British Journal of Educational Psychology (2017)

The British Psychological Society

www.wileyonlinelibrary.com

Revealing hidden talents: The development, use, and benefit of VESPARCH

Julia R. Badger* and Jane Mellanby

Department of Experimental Psychology, University of Oxford, UK

Background. School attainment tests and Cognitive Abilities Tests are used in the United Kingdom to set targets for educational outcome. Whilst these are good predictors, they depend not only on basic ability but also on learnt knowledge and skills, such as reading.

Method and Aims. VESPARCH is an online group test of verbal and spatial reasoning, which we propose gives a measure that more closely approximates to basic ability – fluid intelligence. The verbal test contains highly familiar words, does not require the child to read them, is untimed, and provides detailed feedback on five practice questions for each part of the test. The tests – one suitable and standardized for children aged 7–9 years and one for children aged 10–12 years – have good test–retest reliability and validity and conform to the Rasch model. Comparison of VESPARCH scores with school attainment measures allows identification of those students who are underachieving academically relative to their potential. The matched nature of the verbal and spatial tests allows reasoning ability in the two domains to be compared; those with much higher spatial scores might be expected to do well in STEM subjects.

Conclusion. VESPARCH can be used alongside current school tests to ensure targeted teaching and encouragement for every child.

In the United Kingdom, there are wide disparities in the performance of pupils in national Standard Attainment Tests (SATs) in years 2 and 6, GCSEs in year 11 and A-levels in year 13. For example, the percentage getting 3 As or better at A-level differs widely by region: In 2009–2012, the proportion (of those who applied to university through UCAS, see SDMA, University of Oxford) was 2.4 times higher for those living in the South East compared to the North East. Furthermore, comparison of A-level results by area of socio-economic status (SES) shows that at each level of A-level performance, high deprivation is associated with around a twofold reduction in the proportion of candidates reaching that level (Gill, 2014). Such marked differences will be related to a large number of factors including aspirations and educational experience of parents, financial resources of the school, training and motivation of teachers, proportion of children with English as an additional language (EAL), and the nature of the school built environment. There is no a priori reason to believe that schools where the achievement of pupils is low necessarily have a pupil intake whose average basic ability is below that of the intake of other schools.

*Correspondence should be addressed to Julia R. Badger, Department of Experimental Psychology, University of Oxford, 9 South Parks Road, Oxford OX1 3UD, UK (email: julia.badger@psy.ox.ac.uk).

DOI:10.1111/bjep.12189

So-called elite universities struggle to increase recruitment from lower SES groups because smaller proportions of these pupils reach the required level of attainment in school examinations (GCSE and A-level) judged necessary to cope with the rigorous courses that such universities teach. Oxford and Cambridge, for example, whilst willing to ask for slightly lower grades in school examinations from students who have attended poorly performing schools, still expect that successful applicants will have top grades at GCSE and predictions of top grades at A-level.

How can we improve attainment and access for the clever children that are being missed by school achievement tests? And how might we identify them and give them appropriate encouragement and support to reach their potential at school and later on at university? School tests are heavily dependent on literacy, and indeed, literacy is a most important predictor of school success. Results on SATs, which measure reading, writing, mathematics, and science, influence most teachers' expectations of future progress and so those who perform poorly tend to have their sights already set low. Whilst literacy is important for reaping the benefit of school education, there are also non-verbal skills that are equally important in life outside classrooms but are not really measured in current assessments.

In this study, we present VESPARCH, a test of verbal and spatial reasoning, which we argue to allow better measurement of underlying potential which can then be compared with current school achievement to identify those children who are underachieving relative to this potential.

Reasoning and ability

Reasoning is used throughout schooling and in everyday life and involves problem-solving, thinking logically and identifying and understanding patterns and relationships in novel situations. Reasoning tests differ from school assessment tests because they tap into underlying cognitive abilities without relying on knowledge of a specific curriculum. In this way, they are able to provide a measure of academic potential.

The most commonly used reasoning test in the United Kingdom is the Cognitive Abilities Test, currently the fourth edition (CAT4, GL Assessment, 2012; Thorndike, Hagen, & France, 1986), which comprise reasoning in the verbal, non-verbal, spatial, and quantitative domains. The verbal test is a good predictor of GCSE achievement; however, it requires the child to read, some of the concepts are quite sophisticated, and the scores also relate to understanding of syntax (Mellanby, Anderson, Campbell, & Westwood, 1986). Therefore, the results are dependent not only on fluid intelligence but also on aspects of crystallized intelligence. To some extent, these scores are leading to self-fulfilling prophecies. A better solution would be a group test that will measure potential in the verbal domain, requiring only a level of vocabulary and general knowledge expected to be common to all children in the age group and that is not dependent on reading. To see whether a test can measure fluid intelligence – an indicator of potential – we need to consider theoretical accounts of the structure of human intelligence.

Reasoning ability is considered to be the core of the fluid part of intelligence (Gf; Schneider & McGrew, 2012) proposed by Cattell (1963), whilst acquired knowledge and skills including reading, writing, and vocabulary are part of crystallized intelligence (Gc). Cattell's model has been expanded greatly, firstly by Carroll (1993) as a hierarchical model with g, the general intelligence factor, at the top, with Gf and Gc separately in level two, and with numerous narrower specific abilities at level one. The Cattell–Horn–Carroll (CHC) model (Horn & Noll, 1997; Schneider & McGrew, 2012) also incorporates three

levels. However, there is now no consensus concerning whether it is useful or theoretically necessary to postulate Spearman's 'g' as overarching the structure of intelligence (see, e.g., Horn & Blankson, 2014; Kan, Kievit, Dolan, & van der Maas, 2011), and instead, it is proposed to equate Gf with g and thus produce a 2-level model. In the 1997 model of McGrew, Gf was one of ten second-order factors grouped by factor analysis into three clusters. The more recent conceptual model of Schneider and McGrew (2012) groups those factors which can be considered as being domain-general rather than domain-specific as constituting the various contributors to Gf: Short-term/working memory (executive function capacity) and long-term learning/storage and retrieval efficacy are grouped in one cluster; psychomotor speed, reaction time and speed, and accuracy of body movements are grouped into a second cluster. Gc is then found in a separate part of the model encompassing a cluster containing also Gq (quantitiative), Gkn (domain-specific knowledge), and Gvw (reading and writing). They define Gc as comprehension of 'the depth and breadth of knowledge and skills that are valued by one's society'.

