British Journal of Educational Psychology (2009), 79, 409–418
© 2009 The British Psychological Society

The British Psychological Society

www.bpsjournals.co.uk

Thirty years on – a large anti-Flynn effect? (II): 13- and 14-year-olds. Piagetian tests of formal operations norms 1976–2006/7

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

Background. Shayer, Ginsburg, and Coe (2007) showed that children leaving primary school in Y6 entered secondary school with much lower levels of understanding of the physical conservations than in 1976. It seemed desirable to investigate cognitive development in the first three years of secondary education.

Aims. By using two Piagetian tests of formal operations, one of which had been used in the 1976 CSMS survey, the performance of Y8 and Y9 pupils would be compared with the survey results published in 1978.

Sample. Eight schools were selected as willing to test their Y9 or Y8 classes on either the *Pendulum* (in 2007) or the *Equilibrium in the Balance* tests (in 2006), giving 39 classes on which to make the comparison with 1976 performance.

Method. Regression of the pupils' class mean on each formal test on either the class's 2004 MidYIS or nferCAT standardized scores, and computing the regression at IQ = 100 allows comparison with that found in 1976.

Results. On *Equilibrium in the Balance* the Y8 pupils were down on the proportion at the early formal level or above by −0.61 standard deviations (*SD*) for boys and −0.63 *SD* for girls on the 1976 results. On the *Pendulum* the Y9 boys were down by −0.49 *SD* and the girls by −0.48 *SD* on the proportion at the early formal level or above.

Conclusion. The negative Flynn-effect found on *Volume & Heaviness* for Y7 pupils is paralleled by a similar negative effect on attainment of formal operations by Y8 and Y9, compared with 1976. Yet at the same time the proportion of pupils using the top level of concrete operational thinking has increased on both tests. It seems that there has been a change either in general societal pressures on the individual or in the style of teaching in schools – or both – favouring a lower level of processing of reality.

* *Correspondence should be addressed to Professor Michael Shayer, 16 Fen End, Over, Cambridge, CB24 5NE, UK (e-mail: m.shayer@ukonline.co.uk).*

DOI:10.1348/978185408X383123

Shayer, Ginsburg, and Coe (2007) in this journal reported that on the Piagetian test *Volume & Heaviness,* between 1976 and 2003, there had been a large drop in performance by English 11- / 12-year-olds entering secondary school in Y7. Instead of showing a modal value of concrete generalisation (2B*), as reported in the CSMS[1] survey (Shayer, Küchemann, & Wylam, 1976) by 2003 the modal level was down to middle concrete (2A/2B). The effect sizes for boys were − 1.04 *SD* and for girls − 0.55 *SD*. This called in question whether pupils' development of formal operations during adolescence might also be compromised. To investigate this eight schools were approached and asked to test their current Y8 or Y9 pupils (13; 14 years-old). These were currently taking part in PD (Professional Development) programmes for CASE[2] with their Y7 pupils, but whose current Y8 or 9 classes had no previous CASE or CAME[3] interventions.

In the CSMS survey, in addition to the use of the *Volume & Heaviness* test with 10- to 16-year-olds the *Pendulum* test was also used to assess the development of formal operations. While it would obviously be important to use the same test to compare 2007 norms with 1976, there were reasons to believe – as will be discussed later – that *Pendulum* might now overestimate the number of pupils showing formal operational thinking. Accordingly half of the schools were asked to use another test of formal operations, *Equilibrium in the Balance*, which had been shown in the CSMS research to give, for boys, the same estimate of formal operations as *Pendulum* (Shayer, 1978, p. 227).

Details of the sample

In order to be able to relate the Piagetian test data to National norms, schools in England were chosen knowing that their Y7 pupils had been tested either with the NFER[4] Cognitive Abilities Test (CAT) or with the MidYIS test from the CEM Centre[5], University of Durham. Details of the schools are given in Tables 1 and 2.

