British Journal of Educational Psychology (2007), 77, 25–41
© 2007 The British Psychological Society

The British Psychological Society

www.bpsjournals.co.uk

Thirty years on – a large anti-Flynn effect? The Piagetian test *Volume & Heaviness* norms 1975–2003

Michael Shayer[1]*, Denise Ginsburg[2] and Robert Coe[3]
[1]King's College, University of London, UK
[2]Independent Consultant, Cambridge, UK
[3]University of Durham, UK

Background. *Volume & Heaviness* was one of three Piagetian tests used in the CSMS survey in 1975/76. However unlike psychometric tests showing the Flynn effect – that is with students showing steady improvements year by year requiring tests to be restandardized – it appeared that the performance of Y7 students has recently been getting steadily worse.

Aims. A sample of schools sufficiently large and representative was chosen so that the hypothesis of worsening performance could be tested, and estimated quantitatively.

Sample. Sixty-nine Y7 school year groups containing pupil data on the *Volume & Heaviness* test and the University of Durham CEM Centre MidYIS test were located giving a sample of 10, 023 students covering the years 2000 to 2003.

Method. Regression of the students' school mean on *Volume & Heaviness* on the schools' mean MidYIS 1999 standardized score, and computing the regression at MidYS = 100 allows comparison with that found in 1976.

Results. The mean drops in scores from 1976 to 2003 were boys = 1.13 and girls = 0.6 levels. A differential of 0.50 standard deviations in favour of boys in 1976 had completely disappeared by the year 2002. Between 1976 and 2003 the effect-size of the drop in the boys' performance was 1.04 standard deviations, and for girls was 0.55 standard deviations.

Conclusion. The idea that children leaving primary school are getting more and more intelligent and competent – whether it is viewed in terms of the Flynn effect, or in terms of government statistics on performance in Key Stage 2 SATS in mathematics and science – is put into question by these findings.

There has been much discussion, in recent years, as to whether standards have dropped or have been allowed to drop – whether it be at A-level (Tymms & FitzGibbon, 2001),

*Correspondence should be addressed to Prof. Michael Shayer, 16 Fen End, Over, Cambridge CB4 5NE, UK (e-mail: m.shayer@ukonline.co.uk).

DOI:10.1348/000709906X96987

GCSE, or in the various government statistics on Key Stage 1, 2 and 3 National Curriculum Tests[1] (NCTs).

On the other hand, those concerned with constructing and monitoring tests of general intelligence have found that, on retesting with nationally representative samples, children's responses have been improving – the Flynn effect. Thus, every 15 years or so the tests now have to be restandardized.

One might reasonably argue that, over the years, examination results are subject to pressures that are mostly unperceivable by those setting examinations and evaluating examination pass levels. If more students sit the examination, and the same proportion are given A or C grades and above, on the principle of more means worse (Amis, 1959/1990) the standard must drop, even if the quality of the questions remain the same. Such an argument could partially be countered by consideration of the Flynn effect.

In order to obtain a purchase on both of these kinds of problems it would be necessary to obtain data on a criterion-referenced test that is also tied to a theory-base which at least in principle offers an underlying scale that has measurement rather than relative properties. This article presents such data – originally researched by Piaget and Inhelder (1974) and in 1975 used in the CSMS[2] survey (Shayer, Küchemann, & Wylam, 1976) on a large nationally representative sample. The test *Volume & Heaviness* assesses children's concepts of physical quantities with, at the top end, their ability to use simple mathematical modelling of weight/volume relationships. The comparisons that will be offered here are for students just entering secondary from primary school in Y7, average age 11/7.

The psychometric and Genevan literatures and applied research

In the 1920s two quite different approaches to the description of children's cognitive development were initiated: from Binet's original test developed at the turn of the 20th century the art of psychometric testing grew (Binet, 1975). However, parallel to this Piaget (who had been a student of Binet's), beginning with 5- to 8-year-olds, sought to describe what, under the surface, the 'rules of the game' were for children's thinking. Until well after the Second World War the communication between these two arts – of the psychometric assessment of intelligence and of rich descriptions of the development of intelligence – was virtually zero. Yet they were both attempts to describe the same thing.

The achievement of the psychometric tradition is well summarized in Cattell (1971). Factor analysis on batteries of tests reveals a spectrum of different mental abilities, together with a factor common to all, called *g*, following Spearman (1927). Moreover, Cattell showed that there were two families of tests. Tests of crystallized intelligence related to learned or culturally determined skills and knowledge, whereas fluid intelligence tests were measures of 'here-and-now' thinking.[3] Raven's matrices is the most widely used of these tests of fluid intelligence (Raven, 2000).

After his initial work and publications in the 1920s Piaget restarted his life's programme by 1929–1931, describing the genesis of the first 2 years of his own

[1] *http://www.dfes.gov.uk/*

[2] *CSMS: Concepts in Secondary Science and Mathematics. Research programme funded at Chelsea College by the SSRC 1974–1979.*

[3] *This does not imply superficiality. Cattell (1971, p. 120) has 'fluid ability. . . a general, relation-perceiving span based on the magnitude of a neurologically efficient mass, and appearing as an existing energy in any current behaviour'.*

children's life from 0 days, 2 hours onwards (Piaget, 1953, 1936). In the 1930s he produced books, with many collaborators in Geneva at a rate of almost one a year, containing descriptions of different aspects of thinking of children up to about 11 years of age. These could be considered Piaget's versions of differential mental abilities. During the Second World War, isolated in Switzerland, he worked mainly on the possibility of explaining his findings in terms of logical models of different degrees of complexity (Piaget, 1949). This then made possible, together with Inhelder, the two major works which summed, between them, the whole genesis of children's thinking between the ages of 5 and 16 (Inhelder & Piaget, 1958, 1964).