The relationship between fluid intelligence (Gf) and the acquisition of knowledge and skills is an important one in education. It raises the question of whether Gf continues to be important throughout childhood or whether it becomes irrelevant as the child ages because what determines the further acquisition of knowledge and skills is the current level of knowledge and skills (crystallized intelligence, Gc). Cattell considered that Gf was more important during the early years but continued to have a role throughout childhood. Gustafsson and colleagues in Sweden have tested models in which different importance has been ascribed to Gf at different ages. Kvist and Gustafsson (2008) showed that it is important to use a homogeneous population in order to study this situation. Thorsen, Gustafsson, and Cliffordson (2014) working with a homogeneous sample of 9,000 school children from ages 9 to 10 through to ages 15 to 16 showed that a model in which Gf continued to influence Gc fitted the data better than one which allowed little effect of Gf after the first measures. Discussion of Cattell's investment hypothesis, and the alternative description of the importance of the effect of Gf being confined to the early years, sometimes called the encapsulation hypothesis (introduced by Gustafsson and Carlstedt; see Thorsen, 2014) also requires definition of what Gc actually is (see Ferrer and McArdle (2004), Kan *et al.*, 2011). If it is only a conglomerate of measured knowledge and skills, rather than a domain-general ability to acquire knowledge and skills, then it can be argued that it roughly equates to what we measure as academic attainment. The relationship between the various factors at different stages of development is a matter of more than theoretical importance. An extreme adoption of the view that Gf is only important in early childhood suggests that there is little we can do to improve a child's academic performance in the later years of schooling if s/he has not acquired an adequate level knowledge and skills in the early years. In contrast, the investment theory suggests that it is worth measuring potential (Gf) in later childhood and adolescence to see how much further Gc and academic attainment can be enhanced at later stages of school education.

We are proposing that verbal VESPARCH is approximating a measure of Gf in the verbal domain and this would be predictive of potential academic attainment. We have minimized contribution from general language experience using high-frequency words, familiar concepts, and by teaching the basic logic of the tests through feedback on practice questions. To avoid influence of prior education, items are not reliant on the understanding of abstract concepts and the extensive practice items introduced the types of reasoning required. The test is untimed and so processing speed should not affect

scores. Overall, the test relies heavily on Gf and does not use the other factors, which have been hypothesized to constitute intelligence by, for example, Schneider and McGrew (2012). Spatial VESPARCH is in a format that matches the verbal thus allowing direct comparison. It is intended that this measures Gf in the visuo-spatial domain.

In the next sections, we report the development and characteristics of the VESPARCH tests. We propose that VESPARCH should be used alongside tests of current attainment to identify academic potential, but also identify those children underachieving at school relative to their potential – the academically 'missed' children. It can also pick out those children whose spatial reasoning is much better than their verbal reasoning.

Development of the Verbal and Spatial Reasoning test for Children (VESPARCH)

VESPAR

VESPARCH has been developed in the same format as the verbal and spatial reasoning test for adults (VESPAR; Langdon & Warrington, 1995), which was designed for a clinical population. VESPAR sought to measure reasoning ability in the verbal and spatial domains separately. Thus, VESPAR has the advantage over other ability tests, for example, the Wechsler Adult Intelligence Scale (WAIS; Wechsler, 2008), where the two domains are tested in very different ways (e.g., reconstruction of a cube, or picture completion versus information or vocabulary), or Raven's Progressive Matrices (RPM; Pearson, 2004), which is entirely non-verbal. VESPAR was designed to test verbal reasoning in as culture-fair a way as possible (for native English speakers at least) using high-frequency words (A and AA) and simple concepts. The multiple-choice format is also known to reduce test anxiety (Coughlan & Hollows, 1984), as does the untimed nature of the test. The items all remain for inspection until a choice is made to reduce the load on working memory. As the tests are administered as one-to-one pencil and paper tests, with the administrator reading out the verbal items, there is no need for the patient to read. A further important feature of VESPAR is that it has five practice questions with feedback at the start of each of the subparts of the test to ensure that a patient fully understands what is required.

The VESPAR tests have been successfully administered and informative in a range of research studies, such as those investigating cognitive function in patients with multiple sclerosis (Camp *et al.*, 1999, 2005); those looking into the role of the left and right hemispheres in reasoning with regard to unilateral hemisphere lesion patients (Langdon & Warrington, 2000), and as a part of a battery of tests exploring sequential learning in dyslexia (Kelly, Griffiths, & Frith, 2002). In 1998, Langdon, Rosenblatt and Mellanby used VESPAR to assess reasoning ability in children aged 14–15 and showed that for this group it could give a measure for the full range of ability. However, pilot testing of VESPAR in younger children (aged 11–12) showed that it was too difficult. Therefore, we constructed verbal and spatial reasoning for children (VESPARCH), one for children aged 7–9 (years 3 and 4: hereafter Y3-4), and one for children aged 10–12 (years 6 and 7: hereafter Y6-7).

VESPARCH

We chose these particular age groups because they are points of transition in the United Kingdom state educational system: from infant to junior (following KS1 SATs at the end of year 2) and from primary to secondary (KS2 SATs at the end of year 6). To make the verbal part of the test suitable, we chose words known to children 2 years younger than our target groups and familiar concepts. New items were piloted with many small groups of

children in pencil and paper format, and items were removed and substituted when the concepts were not familiar to the children or the questions were ambiguous. The tests were then computerized, and audio files produced, so that they could be administered to whole classes at once but remain independent of reading ability. The VESPARCH tests, like the VESPAR tests, include five practice questions for each subpart (now only verbal analogy, verbal category, spatial analogy, and spatial category), and the feedback – given through headphones – allows up to two further tries if the answer is wrong and then provides explanations of the correct answers. Each part in the Y3-4 test has a total of 40 questions with five practice questions per subpart; each part in the Y6-7 test has a total of 50 questions with five practice questions per subpart (e.g., see Figure 1). The Y3-4 test has fewer questions overall because we found that some of the younger children were less able to maintain their attention for as long as the older children. The practice questions contain elements of the types of questions which constitute the actual test so that there should be a 'level playing field' for all children even though some may not have previously encountered concepts such as analogies or visual rotation. Thus, the VESPARCH tests are designed to give separate measures of verbal and spatial reasoning, approximating to fluid ability, with less reliance on crystallized ability.

