Published in final edited form as:

Int J Educ Res. 2010 ; 49(2-3): 92–101. doi:10.1016/j.ijer.2010.09.003.

Sex Differences in School Science Performance from Middle Childhood to Early Adolescence

Claire M.A. Haworth^{a,*}, Philip S. Dale^b, and Robert Plomin^a

Claire M.A. Haworth: Claire.Haworth@kcl.ac.uk; Philip S. Dale: dalep@unm.edu; Robert Plomin: Robert.Plomin@kcl.ac.uk

^a Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, King's College London, De Crespigny Park, London, SE5 8AF, UK

^b Department of Speech and Hearing Sciences, University of New Mexico, 1700 Lomas Blvd NE, Albuquerque, NM 87131, USA

Abstract

We investigated whether the sexes differ in science performance before they make important course and career selections. We collected teacher-report data from a sample of children from the Twins Early Development Study (TEDS) assessed at ages 9, 10 and 12 years (N>2500 pairs). In addition we developed a test of scientific enquiry and administered it to a sub-sample of TEDS (n=1135; age=14 years). We found no evidence for mean sex differences in science performance assessed by teachers, or by a test of scientific enquiry, although boys were somewhat more variable. At a time when adolescents are making important course choices, girls are performing just as well as boys.

1. Introduction

Science and technology are becoming increasingly important to society and to the world economy, yet women make up only 24% of the science and technology workforce in the USA (National Science Board, 2002) and similar statistics are found in the UK (25%) and in Europe (23%) (Eurostat, Labour Force Survey). Even those women who do work in science and technology fields are less likely to be made professor (Handelsman et al., 2005). Although there are some areas of science in which few women excel, there are also areas of science in which women outnumber men. Women receive higher degrees as often as men in biological sciences, but achieve far fewer higher degrees in mathematics and engineering (Handelsman et al., 2005). It is likely that the under-representation of women in scientific careers is due to a combination of many factors, but much emphasis has been given to the possible role of sex differences in cognitive abilities and performance. A large body of literature on such cognitive sex differences exists, and most meta-analyses conclude that there are some differences between males and females on certain tests of cognitive ability (see Halpern et al., 2007 for a recent review). But as Halpern (1989) noted in the title of her paper: "What you see depends on where you look". Importantly, Hyde (2005) concluded

*Address for Correspondence: Claire M.A. Haworth, SGDP Centre P080, Institute of Psychiatry, King's College London, De Crespigny Park, London SE5 8AF. Tel: +44 (0) 20 7848 5433, Fax: +44 (0) 20 7848 0866, Claire.Haworth@kcl.ac.uk.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

that males and females are more alike than they are different in her work on the gender similarities hypothesis.

Performance at school is a good predictor of career success (e.g., Baird, 1985), and it is at school that young adolescents make important decisions about course choice that impact on their later professional opportunities. In the U.K. school system, students make their first course choices at age 13–14 years when they select subjects for their General Certificate of Secondary Education (GCSEs, or equivalents). Although science is a compulsory subject, there are a range of science courses available to students. A new science curriculum focusing on scientific literacy and science skills that are essential for public engagement in science-related issues was introduced in 2006 (QCA, 2005). In addition to this compulsory science literacy course, students can also choose to study separate sciences (e.g., Biology, Chemistry and Physics), applied science, or ‘additional’ science in which specific courses depend on staffing availability and schools with more resources (such as private schools) will be more likely to offer these courses. Typically more boys than girls select difficult science and mathematics courses (Radford, 1998; van Langen, Rekers-Mombarg, & Dekkers, 2006). In the present study, we investigated whether girls and boys differ in performance in science before they make these course selections because any differences later in development might reflect differences in courses taken. We also compared results for teacher reports of science performance and test data on science performance.

1.1 Sex differences in Science Performance in Adolescence

In 2005, the Nation’s Report Card (http://nationsreportcard.gov/science_2005/) in the USA reported that boys performed better than girls in science in adolescence. As noted by Hyde and Linn (2006), a closer examination of the scores of males and females indicates that the difference between males and females at all grades tested (9–10 years; 13–14 years and 17–18 years) was small (e.g., effect size (d) for 9–10 years = 0.12). This small difference is significant because of the large sample, but the presentation of results – that boys are outperforming girls – is easy to misinterpret. If all you know is the sex of a child you can predict very little about their performance in science. Highlighting even small mean differences between the sexes in performance could result in continued stereotypes that impact on career choice and career success (Linn, 1998).

The Programme for International Student Assessment (PISA) has collected data on performance in reading, mathematics and science from 15 year olds around the globe. In the most recent round of PISA, the emphasis was on science performance. Results across the Organisation for Economic Co-operation and Development (OECD) countries indicated that in the majority of countries there were no large differences in science performance between males and females (OECD average males: mean = 501, SE = 0.7; females: mean = 499, SE = 0.6). However, this small sex difference ($d = 0.02$) across all the OECD countries was significant because of the very large sample size. Results from the U.K. sample show the largest difference in favor of males of all of the OECD countries (male: mean = 520, SE = 3.0; female: mean = 510, SE = 2.8). Although this difference of 10 points in favor of males is still relatively small, with an effect size (d) of 0.09, it was significant. There was also a mixed pattern of sex differences in the U.K. when the overall science competencies score was broken down into its constituent parts: identifying scientific issues; using scientific evidence; and explaining phenomena scientifically. Females performed slightly better on identifying scientific issues ($d = -0.07$), and males performed slightly better in using scientific evidence ($d = 0.06$, NS). However, males performed much better than females in explaining phenomena scientifically (male mean (SE) = 527 (3.0); female mean (SE) = 506 (2.7)), with an effect size (d) of 0.19. Males were also superior in all content-related knowledge scales (e.g. living systems; physical systems and earth and space systems)

(OECD, 2007). Over 13,000 students participated in the U.K. testing for PISA and therefore the sample has power to detect small sex differences in performance.