From the point of view of applied and applicable research in intervention studies (Belbin, 1979) the psychometric tradition has a severe disadvantage. It rapidly dropped the concept of mental age – with the implication that an absolute scale was possible – to focus its methodology on comparing children with those of the same age in the standardization process, typically to a mean of 100 and standard deviation of 15. Draft test items are kept in, or ejected on the basis of their statistical properties rather than on a theoretical model. Thus, if we know that an 11-year-old, on a test of mathematical ability, has a score of 105, all we know is that he is slightly above average for his age. We do not know what maths he can or cannot understand, except in very broad terms. It is worse still if the question is asked of a score of 135, or of what he might do at 14. Moreover, Embretson (1988) showed that the IQ scale does not reach even the level of being equal-interval as described by Stevens (1946), which creates major problems in interpreting intervention studies where children make big increases.

By contrast, Piaget's work comes very close to having, in principle, a ratio scale as described by Stevens (1946): that is, a set of observationally defined levels and sublevels stretching back from Early Concrete, observed on 5-year-olds, to a hypothetical zero of his first recorded observation at 0 days, 2 hours. Although found very controversial Piaget's descriptions do have an underlying logic-based theoretical model to differentiate different levels of complexity. It also describes the *same* behaviours – for example in the ability to control variables in experimenting – whether the subject is 9 or 16. So any Piaget-based test is, by definition, criterion-referenced, and can be given equal-interval properties. The convenience of this for assessing school performance in science was shown in great detail in Shayer and Adey (1981).

The problem of reconciling the psychometric and developmentalist approaches was addressed in a monograph (Shayer, Demetriou, & Pervez, 1988) by scaling a battery of Piagetian tests taken from the work of Piaget on children from 5 to 10 years of age, in England, Pakistan, Greece and Australia using Rasch analysis (Rasch, 1980, 1960; Wright & Stone, 1970), which confers equal interval properties to the scale. This showed that children's performance could be differentiated into mental abilities similar to those that had earlier been described in the psychometric literature – a phenomenon unhelpfully labelled 'horizontal décalage' in Geneva – as well as the broad logical levels corresponding to the psychometric *g*. This led to two decades of important work by Demetriou, taking from and unifying the two traditions (Demetriou, Christou, Spanoudis, & Platsidou, 2002; Demetriou & Kazi, 2006), and the methods were also used on Piagetian tests used to assess the outcomes of various intervention studies (Shayer, 1999; Shayer & Adey, 1993).

The Flynn effect

Flynn (1987, 1994) surveying data on the standardization and restandardization of various psychometric tests from the late 1940s on, reported gains of about 9 points per

generation for tests of crystallized intelligence, and 15 points or one standard deviation for tests of fluid intelligence (Raven's matrices). This general apparent improvement has entered the literature as the Flynn effect. In a later review (Flynn, 1998), examining evidence going back to the 19th century, he had become sceptical of claims that this meant that people's intelligence – children or adults – was improving generation by generation. He wondered whether the concept of 'intelligence' was still viable. Certainly, people's test-taking ability has improved; but after discounting as minor factors such as nutrition, urbanization and TV, and pointing out that children's gains on arithmetic reasoning, vocabulary, creativity and speed of learning were far less, he was tempted to suggest that school and environmental stimulus in some undefined way were improving children's decontextualized problem-solving skills.

Yet this may be only a historic effect: in Grissmer *et al.*'s chapter in Flynn (1998, pp. 262–265) there is some evidence suggesting that gains are levelling off from the late 1980s. In this country the nferNelson CAT test showed no change in norms between the 1984 and 2000 standardizations (see the Endnote for details). Flynn (2006) wonders whether IQ gains will cease in all highly industrialized nations, citing Sundet, Barlaug, and Torjussen (2004) as evidence that gains have ceased in Norway, and Emanuelsson, Reuterberg, and Svensson (1993) and Teasdale and Owen (2000) that gains have flattened since the late 1980s in Sweden and Denmark, respectively.

The Volume & Heaviness *test*

The Science Reasoning Test II *Volume & Heaviness* (NFER, 1979), used in the original CSMS survey (Shayer *et al.*, 1976) has been used, since the early 1990s, as a pre-test for Y7 pupils entering secondary schools subsequently to be given the CASE intervention[4] (Adey & Shayer, 1990; Shayer, 1999). The median values for all studies in Flynn (1987) was 0.59 point a year for tests of fluid intelligence and 0.38 point a year for tests of crystallized intelligence. Thus in the 30 years since the standardization of the *Volume & Heaviness* test in 1975, the anticipated increase in performance might have been expected to be 1.18 standard deviations were it a test of fluid intelligence, or at least 0.76 standard deviation if a test of crystallized intelligence, if subject to the Flynn effect.

On the other hand, given that it is a criterion-referenced test, it is possible that these considerations do not apply. Although two of the tests used in the CSMS survey – SRT 1, *Spatial Relations* and SRT III, *Pendulum* – test essentially here-and-now thinking (fluid intelligence), SRT II, *Volume & Heaviness* is somewhat different. Ten of the items concern the concrete operational schemata of conservations of quantity of substance (mass), internal volume, external volume, weight, displacement volume and intuitive density. Only four of the items relating to the formal operational level, concerning density as a weight/volume relation, test here-and-now (fluid) thinking, but for these there is probably a crystallized component as well. The crucial period for the attainment of most of the conservations tested is the period from 5 to 8 years of age. If the children have them at 11 + years old they will pass the items without having to think, as 'obvious': the test content mostly relates to children's past experience in school and home.

Increases in achievement reported in government statistics

Whether it be NCTs at Key Stages 2 or 3, or GCSE, or A-levels, each year's published statistics show improvements (with the occasional exception) from one year to the next

[4] *CASE: Cognitive Acceleration through Science Education. Research project funded by the SSRC, 1984–87.*

in science and maths. On the other hand, a study published by the Engineering Council (2000) where teachers at 60 university departments had used their own means of testing their incoming students' mathematical competence, showed a decline in basic mathematical skills in relation to the same A-level grades, the drop starting in 1985. Evidence that just such a lowering of standards has occurred is published in Tymms and FitzGibbon (2001). By using the International Test of Developed Abilities as a benchmark measure it was shown that just between 1996 and 1999 grade inflation in A-level maths was nearly one grade.