The first computerized version of VESPARCH (Mellanby & McElwee, 2009) was used to collect verbal reasoning data on around 1,700 children aged 11, and these children's school careers have been followed up to GCSE (aged 16). The verbal test was a good predictor of overall GCSE grades with comparable correlations to those seen with CAT scores. However, the work showed that there were children who apparently had the potential to do well that were not recognized by their CAT scores. Since 2009, in collaboration with Cambridge Assessment, the VESPARCH tests have been re-designed and converted into an online version. The tests have been subjected to Rasch analysis (Rasch, 1960), to establish that the items and the test as a whole conform to the Rasch model. This has required many iterations of designing new items to replace those that were unsatisfactory, running the revised test with another year group of children and so on. We now have standardized tests for both Y3-4 (age 7–9) and Y6-7 (age 10–12) that conform to the Rasch model (Mellanby, McElwee, & Badger, 2016). The online tests can now be administered to large groups of children. The instructions and the verbal items are read aloud through headphones and simultaneously highlighted on screen which makes the measure of verbal reasoning relatively independent of reading ability. Every screen has a simple multiple-choice format and the option to listen to individual words or instructions as many times as required. Analogical questions follow the format 'A' is to 'B' as 'C' is to 'D', where in the case of our test 'D' must be chosen from four options. Categorical questions follow the 'odd-one-out' format where the child must identify the option out of four choices that does not follow the same categorical rules as the other three (see Figure 1).

Additionally, the matched verbal and spatial tests allow identification of children whose reasoning strengths are greater in the spatial direction. Such children tend to be missed by selection procedures which depend heavily on reading and writing. For example, in the 11 + examination, which is still used in some areas of England for selection to grammar schools, the weighting of the non-verbal reasoning score versus maths and verbal reasoning is such that it only contributes 20% to the total for calculation of pass or fail. Solving spatial problems requires reasoning ability and also visualization and mental rotation. Children with high spatial scores as measured by VESPARCH, even where verbal score is lower, might well be our future engineers, physicists, artists, architects, and

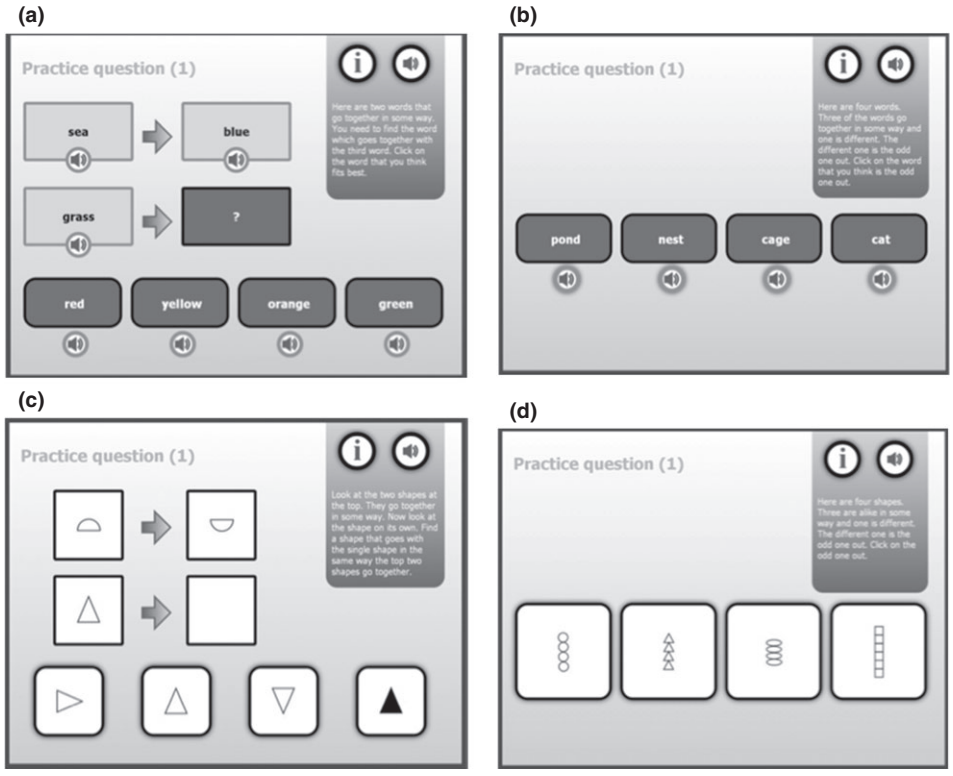

Figure 1. (a) A Y3-4 verbal analogical practice question, the correct answer is green; (b) A Y3-4 verbal categorical practice question, the correct answer is cat; (c) A Y6-7 spatial analogical practice question, the correct answer is 3; and (d) A Y6-7 spatial categorical practice question, the correct answer is 4.

surgeons (see, e.g., Garg, Norman, & Sperotable, 2001; Humphreys, Lubinski, & Yao, 1993).

Administration of VESPARCH

The VESPARCH tests can be administered to individual or large groups of children. When multiple children complete the test at one time, headphones are worn and the computer screens arranged so that children cannot see their neighbour's screen. At the start of the test, children are told that there is no time limit. This means that processing speed should not affect performance. When the first test is finished, the child either clicks to start the next test or clicks to return to the home page and completes the second test another time. The average time taken to complete a single test is 25 min. Teachers log into a private site to download every child's two standardized scores, one verbal and one spatial, and a whole year group scatter plot.

Rasch analysis of the VESPARCH items

Rasch analysis uses a mathematical model (the Rasch model) to 'incorporate a method for ordering persons (e.g., from a sample of school children) according to their ability, and ordering items. . .according to their difficulty' (Bond & Fox, 2007; pp. 10). Using expected

probabilities, the model predicts how items should perform to generate reliable measurement; the analyses examine the degree to which each item fits with the model. The model states that the higher the ability of an individual, the greater their ability to solve any of the items presented. Equally, the more difficult an item is to solve, the lower the probability of any person being able to correctly solve it, compared to an easier item (Rasch, 1960). The Rasch model uses two statistical-fit indicators – the fit residual and the chi-square – to determine how well the set of items fit the requirements of the model and each item's degree of deviation. The fit residual should range between ±2 for a sample of 30–300 (Bond & Fox) and is an indication as to how well an item discriminates between different levels of a construct. The chi-square test of fit should be .05 or greater, to show that individuals of a particular class interval do not significantly deviate from the model's expected mean score for this class interval. Graphical outputs – the item characteristic curves (ICC) and the multiple-choice distracter curves (MC DC) – are also used to identify any ill-fitting or ambiguous items. Figure 2 provides an example of an ICC good-fit item where the line represents the expected response based on ability (person location logits) and the dots represent the class intervals of ability, as calculated by the Rasch model. In this instance, the dots fall almost perfectly in-line with the expected response curve. The Rasch analyses produce two statistical measures of reliability – the person separation index (PSI) and Cronbach's alpha (α) – which is acceptable at .7 (Kline, 1999).