The PISA science tests were administered at age 15 using paper-and-pencil tests during the school day. The PISA science test was not designed to assess performance based on a particular curriculum, but instead focused on scientific literacy, which was defined as: *“Scientific knowledge and use of that knowledge to identify questions, to acquire new knowledge, to explain scientific phenomena, and to draw evidence-based conclusions about science-related issues. Understanding of the characteristic features of science as a form of human knowledge and inquiry. Awareness of how science and technology shape our material, intellectual and cultural environments. Willingness to engage with science-related issues, and with the ideas of science, as a reflective citizen.”* (OECD, 2007; Volume 1, pages 34–35).

This PISA definition of scientific literacy is similar to the domain of scientific enquiry as assessed in the U.K. National Curriculum. Scientific enquiry in the U.K. National Curriculum includes teaching about: *“(1) The interplay between empirical questions, evidence and scientific explanations using historical and contemporary examples. (2) The importance of testing explanations by using them to make predictions and by seeing if evidence matches the predictions. (3) The ways in which scientists work today and how they worked in the past, including the roles of experimentation, evidence and creative thought in the development of scientific ideas.”* (The National Curriculum, 2004, page 72).

If sex differences, albeit small in nature, are present for science performance at the age of 15, when most adolescents have already made important choices about which courses and career paths to take, then this could go some way towards explaining why fewer women enter scientific career paths. However, these sex differences might in fact be due to differences in course choice. In the present study, we investigated whether sex differences for performance in science can be found in younger children, before they make decisions about what courses they will take, which restricts later career paths open to them.

1.2 Sex differences in Variances

When considering sex differences in science performance, it is important to consider not only mean differences but also differences in variance (Feingold, 1992; 1995). This is especially important when exploring the reasons behind the under-representation of women in scientific careers – careers more likely to be pursued by those in the upper tail of the ability distribution. Even with no mean differences, greater variance for males than females could result in more males in the top tail of the distribution (as well as in the bottom tail of the distribution). This phenomenon of greater variance in male scores has been observed for math and science performance (e.g., Feingold, 1995; Hedges & Nowell, 1995; Stanley, Benbow, Brody, Dauber, & Lupkowski, 1992; Strand, Deary, & Smith, 2006). In the PISA (2006) U.K. sample, 5.3% of males and 4.3% of females score in the lowest proficiency level, suggesting a slight over-representation of males at the lowest levels of performance. Moreover, at the high extreme of performance (proficiency levels 5 and 6), a higher proportion of males (16%) than females (11.5%) are found. This is mirrored to a lesser extent in the OECD average (10% versus 7.9%) for these high proficiency levels.

It is not clear why male scores vary more than female scores, but any greater variance for males at an early age could lead to an escalation of the differences between the sexes at the high (and low) extremes later in development.

1.3 The Current Study

The present study asked whether the under-representation of women in scientific careers is preceded by differences in science performance in school-age children. Our goal was to assess whether sex differences in school science performance emerge during the transition from middle childhood to early adolescence. To answer these questions we collected teacher-report data based on U.K. National Curriculum criteria from a representative sample of children from the Twins Early Development Study (TEDS) assessed at ages 9, 10 and 12 years. In addition, because findings may differ by assessment method, we developed a test of scientific enquiry and administered it to a sub-sample of TEDS.

2. Materials and Methods

2.1 Sample

The Twins Early Development Study (TEDS) is a study of twins born in England and Wales between 1994 and 1996 (Oliver & Plomin, 2007). The TEDS sample has been shown to be reasonably representative of the general population in terms of parental education, ethnicity and employment status (Kovas, Haworth, Dale, & Plomin, 2007). One randomly selected member of each twin pair was selected for these analyses; because scores from twin pairs are not independent measurements. Teacher report data are available for more than 2500 pairs of TEDS twins at 9, 10 and 12 years of age, and internet-based test performance is available in a sub-sample (N=1135 individuals) at 14 years of age. This sub-sample consists of the first birth cohort in TEDS (children born in 1994), and is representative of the entire TEDS sample. Exact N values for each measure, split by sex, are shown in Table 1.

2.2 Measures

Two types of performance data were collected: teacher reports based on the UK National Curriculum, and test data from a novel internet-based version of our scientific enquiry test. This internet-based test was developed specifically for this study, and we describe the development and piloting of this test here.

