	0.884	1.000	
Mazes	0.733	0.730	0.729	0.746	0.784	0.770	0.768	0.884	0.759	1.000

Shared environmental correlations	Cross sections	2D drawing	Pattern assembly	Elithorn maze	Mechanical reasoning	Paper folding	3D drawing	Mental rotation	Perspective taking	Mazes
Cross sections	1.000									
2D drawing	0.786	1.000								
Pattern assembly	0.676	0.864	1.000							
Elithorn maze	-0.059	0.164	0.261	1.000						
Mechanical reasoning	0.824	0.635	0.772	0.067	1.000					
Paper folding	0.795	0.606	0.728	-0.335	0.776	1.000				
3D drawing	0.683	0.826	0.551	0.415	0.428	0.213	1.000			
Mental rotation	0.710	0.740	0.868	0.294	0.738	0.689	0.612	1.000		
Perspective taking	0.537	0.194	-0.114	-0.605	0.067	0.412	0.234	0.056	1.000	
Mazes	0.191	0.702	0.730	0.126	0.282	0.313	0.335	0.382	-0.333	1.000

Non-shared environmental correlations	Cross sections	2D drawing	Pattern assembly	Elithorn maze	Mechanical reasoning	Paper folding	3D drawing	Mental rotation	Perspective taking	Mazes
Cross sections	1.000									
2D drawing	0.137	1.000								
Pattern assembly	0.137	0.113	1.000							
Elithorn maze	-0.005	0.113	0.103	1.000						
Mechanical reasoning	0.101	0.072	0.082	0.027	1.000					
Paper folding	0.095	0.161	0.168	0.033	0.105	1.000				
3D drawing	0.141	0.209	0.139	0.066	0.129	0.224	1.000			
Mental rotation	0.112	0.118	0.175	0.071	0.059	0.180	0.177	1.000		
Perspective taking	0.112	0.061	0.064	0.035	0.069	0.055	0.069	0.120	1.000	
Mazes	0.061	0.090	0.132	0.080	0.020	0.090	0.193	0.075	0.083	1.000

Genetic correlation is an index of pleiotropy: the extent to which the same genetic variants influence multiple traits. Importantly, the genetic correlation is estimated independently of the heritabilities of the traits; that is, the genetic correlation between the traits could be high even if the heritabilities of both traits were low. A shared environmental correlation of 1.0 indicates that the same environmental factors that make twins similar on one trait also make twins similar on another trait. Likewise, for non-shared environment (which is not shared between individuals, but may influence multiple traits for each individual), a correlation of zero indicates that completely different non-shared environmental influences affect the two traits. The results of the multivariate analyses shows that genetic correlations between spatial tests is very high, indicating that to a large extent the performance on these spatial tests is influence by the same genetic factors.

Table S10.. Summary of the development of the gamified battery (King's Challenge): a) Feasibility studies; b) TEDS pilot study.

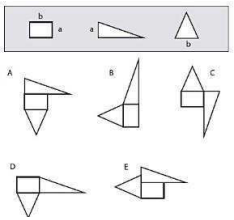
The "King's Challenge" game was constructed after conducting a literature review of the many measures used to test spatial ability, assembling a large variety of measures to test each of the putative components of this cognitive domain. We conducted several feasibility and pilot studies, modifying existing tests and developing some new ones as needed. We started with a paper-and-pencil battery including 27 different tests, and after multiple stages of feasibility and pilot testing (mostly conducted online) ultimately reduced the battery to 10 tests, selected according to the psychometric properties and test-retest reliability of each measure. Here we present:

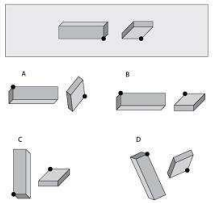
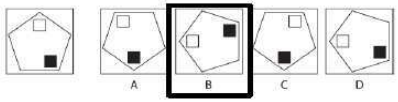
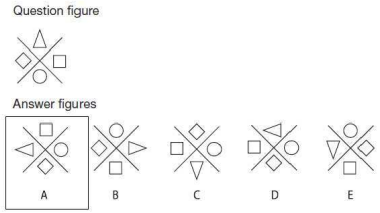
a) the results of two feasibility studies: feasibility 1- the initial paper-and-pencil battery, in which participants were tested in person and were subject to test-level time limits as described in the table; feasibility 2- the first battery administered online (with item-level time limits), from which initial test-retest correlations were obtained (with a 1-week interval between test and retest);

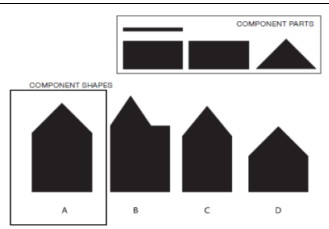
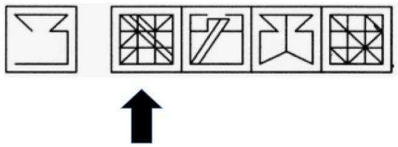
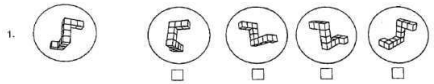
b) the results of the final stage prior to "gamification": a TEDS pilot study with the 10 selected tests. For the latter pilot study, siblings of the TEDS twins were recruited and final test-retest correlations obtained (with a 2-week interval between test and retest).

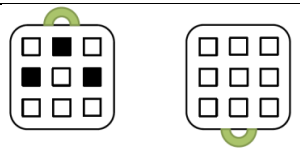
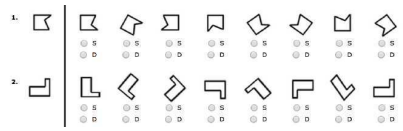
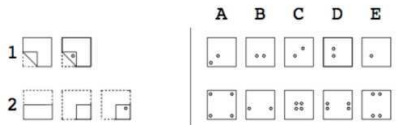
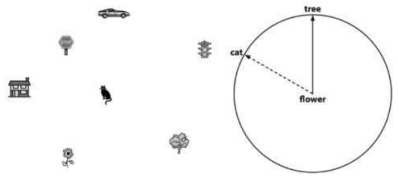
Following the final pilot study, the "gamified" battery was developed. The actual test items were administered in a format identical to those in the final pilot study, but the tests themselves were embedded into an overarching game narrative to encourage participation. This final battery was administered to a large twin sample as described in the manuscript.

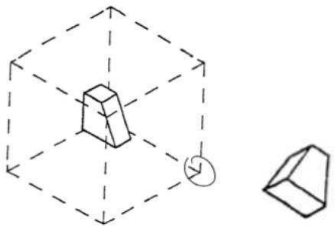
a) Feasibility studies

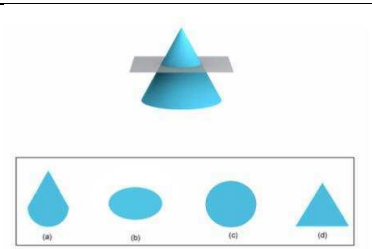
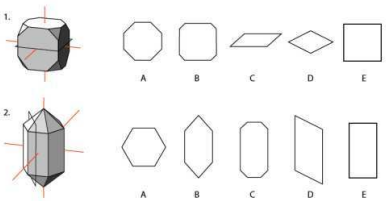
TEST	DESCRIPTION for administration in feasibility 1	REASON FOR KEEPING/DROPPING and ADJUSTMENTS during following stages / Feasibility 2 results	SOURCE
1. Pattern Assembly (1) 	Participants are asked to decide which option (A - E) is made up of the parts presented in the grey box at the top. The test includes **20 items** and participants are allowed **10 minutes** to complete the test.	Included in the second feasibility study (online). The test produced normal distribution and reasonable test-retest reliability. Test-retest (cleaned, standardised): r=0.59, N=40, p<0.001. **Adapted version included in the gamified test (the King's Challenge)**.	Spatial reasoning section 1 in the "How2become" booklet (https://www.how2become.com/testing/spatial-reasoning-tests/)(originally 40 items)
2. 3D rotation (1)	Participants are presented with a pair of three-dimensional objects; one of the corners of each object is marked with a black dot. Participants are asked to imagine which one of the 4 options (A - D) would reflect what the pair of objects would	This task produced a ceiling effect and was dropped after first feasibility study.	Spatial reasoning section 1 in the "How2become" booklet (https://www.how2

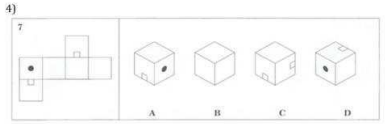
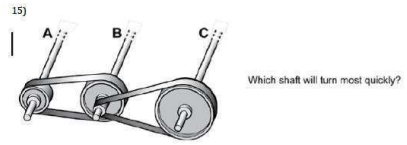
	look like if they were both rotated by the same amount. Participants have **10 minutes** to complete **20 questions**.		become.com/testing/spatial-reasoning-tests/)(originally 40 items)
3. 2D rotation (1) 	Participants are asked to identify which one of the 4 options (A - D) is the same 2D object as the question figure on the left, but rotated. The test includes **20 items** and participants are allowed **10 minutes** to complete them all.	Dropped after the first feasibility study as another task assessing 2D rotation (task 5 below) performed better in terms of distribution and internal reliability.	Spatial reasoning section 1 in the "How2become" booklet (https://www.how2become.com/testing/spatial-reasoning-tests/)
4. Identical shapes 	Participants are asked to identify which two 2D objects (A - E) are identical. The test includes **20 items** and participants have **4 minutes** to complete it.	Dropped after the first feasibility study: too easy, highly skewed distribution.	Spatial reasoning section 1 in the "How2become" booklet (https://www.how2become.com/testing/spatial-reasoning-tests/)
5. 2D rotation (2) 	Participants are asked to identify which one of the answer figures (A - E) is the same object as in the question figure, but rotated. Participants have **7 minutes** to complete **19 items**.	Included in the second feasibility study (online). Produced a normal distribution and good test-retest reliability. Test-retest (cleaned, standardised) r=0.73, N=43, p<0.001. **Adapted version included in the gamified test (the King's Challenge).**	Spatial reasoning section 1 in the "How2become" booklet (https://www.how2become.com/testing/spatial-reasoning-tests/)
6.Pattern assembly (2)	Participants are asked to identify which one of the component shapes (A - D) is made from the component parts displayed in the rectangular box at the top. The test includes **20 items** and participants have **7 minutes** to complete it.	Included in the second feasibility study (online). The test produced a good distribution but very poor test-retest reliability. Test-retest (cleaned, standardised): r=0.26, N=44, p=0.08. The other pattern assembly test (task 1), showed much higher reliability, so was retained instead.	Spatial reasoning section 1 in the "How2become" booklet (https://www.how2become.com/testing/spatial-reasoning-tests/)

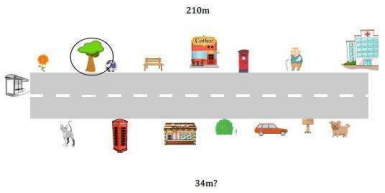
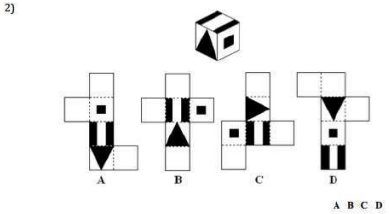
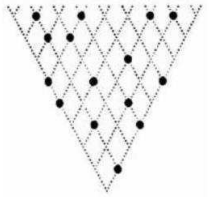
			
7. Embedded figures 	Participants are asked to identify which one of the 4 figures presented on the right-hand side of the page includes the question figure on the left-hand side embedded in its pattern. Participants are given **8 minutes** to complete **25 questions**.	Kept for the second feasibility study (online) including the same 25 items, but subsequently dropped due to relatively poor test-retest reliability and other psychometric properties (other scanning tasks had better psychometric properties): Test-retest (cleaned, standardised; *r*=0.50, N=48, p<0.001.	www.indiabix.com /non-verbal-reasoning/embedded-images
8. 3D mental rotation (2) 	Participants are asked to identify which 2 options (out of the 4 presented on the right-hand side) are rotated versions of the question figure. Only 2 options are correct at all times. The test is divided into **2 parts** and each part includes **10 questions**. Participants have **3 minutes** to complete each part.	An adapted version was retained for the second feasibility study (online), with only one correct answer per item and two incorrect options. Participants discontinued from the test after 4 consecutive incorrect responses. This was subsequently dropped due to very low test-retest reliability. Test-retest (cleaned, standardised): *r*=0.29, N=34, p=0.092.	Shepard &Metzler (Shepard, R and Metzler. J. "Mental rotation of three dimensional objects." *Science* 1971. 171(972):701-3 Adapted by S.G. Vanderberg, University of Colorado, July 15, 1971; Revised instructions by H. Crawford, University of Wyoming, September, 1979; Images digitalized and reprinted by Susanna Douglas, University of Texas, March 1996
9. 2D mental rotation (3) –AKA suitcase task	The task requires participants to mentally rotate the image on the left-hand side, and to colour in the corresponding pattern made up of squares in the figure on the right-hand side. Participants are	Dropped after the first feasibility study, as the task was much too easy –produced a very skewed negative distribution.	Adapted from Tzuriel, D. (1995). The Cognitive Modifiability