For NCTs at Key Stage 2, two studies, Tymms (2004) and Brown, Askew, Millett, and Rhodes (2003) present evidence which is more tricky to interpret. Between 1995, when the National Numeracy and Literacy Projects were introduced and the year 2000, rises of 0.7 standard deviations were reported by the DfES on both English and Maths Key Stage 2 NCTs. However then, post 2000, the levels plateaued. This could be interpreted as being due to schools – themselves subject to extreme pressure to show improvement – professionally teaching to the test and also giving their children test practise. Once this adjustment had taken place generally, the system, post 2000, would have no further scope for showing an increase in standards. However the evidence is different for the two subjects. Tymms and FitzGibbon (2001) show evidence from many large research sources that little or no increase in reading had taken place right from 1975 to 2000, yet in the last 3 years the KS2 English results had shown the rise mentioned above. Tymms continued the story: from 1997 to 2004 on the PIPS test[5] the gain was only 0.16 standard deviation (*SD*) on reading. For maths the evidence conflicts. Tymms, taking the period 1998 to 2002, cites effects sizes of 0.41 *SD* for KS2 maths, 0.2 *SD* for the maths component of PIPS and 0.18 *SD* from Brown *et al.*. This does suggest that at least half of the KS2 improvements in maths represent a real gain for the pupils, possibly due to an increased time of exposure to maths teaching.

The research presented

Since 1995 Ginsburg has offered schools a report service (NFER, 1979) initially for King's College PD only, but later independently for many LEAs in the UK. Schools send her the Y7 *Volume & Heaviness* test results in EXCEL file, and also the Y8 end-of-year tests on the formal operational SRTs *Pendulum* or *Equilibrium in the Balance*, and she gives them a class by class assessment of the effects of 2 years CASE teaching, related to the CSMS norms. She also compares the item analysis for each class with the item pattern from a good research sample, in order to highlight evidence for occasional faulty administration of the test. Her finding that the Y7 means of new schools from recent years have appeared to be unexpectedly low led to the research reported in this article.

Since 1983 the Curriculum, Evaluation and Management Centre (CEM), based at the University of Durham, has offered schools an extensive testing service, enabling many checks to be made on their children's ability, achievement and potential. Durham possesses well-established and up-to-date national norms on various tests, and in particular the MidYIS test, being a general test of developed ability used for Y7 pupils and standardized in 1999. There were 69 matches between schools that have used *Volume & Heaviness* in Y7 and are on Ginsburg's database, and have in the same year used the MidYIS[6] test, with the results on the same children in the CEM database, giving

[5] *PIPS. Performance Indicators in Primary Schools test run by the CEM Centre at Durham University.*
[6] *See www.midysproject.org*

a sample of 10,023 with approximately the same number of pupils in each year from 2000 to 2003. Thus, the MidYIS norms can be used to estimate what would be the national mean on *Volume & Heaviness* for each of these 4 years.

Both the paper and pencil MidYIS test and *Volume & Heaviness* are administered by teachers to whole classes. However, although there is a pupil response sheet, *Volume & Heaviness* consists of a number of demonstrations and all the questions are read out by the teacher, who is encouraged to make sure that all pupils understand what the questions on the paper mean, and have enough time to decide on their responses. Both tests occupy about 50 minutes of pupil time.

Details of the sample

In Table 1 the locations of the schools covered by the research are shown.

Table 1. Types and locations of schools providing test data

	Comprehensive		Independent			
Area	Girls	Mixed	Boys	Girls	Mixed	Total
London and the Home counties	4	4		1	1	10
East Midlands			1			1
West Midlands		11				11
North East					1	1
North West		2			1	3
South		1		1		2
South Wales		9				9
North Wales		2				2
Total	4	29	1	2	3	39

Among the 39 schools in the table, there were 63 Y7 year groups, as some schools provided data on more than one of the 4 years sampled. Comprehensive school year groups range from 160 to 200 in number of pupils, independent schools less. Each year supplied over 2000 pupils, as can be seen in Table 2.

Table 2. Representative sample distributions of *Volume & Heaviness* over 28 years

		6 and above		3.9 to 4.6	
N	Year	Boys (%)	Girls (%)	Boys (%)	Girls (%)
2,350	1975/76	33.4	23.9	12.6	17.3
2,816	2000/01	15.2	8.1	37.3	43.2
2,569	2001/02	10.8	8.4	39.3	46.1
2,456	2002/03	9.9	6.9	41.9	45.2
2,162	2003/04	5.7	4.7	41.8	45.3

Although it cannot be claimed that the sample of schools is in proportional relation to the total population of schools – in particular the south-west is not represented – they do represent both the areas of England and Wales and also state and independent schools.

The analysis of data

A general drop in students' success on VH: 1975–2003

In Figure 1 the distribution of scores of the original CSMS data on all Y7 students are compared with all the students in the 2003 sample, separately for boys and girls. Neither is an exact national sample: The MidYIS mean for 2003/04 is 105, and the NFER Calvert Non-Verbal mean for CSMS 1975/76 is 101.8, and so the difference is slightly underestimated. However the qualitative contrast is clear.

The data in Figure 1 are reported on a Piagetian scale of 3 = 2A, Early Concrete; 4 = 2A/2B, Middle Concrete; 5 = 2B, Mature Concrete; 6 = 2B*, Concrete Generalization; and 7 = 3A, Early Formal. Wylam and Shayer (1980) give the psychometric properties of the *Volume & Heaviness* test, and Shayer and Adey (1981) give information on its school use in relation to science teaching. In order to make a quantitative comparison of the two distributions two score ranges were selected: 6 and above – the 1975/76 modal value for boys – in order to compare high performance, and 3.9 to 4.6 as characteristic of the modal performance of both boys and girls in 2003. The proportions of the students in each range in each school were regressed on the school MidYIS mean, separately for boys and girls, and computed from the regression equation at MidYIS = 100. In Table 2 the proportions in each range are compared year by year.

It can be seen that in 1975/76, as reported in Shayer and Wylam (1978), there was a substantial difference, in favour of boys, in performance on *Volume & Heaviness*, which has disappeared by 2003. In addition, although both boys and girls have shown great drops in performance, the relative drop is greater for boys. Moreover, there was rapid change between 2000 and 2003.