Schools D and E were willing only to test their more able classes: school G tested half the year-group. For the subsequent data-analysis the sampling unit chosen was the class rather than the school, as earlier research, e.g. Rutter, Maughan, Mortimore, Ouston, and Smith (1979); Goldstein (1980), had shown that this picks up more sources of relevant variation.

The analysis of data

The general method of data-analysis is shown in Figure 1. The Piagetian levels are shown on a scale from Mature Concrete (2B) = 5; Concrete Generalisation (2B*) = 6; Early Formal (3A) = 7 and Mature Formal (3B) = 8. Here the mean levels for the classes on the

[1] *CSMS:* Concepts in secondary science and mathematics. *Research programme funded at Chelsea College by the SSRC 1974–1979.*
[2] *CASE:* Cognitive Acceleration through Science Education. *Research project funded by the SSRC, 1984–87. This features a two-year intervention from Years 7 to 8 (Y7/Y8) for pupils between 12 and 14 years of age.*
[3] Cognitive Acceleration in Mathematics Education I *(1993–1995) project funded by the Leverhulme Foundation.* Cognitive Acceleration in Mathematics Education II *(1995–1997) project funded jointly by the ESRC and the Esmée Fairbairn Trust.*
[4] *NFER: National Foundation for Educational Research*
[5] *CEM: Centre for Evaluation and Monitoring – MidYIS is their Middle Years Information System with tests for 12- to 14-year-olds*

Table 1. Sample for Equilibrium in the Balance: Y8 classes, 2006

			Mean IQ		*N*		
School	Classes	Test	Boys	Girls	Boys	Girls	Type
A	5	CAT	–	116.3	–	93	Girls independent
B	5	CAT	106.4	102.4	84	93	Comprehensive
C	7	CAT	–	105.8	–	159	Comprehensive
D	1	MidYIS	106.3	106.5	9	8	Comprehensive
Totals	18				93	353	

Table 2. Sample for Pendulum: Y9 classes, 2007

			Mean IQ		*N*		
School	Classes	Test	Boys	Girls	Boys	Girls	Type
E	2	MidYIS	108.5	112.4	13	33	Comprehensive
F	5	CAT	112.1	111	43	53	Comprehensive
G	6	CAT	106.6	105.5	54	42	Comprehensive
H	8	MidYIS	106.4	–	129	–	Comprehensive
Totals	21				239	128	

Pendulum test – shown separately for boys and girls – are regressed on their mean IQ as assessed by nferCAT or MidYIS.

Here there is almost no difference between the boys and girls scores, and the mean scores are about 0.14 of a level at IQ = 100 above those in 1976. However, in Figure 2 and Table 3 it is shown that the number of pupils at the early formal level or above is well down on the 1976 proportions, with the girls still slightly ahead of boys.

The results of this analysis, with a comparison to 1976, are shown in Table 3.

Thus between 1976 and 2007 both the boys' and the girls' percentages at Early Formal and above have halved (10.4/22.6 for the boys and 12.9/25 for the girls). A similar analysis yields the results for *Equilibrium in the Balance*, in Table 4.

In order to shed light on these results, and to address the reason that the 1976 *Pendulum* results were used for comparison in Table 4, it is necessary to give more details of the CSMS research on formal operations. At that time the team were concerned to investigate whether Inhelder and Piaget's (1958) use of a generalized model of thinking applicable to all contexts was valid. Accordingly they prepared 5 'Class Tasks' from Chapters 3, 4, 5, 7 and 11 using the scoring rules for performances given by Inhelder and Piaget (1958). These were 14 or 15 item tests focused on demonstrations on apparatus that could be administered to pupils in a classroom so that pupils would see changes similar to those that were obtained by individual interview in the original Genevan research. These tests were then given to the same approximately 100 boys and 100 girls in the top four Y9 forms of two Comprehensive schools.

Table 5 is prepared from a table in Shayer (1978, p. 227) displaying the comparisons.