We have used Rasch analysis to identify and retain good items, and identify and remove the ill-fitting items. Here we present representative analysis of 112 Y3-4 children (*M* = 7;8 years), and 250 Y6-7 children (*M* = 11;3). Every item in our four finalized tests has a fit residual between ±2 and a non-significant chi-square; therefore, we can be confident that the items do not significantly differ from the model expectations. All four tests show good reliability of scale, having strong PSI and Cronbach's alpha, and their person-item location fit is good. The Y3-4 spatial test's PSI = .63; α = .68. Item location = 0.0 (*SD* = 0.81; *SE* = .22); person location = −0.52 (*SD* = 0.59; *SE* = .36); and mean SAS = 91 (*SD* = 12.65). The Y3-4 verbal test's PSI = .81; α = .83. Item location = 0.0 (*SD* = 1.15; *SE* = .18); person location = 0.60 (*SD* = 0.95; *SE* = .39); and mean SAS = 94 (*SD* = 13.08). The Y6-7 spatial test's PSI = .86; α = .87. Item location = 0.0 (*SD* = 0.11; *SE* = .16); person location = −0.06 (*SD* = 0.88; *SE* = .34); and

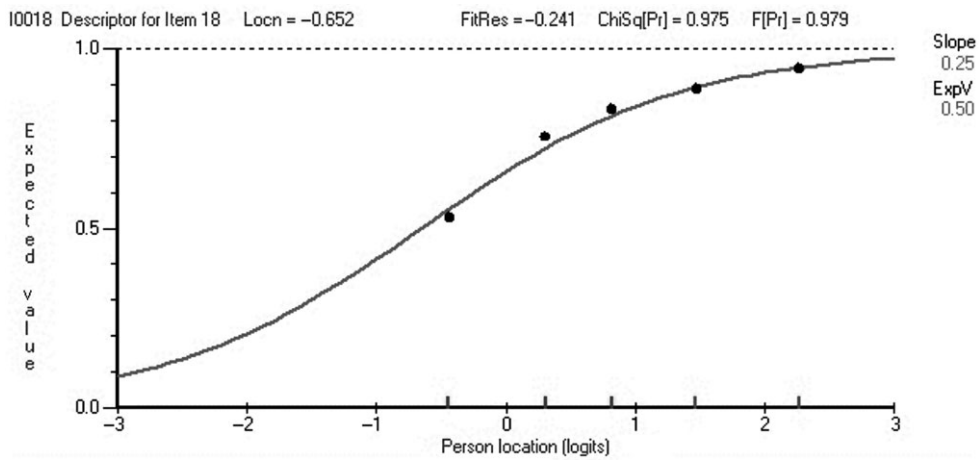

Figure 2. An item characteristic curve of a good-fit item.

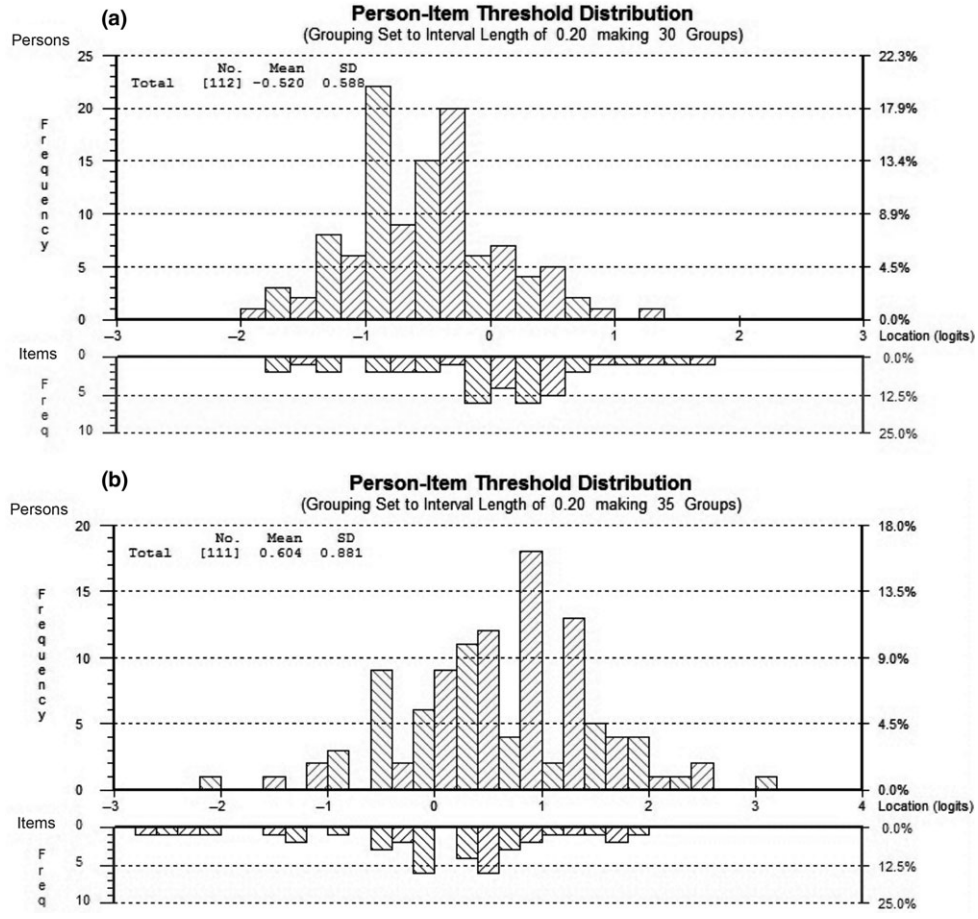

Figure 3. (a) The Y3-4 spatial test example of a person-item threshold distribution. The children are represented by the top histogram; the items are represented by the bottom histogram. Both are placed on a logit scale, determined by ability (children) or difficulty (items). (b) The Y3-4 verbal test example of a person-item threshold distribution. The children are represented by the top histogram; the items are represented by the bottom histogram. Both are placed on a logit scale, determined by ability (children) or difficulty (items).

mean SAS = 93 (13.79). The Y6-7 verbal test's PSI = .87; α = .88. Item location = 0.0 (*SD* = 0.92; *SE* = .15); person location = .23 (*SD* = 0.91; *SE* = .33) and mean SAS = 94 (13.87).