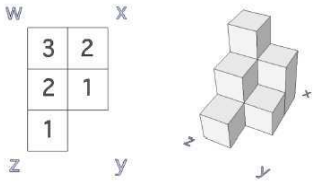
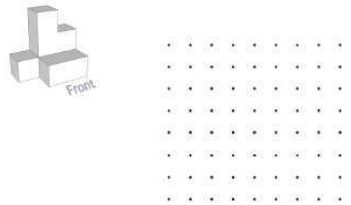
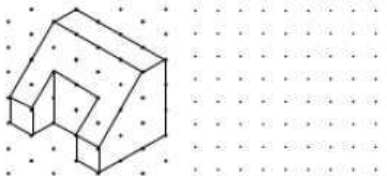
	given **3 minutes** to complete **11 items**.		Battery (CMB). Assessment and intervention: User's manual. Tel Aviv, Israel: School of Education, Bar-Ilan University
10. Card rotation test 	Participants are asked to identify whether each one of the 8 options presented on the right-hand side of the page is the same shape as the one presented on the left-hand side. If participants think the shape is the same shape (but rotated) they should tick the option "s" at the bottom of the answer shape. If they think it's a different shape, then they should select the option "d". The test includes **10 items** to be completed in **3 minutes**.	Dropped after the first feasibility: produced a very skewed distribution, and several other 2D rotation tasks performed better.	French, J., Ekstrom, R., Price, L.: Manual for a kit of reference tests for cognitive factors. Princeton, New Jersey: Educational Testing Service 1963
11. Paper folding test 	On the left-hand side of the page participants are shown a sheet of paper folded following several stages. The last image of the sequence includes a dot. This dot represents a hole that is punched through all the thickness of the paper at that point. Participants are asked to identify which one of the 5 pictures on the right-hand side shows where the holes will be when the paper is completely unfolded again (by reversing the specific steps shown). Participants have **7 minutes** to complete **20 items**.	Included in the second feasibility study (online), with items re-ordered in progressively increasing difficulty (as indicated by scores in the first feasibility study). The resulting test produced a normal distribution and acceptable test-retest reliability. Test-retest (cleaned, standardised): r=0.59, N=44,p<0.001. **Adapted version included in the gamified test (the King's Challenge).**	Adapted from University of Otago, New Zealand http://www.cs.otago.ac.nz/brace/resources/Paper%20Folding%20Test%20Vz-2-BRACE%20Version%2007.pdf
12. Perspective taking (1) –AKA "Point to the cat" 	Participants are first shown the picture on the left-hand side. They are asked to imagine that they are standing in a certain location (one of the shapes), facing another location, and they need to imagine pointing to a third location. They are then asked to draw the direction of their pointing, on the circular diagram shown on the right-hand side. For example: "Imagine you are standing at the flower and facing the tree. Now point to the cat". Participants have **7 minutes** to complete **12 items**.	Dropped after the first feasibility study, due to poor distribution and internal reliability. Another perspective-taking task (task 13 below) performing better psychometrically and was retained instead.	Kozhevnikov, M. & Hegarty, M. (2001). A dissociation between object-manipulation and perspective-taking spatial abilities. *Memory & Cognition*, 29, 745-756. Hegarty. M. & Waller, D. (2004). A dissociation

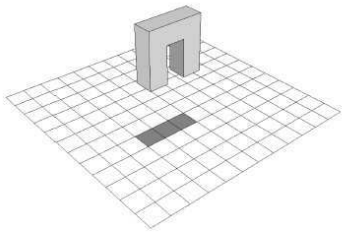
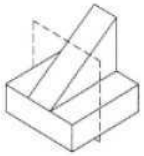
			between mental rotation and perspective-taking spatial abilities. *Intelligence*, 32, 175-191.
13. Perspective taking (2) 	Participants are presented with a transparent cube containing an irregular polygon suspended in the middle of the cube (see example figure). The same polygon is also presented outside the cube from a different viewpoint. Participants are asked to indicate on which corner of the cube they would have to stand in order to see the polygon from the new viewpoint (e.g. the bottom right corner in the example figure). Participants were allowed **8 minutes** to go through **24 questions**.	Included in the second feasibility study (online), with items re-ordered in progressively increasing difficulty. The test produced a normal distribution and very good test-retest reliability. Test-retest (cleaned, standardised): r=0.83, N=40, p<0.001. **Adapted version included in the gamified test (the King's Challenge).**	Adapted from Hegarty, M., Keehner, M., Khooshabeh, P., & Montello, D. R. (2009). How spatial abilities enhance, and are enhanced by, dental education. *Learning and Individual Differences*, 19(1), 61-70. Keehner, M., Hegarty, M., Cohen, C. A., Khooshabeh, P., & Montello, D. R. (2008). Spatial reasoning with external visualizations: What matters is what you see, not whether you interact. *Cognitive Science*, 32(7), 1099–1132.
14. Cut the cross-section (1)	Participants are asked to identify the cross-section of three types of figures: single objects (like the example figure), attached objects, and nested objects (where one object is inside the other). The plane cutting the figure can be vertical, horizontal (like the example) or oblique. Participants are given **7 minutes** to complete **15 items**.	Dropped after the first feasibility study, as its correlation with the other cross-sections test, task 15 (r = .76) was so high as to render it redundant. Participants also preferred the other cross-sections test.	Cohen, C. A. & Hegarty, M. (2007). Sources of difficulty in imagining cross sections of 3D objects. In D. S. McNamara & J. G. Trafton (Eds.),

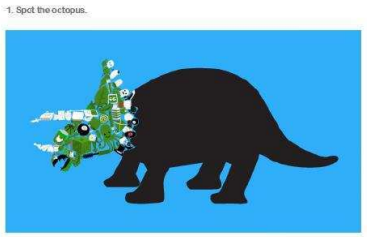
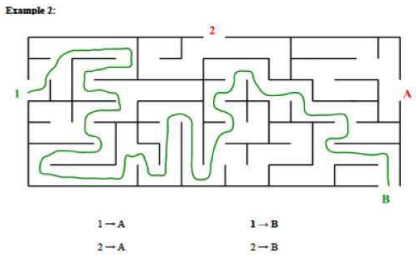
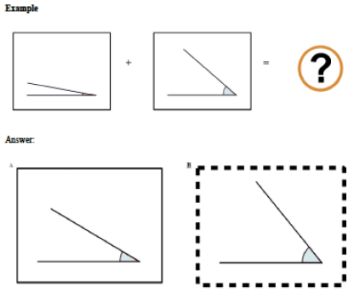
			*Proceedings of the Twenty-Ninth Annual Conference of the Cognitive Science Society* (pp.179-184). Austin TX: Cognitive Science Society. Cohen, C. A. & Hegarty, M. (2012). Inferring cross sections of 3D objects: A new spatial thinking test. *Learning and Individual Differences*, 22(6), 868-874.
15. Cross-section (2) 	Participants are asked to identify the shape that the cutting plane will produce when cutting through several symmetrical solids (see example figures). The plane can cut the solid vertically, horizontally or obliquely. Participants are given **7 minutes** to go through **15 questions**.	Included in the second feasibility study (online) with items re-ordered for progressively increasing difficulty. The test produced a normal distribution and good test-retest reliability. Test-retest (cleaned, standardised): r=0.75, N=43, p<0.001. **Included in the gamified test (the King's Challenge).**	Adapted from Ormand, C. J., Shipley, T. F., Tikoff, B., Manduca, C. A., Dutrow, B., Goodwin, L., Hickson, T., Atit, K., Gagnier, K. M., & Resnick, I. (2013). Improving Spatial Reasoning Skills in the Undergraduate Geoscience Classroom Through Interventions Based on Cognitive Science Research. Talk presented at the AAPG Hedberg Conference on 3D Structural Geologic

			Interpretation.
16. 2D to 3D visualization  	Participants are asked to identify which one of 4 3D shapes could be built from the 2D pattern presented on the left-hand side of the picture. Only one shape out of the 4 is the correct answer. Participants are given **8 minutes** to complete **25 items**.	Kept for the second feasibility study (online) but subsequently dropped due to a high positive skew (i.e., it was too difficult), and very poor test-retest reliability. Test-retest (cleaned, standardised): *r*=0.16, N=30, p=0.41.	Harcourt Assessment (1995), DAT for Selection-Technical Abilities Battery. Pearson Assessment: London
17. Mechanical reasoning  	Participants have **5 minutes** to complete **15 questions** revolving around a common theme: mechanical reasoning. Examples of questions are: “Which shaft will turn more quickly?” (See example picture) and “If only the right oar of the boat is pulled, in which direction will the boat go?”	Included in the second feasibility study (online), with 6 extra items added to the original 15. The test produced a normal distribution and good test-retest reliability: Test-retest (cleaned, standardised): r=0.69, N=46, p<0.001.). In addition to the overall score, the 21 items were grouped thematically into subtests: 5 'pulley' items, 4 'gear' items, and 12 'miscellaneous' items, each with their own subtest score. Following the second feasibility study, the 5 'pulley' items were removed, as this subtest produced poor test re-test reliability (r = 0.39, N=46, p=0.006). **Included in the gamified test (the King’s Challenge).**	Adapted from Harcourt Assessment (1995), DAT for Selection-Technical Abilities Battery. Pearson Assessment: London Wiesen, J. (2009), Barron’s Mechanical Aptitude and Spatial Relations Test, 2nd edition, Barron’s Educational Series Wiesen, J. (2009), Barron’s Mechanical Aptitude and Spatial Relations Test, 2nd edition, Barron’s Educational Series
18. Spatial number line	Participants are shown a strip of street with a number at the top indicating the length of the street. At the bottom of each picture is a number followed by a question mark indicating a specific distance. Participants are asked to decide which landmark is situated at that specific distance. E.g. in the example picture the total length of the street	Kept for the second feasibility study (online) but subsequently dropped despite good test-retest reliability and distribution of scores: Test-retest (cleaned, standardised): *r*=0.67, N=50, p<0.001. This task was included in the initial battery	Adapted from the number line test (Siegler, R. S. and Opfer, J. E. (2006). Representational change and

	is 210 meters and participants are asked to identify which landmark is situated at a distance 34 meters from the beginning of the street located on the left-hand side of the page. The tree is the correct answer in this case. The numerical proportions are taken from those in the number line test (Siegler & Opfer, 2006). Participants have **2 minutes** to complete **9 items**.	experimentally as a 'number line' measure, to assess the relationship with mathematical abilities. Its low correlations with other measures appeared to confirm that this was not a spatial task, and it was dropped accordingly.	children's numerical estimation. *Cognitive Psychology*. doi:10.1016/j.cogpsych.2006.09.002
19. 3D to 2D 	Participants are presented with a cube and 4 unfolded 2D patterns. Participants have to decide which one of the 4 unfolded patterns makes the 3D cube. There is only one correct option. Participants have **9 minutes** to complete **13 items**.	Test kept for the second feasibility study (online). Subsequently dropped as it produced a positively skewed distribution and very poor test-retest reliability. Test-retest (cleaned, standardised): r=0.19, N=35, p=0.27.	http://www.psychometric-success.com/aptitude-tests/spatial-ability-tests-cubes.htm
20. Elithorn Maze 	Participants are asked to trace their route on each one of the grids presented (both triangular and rectangular grids are included in this version of the test). The aim of the task is to trace the route passing through the largest possible number of black dots. Participants are asked to start from the bottom part of the shape (the point of the triangle in this case) and move upwards; they can only move left or right on the grid and cannot go backwards; it is not possible to collect all the dots in the grid. **9 items** should be completed in **4 minutes**.	Included in the second feasibility study (online). This test was a computerised version of the original paper and pencil task, in which (in each item) a line moved upwards at a constant speed through a triangular grid, and the participant could change direction (left/right) at each intersection, in an attempt to collect the largest possible number of dots. This test produced a normal distribution and good test-retest reliability: (cleaned, standardised) r=0.76, N=51, p<0.001. **This adapted version was included in the gamified test (the King's Challenge).**	Adapted from Test of spatial planning ability included as a process subtest of the WISC-IV Integrated. ELITHORN, A. (1955). A preliminary report on a perceptual maze test sensitive to brain damage. J. *neurol. neurosurg. Psychiat*, 18, 287-292.
21. Drawing task	This task is divided into 5 subsections each asking participants to draw (see a description and examples for each subsection below). Participants	See subsections below	Adapted from Engage Students in Engineering

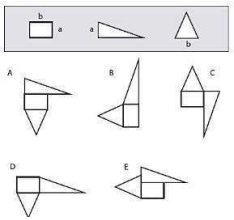
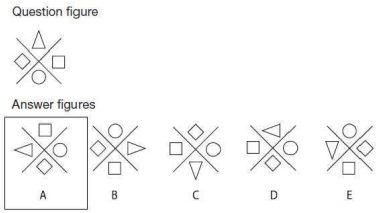
	were given **20 minutes** to complete all **5 subsections**.		site: http://www.wskc.org/documents/281621/307751/ENGAGE_SV_Sample_quiz_on_modules_3_4_and_5.pdf/c3df8086-b535-4ad3-a669-e55be8168820?version=1.0
21.1 2D to 3D drawing 	Participants are presented with the coded plan (see left-hand side of the example picture) and are asked to draw the 3D object corresponding to the plan (like the diagram in the right-hand side of the example picture). This subsection includes **5 items**.	Included in the second feasibility study (online). This test was a computerised version of the original paper-and-pencil task, with participants clicking on dots arranged in an isometric grid to draw lines between them. Showed good distribution and high test-retest reliability. Test-retest (cleaned, standardised): r=0.79, N=37, p<0.001. **Included in the gamified test (the King's Challenge).**	
21.2 3D to 2D viewpoints 	Participants are asked to draw the viewpoint indicated as the 'front' of the picture of the 3D solid (see example figure). The drawing that participants should produce is a 2D viewpoint of the 3D shape. This subsection includes **5 items**.	Included in the second feasibility study (online). This test was a computerised version of the original paper and pencil task, exactly the same as task 21.1, but with the dots arranged in a square rather than an isometric pattern. Showed good distribution and high test-retest reliability. Test-retest (cleaned, standardised): r=0.78, N=47, p<0.001. **Included in the gamified test (the King's Challenge).**	
21.3 Sketch the front and top views 	Participants are asked to sketch the front and top views of the shapes shown on the left-hand side of the grid (see example picture). This subsection included **5 items**.	Dropped after the first feasibility study, due to a highly positively skewed distribution.	
21.4 Draw the reflection	Participants are shown drawings of 3D floating objects and asked to draw the reflection of each	Dropped after the first feasibility study, due to a highly positively skewed	