When did the decline start? Boy/girl differences: 1976 to 2003/04

Figure 2 shows various data. The data points joined by solid lines from the year 2000 to 2003 represent the major data presented in this paper. VH means for each school were regressed on the school MidYIS means, separately for boys and girls, and national values read off at MidYIS = 100. The way in which this, and other data analysis reported was done is illustrated in the Appendix. The 1976 CSMS data points were recalculated from the original CSMS data by regressing the school VH means on the NFER Calvert Non-Verbal Reasoning test means and reading off at Calvert NV = 100. In Shayer and Wylam (1978) the mean levels were reported as

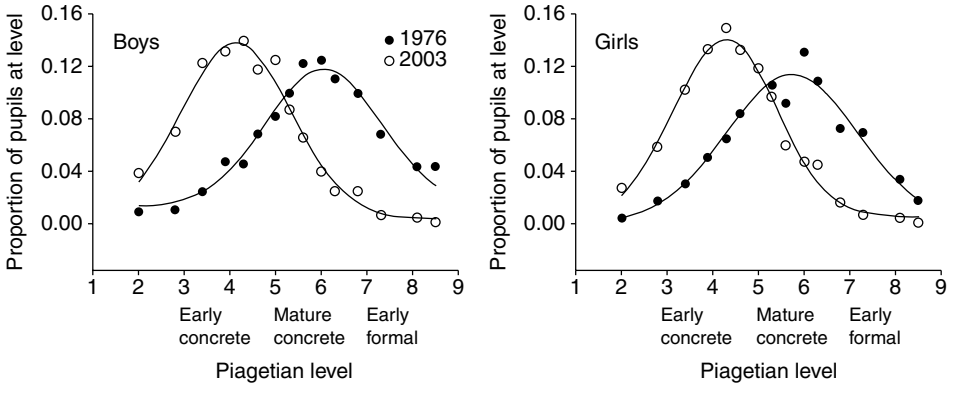

Figure 1. Comparison of performance for boys and girls on *Volume & Heaviness* 1976 to 2003.

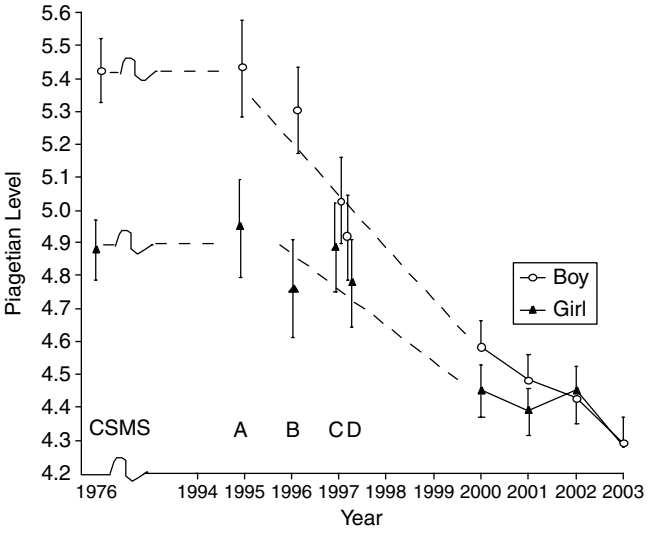

Figure 2. Boys/Girl means on *Volume & Heaviness* in various data 1976 to 2003.

boys = 5.67 and girls = 5.13. However this testing was done in the Spring term, at an average age of 12/1. In order to compare the original CSMS data with those tested on the MidYIS test from 2000 to 2003 with average age 11/7, early in the school year, the CSMS values were corrected for the age difference, using curves for the relation between age and mean score computed from the original CSMS data. The error bars here are the mean regression errors: that is they represent the actual school-to-school variation in the phenomena in question. In assessing the year-on-year fluctuations the standard error of the estimates of the means is around 0.07 level, so although there is no doubt about the significance of the changes from 2000 to 2003, the fluctuations in the girls' means from 2000 to 2002 are within sampling variation. Table 3 shows the data used in Figure 2, and also the effect sizes of changes.

Table 3. Effect-sizes of change on *Volume & Heaviness* from 1975 onwards

	Mean level		Effect-size (*SD*) of drop		
Year	Boys	Girls	Boys	Girls	B/G difference
1975/6	5.42	4.88	–	–	0.5
2000/01	4.59	4.45	0.76	0.39	0.13
2001/02	4.49	4.43	0.85	0.41	0.06
2002/03	4.43	4.45	0.91	0.39	−0.02
2003/04	4.29	4.28	1.04	0.55	0

In order to address the issue of when the decline from 1976 began other data were sought, and are represented on the graph from 1995 to 1997. The error bars here are based on school sample numbers and represent the statistical expectation of variation in the means.

The values for 1995 were obtained from one school in the research reported in Shayer and Adhami (in press). This school had given us nferNelson Cognitive Ability test (CAT) scores as well as the VH scores for their students. The CAT mean percentile was given as 50.8, and the VH mean, based on the original 1976 standardization of the test, as 48.6. Since neither of these were outside sampling variation of the 50th percentile this school (A) was taken as estimating the national average.

The values for the three schools (B, C, D) from 1996 to 1997 were computed from data on the school nferNelson CAT means supplied by Fernandez (see Endnote). In order to represent the data from these four schools – all with below-average intakes – their VH means were adjusted. This adjustment made use of the original Rasch scaling of the CSMS VH data that gives an equal interval exact linear relation between logit percentile and VH scale score. The logit difference between the VH scale value based on the CAT logit percentile, being the best current estimate of the school intake level, and the VH scale value for logit = 0 (50th percentile) was added to the school mean VH level to compare the school's data with other data estimating the national average. The original data are given in Table 4.

Table 4. Data for Figure 2

		VH mean		VH means adjusted		
Year	School	Boys	Girls	Boys	Girls	CAT Mean
1995/96	A	5.44	4.97	–	–	100
1996/97	B	5.27	4.73	5.32	4.77	99.3
1997/98	C	4.43	4.43	5.03	4.9	91.3
1997/98	D	4.49	4.45	4.93	4.78	94.9

The dashed lines on Figure 2 represent an interpretation of the data: they are not regression lines. The error bars allow the reader to gauge the reasonableness of the interpretation. It is suggested that the data indicate that there may have been no change in the VH norms until a year or two of the year 1995 and that the decline from 1995 onwards was consistent with the more accurate estimates from 2000 onwards. Nevertheless, it is acknowledged that this is based only on an opportunity sample: schools that should have been able to provide several years' data had already destroyed their CAT test records.

