Author(s): Strand, Steve; Deary, Ian J.; Smith, Pauline

Article Title: Sex differences in Cognitive Abilities Test scores: A UK national picture

Year of publication: 2006

Link to published version: http://dx.doi.org/ 10.1348/000709905X50906

Publisher statement: None

Sex Differences in Cognitive Abilities Test Scores: A UK National Picture

Steve Strand

University of Warwick, Coventry, UK

Ian J. Deary

University of Edinburgh, Edinburgh, UK

Pauline Smith

nferNelson, London, UK

Please cite the published version, 2006, as detailed below.

British Journal of Educational Psychology, 76, (3), 463-480.

(First submitted, 20th January 2004; Revised and resubmitted, 12th October 2004; Further revised and resubmitted, 31st January 2005; accepted February 2005, in print 2006).

Correspondence to:

Dr Steve Strand, Reader in Education, University of Warwick, Coventry, CV4 7AL

Tel. (024) 7652 2197. e-mail: steve.strand@warwick.ac.uk.

Sex Differences in Cognitive Abilities Test Scores: A UK National Picture

Abstract

Background and aims

There is uncertainty about the extent or even existence of sex differences in the mean and variability of reasoning test scores (Lynn, 1994, 1998; Mackintosh, 1996, Jensen, 1998). This paper analyses the Cognitive Abilities Test (CAT) scores of a large and representative sample of UK pupils to determine the extent of any sex differences.

Sample

A nationally-representative UK sample of 320,000+ school pupils aged 11-12 years was assessed on the CAT (Third Edition) between September 2001 and August 2003. The CAT includes separate nationally standardised tests for Verbal, Quantitative and Non-Verbal reasoning. The size and recency of the sample is unprecedented in research on this issue.

Methods

The sheer size of the sample ensures that any sex difference will achieve statistical significance. Therefore effect sizes(d) and variance ratios (VR) are employed to evaluate the magnitude of sex differences in mean scores and in score variability respectively.

Results

The mean verbal reasoning score for girls was 2.2 standard score points higher than the mean for boys, but only 0.3 standard points in favour of girls for NVR, and 0.7 points in favour of boys for QR. However for all three tests there were substantial sex differences in the standard deviation of scores, with greater variance among boys. Boys

were over-represented relative to girls at both the top and the bottom extremes for all tests, with the exception of the top 10% in verbal reasoning.

Conclusions

Given the small differences in means, explanations for gender differences in wider domains such examination attainment at age 16 need to look beyond conceptions of ‘ability’. Boys tend to be both the lowest and the highest performers in terms of their reasoning abilities, which warns against the danger of stereotyping boys as low achievers.

Sex Differences in Cognitive Abilities Test Scores: A UK National Picture

The question of sex differences in cognitive performance has a long history in psychology and education. The issue has a high profile within the current UK educational context. National testing in England has provided data to show that girls outperform boys in assessments of English at age 7, 11 and 14, although differences in mathematics and science are less clear-cut. In public examinations at age 16, girls again achieve greater success than boys. For example in General Certificate of Secondary Education (GCSE) public examinations in England in 2002, Department for Education and Skills (DfES, 2002) statistics show that 57% of girls, but only 46% of boys, achieved 5 or more higher grade (A*-C) passes. In individual subjects, the proportion of girls achieving A*-C grades exceeded the proportion for boys not just in English but also in subjects where males have traditionally been thought to have an advantage, such as mathematics, business studies, design & technology, science and information technology. The only GCSE subject in which the performance of boys exceeded that of girls was physics, in which 90% of boys, compared with 89% of girls, achieved an A*-C grade.

The public and media are intensely interested in this so-called ‘gender gap’, reflected in headlines such as *‘Failing boys “public burden number one”* (Times Educational Supplement [TES], 27th November 1998); ‘*Gender gap widens to a gulf’* (TES, 29th January 1999); ‘*Bright girls leave boys out-classed’* (TES, 16th June 2000); *‘Boys in crisis’* (UK Daily Mirror, 17th August 2000); *‘The trouble with boys’* (UK Guardian 21 August 2000); ‘*Gender gap continues to grow’* (UK Guardian, 22 August 2002). This concern is not limited to the media. For example Chris Woodhead, the former Chief Inspector of schools in England, described under-achieving boys as “one of the most disturbing problems facing the education system" (TES, 27 November 1998). There has also been a strong political input, involving national strategies, task groups and targets. For example, one of the Welsh National Assembly's targets was that, by the year 2002, the underachievement of boys against girls in national tests and examinations should be cut by 50 per cent as compared with 1996.

We can locate this concern with the 'gender gap' within a long history of investigating sex differences in intellectual abilities. Does the gender gap in examination attainment reflect sex differences in more fundamental cognitive domains such as aspects of psychometric intelligence or reasoning abilities? Do boys and girls differ in their scores on IQ-type or reasoning abilities tests?

Sex Differences in IQ

Early standardisations of the Stanford-Binet and Weschler-Bellevue IQ tests tended to indicate a small score difference favouring females, although these were not considered significant (Mackintosh, 1996, pp. 182-199). However, standardisations of the revised editions of the Wechsler Intelligence Scale for Children (WISC-R) and Wechsler Adult Intelligence Scale (WAIS-R) in the early 1980s showed a small difference, favouring males, of around 1.7 points on the WISC-R and 2.2 points on WAIS-R (Jensen & Reynolds, 1983; Reynolds et al., 1987). The results obtained on recent large, representative population samples are also equivocal. Thus Hernstein & Murray (1994), describing the USA's National Longitudinal Survey of Youth 1979, described the tests scores of some 12,000 teenagers and young adults, and found a difference of 0.9 IQ points in favour of men. But Lubinski & Humphreys (1990) analysed the test scores of 100,000 sixteen-year-old US school students and found a difference of 0.3 IQ points in favour of girls.

There continues to be debate on the extent, or even existence, of sex differences in the mean level of IQ scores (Lynn, 1994, 1998; Mackintosh, 1996; Jensen, 1998; Colom, Juan-Espinosa, Abad & Garcia, 2000; Halpern & LaMay, 2000; Lynn & Irwing, 2004, Lynn, Allik & Irwing, 2004). However, it is apparent in the majority of studies that, even when sex differences in mean IQ scores are found, they tend to be small. Intelligence is not a single homogeneous ability (Carroll, 1993) and IQ tests reflect this. Males tend to perform better on some sub-tests, and females on others; when these results are averaged across sub-tests these differences tend cancel each other out. The main evidence for sex differences tends to come from differential performance in specific abilities.