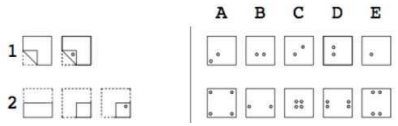
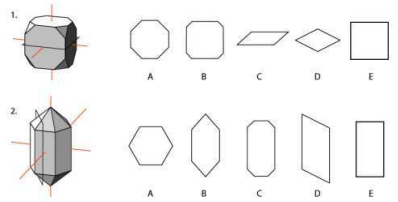
	object on the grid provided (see example). Each grid is like a mirror. This subsection includes **5 items**.	distribution.	
21.5 Sketch the cross-section 	Participants are provided with a grid onto which they need to sketch the cross-section of the objects cut by an imaginary plane shown on the left-hand side of the page (see example figure). This subsection includes **5 items**.	Kept for the second feasibility study (online), including only the easier items from the set. This test was a computerised version of the original paper-and-pencil task, conducted the same way as task 21.2. Dropped after the second feasibility study: it was normally distributed and reliable (test-retest r=0.76, N=47, p<0.001), but highly correlated with task 15 above (cross-sections 2) (r=0.65, N=70, p<0.001), so added little to the battery to justify its long duration compared to other tests.	
22. Water level task 	Participants are presented with water containers of different sizes drawn on the left side of the page. They need to decide which one of the 4 containers on the right side of the page (A, B, C, or D) has the exact same amount of water as that of the first container on the left hand side of the page. Participants are allowed **3 minutes** to complete **9 questions.**	Kept for the second feasibility study (online) as it produced a good distribution in the paper-pencil version. The test was subsequently dropped due to very poor test-retest reliability (cleaned, standardised): *r*=0.13, N=35, p=0.45.	Adapted from Piaget's water level task. Piaget, J., & Inhelder, B. (1956). The child's conception of space. London: Routledge & Kegan Paul.
23. Light bulb task	Participants are presented with a drawing of a car moving on a plane (flat) surface. Inside this car there is a hanging light bulb attached to a string. Participants are then presented with 8 drawings of the same car proceeding on different slopes (uphill and downhill). Their task is to draw the string and the light bulb in the correct inclination for each car, with reference to the angle at which the car is moving uphill or downhill. Participants have **3 minutes** to complete **8 questions**.	Dropped after the paper-pencil feasibility study, as it was much too easy, producing a highly negatively skewed distribution.	Developed by the team

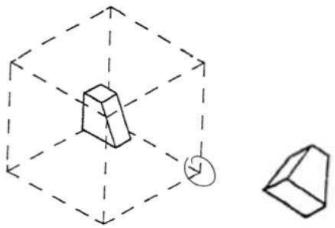
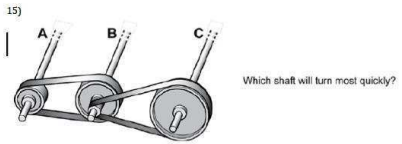
 1. 			
24. Scanning task (aka 'little things') 	Participants are presented with several drawings made of small icons. Their aim is to spot the item indicated at the top of each drawing, hidden within the larger figure (see example figure on the left). In order to test quick scanning skills, parts of the original drawings have been blackened out. In this way participants could focus on a restricted area and proceed as fast as possible. Participants are allowed **4 minutes** to complete **10 questions.**	Dropped after the first feasibility study, as another task assessing spatial scanning (task 25 – the mazes task) performed much better in terms of distribution and reliability.	Taken from an IPhone App "Little Things".
25. Mazes task 	Participants are presented with a series of mazes, each with multiple ways in and out, but with only one valid route connecting one of the entrances to one of the exits. Participants are asked to look at the map (see example picture on the left) and choose from the options available the valid route between a single entrance and exit. The test includes **10 items** with increasing difficulty to be completed in **4 minutes**.	Included in the second feasibility study (online). The test produced a normal distribution and good test-retest reliability: Test-retest (cleaned, standardised): r=0.74, N=42, p< 0.001. **Included in the gamified test (the King's Challenge).**	Developed by the team
26. Angle task 	Participants are presented with a series of angles and a mathematical operation to be performed on those angles (adding or subtracting). From four possible options, participants are asked to choose the angle that most closely represents the correct answer (see the example figure on the left). Participants have **2 minutes** to complete **10 questions** of increasing difficulty.	Kept for the second feasibility study (online) but subsequently dropped due to poor test re-test reliability (cleaned, standardised; *r*=0.41, N=41, p=0.009).	Developed by the team

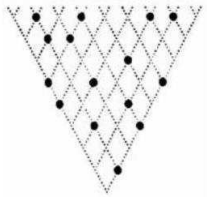
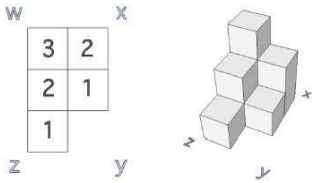
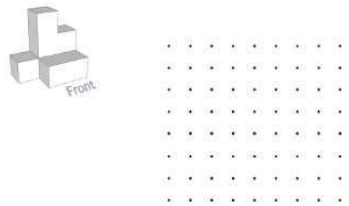
27. Water level task (2) 	Participants are presented with a series of bottles containing some water laying on a plane (flat) surface. Next to each bottle are four empty tilted bottles. Participants are asked to draw a line showing the water level for each tilted bottle as if they were filled with the same amount of water as that in the bottle on the left-hand side. The task includes **5 items** and participants have **2 minutes** to complete it.	Dropped after the paper-pencil feasibility study as it was too easy – negatively skewed distribution.	Adapted from Piaget's water level task Piaget, J., & Inhelder, B. (1956). The child's conception of space. London: Routledge & Kegan Paul.

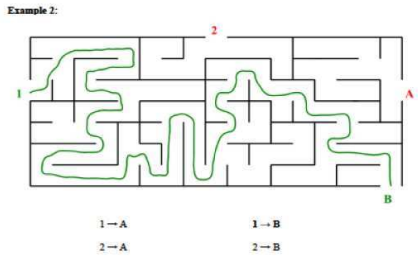
b) The King's Challenge TEDS sibling pilot analyses of 10 tests:

TEST (numbered as above, for reference)	DESCRIPTION	RESULTS	REFERENCE
1. Pattern Assembly (1) 	Participants are asked to decide which option (A - E) is made up of the parts presented in the grey box at the top. The test includes **15 items** each to be completed within a **20 seconds** time frame. Participants are discontinued if they provide 4 consecutive incorrect answers.	**TEDS sibling pilot results:** normally distributed, no floor/ceiling effects, with a mean score of 8.14, SD 2.4, N = 168; test-retest correlation *r*=.56, N = 101, p< .001	Adapted from Spatial reasoning section 1 in the "How2become" booklet (https://www.how2become.com/testing/spatial-reasoning-tests/)(originally 40 items)
5. Shapes rotation (mental rotation) 	Participants are asked to identify which one of the answer figures (A - E) is the same object as in the question figure, but rotated. The test included **15 items** each with a **20 seconds** time limit. Participants are discontinued if they provide 4 consecutive incorrect answers.	**TEDS siblings pilot**: reasonably normally distributed, M = 9.01, SD = 3.30, N = 154. Test-retest *r*= .56, N = 98, p< .001.	Adapted from Spatial reasoning section 1 in the "How2become" booklet (https://www.how2become.com/testing/spatial-reasoning-tests/)
11. Paper-folding test	On the left-hand side of the page participants are	**TEDS sibling pilot:** Normally distributed,	Adapted from

	shown a sheet of paper folded following several stages. The last image of the sequence includes a dot. This dot represents a hole that is punched through all the thickness of the paper at that point. Participants are asked to identify which one of the 5 pictures on the right-hand side shows where the holes will be when the paper is completely unfolded again (by reversing the specific steps shown). The test included **15 items** each to be completed within a **20 second** time limit. Participants are discontinued if they provide 4 consecutive incorrect answers.	N = 166, M = 8.83, SD = 3.3. Test-retest correlation *r* = .58, N = 104, p< .001.	University of Otago, New Zealand http://www.cs.otago.ac.nz/brace/resources/Paper%20Folding%20Test%20Vz-2-BRACE%20Version%2007.pdf
15. Cross-section (2) 	Participants are asked to identify the shape that the cutting plane will produce when cutting through several symmetrical solids (see example figures). The plane can cut the solid vertically, horizontally or obliquely. The test included **15 items** each to be completed within a **20 second** time limit. Participants are discontinued if they provide 4 consecutive incorrect answers.	**TEDS sibling pilot:** normally distributed, M = 7.67, SD = 2.8, N = 159. Test-retest r= .64, N = 91, p< .001.	Adapted from Ormand, C. J., Shipley, T. F., Tikoff, B., Manduca, C. A., Dutrow, B., Goodwin, L., Hickson, T., Atit, K., Gagnier, K. M., & Resnick, I. (2013). Improving Spatial Reasoning Skills in the Undergraduate Geoscience Classroom Through Interventions Based on Cognitive Science Research. Talk presented at the AAPG Hedberg Conference on 3D Structural Geologic Interpretation.
13. Perspective taking (2) –AKA "The cube"	Participants are presented with a transparent cube containing an irregular polygon suspended in the middle of the cube (see example figure). The same polygon is also presented outside the cube from a different viewpoint. Participants are asked to indicate on which corner of the cube they would have to stand in order to see the polygon	**TEDS sibling pilot:** normally distributed, M = 6.61, SD 3.34, N = 147. Test-retest *r*= .56, N = 92, p< .001.	Adapted from Hegarty, M., Keehner, M., Khooshabeh, P., & Montello, D. R. (2009). How spatial abilities

	from the new viewpoint (e.g. the bottom right corner in the example figure). The test included **15 items** each to be completed within a **20 second** time limit. Participants are discontinued if they provide 5 consecutive incorrect answers.		enhance, and are enhanced by, dental education. *Learning and Individual Differences*, 19(1), 61-70. Keehner, M., Hegarty, M., Cohen, C. A., Khooshabeh, P., & Montello, D. R. (2008). Spatial reasoning with external visualizations: What matters is what you see, not whether you interact. *Cognitive Science*, 32(7), 1099–1132.
17. Mechanical reasoning 	Examples of questions are: "Which shaft will turn more quickly?" (See example picture) and "If only the right oar of the boat is pulled, in which direction will the boat go?". The test included **16 items** each to be completed within a **25 second** time limit. Participants are required to complete every item.	**TEDS sibling pilot:** close to normally distributed, M = 9.53, SD 2.25, N = 180. Test retest *r*= .65, N = 113, p< .001.	Adapted from Harcourt Assessment (1995), DAT for Selection-Technical Abilities Battery. Pearson Assessment: London Wiesen, J. (2009), Barron's Mechanical Aptitude and Spatial Relations Test, 2nd edition, Barron's Educational Series Wiesen, J. (2009), Barron's Mechanical

			Aptitude and Spatial Relations Test, 2nd edition, Barron's Educational Series
20. Elithorn Maze 	Participants are asked to trace their route on each one of the triangular grids presented. The aim of the task is to trace the route passing through the largest possible number of black dots. Participants start from the bottom of the triangle and move upwards at a fixed speed; they can only move left or right on the grid, changing direction as desired at each intersectionI, It is not possible to collect all the dots in the grid. The test included **10 items**, each to be completed within **7 second**s. Participants are required to complete every item.	**TEDS sibling pilot**: fairly normally distributed, M = 7.31, SD = 1.94, N = 184. Test-retest *r*= .69, N = 117, p< .001.	Adapted from Test of spatial planning ability included as a process subtest of the WISC-IV Integrated. ELITHORN, A. (1955). A preliminary report on a perceptual maze test sensitive to brain damage. J*. neurol. neurosurg. Psychiat*, 18, 287-292.
21.1 2D to 3D drawing 	Participants are presented with the coded plan (see left-hand side of the example picture) and are asked to draw the 3D object corresponding to the plan (see right-hand side of the example picture), by clicking on dots arranged in an isometric grid. This test included **5 items**, each with a time limit of **70 seconds**. Participants are required to complete every item.	**TEDS sibling pilot:** fairly normal distribution, M = 3.46, SD = 1.69, N = 155. Test-retest *r* = .63, N = 99, p< .001.	Developed by the team
21.2 3D to 2D viewpoints 	Participants are asked to draw the viewpoint indicated in the picture of the 3D solid as the 'front' (see example figure), by clicking on dots arranged in a square grid. The drawing that participants should produce is a 2D viewpoint of the 3D shape. This test included **5 items**, each had a time limit of **45 seconds**. Participants are required to complete every item.	**TEDS sibling pilot:** slightly negatively skewed distribution M = 3.73, SD .99, N = 186. Test-retest correlation *r*= .68, N= 117, p< .001.	

25. Mazes task  	Participants are presented with a series of mazes, each with multiple ways in and out, but with only one valid route connecting one of the entrances to one of the exits. Participants are asked to look at the map (see example picture on the left) and choose from the options available the valid route between a single entrance and exit. The test includes **10 items** with increasing difficulty, each with a **25 second** time limit, discontinuing after 4 consecutive incorrect responses	**TEDS sibling pilot:** Normal distribution, M = 5.92, SD = 1.76, N = 167. Test-retest correlation *r*= .48, N = 106, p< .001.	Developed by the team

Appendix 4

Genetic specificity of face recognition

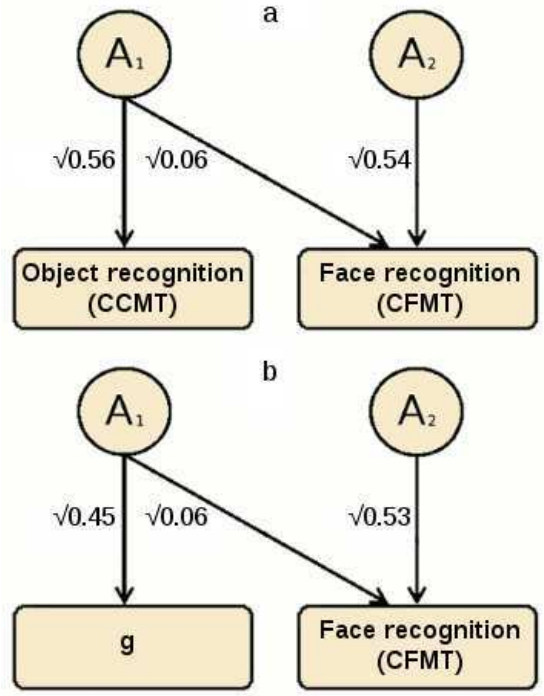
Nicholas G. Shakeshaft and Robert Plomin

Supplementary Information

Figures

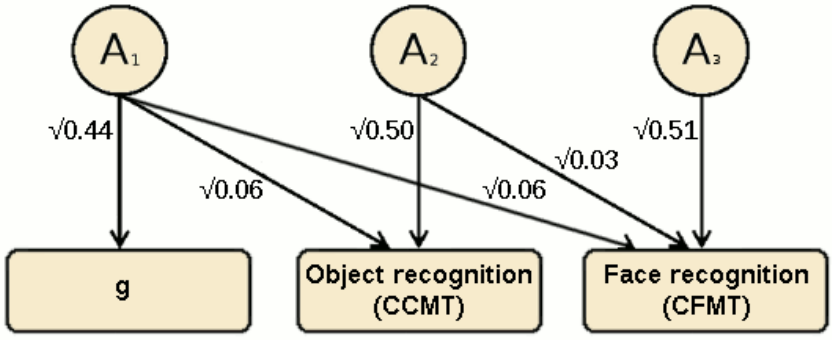
Tables

Fig. S1. Bivariate Cholesky A (additive genetic) path estimates.