Inspection of the concepts involved in the Volume & Heaviness *test*

In Figure 3 the mean facilities of the 15 items in VH obtained in the CSMS survey in 1975/76 are shown. They were determined by regressing the schools' mean facilities on the schools' NFER Calvert Non-Verbal mean score, and computing the facility at Calvert = 100. For the schools from the year 2003/04 their mean facilities on each of the items were regressed on the MidYIS school mean, and their values read off at MidYIS = 100, for comparison with the CSMS survey. A drastic drop in performance can be seen comparing 2003 with 1975, and by comparing 2000/01 with 2003/04 it can be seen that the deterioration was continuing in the 3 years following 2000.

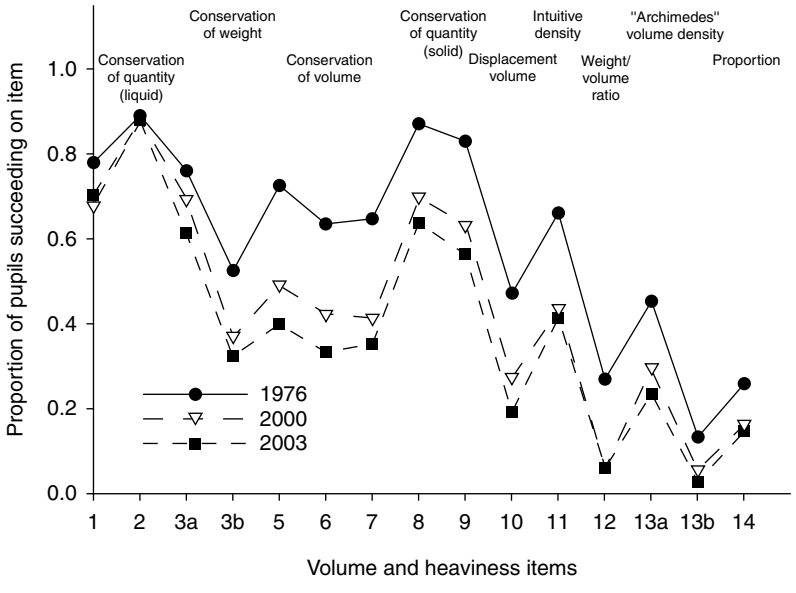

Figure 3. Item facilities on *VH* test for representative samples 1976, 2001 and 2003.

Items 1, 2 and 3a test the earliest of the conservations – that of quantity of liquid material. A narrow cylinder is filled with water and poured into a wider glass cylinder. The narrow cylinder is refilled and children are asked how the quantities of water in the two cylinders compare. It can be seen that there is relatively little change on these items. The majority of 7-year-olds possess this conservation. Item 3b tests the harder conservation of weight.

Items 5, 6 and 7, on the other hand, test the conservation of internal volume. For item 6 students are shown a rectanguloid block of plasticine, dimensions 5 by 4 by 3 cm. A glass cylinder is filled to the brim with water, and the block is lowered in by a thread just under the surface so the students see the water overflow. The cylinder is then refilled, and the students are asked whether more, less or the same amount of water would overflow if the block is lowered in (a) to half-way down and (b) right to the bottom. They are being asked implicitly whether they understand that since the volume of the block has not changed the position of it under the water should make no difference. The drop in performance on these items to half that of 1976 shows students have a much weaker concept of physical quantities.

Items 8 and 9 test a similar concept to items 1, 2 and 3a, except in this case the material is solid, not liquid. The plasticine block is first rolled into a ball, and then into a sausage shape.

The most drastic drop is on item 10, testing the concept of displacement volume. A 5 by 4 by 3 cm block of brass and a block of plasticine of identical dimensions are quickly passed around the class, and students are asked to heft them in their hands. They are then asked whether the metal block would displace more, less or the same amount of water than the plasticine. Here they need to have established that it is the volume of the blocks, not their weight that determines how much water they will push aside. In 1976 many more girls than boys had an implicit model that the weight is the determining cause. The average success rate on this item in 1975/76 was 54% for boys and 27% for girls, and by 2003/04 the differential had completely disappeared, with the success rate for both at 17%.

Item 11 in effect tests conservation of intuitive density, which is that density is a property of the material, not of its size or shape. Students are shown that the plasticine block sinks in water. Part of it is then rolled into a flat disc and they are asked whether it will float. Finally, a piece the size and dimensions of a penny piece is held up, and the same question is asked.

Thus far VH is only testing the conservations and concepts of physical quantities children have already developed over the last 5 or 6 years. In the items that follow they need both to be able to handle the weight/volume relationship as a concept of physical quantities relating to floating and sinking, and also carry out mathematical modelling on the spot. In item 12 they are shown a 10 by 10 by 10 cm box of thin Perspex, open at the top. In one question they are told that full of dry-cleaning fluid it would weigh 1,500 gm, and in the other full of alcohol it weighed 850 gm. They are asked to say whether each would float or sink if lowered into water, and to show how they got their answer. To assist them the same box is drawn on the paper, with another, just twice as tall (10 by 10 by 20 cm) immediately below it, which they are told holds 2,000 gm of water.

For items 13a and 13b the teacher tells them the whole story of Archimedes and the king's crown. They are told that copper is lighter than gold, and that Archimedes first found out the old and the new crown's weights and then their volumes. In 13a they are asked how he got their volume (with a visual hint of him in a full bath), but in 13b they are told that the new crown weighed more than the old, and yet Archimedes still proved there was some lighter metal in it. They have to show what his proof strategy was – a very difficult item! Typical strategies are either to argue from the weight/volume ratios in each, or to say that 'for the same volume the old crown weighed more'.