National Curriculum Teacher Reports—As for all children in the U.K., the twins' academic performance was assessed throughout the school year by their teachers, using the assessment materials of the National Curriculum for England and Wales (NC), the core academic curriculum developed by the Qualifications and Curriculum Authority (QCA). Teachers were contacted when the children were in the second half of their school year so that the teachers would be familiar with the children's performance during the school year. Teachers were sent a covering letter with the background and aims of TEDS, as well as explaining that we had obtained consent from the twins' parents to ask teachers for information about the child's performance at school. For the 9- and 10-year studies, the NC Teacher Assessments at Key Stage 2 were used. These are familiar to teachers and are designed for children age 8 through their sixth year of primary school at age 11. Teachers assessed three broad areas of science ability: Scientific enquiry, Life processes and living things, and Physical processes. A fourth domain of science, Materials and their properties, was accidentally omitted from the assessments at Key Stage 2. However, the science domains are all highly intercorrelated (as reported later), including this additional domain which was assessed at age 12. This is therefore unlikely to impact on the findings of this study. For the 12-year study, the NC Teacher Assessments at Key Stage 3 were used, which are designed for children aged 11–14 years old. At 12 years, teachers assessed four broad areas of science ability: Scientific enquiry, Life processes and living things, Physical processes, and Materials and their properties. Teachers provide a rating from 1–9 to indicate performance (Level 9 = exceptional performance). As with other developmentally sensitive assessments, these levels are linked to defined developmental criteria and therefore not all

levels can be attained at every age. For this reason scores for the Key Stage 2 (ages 9 and 10) assessments range only from levels 1 to 5. By the end of Key Stage 2 (age 11), children are expected to reach level 4, and by the end of Key Stage 3 at age 14, children are expected to reach levels 5/6. For scientific enquiry, example criteria for level 3 (which is the mean level of performance of TEDS twins at age 9) are: (1) Pupils respond to suggestions and put forward their own ideas about how to find the answer to a question. (2) Pupils recognize why it is important to collect data to answer questions. (3) They make relevant observations and measure quantities, such as length or mass, using a range of simple equipment. At 12 years the TEDS twins are performing between levels 4 and 5 on average. For level 4, criteria for scientific enquiry include: (1) Pupils recognize that scientific ideas are based on evidence. (2) In their own investigative work, they decide on an appropriate approach (for example, using a fair test) to answer a question. (3) They begin to plot points to form simple graphs, and use these graphs to point out and interpret patterns in their data. For level 5, criteria for scientific enquiry include: (1) Where appropriate, pupils make predictions based on their scientific knowledge and understanding. (2) They begin to repeat observations and measurements and to offer simple explanations for any differences they encounter. (3) They draw conclusions that are consistent with the evidence and begin to relate these to scientific knowledge and understanding. (4) They make practical suggestions about how their working methods could be improved. Further information about the National Curriculum, including the full level descriptions, is available at <http://curriculum.qca.org.uk>.

Similar to other measures used in the behavioural sciences, this is an ordinal rather than interval scale but it is better than most ordinal Likert scales in that it attempts to specify behavioural criteria at each level rather than merely indicating performance relative to an unspecified average performance. Teachers are familiar with the use of these criteria because they follow the NC for England and Wales; in addition we provide a reminder of the level descriptions with the questionnaire. As discussed below, we supplement the expert teacher ratings with a test of scientific enquiry.

These judgments were not made specifically for the present study, but rather constitute the continuing assessment of each child that ultimately leads to the final NC Teacher Assessment score submitted to the QCA at the end of the Key Stage to indicate the child's academic achievement during that period.

The science scales are highly correlated, with an average inter-correlation of .85 at 9 years, .87 at 10 years, and .93 at 12 years. A factor analysis indicated that the first principal component accounted for 90% of the variance at 9 years, 91% of the variance at 10 years, and 95% of the variance at 12 years. We therefore also present results of a composite science score calculated by taking a mean of the individual scores.

Test Development—To complement our teacher reports of science performance we developed a test of scientific enquiry. We decided to focus on scientific enquiry for two reasons. First, we had a limited amount of testing time available and wanted to assess one domain well rather than trying to assess many domains only superficially. Secondly, we chose scientific enquiry because it can be applied to diverse science topics, would be less influenced by differences in school curriculum (timing and content), and because the concept of scientific enquiry skills fitted better with our conceptualization of science ability than a test of learned facts. In addition, using scientific enquiry allowed us to make more direct comparisons to PISA's similar scientific literacy scale. PISA opted for scientific literacy because the goal was not to assess different school science programs, but rather to assess whether students could use scientific knowledge and techniques to answer questions that are socially and globally relevant, including the ability to interpret and evaluate

evidence (OECD, 2003). These skills are essential for future engagement in science-related public policy.

Items assessing scientific enquiry were drawn from a range of publicly available tests including the NFER science series (Russell & McGuigan, 2001), past papers of the U.K. National Curriculum Tests (available for download at <http://orderline.qca.org.uk/>), past papers of GCSE tests (taken by all 16 year olds in the UK) (available for download at exam board websites – e.g. Edexcel: <http://www.edexcel.org.uk/home/>; OCR: <http://www.ocr.org.uk/pastpapermaterials/>; and AQA: <http://www.aqa.org.uk/index.php>) and items from the TIMMS study (<http://nces.ed.gov/timss/educators.asp>). The PISA scientific literacy items were not available at this time. Items were selected based on their difficulty level. They were also selected on the basis of their suitability for conversion to an internet-based test; for example, items that included figures or tables needed to fit on a single computer screen, and all open-ended items had to be converted to multiple-choice format.

We then invited nine children (mean age = 13.78 years, $sd = .44$) to our laboratory to take part in a focus group to discuss the test items. During this focus group, we administered two forms of the test, each with a core set of items before discussing the items with the teenagers. Information from this focus group allowed us to refine the test –including editing ambiguous questions, removing items that were too easy and increasing the length of the test. The refined test was then piloted, as a paper-and-pencil test on 50 children (56% male; mean age = 12.91 years; $sd = .35$) all in the same grade at one school. The test was administered as part of their normal science lessons during the final term of the academic year.