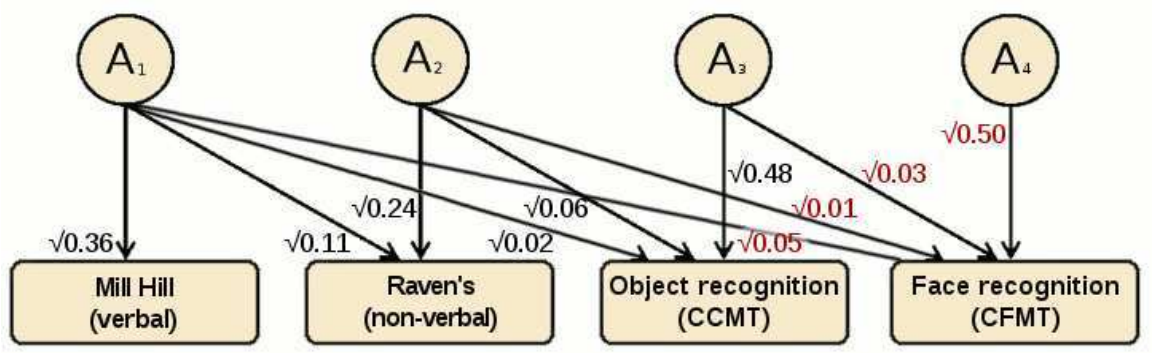
Path estimates for genetic influences on face recognition, shared with and independent from (a) general object recognition, and (b) *g*. Precise heritability estimates differ slightly between models depending on the parameters used.

Fig. S2. Multivariate Cholesky A (additive genetic) path estimates (with *g* composite).



Path estimates for genetic influences on face recognition shared with both object recognition and *g*, those shared only with object recognition (not with *g*), and those unique to face recognition.

Fig. S3. Multivariate Cholesky A (additive genetic) path estimates (with separate *g* components).



Path estimates for shared/unique genetic influences among the Mill Hill Vocabulary Scale and Raven's Progressive Matrices (the two measures forming the *g* composite used in the other analyses presented), object recognition and face recognition measures. Genetic influences on face recognition (shown in red) are those shared with all variables, those shared with Raven's and object recognition independent from Mill Hill, those shared with object recognition independent from both other measures, and finally those unique to face recognition, the latter accounting for the large majority of its genetic variance.

Table S1. Reliability of face and object recognition measures.

	Cronbach's alpha	
	Cambridge Face Memory Test (N=1068)	Cambridge Car Memory Test (N=1042)
Memorization (18 items)	0.650	0.550
Test (clean) (30 items)	0.837	0.782
Test (degraded) (24 items)	0.775	0.752
Full test (72 items)	**0.893**	0.875

Reliability (Cronbach's alpha) for the Cambridge Face Memory Test and Cambridge Car Memory Test, for the complete tasks and separately by test phase (see Methods). The sample is fully independent, with one individual selected randomly from each twin pair.

Table S2. Univariate model-fitting results.

	A	C	E
Face recognition	0.61 (0.52 – 0.66)	*0.00 (0.00 – 0.06)*	0.39 (0.34 – 0.46)
Object recognition	0.56 (0.45 – 0.62)	*0.00 (0.00 – 0.08)*	0.44 (0.38 – 0.51)
g	0.48 (0.30 – 0.64)	*0.10 (0.00 – 0.27)*	0.43 (0.36 – 0.51)

Model-fitting estimates (95% confidence intervals) for additive genetic (A), shared environmental (C) and residual (E; i.e., non-shared environment and error) components of variance. Italicized estimates are non-significant (their confidence intervals include zero).

Table S3. Univariate model fit statistics.

	Model	ep	χ2	df	AIC	Δχ2	Δdf	p
Face recognition	Saturated	10	5866.94	2131	1604.94	-	-	-
	ACE	4	5875.37	2137	1601.37	8.43	6	0.21
Object recognition	Saturated	10	5746.09	2076	1594.09	-	-	-
	ACE	4	5749.48	2082	1585.48	3.38	6	0.76
g	Saturated	10	4131.44	1504	1123.44	-	-	-
	ACE	4	4140.54	1510	1120.54	9.11	6	0.17

Comparison of univariate ACE models to fully saturated models. ep = estimated parameters; χ2 = -2 log-likelihood; df = degrees of freedom, AIC = Akaike information criterion. The p-values indicate no significant deterioration in fit between the saturated and constrained models.

Table S4. Bivariate correlated factors solution.

	Variance component estimates			**Component correlations**		
	A	C	E	rA	rC	rE
Object recognition	0.66 (0.44 – 0.84)	*0.00 (-0.11 – 0.15)*	0.34 (0.19 – 0.51)	0.31 (0.21 – 0.43)	*1.00 (-1.00 – 1.00)*	0.23 (0.13 – 0.32)
g	0.88 (0.34 – 1.42)	*-0.12 (-0.54 – 0.29)*	*0.24 (-0.01 – 0.52)*	0.32 (0.12 – 0.58)	*-1.00 (-1.00 – 1.00)*	*0.11 (-0.01 – 0.22)*

Proportions (with 95% confidence intervals) of phenotypic correlations with face recognition due to additive genetic (A), shared environmental (C) or non-shared environmental/error (E) components, and correlations between the traits for each of these components (rA, rC and rE, respectively). The proportions explained reflect the correlation between the traits for that component, weighted by the two univariate component estimates. For example, the proportion of the phenotypic correlation due to A equals the genetic correlation (rA) weighted by the product of the square roots of the two univariate heritabilities estimated by the model. Italicized estimates are non-significant (their confidence intervals include zero).

Table S5. Bivariate Cholesky decompositions of face recognition.

	Shared	Unique
Predictor: object recognition	0.06 (0.03 – 0.11)	0.54 (0.45 – 0.60)
Predictor: *g*	0.06 (0.01 – 0.16)	0.53 (0.43 – 0.62)

Portions (with 95% confidence intervals) of the heritability of face recognition (estimated at 60% in these models) due to genetic influences shared with, and unique from, each predictor variable.

Table S6. Bivariate Cholesky decomposition of object recognition.

	Shared	Unique
Predictor: *g*	0.06 (0.004 – 0.15)	0.50 (0.30 – 0.58)

Portions (with 95% confidence intervals) of the heritability of object recognition (estimated at 55% in this model) due to genetic influences shared with, and unique from, general cognitive ability.

Table S7. Bivariate model fit statistics.

	Model	ep	χ2	df	AIC	Δχ2	Δdf	p
Object recognition & face recognition	Saturated	28	11450.84	4199	3052.85	-	-	-
	ACE	11	11468.93	4216	3036.93	18.09	17	1.00
g & face recognition	Saturated	28	9943.99	3627	2689.99	-	-	-
	ACE	11	9969.62	3644	2681.62	25.63	17	0.08
g & object recognition	Saturated	28	9842.82	3572	2698.82	-	-	-
	ACE	11	9863.80	3589	2685.80	20.99	17	0.23

Comparison of bivariate ACE models to fully saturated models. ep = estimated parameters; χ2 = -2 log-likelihood; df = degrees of freedom, AIC = Akaike information criterion. The p-values indicate no significant deterioration in fit between the saturated and constrained models.

Table S8. Multivariate Cholesky decomposition of face recognition (with *g* composite).

Shared with *g* and object recognition	Shared with object recognition	Unique to face recognition
0.06 (0.01 – 0.14)	*0.03 (0.00 – 0.07)*	0.51 (0.45 – 0.58)

Portions (with 95% confidence intervals) of the heritability of face recognition (estimated at 60% in this model) due to genetic influences i) shared between all three variables, ii) shared only with object recognition (i.e., independent from *g*), and iii) unique to face recognition (not shared with either variable). The italicized estimate is non-significant (its confidence intervals includes zero).

Table S9. Multivariate model fit statistics (trivariate model with *g* composite).

	Model	ep	χ2	df	AIC	Δχ2	Δdf	p
Trivariate model	Saturated	54	15507.77	5687	4133.77	-	-	-
	ACE	21	15549.45	5720	4109.45	41.69	33	0.14
	ACE sub.	20	15552.61	5721	4110.61	3.16	1	0.08

Comparison of trivariate ACE model to fully saturated model, and a further-constrained ACE sub-model to the primary ACE model. ep = estimated parameters; χ2 = -2 log-likelihood; df = degrees of freedom, AIC = Akaike information criterion. In the sub-model, the non-significant shared genetic path was deleted (constrained to zero). There was no significant deterioration in fit in either comparison.

Table S10. Multivariate Cholesky decomposition of face recognition (with separate *g* components).

Shared with Mill Hill, Raven's and object recognition	Shared with Raven's and object recognition	Shared with object recognition	Unique to face Recognition
0.05 (0.004 – 0.25)	*0.01 (0.00 – 0.43)*	*0.03 (0.00 – 0.59)*	0.50 (*0.00* – 0.58)

Portions (with 95% confidence intervals) of the heritability of face recognition (estimated at 59% in this model) due to genetic influences shared with and independent from the Mill Hill Vocabulary Scale and Raven's Progressive Matrices (the two measures forming the *g* composite used in the other analyses presented), and object recognition. The genetic influences shown are those i) shared between all four variables, ii) shared only with Raven's and object recognition (i.e., independent from Mill Hill), iii) shared only with object recognition, and iv) unique to face recognition (not shared with any other variable). Italicized estimates are non-significant (their confidence intervals include zero) – N.B. The lower bound for the last estimate (genetic influences unique to face recognition) is also only very marginally above zero, perhaps indicating that the sample is underpowered for this model.

Table S11. Multivariate model fit statistics (quadrivariate model with separate *g* components).

	Model	ep	χ2	df	AIC	Δχ2	Δdf	p
Quadrivariate model	Saturated	88	19969.92	7320	5329.92	-	-	-
	ACE	34	20039.97	7374	5291.97	70.05	54	0.07
	ACE sub.	33	20043.20	7375	5293.20	3.23	1	0.07

Comparison of quadrivariate ACE model to fully saturated model, and a further-constrained ACE sub-model to the primary ACE model. ep = estimated parameters; χ2 = -2 log-likelihood; df = degrees of freedom, AIC = Akaike information criterion. In the sub-model, the non-significant shared genetic paths were deleted (constrained to zero). There was no significant deterioration in fit in either comparison.

Table S12. Descriptive statistics for Caucasian subsample.

	N	Mean (SD)
Face recognition	998	54.15 (9.45)
Object recognition	976	50.65 (9.82)
g	712	0.04 (0.94)

Mean scores (standard deviations) for Caucasian participants. N = sample size (sample shown is fully independent, randomly selecting one individual per twin pair).

Table S13. Bivariate Cholesky decompositions for Caucasian subsample.

	Shared	Unique
Predictor: object recognition	0.06 (0.03 – 0.11)	0.55 (0.45 – 0.61)
Predictor: *g*	0.08 (0.01 – 0.24)	0.52 (0.36 – 0.62)

Portions (with 95% confidence intervals) of the heritability of face recognition for Caucasian participants, (estimated at 60% in these models) due to genetic influences shared with, and unique from, each predictor variable.

Table S14. Bivariate model fit statistics for Caucasian subsample.

	Model	ep	χ2	df	AIC	Δχ2	Δdf	p
Object recognition & face recognition	Saturated	28	10693.33	3917	2859.33	-	-	-
	ACE	11	10709.33	3934	2841.33	16.00	17	0.52
g & face recognition	Saturated	28	9249.25	3390	2469.25	-	-	-
	ACE	11	9276.12	3407	2462.12	26.87	17	0.06

Comparison of bivariate ACE models to fully saturated models. ep = estimated parameters; χ2 = -2 log-likelihood; df = degrees of freedom, AIC = Akaike information criterion. The p-values indicate no significant deterioration in fit between the saturated and constrained models.

Appendix 5

STEM is spatial, English is social: genetic dissociations in the prediction of educational achievement

Nicholas G. Shakeshaft, Kaili Rimfield, Margherita Malanchini, Kerry L. Schofield, Maja Rodic, Saskia Selzam & Robert Plomin

Supplementary Information

Table S1. Descriptive statistics.

	N	Whole sample	Males	Females	MZs	DZs	Sex	Zyg	Sex x Zyg	R^2
Bricks	1339	5.52 (1.21)	5.84 (1.24)	5.35 (1.16)	5.52 (1.16)	5.53 (1.24)	61.07 **	0.00	0.01	0.04
King's Challenge	813	0.01 (1.01)	0.50 (0.93)	-0.29 (0.93)	-0.06 (1.03)	0.05 (0.99)	140.14 **	1.54	0.28	0.15
Face recognition	911	54.12 (9.52)	53.33 (9.62)	54.63 (9.44)	53.75 (9.40)	54.35 (9.61)	3.54	1.61	0.10	0.01
Object recognition	901	50.18 (10.20)	53.89 (10.56)	47.78 (9.19)	49.86 (10.25)	50.38 (10.17)	84.85 **	0.04	1.14	0.09
"Pure" face recognition	901	0.00 (0.99)	-0.20 (1.02)	0.13 (0.95)	-0.03 (0.98)	0.02 (1.00)	23.84 **	1.44	0.17	0.03
Verbal ability	1702	15.61 (3.95)	15.67 (3.99)	15.58 (3.94)	15.27 (3.78)	15.83 (4.05)	0.18	7.40 **	1.72	0.01
Overall mean GCSE grade	1934	9.11 (1.10)	9.09 (1.10)	9.13 (1.10)	9.13 (1.10)	9.11 (1.10)	0.81	0.29	0.27	0.00
Number at grade A*-C	1946	8.72 (2.85)	8.66 (2.87)	8.75 (2.84)	8.70 (2.81)	8.73 (2.88)	0.56	0.30	0.96	0.00
"Core" subjects	1928	9.13 (1.14)	9.20 (1.11)	9.09 (1.16)	9.12 (1.15)	9.14 (1.14)	3.27	0.02	0.59	0.00
Humanities	1689	9.21 (1.30)	9.10 (1.34)	9.27 (1.27)	9.24 (1.28)	9.18 (1.31)	7.32 **	0.67	0.35	0.01
English	1931	9.15 (1.14)	9.02 (1.17)	9.24 (1.12)	9.15 (1.17)	9.15 (1.12)	17.18 **	0.01	0.20	0.01
Science	1846	9.19 (1.23)	9.33 (1.17)	9.10 (1.27)	9.20 (1.22)	9.19 (1.24)	15.17 **	0.16	0.40	0.01
Maths	1908	9.14 (1.33)	9.34 (1.23)	9.01 (1.37)	9.10 (1.33)	9.16 (1.32)	26.14 **	0.45	1.16	0.01

Mean scores (standard deviations) for the whole sample, separately by sex, and for MZ and DZ twins, for the five key predictor variables, verbal ability, and the GCSE composites. N = sample size (the sample shown is fully independent, selecting one individual at random per twin pair). ANOVA performed on cleaned, normality-transformed data to test the effects of sex and zygosity. Results = F statistic; ** = $p < 0.01$; R^2 = proportion of variance explained by sex, zygosity and their interaction.