It can be seen that in Table 5 there is no significant difference between the boys' score on all five tests, and no difference between the boys' and girls' scores on *Pendulum*. It was argued that the boys' results showed that the same interpretative

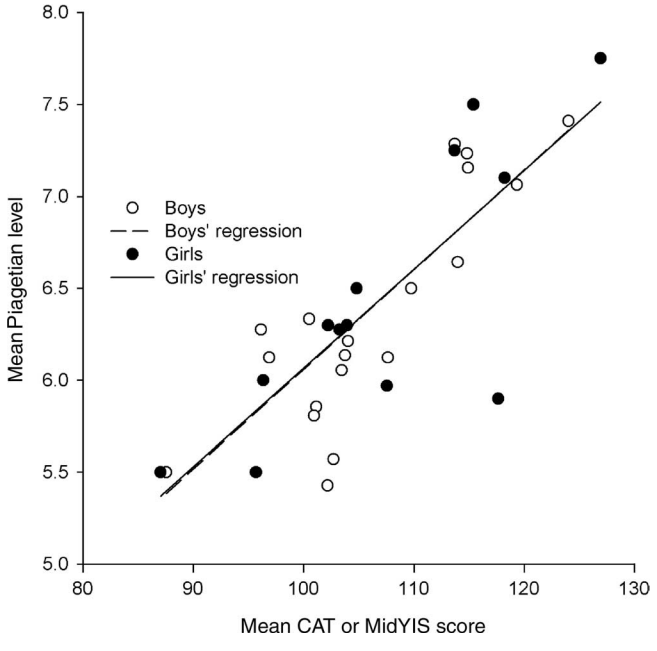

Figure 1. Mean Y9 class Piagetian levels in relation to mean national norms tests (Pendulum).

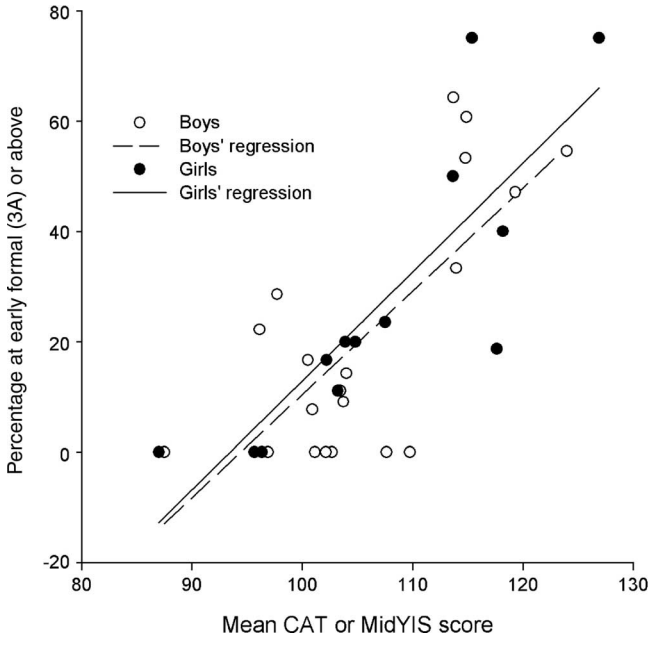

Figure 2. Proportion of pupils in each class showing formal operations skills on Pendulum, 2007.

Table 3. Results for the Pendulum Test: Y9 sample

Year		Boys		Girls		3A and above		3B	
	Mean age	Mean level	*SD*	Mean level	*SD*	Boys	Girls	Boys	Girls
1976 CSMS	14/4	5.88	1.114	5.98	1.066	22.6%	25.0%	7.7%	8.6%
2007	14/2	6.06	1.024	6.04	1.055	10.4%	12.9%	3.4%	2.7%
Change		0.18		0.06		−12.2%	−12.1%	−4.3%	−5.9%
Effect-size		0.17σ		0.06σ		−0.49σ	−0.48σ	–	–