The paper-and-pencil test consisted of 40 questions (some multi-part) comprising 76 separate items, which were in a random order (i.e. not ordered by difficulty). The children were given 50 minutes to complete the test. Not all children completed every test item, so we calculated a mean test score based on the number of items completed. Although the mean test score for individuals who completed all of the items was slightly higher than those who only completed a subset of the items, this difference was not significant. The test had high internal consistency reliability (Cronbach's $\alpha=0.74$).

Example items include: (1) *The primary reason scientists repeat the measurements they take during experiments is so that they can: a) check that the equipment is working; b) list all the results in a table; c) estimate experimental error; d) change the experimental conditions.* This question is adapted from the TIMMS 1999 Scientific enquiry and nature of science questions, and 47% of children in the school sample answered correctly (correct answer = estimate experimental error). (2) *Two open bottles, one filled with vinegar and the other with olive oil, were left on a windowsill in the sun. Several days later it was observed that the bottles were no longer full. What can be concluded from this observation? a) Vinegar evaporates faster than olive oil; b) Olive oil evaporates faster than vinegar; c) both vinegar and olive oil evaporate; d) only liquids containing water evaporate; e) direct sunlight is needed for evaporation.* This question is adapted from the TIMMS 1999 Scientific enquiry and nature of science questions. 80% of the teenagers in the school sample answered correctly (correct answer = both vinegar and olive oil evaporate).

Internet-based Test Performance—Increasing access to inexpensive and fast internet connections in the U.K. makes it possible to test large samples efficiently and economically online. We have previously shown that an internet-based cognitive test battery is a reliable and valid method for collecting cognitive data on children as young as 10 years old (Haworth et al., 2007). Drawing on our experience of creating effective internet-based batteries for the TEDS sample at 10 and 12 years, we created an internet-based version of

the scientific enquiry test to allow administration to the TEDS sample which is spread across the U.K. More than 80% of the TEDS' families have access to the internet at home and most children without access to the internet at home have access in their schools and local libraries.

We found no correlation between internet speed (i.e., broadband versus dial-up connection) and socio-economic status (SES) for our previous internet-based samples at 10 and 12 years. We guarded against potential problems associated with research on the internet; for example, the testing was administered by a secure server in the TEDS office, which also provided a secure site for data storage; identifying information was kept separately from the data. Appropriate safeguards were in place that prevented children from answering the same item more than once. We provided technical support and assigned a caller to each family, who contacted the family at the start of testing and provided support and encouragement throughout testing. Furthermore, our toll-free telephone number was available to parents and children in case of any problems or questions. The test battery is self-paced, and can be completed over a period of several weeks. Each child's performance is monitored online and families are telephoned by their caller to provide support. To create the internet test battery, we worked with Planet Three Publishing (www.planet3.co.uk) and e-Business Systems (www.e-businesssystems.co.uk).

Following piloting, we refined the test and prepared it for administration on the internet. The internet-based test included 24 questions, comprising 39 individual items, with high internal consistency reliability (Cronbach's $\alpha=0.92$). On average, the test takes 22 minutes to complete. A total score was calculated by taking a sum of the items. Example test items are shown in Figures 1 and 2.

2.3 Analyses

To investigate mean sex differences in performance we used an independent samples t-test. We report d as a measure of effect size, and the p -value from both the student's t-test (which assumes equality of variances) and the Welch's t-test (which does not assume equality of variances). To evaluate variance differences between the sexes we used the Levene's homogeneity of variances test. We also calculated the variance ratio (VR), by dividing the male variance by the female variance (Feingold, 1992) to provide an indicator of the effect size of the variance difference. A VR above one indicates greater male variance and a VR below one indicates greater female variance.

3. Results

3.1 Teacher Reports

Teacher report data based on the National Curriculum were collected from the teachers of the children in TEDS at 9, 10 and 12 years of age. The means and standard deviations (SD) for the teacher report measures at 9, 10 and 12 years are presented in Table 1. Males have slightly greater means and variances than females on all of the science scores. The results of the t-tests, shown in Table 1, indicate no significant effects of sex on any of the teacher report measures (e.g., student's t-test results: 9 year Science Composite: $p = .122$; 10 year Science Composite: $p = .305$; 12 year Science Composite: $p = .258$). This was despite the large sample size, which provides 90% power to detect mean differences as small as 0.1 SD (i.e., $d = .1$) that account for just 0.2% of the variance (Cohen, 1988). We repeated the analyses in the remaining twin of each pair and found a similar pattern of results (details available from first author).

3.2 Test Data

Test data were collected from 1135 individuals (from the TEDS sample) via internet-based testing. Means and standard deviations (*SD*) for the test measure, split by sex, are presented in Table 1. Male and female means were almost identical, and males showed only slightly greater variance than females. The results of the t-tests, shown in Table 1, indicate no significant effects of sex on the test scores (student's t-test results: $p = .728$).

3.3 Variance Differences

The magnitude of the variance difference between males and females was calculated using variance ratios (VRs). All variance ratios were greater than one indicating that males had greater variance than females on all of the measures (Table 2). For the teacher report data, males were on average 14% more variable than females across the ages. Results from the test data indicate a similar pattern of greater male variance. Results from the homogeneity of variances test for the teacher report data indicate that there are significant variance differences; eight of the thirteen teacher-report measures show a significant difference ($p < .05$). However the distributions of male and female scores are remarkably similar (Figure 3). Due to the small numbers at the extremes it is not possible to test for an over-representation of males at the high and low extremes.