Table S2. Predictor intercorrelations.

		Bricks	KC	Face recognition	Object recognition	"Pure" faces
Bricks	*r*	1				
	N	*1439*				
King's Challenge	*r*	0.65 **	1			
	N	*692*	*877*			
Face recognition	*r*	0.16 **	*0.07*	1		
	N	*513*	*335*	*1076*		
Object recognition	*r*	0.31 **	0.21 **	0.32 **	1	
	N	*510*	*332*	*1063*	*1064*	
"Pure" face recognition	*r*	*0.06*	*0.01*	0.95 **	*0.00*	1
	N	*510*	*332*	*1063*	*1063*	*1063*

Correlations (Pearson's *r*) between the predictor variables. The sample is fully independent, with one individual selected randomly from each twin pair. N = sample size, ** = $p < 0.01$.

Table S3. Predictor intercorrelations, controlling verbal ability.

		Bricks	KC	Face recognition	Object recognition	"Pure" faces
Bricks	*r*	1				
	df	*0*				
King's Challenge	*r*	0.62 **	1			
	df	*689*	*0*			
Face recognition	*r*	0.15 **	*0.05*	1		
	df	*510*	*332*	*0*		
Object recognition	*r*	0.31 **	0.20 **	0.32 **	1	
	df	*507*	*329*	*765*	*0*	
"Pure" face recognition	*r*	*0.04*	*-0.01*	0.95 **	*-0.00*	1
	df	*507*	*329*	*765*	*765*	*0*

Partial correlations (Pearson's *r*) between the predictor variables, controlling for verbal ability. The sample is fully independent, with one individual selected randomly from each twin pair. df = degrees of freedom, ** = p < 0.01.

Table S4. GCSE intercorrelations.

		GCSE mean	No. A*-C	Core subjects	Humanities	English	Science	Maths
Overall mean GCSE grade	*r*	1						
	N	*6320*						
Number at grade A*-C	*r*	0.75 **	1					
	N	6320	6375					
Core subjects	*r*	0.96 **	0.73 **	1				
	N	*6265*	*6270*	*6270*				
Humanities	*r*	0.84 **	0.61 **	0.75 **	1			
	N	*5356*	*5364*	*5330*	*5364*			
English	*r*	0.88 **	0.69 **	0.89 **	0.72 **	1		
	N	*6262*	*6272*	*6246*	*5336*	*6272*		
Science	*r*	0.90 **	0.65 **	0.92 **	0.69 **	0.73 **	1	
	N	*5825*	*5828*	*5820*	*5005*	*5801*	*5828*	
Maths	*r*	0.84 **	0.65 **	0.90 **	0.63 **	0.69 **	0.79 **	1
	N	*6201*	*6215*	*6185*	*5281*	*6163*	*5745*	*6215*

Correlations (Pearson's *r*) between the GCSE variables. The sample is fully independent, with one individual selected randomly from each twin pair. N = sample size, ** = p < 0.01.

Table S5. GCSE intercorrelations, controlling verbal ability.

		GCSE mean	No. A*-C	Core subjects	Humanities	English	Science	Maths
Overall mean GCSE grade	*r*	1						
	df	*0*						
Number at grade A*-C	*r*	0.70 **	1					
	df	*2259*	*0*					
Core subjects	*r*	0.95 **	0.67 **	1				
	df	*2245*	*2245*	*0*				
Humanities	*r*	0.81 **	0.54 **	0.69 **	1			
	df	*1955*	*1955*	*1955*	*0*			
English	*r*	0.85 **	0.61 **	0.85 **	0.66 **	1		
	df	*2250*	*2250*	*2245*	*1955*	*0*		
Science	*r*	0.88 **	0.57 **	0.90 **	0.63 **	0.66 **	1	
	df	*2128*	*2128*	*2128*	*1955*	*2128*	*0*	
Maths	*r*	0.80 **	0.58 **	0.88 **	0.55 **	0.62 **	0.75 **	1
	df	*2225*	*2225*	*2225*	*1955*	*2225*	*2128*	*0*

Partial correlations (Pearson's *r*) between the GCSE variables, controlling for verbal ability. The sample is fully independent, with one individual selected randomly from each twin pair. df = degrees of freedom, ** = p < 0.01.

Table S6. Correlations between predictors and GCSEs, controlling verbal ability.

		GCSE mean	No. A*-C	Core subjects	Humanities	English	Science	Maths
Bricks	*r*	0.29 **	0.26 **	0.33 **	0.19 **	0.21 **	0.29 **	0.36 **
	df	*1323*	*1333*	*1324*	*1162*	*1325*	*1271*	*1314*
King's Challenge	*r*	0.34 **	0.25 **	0.34 **	0.21 **	0.19 **	0.36 **	0.45 **
	df	*805*	*806*	*802*	*713*	*802*	*783*	*799*
Face recognition	*r*	0.15 **	0.13 **	0.13 **	0.13 **	0.15 **	0.10 **	0.10 **
	df	*771*	*771*	*771*	*771*	*771*	*771*	*771*
Object recognition	*r*	*0.02*	0.10 **	*0.07*	*0.02*	*-0.01*	*0.05*	0.11 **
	df	*765*	*765*	*765*	*762*	*765*	*765*	*765*
"Pure" face recognition	*r*	0.15 **	0.10 **	0.12 **	0.13 **	0.17 **	0.10 **	*0.07*
	df	*765*	*765*	*765*	*762*	*765*	*765*	*765*

Partial correlations (Pearson's *r*) between the predictors and GCSEs, controlling for verbal ability. The sample is fully independent, with one individual selected randomly from each twin pair. df = degrees of freedom ** = p < 0.01, * = p < 0.05.

Table S7. Univariate model-fitting results and sample sizes.

	Variance component estimates			Sample size			
				Paired		Unpaired	
	A	C	E	MZ	DZ	MZ	DZ
Bricks	0.55 (0.43 – 0.60)	*0.00 (0.00 – 0.10)*	0.45 (0.40 – 0.50)	487	679	100	278
King's Challenge	0.69 (0.50 – 0.80)	*0.07 (0.00 – 0.24)*	0.24 (0.20 – 0.29)	209	271	223	458
Face recognition	0.61 (0.52 – 0.66)	*0.00 (0.00 – 0.06)*	0.39 (0.34 – 0.46)	331	476	50	154
Object recognition	0.58 (0.47 – 0.64)	*0.00 (0.00 – 0.09)*	0.42 (0.36 – 0.48)	327	470	52	154
"Pure" face recognition	0.59 (0.50 – 0.65)	*0.00 (0.00 – 0.07)*	0.41 (0.35 – 0.47)	327	470	52	154
Verbal ability	0.45 (0.33 – 0.56)	*0.06 (0.00 – 0.17)*	0.48 (0.44 – 0.53)	675	1072	67	178
Overall mean GCSE grade	0.60 (0.56 – 0.64)	0.29 (0.25 – 0.32)	0.11 (0.10 – 0.12)	803	1326	9	40
Number at grade A*-C	0.60 (0.55 – 0.64)	0.24 (0.19 – 0.28)	0.17 (0.16 – 0.18)	808	1340	6	28
Core subjects	0.62 (0.58 – 0.66)	0.27 (0.22 – 0.30)	0.12 (0.11 – 0.13)	799	1314	11	49
Humanities	0.54 (0.48 – 0.60)	0.21 (0.16 – 0.27)	0.25 (0.24 – 0.27)	653	1029	107	267
English	0.61 (0.56 – 0.65)	0.21 (0.17 – 0.26)	0.18 (0.17 – 0.19)	799	1314	11	49
Science	0.61 (0.57 – 0.66)	0.21 (0.17 – 0.26)	0.17 (0.16 – 0.18)	751	1210	31	131
Maths	0.63 (0.58 – 0.68)	0.19 (0.14 – 0.23)	0.18 (0.17 – 0.19)	788	1303	20	58

Model-fitting estimates (95% confidence intervals) for additive genetic (A), shared environmental (C) and residual (E; i.e., non-shared environment and error) components of variance, for the five key predictor variables, verbal ability, and the GCSE composites. Italicised estimates are non-significant (their confidence intervals include zero). The available sample of monozygotic (MZ) and dizygotic (DZ) twins is shown, following exclusions and data cleaning as described in Methods, for complete pairs and for unpaired individuals – i.e., those whose co-twin either did not provide data or was lost during data cleaning/preparation.

Table S8. Fit statistics: univariate models for each predictor and GCSE variable.

	Model	ep	χ2	df	AIC	Δχ2	Δdf	p
Bricks	Saturated	10	7974.81	2889	2196.80	-	-	-
	ACE	4	7979.20	2895	2189.20	4.39	6	0.62
	AE	3	7979.20	2896	2187.20	0.00	1	1.00
King's Challenge	Saturated	10	4715.08	1742	1231.08	-	-	-
	ACE	4	4721.22	1748	1225.22	6.14	6	0.41
	AE	3	4721.67	1749	1223.67	0.45	1	0.50
Face recognition	Saturated	10	5884.36	2137	1610.36	-	-	-
	ACE	4	5893.41	2143	1607.41	9.05	6	0.17
	AE	3	5893.41	2144	1605.41	0.00	1	1.00
Object recognition	Saturated	10	5819.36	2112	1595.37	-	-	-
	ACE	4	5821.26	2118	1585.26	1.89	6	0.93
	AE	3	5821.26	2119	1583.26	0.00	1	1.00
"Pure" face recognition	Saturated	10	5823.54	2111	1601.54	-	-	-
	ACE	4	5832.72	2117	1598.72	9.18	6	0.16
	AE	3	5832.72	2118	1596.72	0.00	1	1.00
Verbal ability	Saturated	10	14226.58	5142	3942.58	-	-	-
	ACE	4	14240.59	5148	3944.59	14.01	6	0.03
	AE	3	14242.01	5149	3944.01	1.42	1	0.23
Overall mean GCSE grade	Saturated	10	30731.03	12619	5493.03	-	-	-
	ACE	4	30734.85	12625	5484.85	3.81	6	0.70
	AE	3	30893.93	12626	5641.93	159.08	1	0.00
Number at grade A*-C	Saturated	10	32158.13	12736	6686.13	-	-	-
	ACE	4	32163.92	12742	6679.92	5.80	6	0.45
	AE	3	32264.82	12743	6778.82	100.90	1	0.00
Core subjects	Saturated	10	30787.65	12541	5705.65	-	-	-
	ACE	4	30791.32	12547	5697.32	3.67	6	0.72
	AE	3	30923.30	12548	5827.30	131.97	1	0.00
Humanities	Saturated	10	28258.80	10721	6816.80	-	-	-
	ACE	4	28264.39	10727	6810.39	5.59	6	0.47
	AE	3	28315.42	10728	6859.42	51.03	1	0.00
English	Saturated	10	31932.31	12528	6876.31	-	-	-
	ACE	4	31943.92	12534	6875.92	11.61	6	0.07
	AE	3	32020.69	12535	6950.69	76.77	1	0.00
Science	Saturated	10	29647.45	11651	6345.45	-	-	-
	ACE	4	29649.84	11657	6335.84	2.39	6	0.88
	AE	3	29718.68	11658	6402.68	68.84	1	0.00
Maths	Saturated	10	31797.16	12449	6899.16	-	-	-
	ACE	4	31802.34	12455	6892.34	5.18	6	0.52
	AE	3	31859.41	12456	6947.41	57.06	1	0.00

Comparison of univariate model fit statistics. The ACE model is compared to the fully saturated model, and the AE submodel (dropping shared environment) is compared to the ACE model. ep = estimated parameters; χ2 = -2 log-likelihood; df = degrees of freedom, AIC = Akaike information criterion. The p-values indicate whether there is a significant deterioration in fit (i.e., whether the second model is a worse fit than the previous one).

N.B. All of the ACE models fit well apart from the model for verbal ability; this seems likely to be a chance effect, given the large number of models tested. For all predictors, the AE submodel shows no deterioration in fit compared to the ACE model. For all GCSE measures, the full ACE model is by far the better fit.

Table S9. Decomposition of phenotypic correlations among predictor variables.

Variables in model	Variance component estimates	
	A	E
Bricks King's Challenge	0.89 (0.83 – 0.94)	0.11 (0.06 – 0.17)
"Pure" face recognition Bricks	1.00 (0.44 – 1.70)	*-0.00* *(-0.70 – 0.56)*
"Pure" face recognition King's Challenge	*0.06* *(-1.00 – 1.00)*	*0.94* *(-1.00 – 1.00)*

Bivariate correlated factors solutions of three models, representing each pair of predictor variables. Results indicate the phenotypic correlations between the two composites in each model, decomposed into the proportions attributable to additive genetic (A) or non-shared environmental/error (E) components (with 95% confidence intervals). Shared environmental relationships (C) are excluded in the model by design (see Methods). Italicised estimates are non-significant (their CIs include zero).

N.B. The results from the two models with face recognition are included for completeness, but little can be concluded from them as the phenotypic correlations (of which these figures are proportions) are themselves non-significant.

Table S10. Genetic correlations among predictor variables.

	Bricks	KC	Faces
Bricks	1		
King's Challenge	0.98 (0.87 – 1.00)	1	
"Pure" face recognition	0.20 (0.05 – 0.34)	*0.00* *(-0.15 – 0.15)*	1

Genetic correlations (95% confidence intervals) among the predictor variables. Italicised estimates are non-significant (their CIs include zero).

Table S11. Non-shared environmental correlations among predictor variables.