Sex Differences in Specific Abilities

Maccoby & Jacklin (1974) reviewed studies of sex differences published in American journals in the ten years preceding 1974. They concluded that the sexes did not differ consistently in tests of composite abilities such as IQ. However, from adolescence onwards, there was evidence of girls' superiority in a variety of verbal abilities, which continued into adulthood. In contrast, there seemed to be a consistent trend for a male advantage from age 13 onwards in quantitative and visuo-spatial abilities.

Maccoby & Jacklin's book generated considerable debate (see Caplan 1979) and their overall conclusions have been supported in some subsequent research (e.g., Halpern, 1992; Feingold, 1992; Halpern & LaMay, 2000) but not in others. For example, Hyde & Linn (1988) performed a meta-analysis of 165 studies of sex differences in verbal ability. They summarise their results using effect size (d) estimates, which are the difference between the mean scores for boys and girls divided by the pooled standard deviation. They concluded that there was a modest female superiority of d =

0.20 on tests of general verbal ability, d = 0.22 on anagrams, and d = 0.33 for speech production although, paradoxically, they also concluded there was a male advantage of d = 0.16 on verbal analogies, giving an overall verbal effect size of d = 0.11 in favour of girls, which they considered insubstantial. Similarly, for mathematical ability, Hyde, Fennema, and Lamon (1990), performed a meta-analysis of 100 studies and reported an overall effect size of only d = 0.05, and in favour of females. However, the results suggested significant interactions between student age, type of ability and the selectivity of the sample. Thus differences favouring males tended to be restricted to the area of problem solving, emerged only at high school age (15-17 years), and were largest for self-selected samples, such as the US Scholastic Aptitude Test-Maths (SAT-Maths) compared to general population samples.

There is considerable variability in the outcomes of the many small studies included within the Hyde & Linn (1988) and Hyde et al. (1990) meta-analytic reviews. Perhaps the most compelling evidence in relation to sex differences will be found in the analysis of norms from standardised tests, where the sample is large and nationally representative on key demographic, educational and other relevant criteria. Two studies are particularly eminent in meeting these criteria. Feingold (1992) reviewed test norming statistics for four standardisations of the Differential Aptitude Tests (DAT) between 1947 and 1980 with US students aged 14-17+. The results, summarised in Table 1, do not reveal the substantial male advantage in numerical ability, or the female advantage in verbal reasoning, that might be expected from Maccoby & Jacklin's (1974) conclusions, although the female advantage for language and spelling, and the male advantage for spatial relations, are more congruent. A paper by Hedges & Nowell (1995) is also particularly robust in terms of sample size and representativeness. They performed a secondary analysis of six large US national datasets collected between 1960 and 1992. The datasets involved people from age 15 to the early twenties and all were based on large national probability samples. They concluded that females exhibited a slight tendency to perform better on tests of reading comprehension, perceptual speed and associative memory, and males tended to perform better on tests of mathematics and social studies. However with the exception of the male advantage on the vocational aptitude scales the effect sizes were relatively small, less than d = 0.2.

Insert Table 1 about here

Sex Differences in Variability in Test Scores

The majority of studies have only considered sex differences in mean scores. However, in an often overlooked aspect of their review, Maccoby & Jacklin (1974) also concluded that males were more variable than females in mathematical and spatial abilities, although the sexes were equally variable in verbal ability. The issue of increased cognitive variability in males was previously discussed in detail by Heim (1970). Feingold (1992) analysed the results for the national standardisations of the DAT, the SAT, the WAIS, and the California Achievement tests. Males tended to be more variable than females in general knowledge, mechanical reasoning, quantitative ability, spatial visualisation, and spelling. There was little difference in variability for most verbal tests, short-term memory, non-verbal reasoning and perceptual speed (see Table 1 for DAT results). Hedges & Nowell (1995) reported that males had greater variance than females in all but two of the areas they considered, typically in the order of 3%-15% greater variability in boys' scores than in girls' scores. Cole (1997) also reported greater variability in boys' scores on many of the tests analysed. For example, at age 17, males outnumbered females in the top 10% on maths tests by 1.5 to 1, and in science by 2 to 1.

Sex differences in spread or variability are important because they help to explain why males may outnumber females among the highest scoring individuals in tests that show only a small male advantage in mean score (Feingold, 1992; Hedges & Nowell, 1995; Nowell & Hedges, 1998). The obverse was also true: in Hedges & Nowell's (1995) study boys outnumbered girls in the bottom 10% for those tests with only a small female advantage in mean score (e.g., reading comprehension, perceptual speed and associative memory).

Trends Over Time

Hyde & Linn (1988), in their meta-analysis of sex differences in verbal ability, reported a mean effect size of d = 0.23 (favouring girls) for studies conducted before 1973, but a mean effect size in the same direction of only d = 0.10 for studies completed from 1973 onwards. Similarly, for mathematics, Hyde et al. (1990) reported a mean effect size for studies published prior to 1973 of d = 0.31 (favouring boys), but only d = 0.14 in the same direction for the studies completed from 1974 onwards.

Other studies relate to attainment rather than reasoning tests, but suggest a similar trend. Nowell & Hedges (1998) based their assessment of time trends on the US National Assessment of Educational Progress (NAEP) data for 17 year old students, which consists of tests of reading, writing, mathematics and science, similar in its curriculum focus to the England National Curriculum testing programme. They suggested that, over the period 1971-1994, the small sex differences favouring males in mathematics and science scores appeared to have narrowed slightly, but that the relatively large sex differences favouring girls in reading and writing had not. Cole (1997) also reported an analysis of a nationally representative sample of US 15 year olds (Project Talent) revealing an effect size for science that reduced the male advantage from about d = 0.60 to under d = 0.20 from 1960 to 1990, with mathematics showing a similar reduction from d = 0.45 to d = 0.10. However, females sustained their advantage in writing from 1960 to 1990 at approximately d = 0.40.

Studies in the UK

It is interesting that the large meta-analyses undertaken by Hyde & Linn (1988) and Hyde et. al. (1990) specifically excluded all studies from outside the US. UK and other national studies are important because results found in the US are not consistently replicated in other countries (Feingold, 1994). However, very few studies have been completed in the UK that meet the stringent methodological criteria of large and nationally representative samples. Deary, Thorpe, Wilson, Starr, & Whalley (2003) reported an analysis of the Moray House Verbal Reasoning test completed at age 11 by almost all Scottish schoolchildren born in 1921 as part of the 1932 Scottish Mental Survey (N = 87,498). This is probably the only near-complete national examination of a whole year-of-birth cohort. Despite there being about 40,000 boys and girls, they found no sex difference in mean IQ score. However, there was greater variability among boys' scores, such that boys were over-represented relative to girls at both the highest and lowest extremes.