4. Discussion

At a time when adolescents are making their first course and career choices, we find no evidence for mean sex differences in any of the domains of science performance. In addition, we find only slight variance differences, with males showing greater variance in scores than females. Our finding of no mean sex differences includes not only teacher reports of performance but also direct test performance. Finding similar results for the two assessment techniques is notable because previous studies have found that females gain higher grades than males in terms of school grades in adolescence (Dwyer & Johnson, 1997; Kenney-Benson, Pomerantz, Ryan, & Patrick, 2006). Whereas, females tend to score lower than males on college entry tests, particularly when the test is not closely aligned to what has been studied in school (Willingham & Cole, 1997). In our study we find no such difference between teacher reports and test data. It is possible that sex differences may emerge later in development, particularly as girls mature earlier than boys, but any changes in relative performance may, as noted earlier, be due to the courses the adolescents choose to study.

If females are just as good at science as males in early adolescence, why do we see under-representation of females in scientific careers, particularly in the physical and engineering sciences? We could speculate that cognitive factors do not by themselves make a good scientist; other skills and attributes are also important. In addition, although we assume that diverse cognitive skills (e.g., spatial, mathematical and verbal) are required for success in science, we are not aware of any empirical evidence that tests this hypothesis. It is also possible that different fields of science require different complements of skills and attributes, and differentially attract males and females. Another important consideration is motivation and interest in science. Evidence suggests that there is dwindling enthusiasm for school science in early adolescence (Jenkins & Nelson, 2005; Osborne, Simon, & Collins, 2003), an effect that is more marked for females than males (Miller, Slawinski Blessing, & Schwartz, 2006; Osborne et al., 2003; Speering & Rennie, 1996). This effect could be particularly important in the U.K., as adolescents now have more choice in the science topics they study at 15 and 16 years. These ideas are not new: Eccles' (1983) expectancy value model considered both aptitudes and personal beliefs to be important in predicting achievement-related choices. More recent analyses have shown that motivation predicts school course choice above that predicted by grades (Simpkins, vis-Kean, & Eccles, 2006).

Our data cannot speak to the many determinants of course choice and later career choice. Our data only suggest that other factors may be more important than differences in performance between males and females. More information on the other predictors of both course and career selection, such as the effects of attitudes towards science is needed, specifically investigating the antecedents of any sex differences.

In the TEDS sample we have some limited data on interests in science at ages 9 and 12 years. A brief paper-and-pencil questionnaire contained three items that asked children how much they like science-based activities. Responses were made on a 5-point Likert Scale, ranging from “don’t like it at all” to “like it very much”. The three interest items were: “How much do you like... (1) learning about nature and living things? ... (2) testing things out to see what they can do (e.g., magnets)? and ... (3) finding out how things work (e.g., the human body)? A composite measure of interest was created by taking a mean of the item scores. These items (along with items relating to English and Mathematics performance) have been used previously in TEDS to predict academic performance (see Spinath, Spinath, Harlaar and Plomin 2006 for more details).

Males showed higher interest scores at 9 and 12 years, although the effect was small, decreased with age, and was statistically significant only at 9 years (9 year: male mean (sd) = 4.05 (0.90); female mean (sd) = 3.96 (0.87); $p = .004$, $d=0.10$; 12 year: male mean (sd) = 3.77 (0.85); female mean (sd) = 3.73 (0.83); $p = .055$, $d=0.05$). The correlation between interest and performance was low at both 9 and 12 years (9 years: $r = -0.003$; $p = 0.900$; $n = 2304$; 12 years: $r = 0.046$; $p = 0.008$; $n = 3232$).

Even with this very brief questionnaire we find that males are slightly more interested in science as young as 9 years of age, though the effect size is small, as is the relationship with performance. More advanced and age appropriate measurements of science interest are needed to determine the validity and generality of these findings, and to explore how interests impact on later performance in science.

4.1 Limitations

One limitation of the present study is that the data comes from a twin sample. Twins have been found to differ from singletons (Boomsma, Busjahn, & Peltonen, 2002; Martin, Boomsma, & Machin, 1997); for example, twins on average have lower birth weights and are often born three to four weeks prematurely (Plomin, DeFries, McClearn, & McGuffin, 2008). For example, some studies have found that twins have lower IQ scores when compared to singletons (Record, McKeown, & Edwards, 1970), with triplets showing even lower IQ scores than twins. However, those studies that have found differences between twins and singletons have been conducted on young twins, and studies of older twins confirm that these differences have all but disappeared by early to middle childhood (Calvin et al., 2009; Christensen et al., 2006).

Another potential limitation is the use of teacher reports of scientific performance. Although the validity of teacher assessments has been questioned (e.g., Davies & Brember, 1994; Demaray & Elliot, 1998; Glascoe, 2001; Reeves, Boyle, & Christie, 2001), a review of the literature has concluded that on the whole they are valid (Hoge & Coladarci, 1989). Using TEDS data we have shown that teacher assessments of reading performance correlate highly (.68) with a telephone-administered test of word and nonword reading administered at the same age (Dale, Harlaar, & Plomin, 2005). However, because the national curriculum ratings are based on specific criteria (rather than a being a comparison to an un-defined average performance at each age) it is not possible to attain all of the levels at every age. For example, only levels 1 to 5 were attained during Key Stage 2. Although the distributions

remain near-normal (see Figure 3) it should be noted that the restriction of range for the Key Stage 2 assessments could impact on our analysis of variance differences.