Table 4. Results for *Equilibrium in the Balance*: Y8 sample

Year	Mean age	Boys		Girls		3A and above		3B	
		Mean Level	*SD*	Mean level	*SD*	Boys	Girls	Boys	Girls
1976 (Pendulum)	13/4	5.82	1.114	5.82	1.066	19.9	20.8	5.7	5.7
2006	13/4	5.90	0.881	5.72	0.923	4.6%	5.0%	1.2%	0.7%
Change		0.08		−0.1		−15.3%	−15.8%	−4.5%	−5.0%
Effect-size		0.07σ		−0.09σ		−0.61σ	−0.63σ	–	–

Table 5. CSMS evidence on five tests of formal operations (1978)

Task	Boys		Girls		Effect-size of difference
	Mean	SD	Mean	SD	
Pendulum	6.01	1.114	6.02	1.066	0
Equilibrium in the balance	6.11	1.067	5.61	1.133	−0.45 *SD*
Inclined plane	6.00	1.187	5.521	1.143	−0.41 *SD*
Chemical combinations	5.84	1.150	5.91	0.981	+0.07 *SD*
Flexible rods	6.03	1.201	5.72	1.254	−0.25 *SD*

model of thinking withstood the experimental test on all five contexts. This raised the question of the girls' results. As Piaget said (Inhelder & Piaget, 1958, p. 298)

> '. . . one must keep in mind that a concrete structuring of the data is an indispensable prerequisite of the propositional structure.'

That is, the subject first obtains a *description* of the data using concrete operational schemata, and then applies a deeper mode of analysis (formal operations) to produce an *interpretative* model (See Appendix 1 for an explanation of this distinction); but if the initial description is faulty, then the attempt at a deeper model will go astray. The 0.45 *SD* difference between the boys' and girls' scores on *Equilibrium in the Balance* and 0.41 *SD* on *Inclined Plane* is very close to the sex-differential of the 0.5 *SD* quoted in Shayer, Ginsburg, and Coe (2007) for *Volume & Heaviness* in Y7 for 1975. Each of these

sex-differentials was attributed at the time to differential play experiences of boys and girls in their early years. In contrast the *Pendulum* test required no descriptive modelling of the physics of the pendulum: it is just a test of control of variables where the girls' minds could operate just as freely as the boys' without an initial descriptive bias. In the much larger CSMS sample reported in Shayer and Wylam (1978) Y8 and Y9 girls' performance on *Pendulum* was virtually the same as boys' and the same as that of boys on *Volume & Heaviness*.

As in Shayer and Wylam (1978) it can be seen in Table 4 that, although the performance of boys and girls is down substantially on formal operations, the sex differential on *Equilibrium in the Balance* has narrowed from a mean score effect size of −0.45 *SD* to only −0.16 *SD* in the 30 years since 1976. The boys percentage at Early Formal or above has gone down to a quarter – 4.6/19.9 – of what it was in 1976 and the girls the same – 5/20.8 – on the assumption that the *Pendulum* tests conducted in 1976 in the CSMS survey were the best estimate of the formal operational capacity of both boys and girls, unrestricted by any bias due to faulty descriptive modelling. It also appears that any advantage in terms of differential early childhood experience boys may have had over girls in 1976 in terms of the physical world has disappeared, possibly by the boys' declining to that of the girls.

Comparison of Pendulum and Equilibrium in the Balance

As part of the PD for the CASE Project offered from King's College from 1991 onwards schools were offered a service to test whether their teaching had been effective in promoting cognitive development. Schools would give the Piagetian test *Volume & Heaviness* to all their Y7 entrants in the autumn term, and at the end of Y8 the pupils would be tested again on *Pendulum*. By using the CSMS norms for both tests it would be possible to tell if the school, and each class, had shown a value-added increase as indicated by a change in their percentile rankings.