The Present Study

It is important that the abilities assessed in studies are clearly defined. For example, Hyde & Linn (1988) note in their meta-analysis that, "'verbal ability' has been used as a category to include everything from quality of speech in two year-olds, to performance on the Peabody Picture Vocabulary Test (PPVT) at age 5 years, to essay writing by high school students, to solutions to anagrams and analogies". Similarly, 'mathematical ability' has referred to varied measures such as computation, concepts or problem solving (Hyde et. al. 1990; Cole 1997). Many of the measures reported by

Hedges & Nowell (1995) do not focus on reasoning abilities at all but, rather, on vocational aptitude (mechanical reasoning, electronic information, auto & shop information) or school subjects such as science, mathematics and social studies. Performance in these areas might be strongly effected by differential male-female educational experience such as different subject choices and by differential drop-out from school after the compulsory years, particularly for the older students (age 16+) who form the majority of the populations in their study.

The present study reports results from a large and representative UK national sample of 11-12 year olds using the Cognitive Abilities Test (CAT) to address questions of sex differences in specific cognitive abilities scores in the UK. The study has several strengths in relation to previous reports. It focuses on the UK, in contrast to the majority of research that has been conducted within the US. It analyses the results for an extremely large and nationally representative sample of over 320,000 schoolchildren. It focuses on reasoning abilities rather than educational attainment in school subjects or vocational aptitude tests. Reasoning tests should be less affected by subject choices or by differential educational experiences than curriculum-related or vocational tests. It disaggregates verbal, quantitative, and non-verbal reasoning scores to allow a more sophisticated analysis of differences in abilities, in contrast to previous UK studies focussing only on overall IQ (e.g. Deary et al., 2003). It focuses on early secondary school (age 11-12), where all schoolchildren are in compulsory education, and the effects of selective drop out—from education as a whole and from specific subjects—are removed. It reports recent results, from tests completed in 2002 and 2003.

Method

Participants: the CAT3 Data Sample

The total dataset consisted of over 500,000 UK schoolchildren who completed CAT3 between September 2001 and August 2003. The largest proportion of schoolchildren completed level D, designed for 11-12 year-olds in the first year of secondary school. Level D scores were available for over 324,000 schoolchildren from 1,305 schools. The average age of schoolchildren completing Level D was 11 years and 7 months with a standard deviation 4.4 months (these means and SDs were identical for boys and for girls). Two-thirds of schoolchildren were in the age range 11.03 to 12.00. Within the total sample, 49.9% of schoolchildren were boys and 50.1% were girls, equivalent to the 2001 England average for the 11-12 year age range.

The sample of 320,000 participants represents almost half of the UK's population of 11-12 year olds (approximately 700,000 children). However, sheer size does not of itself ensure the absence of selective bias in the sample. Over 84% of the schoolchildren taking CAT3 Level D were drawn from maintained, mainstream secondary schools in England. This sub-sample was compared to national statistics computed for all maintained, mainstream secondary schools in England on a dataset collected by the Department for Education and Skills (DfES) in January 2001. The national dataset was analysed in relation to the selective status of the school (in areas of the country operating selection, grammar schools select the most able pupils and the rest attend secondary modern schools; in non-selective areas all children attended comprehensive schools). In addition, five bands, each containing 20% of all schools nationally, were created to describe the range of school variation on a number of key variables, including entitlement to free school meals (an indication of the economic disadvantage of the school population), the proportion of schoolchildren from ethnic minority groups, and the proportion of schoolchildren with English as an Additional Language. The results are presented in Table 2.

Insert Table 2 about here

The CAT sample of 1,046 schools includes almost one-third (30%) of all maintained, mainstream secondary schools in England. This figure substantially underestimates the total proportion of schools using CAT, since a large minority of English maintained secondary schools were still using CAT Second Edition (CAT2E) during this time period. In comparing the sample against national averages, the sample is broadly representative of all such schools in England. There are only slight variations from the national proportions so that, in terms of selective status, the proportion of schoolchildren entitled to free school meals, the proportion of ethnic minority schoolchildren, and the proportion of schoolchildren with English as an Additional Language, the sample is broadly representative of all schools nationally.

Cognitive Abilities Test Third Edition (CAT3)

The Cognitive Abilities Test Third Edition (CAT3) is the most recent UK version of the CAT, and was published in July 2001 (Lohman et. al., 2001). The CAT provides an assessment of a child's reasoning abilities in the Verbal Reasoning (VR), Quantitative Reasoning (QR) and Non Verbal Reasoning (NVR) domains. Each domain is assessed by a separate battery of three tests. A child's mean score over the three batteries (mean CAT score) is also calculated. The test is divided into eight levels coded as Levels A-H, and is standardised in the UK across the age range 7:6 to 17:0 and above. We next describe the tests within each battery, giving indicative items.

CAT3 Verbal Reasoning (VR) Battery

- *Verbal Classification*. Given three words belonging to one class, select which further word from a list of five belongs to the same class (e.g. eye, ear, mouth : nose, smell, head, boy, speak)
- *Sentence Completion*. Selecting one word from a list of five (e.g., John likes to ____ a football match: eat, help, watch, read, talk)
- *Verbal Analogies*. Given one pair of words, complete a second pair from five possibilities (e.g., bigŁ large; littleŁ ? : boy, small, late, lively, more).

CAT3 Quantitative Reasoning (QR) Battery

- *Number Analogies*. Determine the relationship between the numbers in two example pairs and decide which of five options would complete a third pair in the same way (e.g., [9 Ł 3] [12 Ł 4] [27 Ł ?] : 5, 9, 13, 19, 21)
- *Number Series*. Select one from five possible choices to complete the series (e.g., 2,4,6,8, Ł ? : 9,10,11,12,13)
- *Equation Building*. Select the one answer choice that can be calculated by combining all the given elements to create a valid equation (e.g., 2 2 3 + x : 6, 8, 9, 10, 11) .

CAT3 Non Verbal Reasoning (NVR) Battery

- *Figure Classification*. Given three shapes belonging to one class, select which further shape from five alternatives belongs to the same class.
- *Figure Analogies*. Given one pair of shapes, complete a second pair from five possibilities.

- *Figure Analysis*. A piece of paper is folded and holes are punched through the paper. How will the paper look when it is unfolded?

These sub-tests include item types with a long pedigree (such as classification, analogies and series) as well as relatively more recent forms (such as equation building and figure analysis). The resulting CAT battery scores have extremely high reliability (Strand, 2004) and strong validity correlations with later educational attainment (Smith, Fernandes & Strand, 2001; Strand, 2003, Strand, in press).