Finally, question 14, although involving density as a quantitative concept, is an archetypal mathematical proportion problem, depending more on students' abilities in maths. Two brass blocks are illustrated, one over twice the size of the other. They are told A weighs 60 gm with a volume of 15 cm^3, and B weighs 160 gm. Given that they are made of the same brass, what is the volume of B?

In fact, the pattern of relative success of these last four items is the same in 2003 as it was in 1976 – showing their inter-connection – only now students are coming up from primary schools with far less grasp both of physical causation and the ability to use mathematical modelling of the causal relationships.

However, perhaps the drop for items 5 to 10, testing conservations of quantity, weight and volume is yet more disturbing. In a general way, work in science and mathematics in primary school would be expected to provide the experience base for the development of these concepts, and yet this seems to have happened far less than in 1976.

Interpretation of findings and discussion

Originally VH was chosen as a pre-test for much assessment of the effects of the CASE intervention because it has substantial predictive validity for both science and mathematics achievement (Shayer, 1999). Also its administration in secondary schools was a very effective way of alerting science teachers, who would be administering it, the range of abilities in their pupils – interpreted both in terms of Piagetian levels and also simply in terms of difficulties they encounter with what teachers sometimes see as easy conservation and density concepts.

However the significance of the results presented in this paper can only properly be assessed if this test is located within the whole field of psychometric and other testing of

intelligence and abilities. Since the monograph of Shayer *et al.* (1988), Demetriou has published further research on generalized models of development (Demetriou, 2004; Demetriou *et al.*, 2002) – now strategically divorced from reference to Piaget, but, as with VH, criterion referenced so as to specify the age-related hierarchy of behaviours. From brain research he has taken account of working memory and from information processing he looks at how processing efficiency relates to performance. From differential or psychometric psychology he has refined the concept of mental abilities.

Thus, with five domains as shown in Figure 4 – each specified in terms of behaviours described as typically characteristic of the top 20% of the population – in a table consisting of seven 2-year steps from age 3–4 to age 15–16, he describes the whole course of development from early childhood to the end of adolescence (Demetriou, 2004). Each of the five domains is related to a second order factor, psychometric *g*.

However, then the issue arises, to what extent is *g* a factor-analytical artifact, or does it relate to activity in a separate part of the brain from those where the specifics of the domains are processed? Measures of processing efficiency, including speed of processing and executive control, and working memory account for most of the variance in *g*, and then between 60% and 65% of the variance on the items testing the five domains. This does suggest that *g* may correspond to parts of the brain where information, and also strategies for obtaining information, are more generally and even abstractly processed. Indeed, Duncan *et al.* (2000) has shown, by use of PET scans on adults working on spatial and verbal reasoning tests, each with high *g* loadings, that while each show brain activity in the *different* areas known to relate to the task specifics, nevertheless the *same* area of the left lateral frontal cortex was active during both tests. This is in a part of the brain concerned with concepts of executive control, strategy formation and monitoring the contents of working memory – just those represented by tests in Demetriou's research. Perhaps the two disparate specifics are abstracted to a level at which some kind of abstract decision-tree is tested against the chunks being processed in working memory? It may be that McCullough and Pitts's (1943) original suggestion and Piaget's (1949) symbolic logic model are closer to brain processes than has been thought.

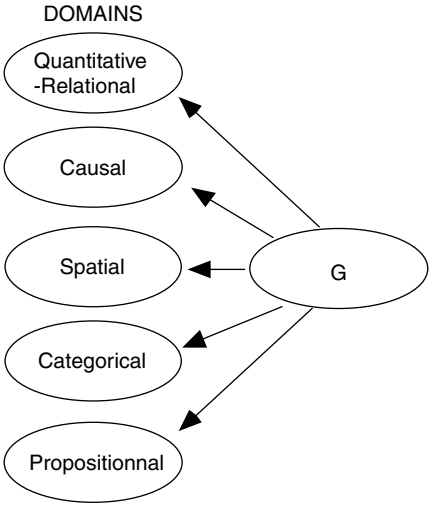

Figure 4. Demetriou's model of mental abilities.

Whatever view is taken on this, it is clear that any use of any general intelligence test (including VH) is conceptually faulty if it is used as implicitly measuring one and only one thing. In terms of Demetriou's model, a general *g* measure, which in some cases is approximated by an overall test-score average, needs to be accompanied by measures on each of the domains sampled by the test items. Some students will always score relatively high on spatial items and low on verbal items, and vice versa (Shayer *et al.*, 1988), and the description of a student needs to be represented accordingly. VH relates to Demetriou's quantitative-relational and causal domains and has substantial predictive validity for science achievement as well as mathematics and so relates to an important part of children's spectrum of learning achievement. Nevertheless, there is a difference between tests based on Piaget's description of cognitive development and other psychometric tests. It may be that they form a third category to Cattell's distinction between fluid and crystallized intelligence. The conservations assessed in VH are not directly teachable, yet they lie deep beneath the surface of what is potentially teachable, or testable by a psychometric test performance item. If a child does not conserve quantity he will not interpret any teaching on measuring volumes (Piaget, 1952, pp. 223–230). Only if a Y7 student does have the concept of displacement volume, thinking about relationships between the weight and the volume of solids or liquids begins to be possible. Most of the test items in VH require neither fluid nor crystallized intelligence, but it could be argued that they are *necessary conditions* for success on tests of crystallized intelligence featuring quantitative reasoning.

Thus the large drops in competence by 11- to 12-year-olds on *Volume & Heaviness* between 1975/76 and 2000/01, and the continuing fast drop in the 3 years following represent an important and objective finding – that is free from any process of adaptation to changing circumstances – that needs to addressed in assessing the overall impact of primary schooling on children. In addition, the much higher drop by boys in this period seems to tie up with other evidence on the deterioration of boys' learning, relative to girls, in schools. It makes it difficult to believe in the validity of the year on year improvements reported nationally on Key Stage 3 NCTs in science and mathematics: if children are entering secondary from primary school less and less equipped with the necessary mental conditions for processing science and mathematics concepts it seems unlikely that the next $2\frac{1}{2}$ years KS3 teaching will have improved so much as more than to compensate for what students of today lack in comparison with 1976.