	Bricks	KC	Faces
Bricks	1		
King's Challenge	0.23 (0.12 – 0.34)	1	
"Pure" face recognition	*-0.00* *(-0.13 – 0.13)*	*0.09* *(-0.09 – 0.27)*	1

Non-shared environmental correlations (95% confidence intervals) among the predictor variables. Italicised estimates are non-significant (their CIs include zero).

Table S12. Bivariate Cholesky decomposition: Bricks, King's Challenge.

	Genetic paths		**Shared environment paths**		**Non-shared environment paths**	
	Bricks	KC	Bricks	KC	Bricks	KC
1. Bricks	0.54		*0.00*		0.46	
	(0.49 – 0.59)		*(constrained)*		(0.41 – 0.51)	
2. KC	0.65	*0.03*	*0.00*	*0.09*	*0.01*	0.22
	(0.57 – 0.73)	*(0.00 – 0.19)*	*(constrained)*	*(0.00 – 0.16)*	*(0.00 – 0.03)*	(0.18 – 0.26)

Path estimates (standardised and squared, with 95% confidence intervals) for bivariate ACE Cholesky decomposition. Influences on the second variable (King's Challenge) are decomposed into influences shared with the first (Bricks), and those unique to the second. Italicised estimates are non-significant (their CIs include zero).

N.B. The independent shared environment path for the second variable was retained in this model for consistency with the other analyses presented, but there are no significant shared environmental influences on either measure.

Table S13. Bivariate Cholesky decomposition: “pure” face recognition, Bricks.

	Genetic paths		**Shared environment paths**		**Non-shared environment paths**	
	Faces	Bricks	Faces	Bricks	Faces	Bricks
1. Faces	0.59		*0.00*		0.41	
	(0.53 – 0.65)		*(constrained)*		(0.35 – 0.47)	
2. Bricks	*0.02*	0.53	*0.00*	*0.00*	*0.00*	0.45
	(0.00 – 0.06)	(0.40 – 0.59)	*(constrained)*	*(0.00 – 0.10)*	*(0.00 – 0.01)*	(0.40 – 0.50)

Path estimates (standardised and squared, with 95% confidence intervals) for bivariate ACE Cholesky decomposition. Influences on the second variable (Bricks) are decomposed into influences shared with the first (“pure” face recognition), and those unique to the second. Italicised estimates are non-significant (their CIs include zero).

N.B. The independent shared environment path for the second variable was retained in this model for consistency with the other analyses presented, but there are no significant shared environmental influences on either measure.

Table S14. Bivariate Cholesky decomposition: “pure” face recognition, King's Challenge.

	Genetic paths		**Shared environment paths**		**Non-shared environment paths**	
	Faces	KC	Faces	KC	Faces	KC
1. Faces	0.59		*0.00*		0.41	
	(0.53 – 0.65)		*(constrained)*		(0.35 – 0.47)	
2. KC	*0.00*	0.69	*0.00*	*0.07*	*0.00*	0.24
	(0.00 – 0.01)	(0.50 – 0.80)	*(constrained)*	*(0.00 – 0.24)*	*(0.00 – 0.02)*	(0.20 – 0.29)

Path estimates (standardised and squared, with 95% confidence intervals) for bivariate ACE Cholesky decomposition. Influences on the second variable (King's Challenge) are decomposed into influences shared with the first (“pure” face recognition), and those unique to the second. Italicised estimates are non-significant (their CIs include zero).

N.B. The independent shared environment path for the second variable was retained in this model for consistency with the other analyses presented, but there are no significant shared environmental influences on either measure.

Table S15. Fit statistics: bivariate models for predictor interrelationships.

Variables in model	Model	ep	χ2	df	AIC	Δχ2	Δdf	p
Bricks, KC	Saturated	28	11954.18	4623	2708.18	-	-	-
	ACE	11	11968.01	4640	2688.01	13.83	17	0.68
	Constrained	9	11968.12	4642	2684.13	0.12	2	0.94
Faces, Bricks	Saturated	28	13784.45	4992	3800.45	-	-	-
	ACE	11	13801.31	5009	3783.32	16.87	17	0.46
	Constrained	9	13801.31	5011	3779.32	0	2	1.00
Faces, KC	Saturated	28	10533.83	3845	2843.83	-	-	-
	ACE	11	10552.46	3862	2828.46	18.63	17	0.35
	Constrained	9	10552.64	3864	2824.64	0.18	2	0.91

Comparison of bivariate model fit statistics. The ACE model is compared to the fully saturated model, and the constrained submodel (with all but one shared environment path constrained to zero; see Methods) is compared to the ACE model. Variables were entered in the order specified. KC = King's Challenge, ep = estimated parameters; χ2 = -2 log-likelihood; df = degrees of freedom, AIC = Akaike information criterion. The p-values indicate whether there is a significant deterioration in fit (i.e., whether the second model is a worse fit than the previous one). All models fit well.

Table S16. Decomposition of phenotypic correlations between predictors and GCSEs.

Variables in model	**Variance component estimates**	
	A	E
Bricks Maths	0.90 (0.85 – 0.96)	0.10 (0.04 – 0.15)
Bricks Science	0.93 (0.86 – 0.99)	0.07 (0.01 – 0.14)
Bricks English	0.88 (0.80 – 0.96)	0.12 (0.04 – 0.20)
Bricks Core subjects	0.91 (0.86 – 0.95)	0.09 (0.05 – 0.14)
King's Challenge Maths	0.91 (0.86 – 0.96)	0.09 (0.04 – 0.14)
King's Challenge Science	0.85 (0.79 – 0.91)	0.15 (0.09 – 0.21)
King's Challenge English	0.92 (0.83 – 1.01)	*0.08* *(-0.01 – 0.17)*
King's Challenge Core subjects	0.90 (0.85 – 0.94)	0.10 (0.06 – 0.15)
"Pure" face recognition Maths	0.92 (0.50 – 1.41)	*0.08* *(-0.39 – 0.71)*
"Pure" face recognition Science	0.87 (0.51 – 1.14)	*0.13* *(-0.14 – 0.49)*
"Pure" face recognition English	0.93 (0.75 – 1.09)	*0.07* *(-0.09 – 0.25)*
"Pure" face recognition Core subjects	0.90 (0.67 – 1.10)	*0.10* *(-0.10 – 0.33)*

Bivariate correlated factors solutions of 12 separate models, representing each pairing of predictor variables and GCSE grades. Results indicate the phenotypic correlations between the two composites in each model, decomposed into the proportions attributable to additive genetic (A) or non-shared environmental/error (E) components (with 95% confidence intervals). Shared environmental relationships (C) are excluded in the model by design (see Methods). Italicised estimates are non-significant (their CIs include zero).

Table S17. Bivariate Cholesky decompositions: predictor variables, Maths GCSE.

	Genetic paths		**Shared environment paths**		**Non-shared environment paths**	
	Predictor	Maths	Predictor	Maths	Predictor	Maths
1. Bricks	0.55 (0.50 – 0.60)		*0.00 (constrained)*		0.45 (0.40 – 0.50)	
2. Maths	0.29 (0.24 – 0.35)	0.35 (0.28 – 0.42)	*0.00 (constrained)*	0.18 (0.14 – 0.22)	*0.00 (0.00 – 0.01)*	0.18 (0.16 – 0.19)
1. KC	0.78 (0.73 – 0.81)		*0.00 (constrained)*		0.22 (0.19 – 0.27)	
2. Maths	0.30 (0.25 – 0.36)	0.34 (0.27 – 0.41)	*0.00 (constrained)*	0.18 (0.14 – 0.22)	*0.01 (0.00 – 0.02)*	0.17 (0.15 – 0.18)
1. Faces	0.59 (0.53 – 0.59)		*0.00 (constrained)*		0.41 (0.35 – 0.47)	
2. Maths	*0.01 (0.00 – 0.02)*	0.62 (0.60 – 0.64)	*0.00 (constrained)*	0.19 (0.19 – 0.20)	*0.00 (0.00 – 0.00)*	0.18 (0.17 – 0.19)

Path estimates (standardised and squared, with 95% confidence intervals) for three bivariate ACE Cholesky decompositions, with each of the three predictors (Bricks, King's Challenge and “pure” face recognition) and Maths GCSE. Influences on the second variable in each model (Maths) are decomposed into influences shared with the first (the predictor), and those unique to the second. Italicised estimates are non-significant (their CIs include zero).

Table S18. Bivariate Cholesky decompositions: predictor variables, Science GCSE.

	Genetic paths		**Shared environment paths**		**Non-shared environment paths**	
	Predictor	Science	Predictor	Science	Predictor	Science
1. Bricks	0.55 (0.49 – 0.60)		*0.00 (constrained)*		0.45 (0.40 – 0.51)	
2. Science	0.25 (0.20 – 0.31)	0.37 (0.30 – 0.44)	*0.00 (constrained)*	0.20 (0.16 – 0.25)	*0.00 (0.00 – 0.01)*	0.17 (0.16 – 0.18)
1. KC	0.76 (0.70 – 0.80)		*0.00 (constrained)*		0.24 (0.20 – 0.30)	
2. Science	0.21 (0.17 – 0.27)	0.41 (0.34 – 0.48)	*0.00 (constrained)*	0.21 (0.16 – 0.25)	0.02 (0.01 – 0.04)	0.15 (0.13 – 0.17)
1. Faces	0.59 (0.53 – 0.65)		*0.00 (constrained)*		0.41 (0.35 – 0.47)	
2. Science	*0.02 (0.00 – 0.04)*	0.60 (0.55 – 0.65)	*0.00 (constrained)*	0.22 (0.17 – 0.26)	*0.00 (0.00 – 0.00)*	0.17 (0.16 – 0.18)

Path estimates (standardised and squared, with 95% confidence intervals) for three bivariate ACE Cholesky decompositions, with each of the three predictors (Bricks, King's Challenge and “pure” face recognition) and Science GCSE. Influences on the second variable in each model (Science) are decomposed into influences shared with the first (the predictor), and those unique to the second. Italicised estimates are non-significant (their CIs include zero).

Table S19. Bivariate Cholesky decompositions: predictor variables, English GCSE.

	Genetic paths		**Shared environment paths**		**Non-shared environment paths**	
	Predictor	English	Predictor	English	Predictor	English
1. Bricks	0.55 (0.50 – 0.60)		*0.00 (constrained)*		0.45 (0.40 – 0.50)	
2. English	0.13 (0.09 – 0.17)	0.49 (0.43 – 0.55)	*0.00 (constrained)*	0.21 (0.16 – 0.25)	*0.00 (0.00 – 0.01)*	0.18 (0.16 – 0.19)
1. KC	0.76 (0.71 – 0.80)		*0.00 (constrained)*		0.24 (0.20 – 0.29)	
2. English	0.10 (0.07 – 0.14)	0.51 (0.45 – 0.57)	*0.00 (constrained)*	0.21 (0.16 – 0.25)	*0.00 (0.00 – 0.01)*	0.18 (0.16 – 0.19)
1. Faces	0.59 (0.53 – 0.65)		*0.00 (constrained)*		0.41 (0.35 – 0.47)	
2. English	0.05 (0.02 – 0.08)	0.56 (0.50 – 0.61)	*0.00 (constrained)*	0.22 (0.17 – 0.26)	*0.00 (0.00 – 0.00)*	0.18 (0.17 – 0.19)

Path estimates (standardised and squared, with 95% confidence intervals) for three bivariate ACE Cholesky decompositions, with each of the three predictors (Bricks, King's Challenge and “pure” face recognition) and English GCSE. Influences on the second variable in each model (English) are decomposed into influences shared with the first (the predictor), and those unique to the second. Italicised estimates are non-significant (their CIs include zero).

Table S20. Bivariate Cholesky decompositions: predictor variables, “core” GCSE subjects.

	Genetic paths		**Shared environment paths**		**Non-shared environment paths**	
	Predictor	Core GCSEs	Predictor	Core GCSEs	Predictor	Core GCSEs
1. Bricks	0.55 (0.49 – 0.59)		*0.00 (constrained)*		0.45 (0.41 – 0.51)	
2. Core GCSEs	0.27 (0.22 – 0.32)	0.36 (0.30 – 0.43)	*0.00 (constrained)*	0.25 (0.21 – 0.29)	*0.00 (0.00 – 0.01)*	0.12 (0.11 – 0.12)
1. KC	0.77 (0.72 – 0.81)		*0.00 (constrained)*		0.23 (0.19 – 0.28)	
2. Core GCSEs	0.25 (0.20 – 0.30)	0.38 (0.31 – 0.44)	*0.00 (constrained)*	0.26 (0.22 – 0.29)	*0.01 (0.00 – 0.02)*	0.11 (0.10 – 0.12)
1. Faces	0.59 (0.53 – 0.65)		*0.00 (constrained)*		0.41 (0.35 – 0.47)	
2. Core GCSEs	0.02 (0.01 – 0.05)	0.59 (0.54 – 0.64)	*0.00 (constrained)*	0.27 (0.23 – 0.31)	*0.00 (0.00 – 0.00)*	0.12 (0.11 – 0.13)

Path estimates (standardised and squared, with 95% confidence intervals) for three bivariate ACE Cholesky decompositions, with each of the three predictors (Bricks, King's Challenge and “pure” face recognition) and the “core” GCSE subjects composite. Influences on the second variable in each model (the GCSE composite) are decomposed into influences shared with the first (the predictor), and those unique to the second. Italicised estimates are non-significant (their CIs include zero).

Table S21. Fit statistics: bivariate models for predictors and GCSEs.