Results

Table 3 presents, for boys and girls separately, the mean score, standard deviation and sample size for standard age scores on each of the three batteries and the mean CAT score. All gender comparisons are highly statistically significant: girls had a higher mean score than boys on the verbal battery, the non-verbal battery and mean CAT score, and boys had a significantly higher mean score than girls on the quantitative battery. Boys had significantly greater variance on all four CAT measures. However, with samples of this size, even very small absolute differences are likely to be statistically significant at conventional *p* values. We therefore need to examine the magnitude of the sex differences in mean scores and in score variability by considering the effect size (d) and the variance ratio (VR).

Insert Table 3 about here

Sex Differences in Mean Scores

The effect size is the difference between the mean scores for boy and girls divided by the pooled standard deviation. Thus the sex difference in mean verbal reasoning score of 2.2 standard score points equates to an effect size (d) of 0.15. Following Cohen (1977) in psychological research 0.2 is considered a small effect, 0.5 is considered medium and effects sizes above 0.80 are considered large. Too rigid an interpretation of these thresholds can be limiting since an interpretation of ‘large’ depends on a number of factors such as the costs of implementing an intervention, the benefits associated with the difference produced, the value attached to the benefits etc (See Coe, 2004 for a further discussion). However it is clear that the effect size for verbal reasoning is very small, and the difference in quantitative and non verbal reasoning means are negligible.

We can explore sex differences in more detail by considering performance on each of the nine CAT sub-tests, three within each battery. The mean and standard deviations of raw scores on each Level D sub-test were used to calculate the effect sizes, as shown in Table 4. The higher female mean is consistent over all three verbal tests. There appear to be almost no sex difference at all on any of the three quantitative tests (all effect sizes are less than 0.05). For non-verbal reasoning the overall equality between boys and girls appears to reflect some averaging of a small female advantage in figure classification and a small male advantage in figure analysis. The later test is designed to tap elements of spatial ability on which males often achieve higher scores, although not typically until late adolescence. However the effect size (d=-0.09) is too small to warrant extended discussion.

Insert Table 4 about here

Sex Differences in Score Variance

The variance ratio is the ratio of male score variance to female score variance. A variance ratio > 1 indicates the variance is greater for boys than for girls, while a variance ratio less than one indicates greater variability in the scores for girls. When viewed as a descriptive statistic, Feingold (1992) suggested that a variance ratio of 1.10 is probably the smallest meaningful effect. In these terms the sex difference in variability borders the threshold for verbal reasoning, and exceeds this level for both non verbal and quantitative reasoning (see Table 3). In percentage terms boys' scores are 9% more variable than girls on verbal reasoning, 13% more variable on non-verbal reasoning, 18% more variable on quantitative reasoning and 13% more variable on mean CAT score. A similar picture of greater variability in the scores for boys is apparent for all nine of the sub-tests (Table 4). Male performance is more variable than female, with variance ratios in excess of 1.10 on eight of the nine subtests, the exception being figure analysis where the variance ratio was 1.07.

Figure 1 presents a graphical illustration of the percentage of boys and girls within each of nine score bands. These score bands (stanines) split the national distribution into nine bands that approximate the normal curve, as shown in Appendix 1. The full data giving participant numbers and percentages are included in Appendix 2. Stanines have been selected because this is one of the forms in which CAT scores are routinely reported to test users. The differences in variability are not huge. For example, about sixty per cent of the pupils scoring in the bottom 5% of the VR range, and 60% of those in the top 5% of the QR range, were boys, giving ratios of 1.5:1 and indicating that three of every five pupils identified with these extreme scores will be boys. Differences in the top and bottom 5% of scores for NVR are smaller, with around 55% boys or a ratio of 1.25:1, indicating that five of every nine children identified at these extremes were boys.

Insert Figure 1 about here

For comparison with previous papers, differences in the number of boys/girls with extreme scores are also shown as the ratios of the number of boys to the number of girls who score in the bottom 5%, bottom 10%, top 10% and top 5% of the score distributions (see Table 3). As with the variance ratios, a positive ratio indicates a greater proportion of boys than girls. Males are over-represented in all extremes, with the exception of the upper tails of the verbal reasoning test. This later result reflects the sex difference in mean score, although it is interesting to note that the under-representation of boys in the top 5% would be even larger (0.73:1) were it not for the greater male variance[i].

DISCUSSION

Sex differences in mean reasoning scores are very small. The only non-trivial sex difference in mean scores was for verbal reasoning, where girls scored on average 2.2 standard age score points above boys. Even here the effect size was only 0.15, compared to a traditional threshold of 0.20 to infer a small effect size (Cohen, 1977). Comparing the magnitude of the VR-NVR score difference for boys and girls suggests the sex difference in verbal scores more strongly reflects a relative under-performance by boys in the verbal domain (average VR-NVR difference is -1.3 points) than an over-performance by girls (average VR-NVR difference is 0.4 points). There were significant differences in the standard deviation of scores between the sexes. Boys' scores were 9%, 13% and 18% more variable than girls' scores for VR, NVR and QR respectively. Boys were over-represented relative to girls in the bottom 5% in verbal reasoning, and at both the bottom and top 5% for quantitative and non-verbal reasoning scores by ratios between 1.2:1 and 1.5:1.

Are the Results an Artefact of Test Construction?

It is sometimes argued in relation to sex differences in IQ that the two sexes have been defined rather than discovered to have equal IQ. If test constructors expect equal performance from boys and girls, they might remove items on which girls show better performance and substitute ones that boost the boys, or vice versa, with the result that both sexes obtain the same mean IQ. However a review of early studies by Mackintosh suggests that IQ tests were not designed from the outset to yield equal scores for the two sexes, and that early test developers did genuinely discover only small sex differences in mean scores (Mackintosh, 1996, pp. 559-560).

It is true that, guided by the early findings of no significant sex differences, modern IQ and reasoning tests do routinely employ differential item functioning (dif) analyses to reject items with extreme sex differences. However, dif analyses are generally assumed to increase the fairness of the test by removing items where the content is better known by one group than another, and therefore confounds content knowledge and reasoning ability. The dif procedure will eliminate question specific dif from the test, and may thereby reduce overall score differences, but it will not eliminate any general strength or weakness across all questions, so group differences in overall score will remain. In our view, the absence of substantial sex differences in the mean scores on the CAT is unlikely to be due to test construction.

It is difficult to see how test construction issues could account for the observed greater variability in boys' scores. One possibility would be a sex difference in speed-accuracy trade-off in a timed test such as the CAT. If boys worked faster but less accurately than girls, they would be more likely than girls to attempt the harder items at the end of the test, and also less likely to be successful on the easiest items at the start. In such a scenario mean scores for boys and girls might be identical, but less able boys might obtain a higher proportion of low scores than girls of otherwise similar ability, while more able boys might gain a higher proportion of high scores than girls of otherwise similar ability.