On the reasons for the decline reported in this article one can only speculate. Piaget believed that it was the whole everyday environmental experience of the child that drove cognitive development, with schooling possibly playing only a minor part in the process – Vygotsky believed that schooling *should* change to play a major part (Shayer, 2003). Passive exposure to many hours of television a week has increased since the 1960s when 1975 CSMS students entered primary school. Computer games may have usurped what might have been, for boys, many hours playing outside with friends with things, tools and mechanisms of various kinds rather than virtual reality. However, it is possible that a decline in the use of activity methods in the early years of primary schools, in favour of an increased proportion of the time dedicated to the 3Rs as instanced by the National Numeracy and Literacy projects, may be partly responsible for the continuing decline from 2000 to 2003.

In saying that the conservations are not teachable, it is not implied that schooling cannot affect their development. Indeed, although primary schooling may only be a small factor in our reported decline, perhaps only the primary school can begin to remedy this major problem. There is recent applicable research addressing the learning of children in

the first 2 years of primary education (Adey, Robertson, & Venville, 2002; Fawcett & Garton, 2005; Shayer, 2005), but much more than this is required. Perhaps the next major government objective in education should be to address the question: in focusing teachers' attention on the specifics of the 3Rs only, What has been lost from the earlier primary practice of attending to the development of the whole person of the child?

Acknowledgements

We are grateful to Dr Cres Fernandez of nferNelson for searching his database for matching school files of the Cognitive Abilities Test. We acknowledge also the work of Nicola Forster of CEM, University of Durham both for finding the 69 matches between CEM and our own data-files, and also getting them safely to our database.

References

Adey, P., Robertson, A., & Venville, G. (2002). Effects of a cognitive acceleration programme on Year I pupils. *British Journal of Educational Psychology*, *72*, 1–25.

Adey, P. S., & Shayer, M. (1990). Accelerating the development of formal thinking in middle and high school pupils. *Journal of Research in Science Teaching*, *27*(3), 267–285.

Amis, K. (1990). *The Amis collection: Selected non-fiction*. London: Penguin Books (Original work published 1959).

Belbin, E. (1979). Applicable psychology and some national problems. *British Journal of Psychology*, *70*(2), 187–197.

Binet, A. (1975). *Modern ideas about children*. Menlo Park, CA: Suzanne Heisler.

Brown, M., Askew, M., Millett, A., & Rhodes, V. (2003). The key role of educational research in the development and evaluation of the National Numeracy Strategy. *British Educational Research Journal*, *29*(5), 655–672.

Cattell, R. B. (1971). *Abilities: Their structure, growth and action*. Boston: Houghton Mifflin.

Demetriou, A. (2004). Mind, intelligence and development: A cognitive, differential, and developmental theory of intelligence. In A. Demetriou & A. Raftopoulos (Eds.), *Cognitive developmental change: Models, methods, and measurement* (pp. 21–73). Cambridge: Cambridge University Press.

Demetriou, A., Christou, C., Spanoudis, G., & Platsidou, M. (2002). The development of mental processing: Efficiency, working memory, and thinking. *Monographs of the Society of Research in Child Development*, 67(Serial No. 268).

Demetriou, A., & Kazi, S. (2006). Self-awareness in g (with processing efficiency and reasoning). *Intelligence, 34*, 297–317.

Duncan, J., Seitz, R. J., Kolodny, J., Bor, D., Herzog, H., Ahmed, A., Newell, F. N., & Emslie, H. (2000). A neural basis for general intelligence. *Science*, *289*, 457–460.

Emanuelsson, I., Reuterberg, S. -E., & Svensson, A. (1993). Changing differences in intelligence? Comparisons between groups of thirteen-year-olds tested from 1960 to 1990. *Scandinavian Journal of Educational Research*, *37*, 259–277.

Embretson, S. E. (1988). Diagnostic testing by measuring learning processes: Psychometric considerations for dynamic testing. In N. Frederiksen, *et al.* (Eds.), *Test design: Developments in psychology and psychometrics*. Hillsdale, NJ: Erlbaum.

Engineering Council. (2000). *Measuring the mathematics problem*. London: Author.

Fawcett, L. M., & Garton, A. F. (2005). The effect of peer collaboration on children's problem-solving ability. *British Journal of Educational Psychology*, *75*, 157–169.

Flynn, J. R. (1987). Massive IQ gains in 14 Nations: What IQ tests really measure. *Psychological Bulletin*, *101*, 171–191.

Flynn, J. R. (1994). IQ gains over time. In R. J. Sternberg (Ed.), *Encyclopedia of human intelligence* (pp. 617–623). New York: Macmillan.

Flynn, J. R. (1998). IQ gains over time: Toward finding the causes. In U. Neisser (Ed.), *The rising curve: Long-term gains in IQ and related measures* (pp. 25–66). Washington, DC: American Psychological Association.

Flynn, J. R. (2006). O efeito Flynn: Repensando a inteligência e seus efeitos [The Flynn Effect: Rethinking intelligence and what affects it]. In C. Flores-Mendoza & R. Colom (Eds.), *Introdução à Psicologia das Diferenças Individuais* [Introduction to the psychology of individual differences] (pp. 387–411). Porto Alegre, Brazil: ArtMed.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence*. London: Routledge.

Inhelder, B., & Piaget, J. (1964). *The early growth of logic in the child*. London: Routledge.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, *5*, 115–133.

NFER. (1979). *Science Reasoning Tasks*. Windsor: National Foundation for Educational Research (now available as *Piagetian Reasoning Tasks* from Science Reasoning, 16 Fen End, OVER, Cambridge CB4 5NE, e-mail d.ginsburg@ukonline.co.uk).

Piaget, J. (1949). *Traité de logique: Essai de logistique opératoire*. Paris: Colin.

Piaget, J. (1952). *The child's conception of number*. London: Routledge.

Piaget, J. (1953). *The origin of intelligence in the child*. London: Routledge (Original work published in French in 1936).

Piaget, J., & Inhelder, B. (1974). *The child's construction of quantities*. London: Routledge and Kegan Paul (Original work published in French in 1941).

Rasch, G. (1980). *Probabalistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press (Original work published in Denmark in 1960).