The use of teacher assessments of science performance may be a strength rather than a limitation of our study, since there is some evidence to support the hypothesis that teacher assessments are likely to add to achievement tests in predicting long-term outcomes. For example, after controlling for socioeconomic status, preschool teachers' overestimates and underestimates of performance relative to IQ scores at 4 years significantly predicted high school grades and Scholastic Aptitude Test results 14 years later (Alvidrez & Weinstein, 1999). A similar study of teacher assessments of underachieving students predicted long-term educational attainment and career outcomes (McCall, Evahn, & Kratzer, 1992). In addition it should be noted that test scores are not without their own limitations (Good & Salvia, 1988; Livingstone, 1995; Marks, 1990; Sharpley & Edgar, 1986).

Finally, it would have been beneficial to collect test data on different aspects of science performance, rather than just focusing on scientific enquiry. As explained in the methods section, we were only able to assess scientific enquiry; however, we feel that scientific enquiry better encapsulates our definition of science, and certainly relates more to science use in society – this is in parallel with the explanations given by PISA for targeting scientific literacy.

4.2 Conclusions

We find no evidence for mean sex differences in science performance in middle childhood and early adolescence. This finding is consistent across the different domains of science assessed, as well as across both expert teacher ratings and test data. Greater variance in males could result in more males at the high and low extremes even though there is no overall mean difference. Males were more variable than females, and this finding should be followed up longitudinally to assess whether these small differences in early adolescence increase in adulthood. At a time when adolescents are making important career decisions, girls are performing just as well as boys at school science.

Acknowledgments

We gratefully acknowledge the ongoing contribution of the parents and children in the Twins Early Development Study (TEDS). TEDS is supported by a program grant (G0500079) from the U.K. Medical Research Council; our work on school environments and academic achievement is also supported by grants from the US National Institutes of Health (HD44454 and HD46167).

Abbreviations

TEDS	Twins Early Development Study
GCSE	General Certificate of Secondary Education
PISA	Programme for International Student Assessment
OECD	Organisation for Economic Co-operation and Development
NC	National Curriculum for England and Wales
VR	variance ratio

6. Reference List

Alvidrez J, Weinstein RS. Early teacher perceptions and later student academic achievement. *Journal of Educational Psychology*. 1999:731–746.

- Baird LL. Do grades and tests predict adult accomplishment? *Research in Higher Education*. 1985; 23:3–85.
- Boomsma D, Busjahn A, Peltonen L. Classical twin studies and beyond. *Nature Reviews Genetics*. 2002; 3:872–882.
- Calvin C, Fernandes C, Smith P, Visscher PM, Deary IJ. Is there still a cognitive cost of being a twin in the UK? *Intelligence*. 2009; 37:243–248.
- Christensen K, Petersen I, Skytthe A, Herskind AM, McGue M, Bingley P. Comparison of academic performance of twins and singletons in adolescence: follow-up study. *British Medical Journal*. 2006; 333:1079–1130. [PubMed: 17124199]
- Cohen, J. *Statistical power analysis for the behavioral sciences*. 2. Hillsdale, New Jersey: Lawrence Erlbaum Associates; 1988.
- Dale P, Harlaar N, Plomin R. Telephone testing and teacher assessment of reading skills in 7-year-olds: I. Substantial correspondence for a sample of 5808 children and for extremes. *Reading and Writing: An Interdisciplinary Journal*. 2005:385–400.
- Davies J, Brember I. The first mathematics standard assessment tasks at KSI: issues by a five school study. *Berkeley EECS Research Journal*. 1994; 20:35–40.
- Demaray MK, Elliot SN. Teachers' judgement of students' academic functioning: A comparison of actual and predicted performances. *School Psychology Quarterly*. 1998; 13:8–24.
- Dwyer, CA.; Johnson, LM. Grades, accomplishments, and correlates. In: Willingham, WW.; Cole, NS., editors. *Gender and fair assessment*. Mahwah, NJ: Erlbaum; 1997. p. 127-156.
- Eccles, JS.; Adler, TF.; Futterman, R.; Goff, SB.; Kaczala, CM.; Meece, JL., et al. Expectancies, values, and academic behaviors. In: Spence, JT., editor. *Perspective on achievement and achievement motivation*. San Francisco: Freeman; 1983. p. 75-146.
- Eurostat, ELFS. EU Labour Force Survey. 2008. Available at: <http://www.ukrc4setwomen.org/html/research-and-statistics/statistics/eu-comparisons/>
- Evans DM, Martin NG. The validity of twin studies. *Genescreen*. 2000; 1:77–79.
- Feingold A. Sex differences in variability in intellectual abilities: A new look at an old controversy. *Review of Educational Research*. 1992; 62:61–84.
- Feingold A. The additive effects of differences in central tendency and variability are important in comparisons between groups. *American Psychologist*. 1995; 50:13.
- Glascocoe FP. Can teachers' global ratings identify children with academic problems? *Developmental and Behavioral Pediatrics*. 2001; 22:163–68.
- Good RH, Salvia J. Curriculum bias in published, norm- referenced reading tests: Demonstrable effects. *School Psychology Review*. 1988; 17:51–60.
- Halpern DF. The disappearance of cognitive gender differences: What you see depends on where you look. *American Psychologist*. 1989; 44:1156–1158.
- Halpern DF, Benbow CP, Geary DC, Gur RC, Hyde JS, Gernsbacher MA. The Science of Sex Differences in Science and Mathematics. *Psychological Science in the Public Interest*. 2007; 8:1–51.
- Handelsman J, Cantor N, Carnes M, Denton D, Fine E, Grosz B, et al. More women in science. *Science*. 2005; 309:1190–1191. [PubMed: 16109868]
- Haworth CMA, Dale PS, Plomin R. A twin study into the genetic and environmental influences on academic performance in science in 9-year-old boys and girls. *International Journal of Science Education*. 2008; 30:1003–1025. [PubMed: 20111670]
- Haworth CMA, Harlaar N, Kovas Y, Davis OSP, Oliver BR, Hayiou-Thomas ME, et al. Internet cognitive testing of large samples needed in genetic research. *Twin Research and Human Genetics*. 2007; 10:554–563. [PubMed: 17708696]
- Hedges LV, Nowell A. Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science*. 1995; 269:41–45. [PubMed: 7604277]
- Hoge RD, Coladarci T. Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research*. 1989; 59:297–313.
- Hyde JS. The Gender Similarities Hypothesis. *American Psychologist*. 2005; 60:581–592. [PubMed: 16173891]