Item data for a nationally representative stratified sample of over 2,000 pupils taking Level D in the summer 2000 UK standardisation of CAT3 were examined to test this hypothesis. The data indicate that a speed-accuracy trade-off might operate in the QR tests, as 11 of the 14 questions showing significant dif were within five questions of the beginning or end of one or other of the three subtests; the three items favouring girls were located within the first five items of a subtest, and 8 of the 11 items favouring boys were in the last five of a subtest. This might account for some part of the particularly large sex difference in variance for QR scores. On the other hand, the fact that greater male variability is also reported on the quantitative measures of the WAIS (Feingold, 1992), an untimed, individually administered, graded response test, suggests the result is not simply a product of the multiple-choice timed format of CAT. There was also no observable pattern of sex differences related to item difficulty or item position for the VR or NVR tests. Overall, it seems unlikely therefore that the greater variability in boys' scores is simply an artefact of the test.

Differences in Mean and Variability of Scores

In contrast to the widely cited sex differences in mean scores reported by Maccoby & Jacklin (1974), the current results support later studies which suggest that sex differences in mean reasoning scores are small or non-significant (Feingold, 1992, Mackintosh, 1996). The current study focused on schoolchildren aged 11-12 and some authors argue that, due to the faster maturation of girls, sex differences will be obscured up to age 16, appearing only in late adolescence or young adulthood (Lynn, 1994). However, decisions about whether to remain in education are also made at around age 16 and school/college populations becomes increasingly self-selected, with males more likely to drop out of education, such that results cannot be generalized to the broader population (Hyde et al., 1990, p150). Even before this age there is selection into school subjects that shows sex bias, something which has not occurred at age 11-12 years. Studies that have been able to analyse large and nationally representative populations in the 14-17 age range have failed to report large sex differences (Hedge & Nowell, 1995).

In relation to sex differences in variability, the current results support the general finding of greater male variability. The sex difference in variability was least pronounced for verbal reasoning, and most pronounced for quantitative reasoning, congruent with Macoby & Jacklin (1974), Feingold (1992) and Hedges & Nowell (1995). However, the results also show significantly greater male variability for non-verbal reasoning, not previously reported. Most importantly, the current results for quantitative reasoning and non-verbal reasoning show boys simultaneously over-represented in both the very low and the very high score groups, while Hedges & Nowell (1995) show boys over-represented either at the lower or the upper score range depending on the particular test. Our result is congruent with the other large UK study which found an excess of boys with both extreme low and high scores (Deary et. al., 2003).

Implications of the Results

Reasoning scores at age 11 are strongly correlated with subsequent educational attainment in national tests of English, mathematics and science at age 14 in England, and with public examination results at age 16 in England and Scotland; such validity data is particularly strong for the CAT (Strand, 2003; Smith et. al., 2001). Given this close association, the lack of substantial sex differences in reasoning scores suggests there is no *a priori* rationale, based on mental ability differences, to expect a large gender gap in subsequent test or examination attainment at age 16. If we wish to look for explanations of the gender gap at GCSE we must look beyond conceptions of ability.

Despite the prominent media and government focus on the gender gap, educators must be careful to avoid general conceptions of boys as underachievers. It is clear from the current study that boys are slightly more likely to be over-represented relative to girls at the high as well as the low extremes of reasoning scores. The differential is not great, but we might hypothesise that there might be a greater proportion of boys within some of the programmes aimed at addressing the needs of the more able students, such as the gifted & talented strand of the Excellence in Cities programme in England. Of course, the degree of overlap in the score distributions of the sexes is vastly greater than the differences between them, and individual pupils should always be considered on the basis of their actual scores rather than their group membership.

The greater variability in boys' reasoning scores may explain to some extent their greater representation within populations with Special Educational Needs and among those who fail to achieve any GCSE or equivalent passes (6.4% of boys versus 4.3% of girls at age 15+, a ratio of 1.49:1 (DfES 2002). However, boys do not appear to be over-represented at the higher end of GCSE performance. In 2002, only 2.8% of boys' GCSE entries were grade A*, compared to 4.4% of girls entries, a ratio of 0.64:1. Only in economics, mathematics and physics did boys exceed girls in the proportion of A* grades awarded (OFSTED, 2002). To this extent, it may be valid to speak of a degree of underachievement, particularly among more cognitively able boys. It is possible though that sex differences in GCSE examinations reflect wider factors related to motivation and effort, such as girls greater likelihood to complete and submit coursework (OHMCI, 1997), gendered patterns of subject choice (Arnot et al, 1996) or gendered allocation to tiered subjects (Elwood, 1995). Salisbury, Rees & Gorard (1999) provide a good review of the literature in this area.

Some authors (Heim, 1970; Deary et. al., 2003) have also suggested that the differences in variability between the sexes in IQ might account for some of the variability in long term life outcomes related to cognition, for example the fact that men are slightly more likely to achieve third class and first class university degrees, and less likely to achieve second class degrees, than women (Smith & Naylor, 2001). While degree classifications are not simply a product of the subject studied or student's social class (Smith & Naylor, 2001) the many social variables intervening between cognitive abilities at age 11 and later adulthood indicate the need for a complex analysis of such outcomes.

GCSE public examinations rely heavily on essays and other modes of assessment requiring extended writing. We also know that largest sex differences reflect girls' superiority in the area of writing. For example, Cole (1997) reports an analysis of multiple US national datasets for school students assessed at age 9, 14 and 17 on a wide variety of tests. There was no sex difference on Verbal reasoning/vocabulary (0.05), a small advantage in verbal-reading (0.20) a medium female advantage in verbal-language use (0.40) and the largest difference for verbal-writing (0.60). If there is a desire to circumvent such sex-based superiority in writing skill, it is important that public examinations continue to utilise a range of assessment methodologies, including non-discursive modes.

Finally, it is worth remembering that the gender gap in performance at all ages, even in GCSE at age 16, is extremely small relative to differences associated with, for example, socio-economic circumstances (e.g., Strand, 1999, Demack et. al., 2000).

The high media attention given to the gender gap should not distract policy makers from attempting to ameliorate other, more sizeable gaps.