Raven, J. (2000). The Raven's progressive matrices: Change and stability over culture and time. *Cognitive Psychology*, *41*, 1–48.

Shayer, M. (1999). Cognitive acceleration through science education: II. Its effects and scope. *International Journal of Science Education*, *21*(8), 883–902.

Shayer, M. (2003). Not just Piaget; not just Vygotsky, and certainly not Vygotsky as alternative to Piaget. *Learning and Instruction*, *13*, 465–485.

Shayer, M. (2005). *Report of the* Realising the cognitive potential of children 5 to 7 through their Mathematics learning *project*. Available from http://www.esrcsocietytoday.ac.uk/ESRCInfoCentre/ViewAwardPage.aspx?AwardId = 1568.

Shayer, M., & Adey, P. S. (1981). *Towards a science of science teaching*. London: Heinemann Educational Books.

Shayer, M., & Adey, P. S. (1993). Accelerating the development of formal thinking in middle and high school students: IV. Three years on after a two-year intervention. *Journal of Research in Science Teaching*, *30*(4), 351–366.

Shayer, M. & Adhami, M. (in press). Fostering cognitive development through the context of Mathematics: Result of the CAME Project. *Educational Studies in Mathematics*.

Shayer, M., & Wylam, H. (1978). The distribution of Piagetian stages of thinking in British middle and secondary school children: II. 14–16 year-olds and sex differentials. *British Journal of Educational Psychology*, *48*, 62–70.

Shayer, M., Demetriou, A., & Pervez, M. (1988). The structure and scaling of concrete operational thought: Three studies in four countries. *Genetic, Social and General Psychological Monographs*, 309–375.

Shayer, M., Küchemann, D. E., & Wylam, H. (1976). The distribution of Piagetian stages of thinking in British middle and secondary school children. *British Journal of Educational Psychology*, *46*, 164–173.

Spearman, C. (1927). *The abilities of man*. London: Macmillan.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*, 677–680.

Sundet, J. M., Barlaug, D. G., & Torjussen, T. M. (2004). The end of the Flynn effect? A study of secular trends in mean intelligence test scores of Norwegian conscripts during half a century. *Intelligence*, *32*(4), 349–362.

Teasdale, T. W., & Owen, D. R. (2000). Forty-year secular trends in cognitive abilities. *Intelligence*, *28*, 115–120.
Tymms, P. (2004). Are standards rising in English primary schools? *British Educational Research Journal*, *30*(4), 477–494.
Tymms, P., & FitzGibbon, C. T. (2001). Standards, achievement and educational performance: A cause for celebration? In R. Philips & J. Furlong (Eds.), *Education, reform and the state*. London: Routledge.
Wright, B. D., & Stone, M. (1970). *Best test design: Rasch measurement*. Chicago: MESA Press.
Wylam, H., & Shayer, M. (1980). *CSMS science reasoning tasks: General guide*. Windsor: NFER.

Received 16 October 2005; revised version received 29 December 2005

Appendix

The method of data analysis used for relating the MidYIS standardized scores to the *Volume & Heaviness* scores is instanced for the boys from year 2002, and relies on the substantial correlation between the two tests ($r = .78$ for Pearson's correlation between the school means). Figure A1 shows the regression line linking the two mean test scores.

In effect the regression line establishes a running mean of the scores when *Volume & Heaviness* is predicted by MidYIS. The line is computed such that the mean of all the squares of all the vertical distances between the line and the data points has become zero. Thus, reading off the *V&H* score from the line when MidYIS = 100 estimates the mean national average for *V&H* since 100 is the defined national average for MidYIS.

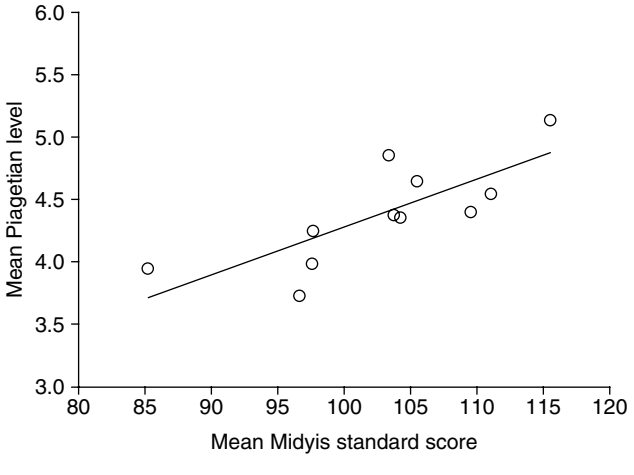

Figure A1. Year 2002: Regression line for estimating national mean for boys on *Volume and Heaviness* from schools using the the MidYIS test.

Endnote

Dr Cres Fernandez, nferNelson: personal communication. The second edition of the Cognitive Abilities Test (CAT2) was standardized in autumn 1984. The latest version (CAT3) was standardized in autumn 2000. An equivalence study was conducted to see the relationship between the CAT3 and CAT2 scores. Schools in the equivalence study were asked to administer both versions of one battery, with the order of administration being counter-balanced within each battery and level. That is, half the schools were asked to give the *CAT2* version first and half the *CAT3* version first. In 422 schools 10,240 pupils took part in the equivalence study. The number of pupils tested in each battery and year group varied between 250 and 850 pupils. The overall results from the equating study shows that an average performing pupil in 2000 with a standardized score of 100 in CAT3 Verbal would have expected to score 97 on CAT2 Verbal. Similarly, a pupil scoring 100 in CAT3 Quantitative would have expected to scored 99 on CAT2 Quantitative. A pupil scoring 100 in CAT3 Non-Verbal would have expected to score 98 on CAT2 Non-Verbal. The language used in a few questions for the CAT 2 Verbal battery may have dated over the period and may have contributed partly to the decline in the verbal scores. One other possible explanation is that the overall ability of the population of pupils in mainstream schools now is slightly lower than before as it includes many SEN pupils that used to be in special schools. More details are given in the technical manual: Smith, P., Fernandes, C., & Strand, S. (2001). *Cognitive Abilities Test Third Edition: Technical manual*. Windsor, UK: nferNelson.