- Hyde JS, Linn MC. Gender similarities in mathematics and science. *Science*. 2006; 314:599–600. [PubMed: 17068246]
- Jenkins EW, Nelson NW. Important but not for me: students' attitudes towards secondary school science in England. *Research in Science & Technological Education*. 2005; 23:41–57.
- Kenney-Benson GA, Pomerantz EM, Ryan AM, Patrick H. Sex differences in math performance: The role of children's approach to schoolwork. *Developmental Psychology*. 2006; 42:11–26. [PubMed: 16420115]
- Kovas Y, Haworth CMA, Dale PS, Plomin R. The genetic and environmental origins of learning abilities and disabilities in the early school years. *Monographs of the Society for Research in Child Development*. 2007; 72:vii-160.
- Linn MC. When Good Intentions and Subtle Stereotypes Clash: The Complexity of Selection Decisions. *Educational Researcher*. 1998; 27:15–17.
- Livingstone DW. For whom The Bell Curve tolls. *Alberta Journal of Educational Research*. 1995; 41:335–341.
- Marks D. Cautions in interpreting district-wide standardized mathematics achievement test results. *Journal of Educational Research*. 1990; 83:349–354.
- Martin N, Boomsma DI, Machin G. A twin-pronged attack on complex trait. *Nature Genetics*. 1997; 17:387–392. [PubMed: 9398838]
- McCall, RB.; Evahn, C.; Kratzer, L. High school underachievers: What do they achieve as adults?. Pittsburgh: Sage; 1992.
- Miller PH, Slawinski Blessing J, Schwartz S. Gender differences in high-school students' views about science. *International Journal of Science Education*. 2006; 28:363–381.
- National Science Board. Science and Engineering Indicators - 2002. Arlington, VA: National Science Foundation; 2002.
- OECD. The PISA 2003 Assessment Framework. OECD; 2003.
- OECD. PISA 2006 Science competencies for tomorrow's world. OECD; 2007.
- Oliver BR, Plomin R. Twins Early Development Study (TEDS): A multivariate, longitudinal genetic investigation of language, cognition and behavior problems from childhood through adolescence. *Twin Research and Human Genetics*. 2007; 10:96–105. [PubMed: 17539369]
- Osborne J, Simon S, Collins S. Attitudes towards science: a review of the literature and its implications. *International Journal of Science Education*. 2003; 25:1049–1079.
- Plomin, R.; DeFries, JC.; McClearn, GE.; McGuffin, P. Behavioral genetics. 5. New York: Worth Publishers; 2008.
- QCA. Changes to the curriculum from 2006 for key stage 4. Qualifications and Curriculum Authority; 2005. <https://www.orderline.qca.org.uk/gempdf/1847210686.pdf>
- Radford, J. Gender and choice in education and occupation. London: Routledge; 1998.
- Record RG, McKeown T, Edwards JH. An investigation of the differences in measured intelligence between twins and single births. *Annals of Human Genetics*. 1970:11–20.
- Reeves DJ, Boyle WF, Christie T. The relationship between teacher assessments and pupil attainments in standard test tasks at Key Stage 2, 1996–98. *British Educational Research Journal*. 2001; 27:141–160.
- Russell, T.; McGuigan, L. Science Assessment Series 2. London: nferNELSON Publishing Company Ltd; 2001.
- Sharpley CF, Edgar E. Teachers ratings vs standardized tests: An empirical investigation of agreement between two indices of achievement. *Psychology in the Schools*. 1986; 23:106–111.
- Simpkins SD, vis-Kean PE, Eccles JS. Math and Science Motivation: A Longitudinal Examination of the Links between Choices and Beliefs. *Developmental Psychology*. 2006; 42:14.
- Speering W, Rennie L. Students' perceptions about science: The impact of transition from primary to secondary school. *Research in Science Education*. 1996; 26:283–298.
- Spinath B, Spinath FM, Harlaar N, Plomin R. Predicting school achievement from general cognitive ability, self-perceived ability and intrinsic value. *Intelligence*. 2006; 34:363–374.

- Stanley, JC.; Benbow, CP.; Brody, LE.; Dauber, S.; Lupkowski, A. Gender differences on eighty-six nationally standardized aptitude and achievement tests. In: Colangelo, N.; Assouline, SG.; Ambrosion, DL., editors. Talent Development. New York: Trillium Press; 1992. p. 42-65.
- Strand S, Deary IJ, Smith P. Sex differences in Cognitive Abilities Test scores: A UK national picture. *British Journal of Educational Psychology*. 2006; 76:463–480. [PubMed: 16953957]
- The National Curriculum. The National Curriculum. Handbook for secondary teachers in England. QCA; 2004.
- van Langen A, Rekers-Mombarg L, Dekkers H. Sex-related differences in the determinants and process of science and mathematics choice in pre-university education. *International Journal of Science Education*. 2006; 28:71–94.
- Willingham, WW.; Cole, NS. Gender and fair assessment. Mahwah, NJ: Erlbaum; 1997.