Variables in model	Model	ep	χ2	df	AIC	Δχ2	Δdf	p
Bricks, Maths	Saturated	28	39293.75	15330	8633.75	-	-	-
	ACE	11	39306.43	15347	8612.43	12.68	17	0.78
	Constrained	9	39308.39	15349	8610.40	1.97	2	0.37
KC, Maths	Saturated	28	36113.10	14183	7747.10	-	-	-
	ACE	11	36128.18	14200	7728.18	15.08	17	0.59
	Constrained	9	36129.85	14202	7725.85	1.66	2	0.44
Faces, Maths	Saturated	28	37604.15	14552	8500.15	-	-	-
	ACE	11	37626.97	14569	8488.97	22.81	17	0.16
	Constrained	9	37627.05	14571	8485.05	0.08	2	0.96
Bricks, Science	Saturated	28	37241.85	14532	8177.85	-	-	-
	ACE	11	37259.65	14549	8161.65	17.81	17	0.40
	Constrained	9	37266.42	14551	8164.42	6.76	2	0.03
KC, Science	Saturated	28	34041.36	13385	7271.36	-	-	-
	ACE	11	34056.45	13402	7252.45	15.09	17	0.59
	Constrained	9	34058.94	13404	7250.94	2.49	2	0.29
Faces, Science	Saturated	28	35451.95	13754	7943.95	-	-	-
	ACE	11	35465.07	13771	7923.07	13.12	17	0.73
	Constrained	9	35465.46	13773	7919.46	0.39	2	0.82
Bricks, English	Saturated	28	39686.65	15409	8868.65	-	-	-
	ACE	11	39707.94	15426	8855.94	21.29	17	0.21
	Constrained	9	39712.82	15428	8856.82	4.87	2	0.09
KC, English	Saturated	28	36520.03	14262	7996.03	-	-	-
	ACE	11	36543.70	14279	7985.70	23.67	17	0.13
	Constrained	9	36547.61	14281	7985.62	3.91	2	0.14
Faces, English	Saturated	28	37702.50	14631	8440.50	-	-	-
	ACE	11	37729.65	14648	8433.65	27.16	17	0.06
	Constrained	9	37729.70	14650	8429.70	0.05	2	0.98
Bricks, Core GCSEs	Saturated	28	38317.85	15422	7473.85	-	-	-
	ACE	11	38330.79	15439	7452.79	12.94	17	0.74
	Constrained	9	38339.17	15441	7457.17	8.38	2	0.02
KC, Core GCSEs	Saturated	28	35155.54	14275	6605.54	-	-	-
	ACE	11	35171.88	14292	6587.89	16.35	17	0.50
	Constrained	9	35176.12	14294	6588.12	4.24	2	0.12
Faces, Core GCSEs	Saturated	28	36580.98	14644	7292.98	-	-	-
	ACE	11	36598.82	14661	7276.82	17.84	17	0.40
	Constrained	9	36598.83	14663	7272.83	0.01	2	1.00

Comparison of bivariate model fit statistics. The ACE model is compared to the fully saturated model, and the constrained submodel (with all but one shared environment path constrained to zero; see Methods) is compared to the ACE model. Variables were entered in the order specified. KC = King's Challenge, ep = estimated parameters; χ2 = -2 log-likelihood; df = degrees of freedom, AIC = Akaike information criterion. The p-values indicate whether there is a significant deterioration in fit (i.e., whether the second model is a worse fit than the previous one).

N.B. Almost all models fit well, but two of the constrained models (Bricks, Science; Bricks, Core GCSEs) deteriorated in fit in comparison to the full ACE models. Since the Bricks measures have been shown to have no shared environmental influences (see Methods and Supplementary Table S7), these seem most likely to be chance effects, given the large number of models tested.

Table S22. Trivariate Cholesky decompositions: verbal ability, predictor variables, Maths GCSE.

	Genetic paths			**Non-shared environment paths**		
	Verbal	Predictor	Maths	Verbal	Predictor	Maths
1. Verbal	0.54 (0.50 – 0.57)			0.46 (0.43 – 0.50)		
2. Bricks	0.12 (0.08 – 0.16)	0.44 (0.38 – 0.50)		*0.00 (0.00 – 0.01)*	0.44 (0.39 – 0.49)	
3. Maths	0.31 (0.26 – 0.35)	0.11 (0.07 – 0.15)	0.24 (0.17 – 0.30)	*0.00 (0.00 – 0.00)*	*0.00 (0.00 – 0.01)*	0.17 (0.16 – 0.19)
1. Verbal	0.53 (0.49 – 0.57)			0.47 (0.43 – 0.51)		
2. KC	0.13 (0.09 – 0.19)	0.65 (0.58 – 0.70)		*0.00 (0.00 – 0.01)*	0.22 (0.18 – 0.26)	
3. Maths	0.30 (0.26 – 0.35)	0.13 (0.09 – 0.17)	0.22 (0.15 – 0.28)	*0.00 (0.00 – 0.01)*	*0.01 (0.00 – 0.02)*	0.17 (0.15 – 0.18)
1. Verbal	0.54 (0.49 – 0.57)			0.46 (0.43 – 0.51)		
2. Faces	*0.02 (0.00 – 0.06)*	0.57 (0.50 – 0.63)		*0.00 (0.00 – 0.01)*	0.41 (0.35 – 0.47)	
3. Maths	0.30 (0.26 – 0.35)	*0.00 (0.00 – 0.01)*	0.34 (0.27 – 0.40)	*0.00 (0.00 – 0.01)*	*0.00 (0.00 – 0.00)*	0.18 (0.17 – 0.19)

Path estimates (standardised and squared, with 95% confidence intervals) for three trivariate ACE Cholesky decompositions, including verbal ability, one of the three predictors (Bricks, King's Challenge and “pure” face recognition) and Maths GCSE. Influences on Maths are decomposed into those i) shared both with verbal ability and with the other predictor, ii) shared only with the latter, independent of verbal ability, and iii) unique to Maths. Italicised estimates are non-significant (their CIs include zero).

Table S23. Trivariate Cholesky decompositions: verbal ability, predictor variables, Science GCSE.

	Genetic paths			**Non-shared environment paths**		
	Verbal	Predictor	Science	Verbal	Predictor	Science
1. Verbal	0.53 (0.49 – 0.57)			0.47 (0.43 – 0.51)		
2. Bricks	0.11 (0.08 – 0.16)	0.44 (0.38 – 0.49		*0.00 (0.00 – 0.01)*	0.45 (0.40 – 0.50)	
3. Science	0.37 (0.32 – 0.42)	0.07 (0.04 – 0.11)	0.20 (0.13 – 0.27)	*0.00 (0.00 – 0.00)*	*0.00 (0.00 – 0.01)*	0.17 (0.15 – 0.18)
1. Verbal	0.53 (0.49 – 0.57)			0.47 (0.43 – 0.51)		
2. KC	0.13 (0.08 – 0.18)	0.63 (0.57 – 0.69)		*0.00 (0.00 – 0.01)*	0.23 (0.19 – 0.28)	
3. Science	0.37 (0.32 – 0.42)	0.06 (0.03 – 0.10)	0.21 (0.14 – 0.28)	*0.00 (0.00 – 0.00)*	0.02 (0.01 – 0.03)	0.15 (0.13 – 0.17)
1. Verbal	0.53 (0.49 – 0.57)			0.47 (0.43 – 0.51)		
2. Faces	*0.02 (0.00 – 0.06)*	0.57 (0.50 – 0.63)		*0.00 (0.00 – 0.01)*	0.41 (0.35 – 0.47)	
3. Science	0.37 (0.32 – 0.42)	*0.00 (0.00 – 0.01)*	0.26 (0.19 – 0.33)	*0.00 (0.00 – 0.00)*	*0.00 (0.00 – 0.00)*	0.17 (0.16 – 0.18)

Path estimates (standardised and squared, with 95% confidence intervals) for three trivariate ACE Cholesky decompositions, including verbal ability, one of the three predictors (Bricks, King's Challenge and “pure” face recognition) and Science GCSE. Influences on Science are decomposed into those i) shared both with verbal ability and with the other predictor, ii) shared only with the latter, independent of verbal ability, and iii) unique to Science. Italicised estimates are non-significant (their CIs include zero).

Table S24. Trivariate Cholesky decompositions: verbal ability, predictor variables, English GCSE.

	Genetic paths			**Non-shared environment paths**		
	Verbal	Predictor	English	Verbal	Predictor	English
1. Verbal	0.53 (0.49 – 0.57)			0.47 (0.43 – 0.51)		
2. Bricks	0.11 (0.08 – 0.16)	0.45 (0.39 – 0.50)		*0.00 (0.00 – 0.01)*	0.44 (0.39 – 0.49)	
3. English	0.39 (0.34 – 0.44)	*0.01 (0.00 – 0.02)*	0.23 (0.22 – 0.30)	*0.00 (0.00 – 0.01)*	*0.00 (0.00 – 0.01)*	0.17 (0.16 – 0.18)
1. Verbal	0.53 (0.49 – 0.57)			0.47 (0.43 – 0.51)		
2. KC	0.11 (0.06 – 0.16)	0.66 (0.60 – 0.72)		*0.01 (0.00 – 0.02)*	0.23 (0.19 – 0.27)	
3. English	0.39 (0.34 – 0.44)	*0.01 (0.00 – 0.03)*	0.22 (0.16 – 0.29)	*0.00 (0.00 – 0.01)*	*0.00 (0.00 – 0.01)*	0.17 (0.16 – 0.18)
1. Verbal	0.53 (0.49 – 0.57)			0.47 (0.43 – 0.51)		
2. Faces	0.02 (0.00 – 0.05)	0.57 (0.51 – 0.63)		*0.00 (0.00 – 0.01)*	0.41 (0.35 – 0.47)	
3. English	0.39 (0.35 – 0.44)	*0.01 (0.00 – 0.04)*	0.22 (0.15 – 0.28)	*0.00 (0.00 – 0.01)*	*0.00 (0.00 – 0.00)*	0.17 (0.16 – 0.19)

Path estimates (standardised and squared, with 95% confidence intervals) for three trivariate ACE Cholesky decompositions, including verbal ability, one of the three predictors (Bricks, King's Challenge and "pure" face recognition) and English GCSE. Influences on English are decomposed into those i) shared both with verbal ability and with the other predictor, ii) shared only with the latter, independent of verbal ability, and iii) unique to English. Italicised estimates are non-significant (their CIs include zero).

Table S25. Trivariate Cholesky decompositions: verbal ability, predictor variables, "core" GCSE subjects.

	Genetic paths			**Non-shared environment paths**		
	Verbal	Predictor	Core GCSEs	Verbal	Predictor	Core GCSEs
1. Verbal	0.53 (0.49 – 0.57)			0.47 (0.43 – 0.51)		
2. Bricks	0.11 (0.07 – 0.15)	0.44 (0.38 – 0.50)		*0.00 (0.00 – 0.01)*	0.45 (0.40 – 0.50)	
3. Core GCSEs	0.43 (0.38 – 0.47)	0.07 (0.04 – 0.10)	0.16 (0.13 – 0.22)	*0.00 (0.00 – 0.01)*	*0.00 (0.00 – 0.01)*	0.11 (0.10 – 0.12)
1. Verbal	0.53 (0.49 – 0.56)			0.47 (0.44 – 0.51)		
2. KC	0.12 (0.08 – 0.18)	0.65 (0.58 – 0.70)		*0.00 (0.00 – 0.01)*	0.23 (0.19 – 0.27)	
3. Core GCSEs	0.43 (0.38 – 0.47)	0.07 (0.04 – 0.10)	0.15 (0.09 – 0.21)	*0.00 (0.00 – 0.01)*	*0.01 (0.00 – 0.02)*	0.11 (0.10 – 0.12)
1. Verbal	0.53 (0.49 – 0.57)			0.47 (0.43 – 0.51)		
2. Faces	*0.02 (0.00 – 0.05)*	0.57 (0.50 – 0.63)		*0.00 (0.00 – 0.01)*	0.41 (0.35 – 0.47)	
3. Core GCSEs	0.43 (0.38 – 0.48)	*0.00 (0.00 – 0.01)*	0.21 (0.14 – 0.27)	*0.00 (0.00 – 0.01)*	*0.00 (0.00 – 0.00)*	0.12 (0.11 – 0.12)

Path estimates (standardised and squared, with 95% confidence intervals) for three trivariate ACE Cholesky decompositions, including verbal ability, one of the three predictors (Bricks, King's Challenge and "pure" face recognition) and the "core" GCSE subjects composite. Influences on the GCSE composite are decomposed into those i) shared both with verbal ability and with the other predictor, ii) shared only with the latter, independent of verbal ability, and iii) unique to the GCSE composite. Italicised estimates are non-significant (their CIs include zero).

Table S26. Quadrivariate Cholesky decomposition: verbal, face recognition, spatial, Maths GCSE.

		Verbal	Faces	KC	Maths
Genetic paths	1. Verbal	0.54 (0.49 – 0.57)			
	2. Faces	*0.02 (0.00 – 0.06)*	0.57 (0.50 – 0.63)		
	3. KC	0.13 (0.09 – 0.19)	*0.00 (0.00 – 0.01)*	0.65 (0.58 – 0.70)	
	4. Maths	0.30 (0.26 – 0.35)	*0.00 (0.00 – 0.01)*	0.13 (0.09 – 0.17)	0.22 (0.15 – 0.28)
Non-shared environmental paths	1. Verbal	0.46 (0.43 – 0.51)			
	2. Faces	*0.00 (0.00 – 0.01)*	0.41 (0.35 – 0.47)		
	3. KC	*0.00 (0.00 – 0.01)*	*0.00 (0.00 – 0.01)*	0.21 (0.18 – 0.26)	
	4. Maths	*0.00 (0.00 – 0.01)*	*0.00 (0.00 – 0.00)*	*0.01 (0.00 – 0.02)*	0.17 (0.15 – 0.18)

Path estimates (standardised and squared, with 95% confidence intervals) for quadrivariate ACE Cholesky decomposition. The last rows indicate the genetic/non-shared environmental influences on Maths i) shared with verbal ability, “pure” face recognition and spatial ability (King's Challenge); ii) shared with both face recognition and spatial ability, independent of verbal ability; iii) shared only with spatial ability, independent of the other predictors; and finally iv) unique to Maths. Italicised estimates are non-significant (their CIs include zero).

Table S27. Quadrivariate Cholesky decomposition: verbal, face recognition, spatial, Science GCSE.

		Verbal	Faces	KC	Science
Genetic paths	1. Verbal	0.53 (0.49 – 0.57)			
	2. Faces	*0.02 (0.00 – 0.06)*	0.57 (0.50 – 0.63)		
	3. KC	0.13 (0.08 – 0.18)	*0.00 (0.00 – 0.02)*	0.63 (0.57 – 0.69)	
	4. Science	0.37 (0.32 – 0.42)	*0.00 (0.00 – 0.01)*	0.06 (0.03 – 0.10)	0.21 (0.14 – 0.28)
Non-shared environmental paths	1. Verbal	0.47 (0.43 – 0.51)			
	2. Faces	*0.00 (0.00 – 0.01)*	0.41 (0.35 – 0.47)		
	3. KC	*0.00 (0.00 – 0.01)*	*0.00 (0.00 – 0.01)*	0.23 (0.19 – 0.28)	
	4. Science	*0.00 (0.00 – 0.00)*	*0.00 (0.00 – 0.00)*	0.02 (0.01 – 0.03)	0.15 (0.13 – 0.17)

Path estimates (standardised and squared, with 95% confidence intervals) for quadrivariate ACE Cholesky decomposition. The last rows indicate the genetic/non-shared environmental influences on Science i) shared with verbal ability, “pure” face recognition and spatial ability (King's Challenge); ii) shared with both face recognition and spatial ability, independent of verbal ability; iii) shared only with spatial ability, independent of the other predictors; and finally iv) unique to Science. Italicised estimates are non-significant (their CIs include zero).