Figure 3a shows that the Y3-4 spatial test discriminates between the highest ability children, whereas Figure 3b shows that the Y3-4 verbal test does not discriminate between the top 5% of children. This is also the case for 8% in the Y6-7 verbal test and 5% in the Y6-7 spatial test.

Reliability of the VESPARCH tests

A total of 73 children – 34 males and 39 females – completed the online Y3-4 verbal VESPARCH test whilst in year 3 (mean = 7;9 years). Testing took approximately 30 min

Table 1. Correlational analysis of test-retest verbal VESPARCH data

Verbal VESPARCH SAS in year 3	Verbal VESPARCH SAS in year 4	
Test–retest when in year 3 and year 4		
103 (13.98)	104 (14.68)	$r = .856$, $n = 73$; $p < .001$
Verbal VESPARCH SAS in year 4	**Verbal VESPARCH SAS in year 6**	
Test–retest when in year 4 and year 6		
104 (12.65)	108 (12.06)	$r = .846$, $n = 75$; $p < .001$

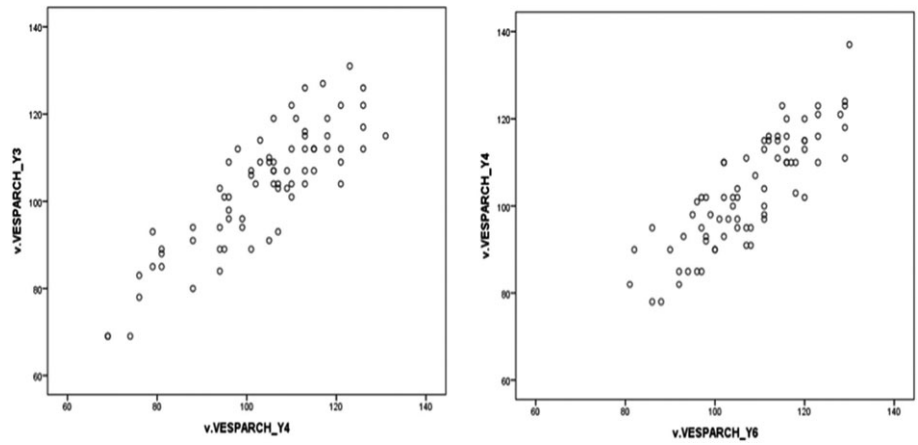

Figure 4. Correlation between SAS on completing the Y3-4 verbal test whilst in year 3, and then again in year 4 (left plot). Correlation between SAS on completing the Y3-4 verbal test whilst in year 4, and completing the Y6-7 verbal test whilst in year 6 (right plot).

and was completed in a whole-class setting with individual computers and headphones. The same children completed the test again a year later, maintaining the same testing condition as before (mean = 8;10 years). A strong positive correlation was found, see Table 1 for the data and Figure 4 for the correlational output. Following the same overall procedure as above, a total of 75 children – 38 males and 37 females – completed the online Y3-4 verbal VESPARCH test whilst in year 4 (mean = 9;3 years) and the Y6-7 verbal VESPARCH test 2 years later whilst in year 6 (mean = 10;9 years). A strong positive correlation was found, see Table 1 for the data and Figure 4 for the correlational output.

Validity of the VESPARCH tests

Assured that our VESPARCH tests contain only unidimensional, discriminative reasoning items and has strong test–retest reliability, we wanted to demonstrate its validity as a measure of cognitive ability. We correlated VESPARCH data against a well-established standardized cognitive test: the Cognitive Abilities Test 4 (CAT4; GL Assessment, 2012). A total of 108 primary school children and 120 secondary school children completed both the VESPARCH and CAT4 online tests (Y3-4 M age = 7;9; Y6-7 M age = 11;6). SAS for both the VESPARCH and the CAT4 were generated. For

Table 2. Distribution data for verbal VESPARCH and verbal CAT, split by year groups

	Mean (*SD*)	Skewness (*SE*)	Kurtosis
Verbal VESPARCH Y3-4	96 (12.68)	.073 (.234)	−.820 (.463)
Verbal CAT Y3-4	102 (11.83)	.224 (.233)	−.015 (.461)
Verbal VESPARCH Y6-7	105 (13.45)	.041 (.221)	.003 (.438)
Verbal CAT Y6-7	103 (14.57)	−.075 (.221)	.407 (.438)

Table 3. Correlational analysis of VESPARCH and CAT standardized age scores for children in year 3 and year 4 and children in year 6 and year 7

Years 3 and 4		
Verbal VESPARCH	Verbal CAT	
96 (12.68)	102 (11.82)	$r = .509$, $n = 108$ $p < .001$ Shared variance = 25.9%
Spatial VESPARCH	Spatial non-verbal CAT	
97 (12.91)	98 (12.20)	$r = .528$, $n = 108$ $p < .001$ Shared variance = 27.9%
Years 6 and 7		
Verbal VESPARCH	Verbal CAT	
105 (13.45)	103 (14.57)	$r = .793$, $n = 120$ $p < .001$ Shared variance = 62.9%
Spatial VESPARCH	Spatial non-verbal CAT	
103 (14.50)	104 (13.56)	$r = .752$, $n = 120$ $p < .001$ Shared variance = 56.6%

the following correlational analyses, verbal VESPARCH was directly compared with the verbal CAT4 (see Table 2 for the similar distribution of scores across the verbal four tests). The spatial VESPARCH has overlapping similarities with both the spatial CAT4 in terms of recognizing and identifying shapes, and the non-verbal CAT4 in terms of categorical reasoning with geometric shapes; therefore, it was compared to a composite score of the spatial and non-verbal CAT4. Data in Table 3 show that the correlations are strong between the Y6-7 VESPARCH and CAT4 tests and moderate between the Y3-4 VESPARCH and CAT4 tests. We would not of course expect fluid ability (as we postulated to be approximated by VESPARCH) to correlate 'perfectly' with CAT4 scores as the latter contains elements of crystallized intelligence.