This procedure worked satisfactorily until about 1995, but by 1997/1998 it was becoming apparent that the estimated school gains were too large. Examination of the item-analysis of *Pendulum* indicated that pupils' performance on some of the items – namely those on control of variables strategies – had changed since the original research reported by Inhelder in chapter 3 of Inhelder and Piaget, 1978. In a typical question pupils would be shown an experiment in which the length was *short*, the weight was *heavy* and the push was *gentle*. They were then asked the question:

> 'Which other arrangements would you use to test the effect that LENGTH has on the number of swings?'

together with the proviso 'Please use as few arrangements as possible; put a star (*) next to any arrangements that you don't really need'.

In both the original Genevan research and in the responses of pupils in 1976 subjects at the concrete generalisation level (2B*) would give a long list of most of the possible combinations of values and variables that were possible without ever isolating the correct early formal response of *long, heavy, gentle*. But following the introduction of National Curriculum in 1988 and particularly after its revision in 1995 and the introduction of KS3[6] testing, science teachers found that one obvious way they could improve their school's

[6] *At the end of KS3 (Key Stage 3) – the first three years of secondary school – Y7 to 9 – all pupils in state schools are given national SATs (Standard Assessment Tests) in Science, Mathematics and English.*

SATs scores was to teach the Science 1 Enquiry strategy of control of variables in experimentation. Pupils would be given the algorithm 'Keep all variables the same except for one'. By 1997 the long lists of combinations in the three control of variables questions had almost disappeared from pupils' scripts. They didn't always control the *right* variables (they might keep the value of the dependent variable the same), but most were now using the algorithm they had been taught, increasing the probability that they would be marked right on these questions. The discrimination level of the items had moved down from a low early formal (3A) to the concrete generalisation level (2B*), whereas the levels of the mature formal (3B) exclusion of irrelevant variables items, and other 3A items, had not changed. Accordingly in 1998 the Y8 post-test for pupils was changed from *Pendulum* to *Equilibrium in the Balance* on the grounds that no relevant algorithm existed for the latter context so that a more honest estimate of gains, if any, might be estimated for the school.

Bearing this information in mind it was thought that, for the purposes of the data-analysis of the *Pendulum* results, it would be better to use the original scoring rules used in 1976 for the CSMS survey, rather than the Piagetian scale established by Rasch analysis later for the CASE PD offered by King's College. These original rules, based on the Inhelder and Piaget (1958) criteria, assigned a subject to the highest level at which they achieved a 67% success, and were modified to take account of the fact that the three control of variable items had dropped to the concrete generalisation level (2B*).

It needs to be pointed out here that there was also another contributor to the inflation of gains that the team offering PD from King's College were unaware of at that time: the decrease in the Y7 *Volume & Heaviness* performance reported recently in Shayer, Coe, and Ginsburg (2007). Estimating the CSMS norms level of the Y7 entrants too low would also increase the apparent gains by the end of Y8. Fortunately all the reported effects of CASE and CAME in publications (e.g. Adey & Shayer 1994; Shayer, 1999; Shayer & Adhami, 2007) were done in terms of comparisons with control schools, so there these considerations do not apply.

Discussion

These data suggest that discussion of the Flynn effect may require a deeper analysis of test data than just examining standardized scores of psychometric tests, although Sundet, Barlaug, and Torjussen (2004) have shown IQ scores to have levelled in Norway, and Teasdale and Owen (2007) have reported recent small declines in Denmark. In terms of mean scores the results for *Equilibrium in the Balance* indicate little change for the boys since 1976 (0.08 *SD*), and a relative improvement for girls (from a deficit of 0.45 *SD* in relation to boys to only − 0.16 *SD* in 2006). On *Pendulum* both boys' and girls' scores have shown increases of the order of only 0.12 *SD* Yet the proportions on both tests showing the higher level thinking of formal operations are down radically from what they were in 1976.