Explanations for Greater Male Variability in Reasoning Scores

The authors of the earliest research suggesting greater male variability championed biological interpretations (Feingold, 1992, p63). Evolutionary explanations of greater male variability in intellectual abilities continue to abound (e.g., Archer & Mehdikhani, 2003). For a long time any research reporting sex differences in variability was taken to support the variability hypothesis and automatically assumed to be consistent with an innate explanation. However, findings of sex differences in variability are not inconsistent with cultural or environmental explanations. For example, Hollingworth (1922) argued that men's occupational roles were less constraining than women's, affording men greater diversity in educational and environmental experiences, which could engender greater male variability. Noddings (1992) argued that, "Girls who remain in school will for the most part listen to the teacher, do at least some assignments, and generally conform sufficiently to avoid landing at the very bottom of any distribution. Similarly many of the brightest girls still feel pressed not to exhibit or actively enhance their superior test-taking and scoring capabilities." (p88). Just as it is widely accepted that differences in reasoning scores are a result of both genetic and environmental influences, so might differences in variability be a consequence of the interaction of such influences.

References

Arnot, M., David, M., & Weiner, G. (1996). *Educational reform and gender equality in schools*. Manchester: Equal Opportunities Commission.

Caplan, P. J. (1979). Beyond the box score: A search for boundary conditions in aggression and achievement behaviour. In *Maher, B. A. (Ed.), Progress in Experimental Personality Research (Vol. 9, p41-87).* New York: Academic Press

Carroll, J. B. (1993). *Human cognitive abilities: a survey of factor analytic studies*. Cambridge, UK: Cambridge University Press.

Coe, R. (2004). Issues arising from the use of effect sizes in analysing and reporting research. *in Schagen, I. & Elliott, K. (Eds.) (2004). But what does it mean? The use of effect sizes in educational research*. Slough: National Foundation for Educational Research.

Cohen, J. (1977). *Statistical power analysis for the behavioural sciences*. New York, NY: Academic Press.

Cole, N. S. (1997). *The ETS Gender study: How females and males perform in educational settings*. Princeton, NJ: Education Testing Service.

Deary, I., Thorpe, G., Wilson, V., Starr, J. M., & Whalley, L. J. (2003). Population sex differences in IQ at age 11: The Scottish mental survey 1932. *Intelligence, 31,* 533-542.

Demack, S., Drew. D., & Grimsley, M. (2000). Minding the gap: ethnic, gender and social class differences in attainment at 16. *Race, Ethnicity & Education, 3,* 117-143.

Department for Education and Skills (DfES) (2002). *Autumn Package 2002*. Latest version available online at http://www.standards.dfes.gov.uk/performance/ap/ [Accessed 12 October, 2004).

Elwood, J. (1995).Undermining gender stereotypes: examination and coursework performance in the UK at 16. *Assessment in Education, 2, (3) 283-303*.

Feingold, A. (1992). Sex differences in variability in intellectual abilities: A new look at an old controversy. *Review of Educational Research, 62,* 61-84.

Feingold, A. (1994). Gender differences in variability in intellectual abilities: A cross-cultural perspective. *Sex Roles*, *30*, 81-92.

Halpern, D. E. (1992). Sex differences in cognitive abilities. Hillsdale, NJ : Lawrence Erlbaum.

Hedges, L. V., & Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science*, *269*, 41-45.

Heim, A. (1970). *Intelligence and personality*. Harmondsworth: Penguin.

Hernstein, R. J., & Murray, C. (1994). *The bell curve*. New York, NY: Free Press.

Hollingworth (1922). Differential action upon the sexes of forces which tend to segregate the feebleminded. *Journal of abnormal and social psychology, 17*, 35-57.

Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin, 104,* 53-69.

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin, 107,* 139-155.

Jensen, A. R. (1998). *The* g *factor*. Westport, CT: Praeger.

Jensen, A. R., & Reynolds, C. R. (1983).Sex differences on the WISC-R. *Personality and individual differences, 4*, 223-226.

Lohman, D. F., Thorndike, R. L., Hagen, E P., Smith, P., Fernandes, C., & Strand, S. (2001). *Cognitive Abilities Test Third Edition*. London, UK: nferNelson.

Lubinski, D., & Humphries, L. G. (1990). A broadly based analysis of mathematical giftedness. *Intelligence*, *14*, 327-355.

Lynn, R. (1994). Sex differences in intelligence and brain size: a paradox resolved. *Personality and Individual Differences*, *17*, 257-271.

Lynn, R. (1998). Sex differences in intelligence: A rejoinder to Mackintosh. *Journal of Biosocial Science*, *30*, 529-532.

Mackintosh, N. J. (1996). Sex differences and IQ. *Journal of Biosocial Science*, *28*, 559-571.

Maccoby, E. E., & Jacklin, C. N. (1974). *The psychology of sex differences*. Stanford, CA: Stanford University Press.

Noddings, N. (1992). Variability: A pernicious hypothesis. *Review of Educational Research, 62,* 85-88.

Nowell, A., & Hedges, L. V. (1998). Trends in gender differences in academic achievement from 1960 to 1994: An analysis of differences in mean, variance and extreme scores. *Sex Roles, 39,* 21-43.

Office of Her Majesty's Chief Inspector (OHMCI) (1997*). The relative performance of boys and girls.* Cardiff: OHMCI.

Office for Standard in Education (OFSTED) (2002). *National Summary Data Report for Secondary Schools (PANDA Annexes for 2002/03). Available online at* http://www.ofsted.gov.uk/publications/index.cfm?fuseaction=pubs.summary&id=3120 *[Accessed 12th October, 2004].*

Reynolds, C. R., Chastain, R. L., Kaufman, A. S., & McClean, J. E. (1987). Demographic characteristics and IQ among adults: analysis of the WAIS-R standardisation sample as a function of stratification variables. Journal of School Psychology, 25, 323-342.

Salisbury, J., Rees, G., & Gorard, S. (1999). Accounting for the differential attainment of boys and girls at school. *School Leadership & Management, 19, (4), 403-426.*

Smith, J., & Naylor, R. (2001). Determinants of degree performance in UK universities: a statistical analysis of the 1993 student cohort. Oxford Bulletin of Economics and Statistics, 63, 29-60.

Smith, P., Fernandes, C., & Strand, S. (2001). *Cognitive Abilities Test 3: Technical Manual*. London, UK: nferNelson.

Strand, S. (1999). Ethnic group, sex and economic disadvantage: associations with pupils' educational progress from Baseline to the end of KS1. *British Educational Research Journal, 25,* 179-202.

Strand, S. (2003). *Getting the best from CAT: A practical guide for secondary schools.* London, UK: nferNelson.

Strand, S. (2004). Consistency in reasoning test scores over time. British Journal of Educational Psychology, 74, (4), 617-631.