A total of 84 year 3 and year 4 children (M = 7;3 years) and 99 year 6 and year 7 children (M = 12; 5 years) completed both the spatial VESPARCH and NNAT2 tests (a well-established non-verbal test: the Naglieri Nonverbal Ability Test Second Edition; Pearson, 2007). For both the Y3-4 and Y6-7 children's data, moderate correlations were found: $r = .505$, $n = 84$, $p < .001$; $r = .574$, $n = 99$, $p < .001$, respectively. Verbal VESPARCH measured in year 7 has also been shown to be a good predictor ($r = .6$–$.7$) of GCSE total score in year 11 (J. Mellanby & S. McElwee, unpublished).

Application of the VESPARCH tests

Comparing verbal and spatial reasoning scores

For the present paper, we tested 1,003 children on the VESPARCH tests (515 Y3-4; $M = 7;11$ and 488 Y6-7; $M = 11;2$). Males = 524. The only significant difference between

males and females on scores of either verbal or spatial VESPARCH can be found between children completing the Y3-4 verbal test, where females score higher than males (see Table 4). Therefore, we collapsed data across sex. We have previously shown no sex differences in much larger samples of children.

We calculated the difference between the two scores (spatial SAS minus verbal SAS) for both the Y3-4 and Y6-7 children. The Y3-4 verbal SAS mean was 93 (*SD* = 14.00), and the spatial SAS mean was 96 (*SD* = 13.93), leading to a positive correlation: $r = .603$, $n = 515$; $p < .001$ (see Figure 5a). There was also a positive correlation between the Y6-7 verbal and spatial reasoning tests: $r = .710$, $n = 488$; $p < .001$ (see Figure 5b). The verbal mean was 97 (*SD* = 13.44), and the spatial mean was 97 (*SD* = 13.81). These analyses show that the two measures share a substantial amount of their variance (36% for Y3-4 and about 50% for Y6-7). However, plots of the distribution of scores show that there are some children whose discrepancy is more than 1 *SD*.

Table 5 shows the numbers and proportions of children in this sample whose verbal–spatial discrepancy (strength) is 1 *SD* or 2 *SD* above the average discrepancy. These data are illustrated in Figure 5a and b. The data in Table 5 show that it is evident that about the

Table 4. Mean VESPARCH standardized age scores split by gender

	Verbal VESPARCH (mean SAS)	Spatial VESPARCH (mean SAS)
Years 3 and 4		
Males	92 (13.67)	96 (13.83)
Females	97 (14.29)	97 (14.04)
	F (1, 514) = 4.97; p = .026	F (1, 514) = 1.39; p = .238
Years 6 and 7		
Males	96 (14.56)	96 (14.56)
Females	98 (12.13)	97 (13.00)
	F (1, 487) = 3.30; p = .070	F (1, 487) = .98; p = .323

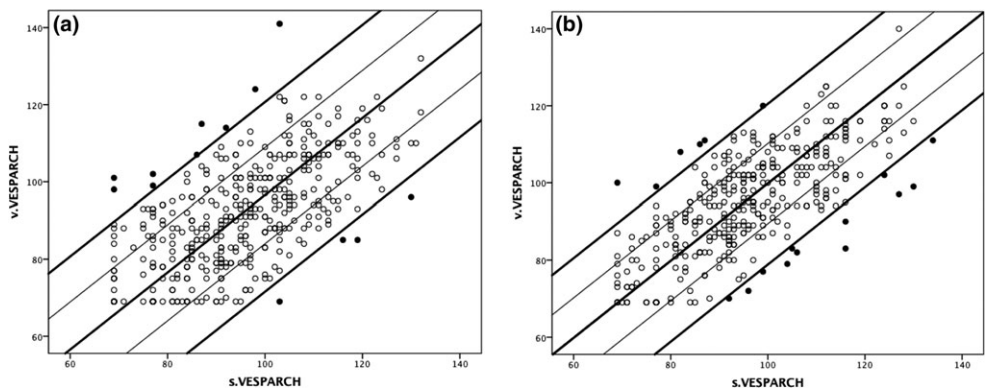

Figure 5. (a) Distribution, with 1 *SD* and 2 *SD* reference lines, of Y3-4 children with a balanced verbal–spatial reasoning profile (white dots), and those with an extreme strength (2 *SD*; black dots. Note: some children have the same scores which results in seemingly fewer black dots than expected). (b) Distribution, with 1 *SD* and 2 *SD* reference lines, of Y6-7 children with a balanced verbal–spatial reasoning profile (white dots), and those with an extreme strength (2 *SD*; black dots Note: some children have the same scores which results in seemingly fewer black dots than expected).

Table 5. Number of children showing a verbal or spatial strength with either a 1 *SD* or 2 *SD* discrepancy

	Y3-4		Y6-7	
	1 *SD* above the mean (%)	2 *SD* above the mean (%)	1 *SD* above the mean (%)	2 *SD* above the mean (%)
Verbal strength	75 (15)	11 (2)	61 (13)	14 (3)
Spatial strength	70 (14)	9 (2)	74 (15)	15 (3)

same number of children completing the Y3-4 and the Y6-7 tests show a verbal strength (1 *SD* or 2 *SD*) as those who show a spatial strength (1 *SD* or 2 *SD*).

Identifying Underachievement Relative to Potential (URP)

We have identified those underachieving at school relative to their potential by comparing school English attainment data with verbal VESPARCH SAS. Of the 1,003 children's data used in the previous section, we were given the most recent attainment data for 970 (498 were Y3-4), which we converted to a point score to allow quantitative comparison. Both the verbal VESPARCH and the attainment points were transformed into *z*-scores. We have arbitrarily defined underachievement (relative to potential) as scoring at least 1.5 *SD* higher on the verbal VESPARCH test than in school achievement. Using this criterion, 24 Y3-4 children (5%) and 35 Y6-7 children (6%) were identified as underachieving. When a child is identified as URP, we administer additional short tests to identify specific weaknesses in factors that are not involved in VESPARCH, but are important in regular school education. Factors include the following: the acquisition of complex grammar, reading efficiency, and short-term memory. In a recent study, we identified 127 children URP and followed up with them and 192 controls, on the above factors (Badger & Mellanby, 2016). The individual paper-based testing took between 7 and 10 min per child. We also asked children to complete a 5-min psychosocial questionnaire considering difficulties with peers, emotions, behaviour, and hyperactivity. We found that those children identified as URP were more likely to show multiple difficulties compared to children non-URP, struggle with reading and show higher levels of hyperactivity/inattention. Follow-up testing can help identify individual targeted support.