Figure 1.

Example Item from Internet-Based Test of Scientific Enquiry

Note. This question was adapted from an Edexcel GCSE paper. 95% of the internet sample answered this question correctly. The testee must click on the correct region of the graph. (correct answer: C).

Figure 3. Percentage of Males and Females across the Scientific Enquiry National Curriculum Levels for Over 2500 Individuals at 9, 10 and 12 years, Indicating Similarity in the Male and Female Distributions and Showing the Small Variance Differences between the Sexes

Table 1

Means (Standard Deviations) and T-test Results for Science Performance in Boys and Girls Assessed by Teachers, and with an Internet-Based Test

Age	Measure	Teacher Report (National Curriculum)					Student's T-test	Welch's T-test
		Mean (sd)	Male Mean (sd)	Female Mean (sd)	d			
9	Scientific Enquiry	2.96 (0.63) n = 2729	2.98 (0.65) n = 1289	2.93 (0.60) n = 1440	0.08		p = .061	p = .062
	Life Processes and Living Things	3.01 (0.59) n = 2713	3.02 (0.62) n = 1280	3.00 (0.56) n = 1433	0.03		p = .412	p = .415
	Physical Processes	2.98 (0.60) n = 2693	3.00 (0.63) n = 1271	2.96 (0.58) n = 1422	0.07		p = .092	p = .094
	Science Composite	2.98 (0.58) n = 2735	3.00 (0.60) n = 1291	2.97 (0.55) n = 1444	0.05		p = .122	p = .124
10	Scientific Enquiry	3.32 (0.71) n = 2858	3.34 (0.73) n = 1347	3.31 (0.69) n = 1511	0.04		p = .255	p = .257
	Life Processes and Living Things	3.42 (0.69) n = 2833	3.42 (0.71) n = 1334	3.41 (0.67) n = 1499	0.01		p = .784	p = .785
	Physical Processes	3.37 (0.69) n = 2805	3.40 (0.71) n = 1323	3.35 (0.66) n = 1482	0.07		p = .091	p = .093
	Science Composite	3.37 (0.66) n = 2864	3.39 (0.69) n = 1350	3.36 (0.64) n = 1514	0.05		p = .305	p = .307
12	Scientific Enquiry	4.39 (0.98) n = 3801	4.42 (1.01) n = 1792	4.36 (0.96) n = 2009	0.06		p = .085	p = .086
	Life Processes and Living Things	4.50 (0.96) n = 3762	4.51 (0.99) n = 1774	4.49 (0.93) n = 1988	0.02		p = .575	p = .576
	Physical Processes	4.46 (0.96) n = 3749	4.48 (0.99) n = 1766	4.44 (0.94) n = 1983	0.04		p = .234	p = .236
	Materials and their Properties	4.47 (0.96) n = 3761	4.49 (0.98) n = 1768	4.45 (0.94) n = 1993	0.04		p = .234	p = .236
	Science Composite	4.46 (0.94) n = 3815	4.47 (0.97) n = 1797	4.44 (0.91) n = 2018	0.03		p = .258	p = .260
Test Data								
14	Scientific Enquiry	30.64 (10.79) n = 1135	30.77 (11.03) n = 474	30.54 (10.63) n = 661	0.02		p = .728	p = .729

Note. Analyses with twin data were done on one randomly selected member of each twin pair. NB. 9 year data presented previously (Haworth et al., 2008). Results from the Welch's t-test are included because this test does not assume equality of variances.

The assessments at ages 9, 10, 12 and 14 years correspond to years 4, 5, 7, and 9 respectively in the UK school system. The timing of these assessments was due to the timing of the larger TEDS study, which is why these assessments do not all fall at the end of Key Stages.

Variance Ratios between Males and Females and Results of Levene's Homogeneity of Variance Test for Science Measures at 9, 10, 12 and 14 Years

Table 2

Age	Measure	Male Variance	Female Variance	Variance Ratio	Levene's F	p
Teacher Report (National Curriculum)						
9	Scientific Enquiry	0.42	0.36	1.17	.604	.437
	Life Processes and Living Things	0.38	0.31	1.23**	12.687	<.001
	Physical Processes	0.40	0.34	1.18	.595	.440
10	Science Composite	0.36	0.30	1.20	1.964	.161
	Scientific Enquiry	0.53	0.48	1.10**	10.898	.001
	Life Processes and Living Things	0.50	0.45	1.11**	7.085	.008
12	Physical Processes	0.50	0.44	1.14**	13.316	<.001
	Science Composite	0.48	0.41	1.17*	5.696	.017
	Scientific Enquiry	1.02	0.92	1.11*	4.581	.032
14	Life Processes and Living Things	0.98	0.86	1.14*	5.667	.017
	Physical Processes	0.98	0.88	1.11	3.424	.064
	Materials and their Properties	0.96	0.88	1.09	3.482	.062
	Science Composite	0.94	0.83	1.13*	4.184	.041
Test Data						
14	Scientific Enquiry	121.66	113.00	1.08	1.062	.303

* significant at alpha <.05;

** significant at alpha <.01.