Strand, S. (in press). Comparing the predictive validity of reasoning tests and national end of Key Stage 2 tests: Which tests are the best? *British Educational Research Journal, in press.*

Author Note

Steve Strand, nferNelson, London, UK;

Ian J. Deary, Department of Psychology, University of Edinburgh, UK;

Pauline Smith, nferNelson, London, UK

These data and results were presented in a paper by the first author to the Annual Conference of the British Educational Research Association, 11th-13th September 2003 at Heriot-Watt University, Edinburgh, Scotland.

Communication to:

Steve Strand, Reader in Education, University of Warwick, Coventry, CV4 7AL

Tel (024) 7652 2197: e-mail: steve.strand@warwick.ac.uk.

Table 1
Sex Differences in Mean Score and Score Variability Averaged Over Four Standardisations of the Differential Aptitude Tests (DAT) Between 1947-1980 With Students Aged 14-18+ (Feingold, 1992)

DAT Battery	Effect size (d)	Variance Ratio (VR)
Numerical Ability	.05	1.11
Mechanical Reasoning	.98	1.28
Space Relations	.24	1.21
Spelling	-.50	1.12
Verbal Reasoning	.05	0.96
Abstract Reasoning	.08	1.01
Language	-.43	0.99
Clerical speed and accuracy	-.03	0.94

Table 2
Comparison of Sample Against Averages for All Maintained Mainstream Secondary Schools in England

		All maintained mainstream secondary schools in England January 2001(a)		Maintained mainstream secondary schools in England using CAT3 with Y7 (b)	
Number of schools		3,481		1,046	30%
Selective Status	Comprehensive	3140	90.2%	941	90.0%
	Secondary Modern	145	4.2%	53	5.1%
	Selective Grammar	159	4.6%	41	3.9%
	Other	37	1.1%	11	1.1%
Entitlement to Free School Meals	Botttom 20%	710	20.4%	179	17.1%
	Low-Middle 20%	697	20.0%	209	20.0%
	Middle 20%	682	19.6%	191	18.3%
	Middle-High 20%	696	20.0%	220	21.0%
	Top 20%	696	20.0%	247	23.6%
Ethnicity	Botttom 20%	740	21.3%	230	22.1%
	Low-Middle 20%	663	19.0%	184	17.6%
	Middle 20%	692	19.9%	191	18.3%
	Middle-High 20%	684	19.6%	186	17.8%
	Top 20%	693	19.9%	252	24.2%
English as Additional Language	Botttom 20%	732	21.1%	232	22.2%
	Low-Middle 20%	674	19.4%	194	18.5%
	Middle 20%	688	19.8%	179	17.1%
	Middle-High 20%	692	19.9%	199	19.0%
	Top 20%	695	20.0%	242	23.1%

Note. (a) includes middle deemed secondary schools. (b) For the purpose of this analysis, CAT results from 292 schools in Scotland, Wales and Northern Ireland, and independent and special schools in England, have been excluded.

Table 3

Mean Standard Age Score, Standard Deviation and Sample Size for Boys and Girls on CAT3 Level D with statistical significance, effect size, variance ratios and tail proportion ratios for each CAT battery

CAT Battery	Statistic	Boys	Girls	Significance	Effect Size (d)	Variance Ratio (VR)	Tail proportion ratios			
							Lowest 5%	Lowest 10%	Top 10%	Top 5%
Verbal	Mean	98.4	100.6	P<.0001	0.15	1.09	1.53	1.42	0.86	0.86
	SD	15.1	14.5	P<.0001						
	N	158,093	158,457							
Quantitative	Mean	99.4	98.9	P<.0001	-0.03	1.18	1.30	1.19	1.34	1.46
	SD	15.0	13.8	P<.0001						
	N	157,862	158,406							
Non-Verbal	Mean	99.7	100.2	P<.0001	0.03	1.13	1.36	1.23	1.09	1.17
	SD	14.8	13.9	P<.0001						
	N	157,830	158,299							
Mean CAT	Mean	99.1	99.9	P<.0001	0.05	1.13	1.38	1.29	1.07	1.10
	SD	13.5	12.7	P<.0001						
	N	156,556	157,258							

Note. Positive effect size (d) indicates the female mean greater than the male mean. Variance Ratios>1 indicates male variance is greater than female variance. Tail probability ratios >1 indicate higher proportion of boys than girls.

Table 4
Effect Size and Variance Ratios for Sex Differences in Raw Scores on Each of the Nine CAT Level D Sub-Tests

CAT sub-test	Effect Size (d)	Variance Ratio (VR)
Verbal Classification	0.16	1.15
Sentence Completion	0.12	1.13
Verbal Analogies	0.15	1.13
Number Analogies	-0.04	1.10
Number Series	0.00	1.12
Equation Building	0.04	1.16
Figure Classification	0.14	1.10
Figure Analogies	0.07	1.17
Figure Analysis	-0.09	1.07

Note. Positive effect size indicates female mean greater than male mean. VR greater than 1 indicates male variance greater than female variance.

Figure caption

Figure 1. Percentage of boys and girls within each stanine score band for the CAT3's three battery scores and the overall mean CAT3 score. Boys' data are closed circles, girls' are open circles.