Identifying underachievement relative to potential: VESPARCH vs. CAT

Using our CAT data sample – 108 primary school children and 120 secondary school children – we decided to compare VESPARCH and CAT data when identifying URP. In this sample, the VESPARCH tests identified 7 Y3-4 children (6%) and 8 Y6-7 children (7%) as underachieving; CATs identified 4 Y3-4 children (4%) and 8 Y6-7 children (7%) as underachieving. There is some overlap of identification, but it is important to note that many of the children are not identified by both measures, thus highlighting the difference and importance of both tests.

Discussion

VESPARCH can be used to reveal the hidden talents of children that have not been identified by current school tests. We argue that the scores give a measure of fluid

intelligence. We propose that using VESPARCH as an approximation to Gf and comparing the scores with school attainment allows identification of those children who are underachieving relative to their potential. In this way, we can start to identify the 'missed children'. If we subscribe to the Investment theory of the development of abilities (Cattell, 1987; Thorsen *et al.*, 2014), then Gf should impact on crystallized intelligence, and hence academic achievement, throughout childhood. This is important in order that these children can be given the necessary educational opportunities and stimulus to reach their potential. The VESPARCH objective is to fairly measure reasoning across a wide spread of abilities by significantly reducing the reliance on language, memory, knowledge, or processing speed: Children do not need to be able to read as everything is read aloud whilst simultaneously presented on-screen; the stimuli remain on screen until the child makes a choice; a speaker icon allows children to listen to instructions or words as many times as necessary; all concepts and words are highly familiar; detailed practice questions are presented; and there is no time limit. It is not intended as a 'high stakes' test for selection at the top end of the ability range. If educators require this sort of discrimination, they might use VESPARCH to identify the top 5 or 10% of pupils, and then, administer tests specifically intended for the very top of the ability range.

VESPARCH has been shown to have good test–retest reliability, and its correlation with CAT4 and NNAT2 scores provides evidence for its validity. However, it is interesting to note that the VESPARCH tests and the CATs showed a stronger correlation for the Y6-7 children compared to the Y3-4 children. The CAT suitable for Y3-4 children is a much shorter version, whereas the VESPARCH test is only slightly shorter for the Y3-4 children than it is for the Y6-7 children and the format is identical. It is possible that the Y3-4 CAT is not able to assess reasoning as accurately as the Y3-4 VESPARCH test. The VESPARCH tests and CATs provide more useful information when used together; it is not our intention to suggest that VESPARCH should replace CATs. However, the differences in format and design means that certain children 'missed' in CAT testing can be identified during VESPARCH testing, and vice versa.

Rasch analysis shows that the items and whole tests have been designed to conform successfully with the Rasch model and it is evident that discrimination is good across the range of difficulty of the questions. The matched nature of the verbal and spatial tests allows reasoning ability in the two domains to be directly compared. We find that there are roughly the same numbers of children with much higher spatial than verbal scores as there are in the opposite direction. This contrasts with findings for CAT scores, particularly with those from underperforming schools, where it is more common for spatial scores to exceed verbal (see, e.g., Langdon, Rosenblatt, & Mellanby, 1998; Mellanby *et al.*, 1986). Furthermore, similar imbalance between verbal and performance scores in IQ tests such as the WISC-R has been reported for decades for children from relatively deprived backgrounds (see, e.g., Berk, 1982; Moffitt & Silva, 1987; Whittington, 1988). This is unsurprising as the verbal parts of these tests involve reading and also considerable social/educational experience, factors that are not influential in VESPARCH.

We suggest that a large discrepancy between verbal and spatial VESPARCH, in either direction, can show where educational strengths and weaknesses may lie. Those with relatively strong spatial reasoning ability can be encouraged to consider school courses with a high spatial content, which may lead to later entry into training in physics, engineering, art, medicine, and architecture. High ability in the spatial domain relative to verbal is likely to affect an individual's preference for school subjects with a spatial component, over ones with a high verbal content, such as history and literature. It is this relative spatial versus verbal strength, and personal belief in relative abilities, that

determines course selection rather than just the level of either area of ability (Nosek & Smyth, 2011).

There is an extensive literature purporting to show with post-pubertal children and adults that males are better at spatial tasks and maths than females (see Mellanby & Theobald, 2014; chapter 6). There is controversy as to whether such differences are still present when stereotype threat (Rydell, McConnell, & Beilock, 2009; Steele, 1997) has been taken into account. Our tests did not find any significant sex differences for spatial reasoning. It is also well known that girls are less likely than boys to choose STEM subjects in school and later education and careers, and this is often ascribed to lower spatial ability. However, irrespective of the cause of any difference later in spatial scores, girls' spatial VESPARCH scores could point to those for whom nurturing spatial skills might be especially beneficial from the point of view of encouraging the girls to choose STEM subjects later. Early identification of girls with high spatial VESPARCH scores could help to improve the percentage of women entering STEM subjects.

In conclusion, we propose that VESPARCH is a useful part of the 'armoury' of tests available to schools to aid in the provision of teaching suitable for every child, which should lead to improvement in academic performance.

Acknowledgement

We would like to thank Professor Dawn Langdon for helpful discussions, and the Realizing Potential Trust for their financial support. We are also very grateful to all the schools and children that took part in this research.

References

Badger, J. R., & Mellanby, J. (2016). Academic underachievement: Who and why? *The Quarterly Journal of Experimental Psychology: Proceedings of the Experimental Psychology Society*, *2016*(69), 2487–2502. https://doi.org/10.1080/17470218.2016.1237416

Berk, R. A. (1982). Verbal-performance IQ discrepancy score: A comment on reliability, abnormality and validity. *Journal of Clinical Psychology*, *38*, 638–641. https://doi.org/10.1177/073724778400900303

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. New York, NY: Routledge.

Camp, S. J., Stevenson, V. L., Thompson, A. J., Ingle, G. T., Miller, D. H., Borras, C., . . . Langdon, D. W. (2005). A longitudinal study of cognition in primary progressive multiple sclerosis. *Brain*, *128*, 2891–2898. https://doi.org/10.1093/brain/awh602

Camp, S. J., Stevenson, V. L., Thompson, A. J., Miller, D. H., Borras, C., Auriacombe, S., . . . Langdon, D. W. (1999). Cognitive function in primary progressive and transitional progressive multiple sclerosis: A controlled study with MRI correlates. *Brain*, *122*, 1341–1348. https://doi.org/10.1093/brain/122.7.1341

Carroll, J. B. (1993). *Human cognitive abilities. A survey of factor-analytic studies*. New York, NY: Cambridge University Press.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, *54*, 1–22.