Table S28. Quadrivariate Cholesky decomposition: verbal, face recognition, spatial, English GCSE.

		Verbal	Faces	KC	English
Genetic paths	1. Verbal	0.53 (0.49 – 0.57)			
	2. Faces	*0.02 (0.00 – 0.05)*	0.57 (0.51 – 0.63)		
	3. KC	0.11 (0.06 – 0.16)	*0.00 (0.00 – 0.01)*	0.66 (0.60 – 0.72)	
	4. English	0.39 (0.34 – 0.44)	*0.01 (0.00 – 0.04)*	*0.01 (0.00 – 0.03)*	0.21 (0.14 – 0.28)
Non-shared environmental paths	1. Verbal	0.47 (0.43 – 0.51)			
	2. Faces	*0.00 (0.00 – 0.01)*	0.41 (0.35 – 0.47)		
	3. KC	*0.01 (0.00 – 0.02)*	*0.00 (0.00 – 0.01)*	0.23 (0.19 – 0.27)	
	4. English	*0.00 (0.00 – 0.01)*	*0.00 (0.00 – 0.00)*	*0.00 (0.00 – 0.01)*	0.17 (0.16 – 0.18)

Path estimates (standardised and squared, with 95% confidence intervals) for quadrivariate ACE Cholesky decomposition. The last rows indicate the genetic/non-shared environmental influences on English i) shared with verbal ability, “pure” face recognition and spatial ability (King's Challenge); ii) shared with both face recognition and spatial ability, independent of verbal ability; iii) shared only with spatial ability, independent of the other predictors; and finally iv) unique to English. Italicised estimates are non-significant (their CIs include zero).

Table S29. Quadrivariate Cholesky decomposition: verbal, face recognition, spatial, “core” GCSE subjects.

		Verbal	Faces	KC	Core GCSEs
Genetic paths	1. Verbal	0.53 (0.49 – 0.57)			
	2. Faces	*0.02 (0.00 – 0.05)*	0.57 (0.50 – 0.63)		
	3. KC	0.13 (0.08 – 0.18)	*0.00 (0.00 – 0.02)*	0.65 (0.58 – 0.70)	
	4. Core GCSEs	0.42 (0.38 – 0.47)	*0.00 (0.00 – 0.01)*	0.07 (0.04 – 0.10)	0.15 (0.09 – 0.21)
Non-shared environmental paths	1. Verbal	0.47 (0.43 – 0.51)			
	2. Faces	*0.00 (0.00 – 0.01)*	0.41 (0.35 – 0.47)		
	3. KC	*0.00 (0.00 – 0.01)*	*0.00 (0.00 – 0.01)*	0.22 (0.19 – 0.27)	
	4. Core GCSEs	*0.00 (0.00 – 0.01)*	*0.00 (0.00 – 0.00)*	*0.01 (0.00 – 0.02)*	0.11 (0.10 – 0.12)

Path estimates (standardised and squared, with 95% confidence intervals) for quadrivariate ACE Cholesky decomposition. The last rows indicate the genetic/non-shared environmental influences on the “core” GCSE subjects composite i) shared with verbal ability, “pure” face recognition and spatial ability (King's Challenge); ii) shared with both face recognition and spatial ability, independent of verbal ability; iii) shared only with spatial ability, independent of the other predictors; and finally iv) unique to the GCSE composite. Italicised estimates are non-significant (their CIs include zero).

Table S30. Fit statistics: trivariate models for verbal ability, predictors and GCSEs.

Variables in model	Model	ep	χ2	df	AIC	Δχ2	Δdf	p
Verbal ability, Bricks, Maths	Saturated	54	52698.03	20456	11786.03	-	-	-
	ACE	21	52737.71	20489	11759.71	39.68	33	0.20
	Constrained	16	52744.96	20494	11756.96	7.25	5	0.20
Verbal ability, KC, Maths	Saturated	54	49542.98	19309	10924.98	-	-	-
	ACE	21	49581.11	19342	10897.11	38.12	33	0.25
	Constrained	16	49586.95	19347	10892.95	5.85	5	0.32
Verbal ability, Faces, Maths	Saturated	54	51030.94	19678	11674.94	-	-	-
	ACE	21	51087.53	19711	11665.53	56.6	33	0.01
	Constrained	16	51092.65	19716	11660.65	5.11	5	0.40
Verbal ability, Bricks, Science	Saturated	54	50524.86	19658	11208.86	-	-	-
	ACE	21	50574.64	19691	11192.64	49.78	33	0.03
	Constrained	16	50589.35	19696	11197.35	14.71	5	0.01
Verbal ability, KC, Science	Saturated	54	47362.69	18511	10340.69	-	-	-
	ACE	21	47408.96	18544	10320.96	46.27	33	0.06
	Constrained	16	47418.90	18549	10320.90	9.94	5	0.08
Verbal ability, Faces, Science	Saturated	54	48781.12	18880	11021.12	-	-	-
	ACE	21	48832.16	18913	11006.16	51.04	33	0.02
	Constrained	16	48840.60	18918	11004.60	8.44	5	0.13
Verbal ability, Bricks, English	Saturated	54	52742.89	20535	11672.89	-	-	-
	ACE	21	52794.07	20568	11658.07	51.18	33	0.02
	Constrained	16	52805.28	20573	11659.28	11.2	5	0.05
Verbal ability, KC, English	Saturated	54	49604.61	19388	10828.61	-	-	-
	ACE	21	49654.51	19421	10812.51	49.9	33	0.03
	Constrained	16	49663.41	19426	10811.41	8.9	5	0.11
Verbal ability, Faces, English	Saturated	54	50816.98	19757	11302.98	-	-	-
	ACE	21	50877.32	19790	11297.32	60.34	33	0.003
	Constrained	16	50883.94	19795	11293.94	6.62	5	0.25
Verbal ability, Bricks, Core GCSEs	Saturated	54	51333.00	20548	10237.00	-	-	-
	ACE	21	51379.59	20581	10217.59	46.58	33	0.06
	Constrained	16	51399.13	20586	10227.13	19.55	5	0.002
Verbal ability, KC, Core GCSEs	Saturated	54	48189.67	19401	9387.67	-	-	-
	ACE	21	48239.37	19434	9371.37	49.71	33	0.03
	Constrained	16	48251.61	19439	9373.61	12.24	5	0.03
Verbal ability, Faces, Core GCSEs	Saturated	54	49612.95	19770	10072.95	-	-	-
	ACE	21	49671.48	19803	10065.48	58.53	33	0.004
	Constrained	16	49681.30	19808	10065.30	9.82	5	0.08

Comparison of trivariate model fit statistics. The ACE model is compared to the fully saturated model, and the constrained submodel (with all but one shared environment path constrained to zero; see Methods) is compared to the ACE model. Variables were entered in the order specified. KC = King's Challenge, ep = estimated parameters; χ2 = -2 log-likelihood; df = degrees of freedom, AIC = Akaike information criterion. The p-values indicate whether there is a significant deterioration in fit (i.e., whether the second model is a worse fit than the previous one).

N.B. Some of the ACE models show deteriorations in fit compared to the saturated model – see Methods for discussion. The drop in fit for some of the constrained submodels could perhaps suggests that there *are* minor shared environmental influences for some of the predictors and their relationships, even though they are so small that even this (reasonably large) sample is underpowered to detect them. In any case, since the drop in fit is small, these simpler submodels are still preferred, both for parsimony and for consistency with the other models tested.

Table S31. Fit statistics: quadrivariate models for verbal ability, predictors and GCSEs.

Variables in model	Model	ep	χ2	df	AIC	Δχ2	Δdf	p
Verbal ability, Faces, KC, Maths	Saturated	88	55315.06	21396	12523.06	-	-	-
	ACE	34	55387.15	21450	12487.15	72.09	54	0.05
	Constrained	25	55395.32	21459	12477.32	8.17	9	0.52
Verbal ability, Faces, KC, Science	Saturated	88	53134.38	20598	11938.38	-	-	-
	ACE	34	53208.01	20652	11904.01	73.63	54	0.04
	Constrained	25	53220.22	20661	11898.22	12.21	9	0.20
Verbal ability, Faces, KC, English	Saturated	88	55352.69	21475	12402.69	-	-	-
	ACE	34	55431.64	21529	12373.64	78.94	54	0.02
	Constrained	25	55443.17	21538	12367.17	11.53	9	0.24
Verbal ability, Faces, KC, Core GCSEs	Saturated	88	53953.44	21488	10977.44	-	-	-
	ACE	34	54034.92	21542	10950.92	81.48	54	0.01
	Constrained	25	54049.49	21551	10947.49	14.57	9	0.10

Comparison of quadrivariate model fit statistics. The ACE model is compared to the fully saturated model, and the constrained submodel (with all but one shared environment path constrained to zero; see Methods) is compared to the ACE model. Variables were entered in the order specified. KC = King's Challenge, ep = estimated parameters; χ2 = -2 log-likelihood; df = degrees of freedom, AIC = Akaike information criterion. The p-values indicate whether there is a significant deterioration in fit (i.e., whether the second model is a worse fit than the previous one).

N.B. Some of the ACE models show deteriorations in fit compared to the saturated model – see Methods for discussion. The constrained submodels all fit well compared to the full ACE models.

Table S32. Trivariate Cholesky decompositions: verbal-regressed predictors, GCSEs.

	Genetic paths			**Non-shared environment paths**		
	Faces (no verbal)	Spatial (KC, no verbal)	GCSE	Faces (no verbal)	Spatial (KC, no verbal)	GCSE
1. Faces (no verbal)	0.57 (0.49 – 0.64)			0.43 (0.36 – 0.51)		
2. KC (no verbal)	*0.00 (0.00 – 0.02)*	0.74 (0.69 – 0.79)		*0.00 (0.00 – 0.01)*	0.25 (0.21 – 0.31)	
3. Maths	*0.00 (0.00 – 0.01)*	0.23 (0.18 – 0.29)	0.40 (0.33 – 0.47)	*0.00 (0.00 – 0.01)*	*0.00 (0.00 – 0.01)*	0.17 (0.16 – 0.19)
1. Faces (no verbal)	0.57 (0.49 – 0.63)			0.43 (0.37 – 0.51)		
2. KC (no verbal)	*0.00 (0.00 – 0.03)*	0.73 (0.66 – 0.78)		*0.00 (0.00 – 0.01)*	0.27 (0.22 – 0.33)	
3. Science	*0.00 (0.00 – 0.02)*	0.15 (0.10 – 0.20)	0.47 (0.40 – 0.54)	*0.00 (0.00 – 0.01)*	*0.01 (0.00 – 0.02)*	0.16 (0.14 – 0.17)
1. Faces (no verbal)	0.57 (0.50 – 0.64)			0.43 (0.36 – 0.50)		
2. KC (no verbal)	*0.00 (0.00 – 0.02)*	0.73 (0.68 – 0.78)		*0.00 (0.00 – 0.02)*	0.26 (0.22 – 0.32)	
3. English	0.02 (0.01 – 0.06)	0.04 (0.02 – 0.08)	0.54 (0.48 – 0.60)	*0.00 (0.00 – 0.00)*	*0.00 (0.00 – 0.01)*	0.18 (0.17 – 0.19)
1. Faces (no verbal)	0.57 (0.49 – 0.64)			0.43 (0.36 – 0.51)		
2. KC (no verbal)	*0.00 (0.00 – 0.03)*	0.73 (0.68 – 0.78)		*0.00 (0.00 – 0.01)*	0.26 (0.22 – 0.32)	
3. Core GCSEs	*0.01 (0.00 – 0.03)*	0.16 (0.12 – 0.21)	0.45 (0.39 – 0.51)	*0.00 (0.00 – 0.00)*	*0.00 (0.00 – 0.01)*	0.11 (0.10 – 0.12)

Path estimates (standardised and squared, with 95% confidence intervals) for four trivariate ACE Cholesky decompositions, including “pure” face recognition and spatial ability (King's Challenge), both regressed phenotypically on verbal ability, then finally the GCSE measure. Influences on the GCSE measures are decomposed into those i) shared both with face recognition and with spatial ability, ii) shared only with spatial ability, independent of face recognition, and iii) unique to the GCSE measure. Italicised estimates are non-significant (their CIs include zero).

Table S33. Fit statistics: trivariate models for verbal-regressed predictors and GCSEs.

Variables in model	Model	ep	χ2	df	AIC	Δχ2	Δdf	p
Faces (no verbal), KC (no verbal), Maths	Saturated	54	40252.94	15608	9036.94	-	-	-
	ACE	21	40284.90	15641	9002.90	31.96	33	0.52
	Constrained	16	40288.58	15646	8996.58	3.68	5	0.60
Faces (no verbal), KC (no verbal), Science	Saturated	54	38177.42	14810	8557.42	-	-	-
	ACE	21	38207.30	14843	8521.30	29.89	33	0.62
	Constrained	16	38211.30	14848	8515.30	4.00	5	0.55
Faces (no verbal), KC (no verbal), English	Saturated	54	40577.45	15687	9203.45	-	-	-
	ACE	21	40617.17	15720	9177.17	39.72	33	0.20
	Constrained	16	40622.17	15725	9172.17	5.00	5	0.42
Faces (no verbal), KC (no verbal), Core GCSEs	Saturated	54	39305.78	15700	7905.77	-	-	-
	ACE	21	39337.82	15733	7871.82	32.05	33	0.51
	Constrained	16	39342.60	15738	7866.60	4.78	5	0.44

Comparison of trivariate model fit statistics. The ACE model is compared to the fully saturated model, and the constrained submodel (with all but one shared environment path constrained to zero; see Methods) is compared to the ACE model. Variables were entered in the order specified. KC = King's Challenge, ep = estimated parameters; χ2 = -2 log-likelihood; df = degrees of freedom, AIC = Akaike information criterion. The p-values indicate whether there is a significant deterioration in fit (i.e., whether the second model is a worse fit than the previous one). All models fit well.