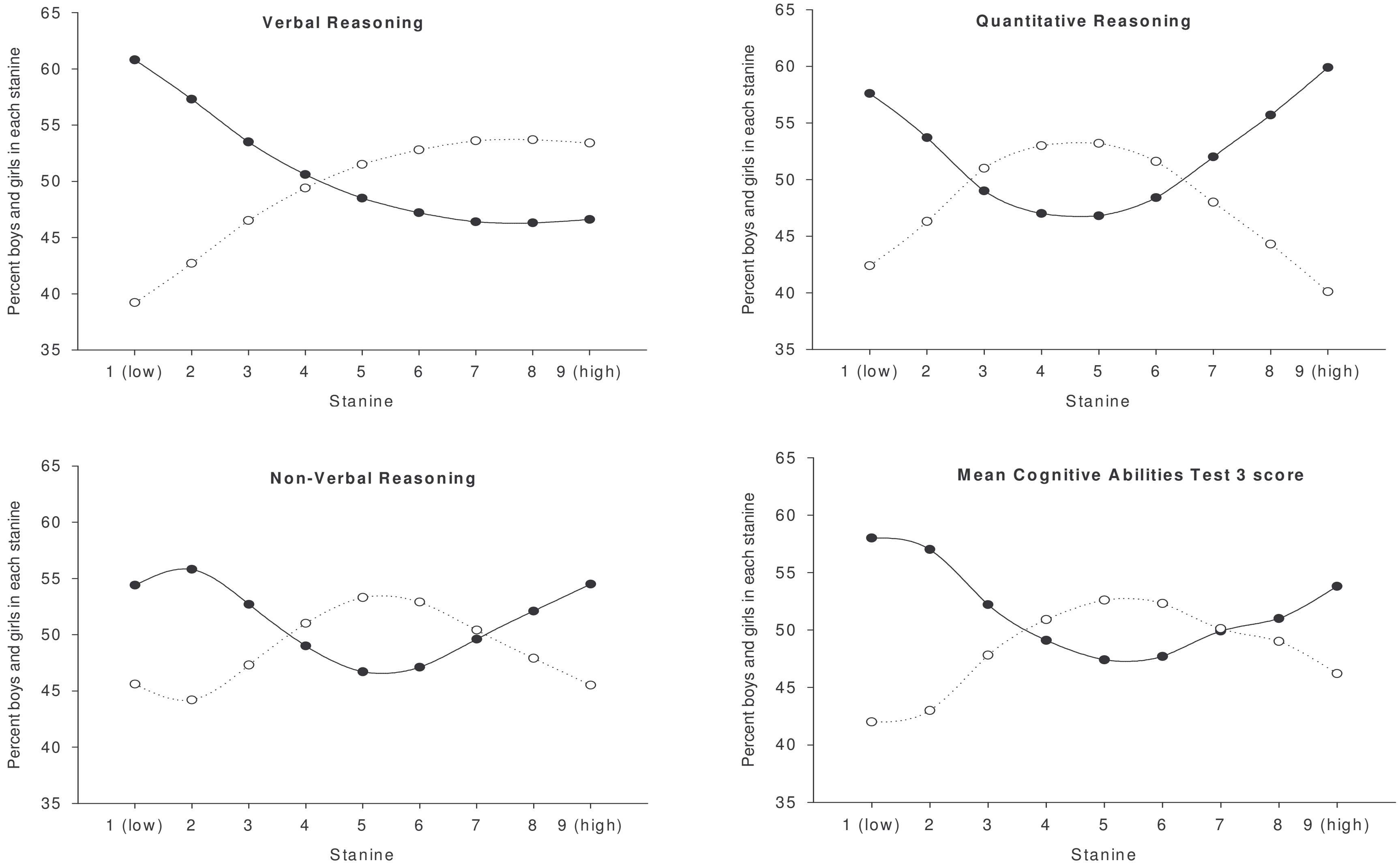
Verbal Reasoning
Quantitative Reasoning
Non-Verbal Reasoning
Mean Cognitive Abilities Test 3 score
Percent boys and girls in each stanine
Stanine
35
40
45
50
55
60
65
1 (low)
2
3
4
5
6
7
8
9 (high)

Figure 1

Appendices

Appendix Table 1

Percentage of pupils in Each Stanine Band

Stanine	National percentage of pupils (a)	Corresponding SAS
9	4	127 and above
8	7	119-126
7	12	112-118
6	17	104-111
5	20	97-103
4	17	89-96
3	12	82-88
2	7	74-81
1	4	73 and below

Note. (a) Percentages have been rounded to the nearest whole number.

Appendix Table 2

Numbers and Percentages of Male and Female pupils in Each Stanine on each CAT3 Battery and for Mean CAT3 Score

		Verbal Stanine									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	Total
Boys	N	8455	14171	17596	29308	30490	27544	16037	9857	4635	158093
	%	**5.3%**	**9.0%**	**11.1%**	**18.5%**	**19.3%**	**17.4%**	**10.1%**	**6.2%**	**2.9%**	
Girls	N	5448	10570	15312	28591	32385	30830	18557	11443	5321	158457
	%	**3.4%**	**6.7%**	**9.7%**	**18.0%**	**20.4%**	**19.5%**	**11.7%**	**7.2%**	**3.4%**	
Total	N	13903	24741	32908	57899	62875	58374	34594	21300	9956	316550
	%	**4.4%**	**7.8%**	**10.4%**	**18.3%**	**19.9%**	**18.4%**	**10.9%**	**6.7%**	**3.1%**	

		Quantitative Stanine									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Total**
Boys	N	3138	19634	18258	29037	23255	30376	16504	12565	5095	157862
	%	**2.0%**	**12.4%**	**11.6%**	**18.4%**	**14.7%**	**19.2%**	**10.5%**	**8.0%**	**3.2%**	
Girls	N	2313	16905	19002	32707	26438	32413	15215	10007	3406	158406
	%	**1.5%**	**10.7%**	**12.0%**	**20.6%**	**16.7%**	**20.5%**	**9.6%**	**6.3%**	**2.2%**	
Total	N	5451	36539	37260	61744	49693	62789	31719	22572	8501	316268
	%	**1.7%**	**11.6%**	**11.8%**	**19.5%**	**15.7%**	**19.9%**	**10.0%**	**7.1%**	**2.7%**	

		Non-Verbal Reasoning Stanine									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Total**
Boys	N	1390	18144	20713	29245	25720	27077	18095	11369	6077	157830
	%	**0.9%**	**11.5%**	**13.1%**	**18.5%**	**16.3%**	**17.2%**	**11.5%**	**7.2%**	**3.9%**	
Girls	N	1165	14370	18564	30488	29342	30458	18387	10450	5075	158299
	%	**0.7%**	**9.1%**	**11.7%**	**19.3%**	**18.5%**	**19.2%**	**11.6%**	**6.6%**	**3.2%**	
Total	N	2555	32514	39277	59733	55062	57535	36482	21819	11152	316129
	%	**0.8%**	**10.3%**	**12.4%**	**18.9%**	**17.4%**	**18.2%**	**11.5%**	**6.9%**	**3.5%**	

		Mean CAT3 Score Stanine									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Total**
Boys	N	2505	14505	19556	29917	29607	30327	17960	9392	2787	156556
	%	**1.6%**	**9.3%**	**12.5%**	**19.1%**	**18.9%**	**19.4%**	**11.5%**	**6.0%**	**1.8%**	
Girls	N	1813	10927	17872	31059	32867	33269	18016	9041	2394	157258
	%	**1.2%**	**6.9%**	**11.4%**	**19.8%**	**20.9%**	**21.2%**	**11.5%**	**5.7%**	**1.5%**	
Total	N	4318	25432	37428	60976	62474	63596	35976	18433	5181	313814
	%	**1.4%**	**8.1%**	**11.9%**	**19.4%**	**19.9%**	**20.3%**	**11.5%**	**5.9%**	**1.7%**	

FOOTNOTES

[i]. The ratios in Table 3 were empirically determined from the data. However the test scores were all normally distributed so the tail proportion ratios could be accurately estimated using the mean and SD for each sex together with the standard normal cumulative distribution function. This allowed modelling of the separate effects of mean and variance differences.