Cattell, R. B. (1987). *Intelligence: Its structure, growth and action*. Amsterdam, The Netherlands: North Holland.

Coughlan, A. K., & Hollows, S. E. (1984). Use of memory tests in differentiating organic disorder from depression. *British Journal of Psychiatry*, *145*, 164–167. https://doi.org/10.1192/bjp.145.2.164

Ferrer, E., & McArdle, J. J. (2004). An experimental analysis of dynamic hypotheses about cognitive abilities and achievement from childhood to early adulthood. *Developmental Psychology*, *40*, 935–952. https://doi.org/10.1192/bjp.145.2.164

Garg, A. X., Norman, G., & Sperotable, L. (2001). How medical students learn spatial anatomy. *Lancet*, *357*, 363–364. https://doi.org/10.1016/s0140-6736(00)03649-7

Gill, T. (2014). Statistics report 62, Research Division, Cambridge Assessment.

GL Assessment (2012). *Cognitive abilities test* (4th ed.). London, UK: GL Assessment.

Horn, J. L., & Blankson, A. N. (2014). Foundation for better understanding of cognitive abilities. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment theories and issues* (3rd ed.) (pp. 73–98). London, UK: Guilford Press.

Horn, J. L., & Noll, J. (1997). Human Cognitive capabilities. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment theories and issues* (1st ed.) (pp. 53–91). London, UK: Guilford Press.

Humphreys, L. G., Lubinski, D., & Yao, G. (1993). Utility of predicting group membership and the role of spatial visualization in becoming an engineer, physical scientist or artist. *Journal of Applied Psychology*, *78*, 250–261.

Kan, K.-J., Kievit, R. A., Dolan, C., & van der Maas, H. (2011). On the interpretation of the Cattell-Horn-Carroll factor Gc. *Intelligence*, *39*, 292–302. https://doi.org/10.1016/j.intell.2011.05.003

Kelly, S. W., Griffiths, S., & Frith, U. (2002). Evidence for implicit sequence learning in dyslexia. *Dyslexia*, *8*, 43–52. https://doi.org/10.1002/dys.208

Kline, P. (1999). *The handbook of psychological testing* (2nd ed.). London, UK: Routledge.

Kvist, A. V., & Gustafsson, J.-E. (2008). The relation between fluid intelligence and the general factor as a function of cultural background: A test of Cattell's Investment Theory. *Intelligence*, *36*, 422–436. https://doi.org/10.1016/j.intell.2007.08.004

Langdon, D. W., Rosenblatt, N., & Mellanby, J. H. (1998). Discrepantly poor verbal skills in poor readers: A failure of learning or ability? *British Journal of Psychology*, *89*, 177–190. https://doi.org/10.1111/j.2044-8295.1998.tb02679.x

Langdon, D. W., & Warrington, E. K. (1995). *Verbal and spatial reasoning test*. Hove: Psychology Press.

Langdon, D. W., & Warrington, E. K. (2000). The role of left hemisphere in verbal and spatial reasoning tasks. *Cortex*, *36*, 691–702.

McGrew, K. S. (1997). Analysis of the major intelligence batteries according to a proposed Gf-Gc framework. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment theories and issues* (1st ed.) (pp. 151–179). London, UK: Guilford Press.

Mellanby, J., Anderson, R., Campbell, B., & Westwood, E. (1986). Cognitive determinants of verbal underachievement at secondary school. *British Journal of Educational Psychology*, *66*, 483–500. https://doi.org/10.1111/j.2044-8279.1996.tb01214.x

Mellanby, J., & McElwee, S. (2009). *Verbal reasoning for children*. Cambridge, UK: Cambridge Assessment.

Mellanby, J., McElwee, S., & Badger, J. R. (2016). *Verbal and spatial reasoning test for children: VESPARCH*. Cambridge, UK: Cambridge Assessment.

Mellanby, J., & Theobald, K. (2014). *Education and learning: An evidence-based approach*. Oxford, UK: Wiley Blackwell.

Moffitt, T. E., & Silva, P. A. (1987). WISC-R verbal and performance IQ discrepancy in an unselected cohort: Clinical significance and longitudinal stability. *Journal of Consulting and Clinical Psychology*, *55*(5), 768–774. https://doi.org/10.1.1.626.9457

Nosek, B. A., & Smyth, F. I. (2011). Implicit social cognitions predict sex differences in math engagement and achievement. *American Educational Research Journal.*, *48*, 1125–1156. https://doi.org/abs/10.3102/0002831211410683

Pearson Assessment (2004). *Raven's progressive matrices*. San Antonio, TX: Author.

Pearson Assessment (2007). *Naglieri nonverbal ability test: Second Edition (NNAT2)*. San Antonio, TX: Author.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press.

Rydell, R. J., McConnell, A. R., & Beilock, S. I. (2009). Multiple social identities and stereotype threat: Imbalance, accessibility and working memory. *Journal of Personality and Social Psychology*, *96*, 949–969. https://doi.org/10.1037/a0014846

Schneider, W. J., & McGrew, K. S. (2012). The Cattell-Horn-Carroll model of intelligence. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment theories and issues* (3rd ed.) (pp. 99–114). London, UK: Guilford Press.

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist*, *52*, 613–629.

Thorndike, R. L., Hagen, E., & France, N. (1986). *The cognitive abilities test* (Revised ed.). Windsor, ON: NFER-Nelson.

Thorsen, C. (2014). Dimensionality and predictive validity of school grades: The relative influence of cognitive and social behavioural aspects. Gothenburg studies in educational sciences, 356.

Thorsen, C., Gustafsson, J.-E., & Cliffordson, C. (2014). The influence of fluid and crystallized intelligence on the development of knowledge and skills. *British Journal of Educational Psychology*, *84*, 556–570. https://doi.org/10.1111/bjep.12041

Wechsler, D. (2008) *Wechsler adult intelligence scale – IV*. San Antonio, TX: Pearson Education.

Whittington, J. (1988). Large verbal-non-verbal ability differences and underachievement. *British Journal of Educational Psychology*, *58*, 205–211. https://doi.org/10.1111/j.2044-8279.1988.tb00894.x

Received 9 March 2017; revised version received 14 July 2017