On the meaning of the difference between the results from the two tests, an interpretation is offered based on the content of Table 5. There, for girls, an effect size difference in 1978 in relation to that of boys of − 0.45 *SD* was noted on *Equilibrium in the Balance*, but none for *Pendulum*. As in 1976 there is in 2007 almost no difference between the boys and the girls' scores on *Pendulum*. Thus these data are not correlated with the general drop in boys' scholastic attainment in schools in recent years: the boys are just as able to reason as girls, as they were in 1976. *Pendulum* is thought to test subjects' general

ability to handle complex information, whereas *Equilibrium in the Balance*, like *Volume & Heaviness,* does require adequate descriptive modelling of the physical world before the deeper analysis of formal operations can be effected. It is suggested that, whatever relative deficits in early childhood experiences may have been responsible for girl's poorer performance on both *Equilibrium in the Balance* and *Volume & Heaviness* in 1976 (Shayer & Wylam, 1978, p. 65), they are now experienced by the boys as well.

It is not suggested that these data reflect badly on the performance of English secondary schools. On the contrary – in Shayer, Ginsburg, and Coe (2007) it is reported that the mean score on *Volume & Heaviness* for 12-year-olds has slipped for boys from 5.42 (Mature Concrete) in 1976 to 4.29 (Middle Concrete) in 2003, and for girls from 4.88 to 4.28. If these data truly estimate children's level of thinking in Piagetian terms, then in the first three years of secondary education pupils' mean level of thinking has gone up by 1.77 sub-levels (6.06 from 4.29 for boys). In the original CSMS data from 1975/76 the mean increase from Y7 to Y9 was reported in Shayer & Wylam, 1978 as only 0.5 sub-levels. The range of performance of 13- and 14-year-olds has narrowed: many more pupils are now performing at the concrete generalisation (2B*) level than in 1976 – it can therefore be argued that a great amount of development is taking place as a result of secondary schooling, but that starting from a lower level than in 1976 the schools have done rather well. Compared with 1976 more pupils now develop and complete the *descriptive* thinking characteristic of concrete operations by the age of 14, but far fewer go on to develop the *interpretative* and *evaluative* level of thinking characteristic of formal operations. But it is also possible that, faced with public pressure to improve SATs scores, teachers may have responded to the National Curriculum by recognising that the performance of the majority of their pupils could better be improved by use only of thinking at the concrete operational level and simple cause-and-effect models.

If there were any bias in the selection of schools it would have been expected to favour relatively higher performance in the pupils, rather than lower. School A is a well-regarded girls independent school; School F is a school from the south of England assessed in 2005 as 'one of the 30 particularly successful schools' in the National list by Ofsted[7]; School G is a Comprehensive with the best GCSE results in its county.

This paper suggests further questions which it is hoped other workers will address. Do schools from Y7 to Y9 remove the deficit on descriptive models of conservations shown for Y7 entrants on *Volume & Heaviness* in Shayer, Ginsburg, and Coe (2007)? Has cognitive development on formal operations now been deferred until ages 14 to 16, or will the 16-year-olds be found to be as far behind the 1976 CSMS pupils as the 13- and 14-year-olds have been found to be in this paper? Do pupils, in the most selective Public Schools that ignore the National Curriculum in favour of teaching their pupils to go as deeply as possible into the subjects they study, develop still as pupils used to develop in 1976? Were this, and the Shayer, Ginsburg, and Coe (2007) study to be replicated in other countries, it would be possible to test the hypothesis that the change is due to general changes in the social environment, as against the counter-hypothesis that it is just due to a specifically English change in primary and secondary school teaching practice.

[7] *Ofsted: Office for Standards in Education, the organisation charged by the government to carry out frequent inspections of schools and to make public their reports.*

References

Adey, P., & Shayer, M. (1994). *Really raising standards*. London: Routledge.

Goldstein, H. (1980). 'Fifteen thousand hours: A review of the statistical procedures'. *Journal of Child Psychology and Psychiatry*, *21*, 363–369.

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

Rutter, M., Maughan, B., Mortimore, P., Ouston, J., & Smith, A. (1979). *Fifteen thousand hours: Secondary schools and their effects on children*. Somerset: Open Books.

Shayer, M. (1978). A test of the validity of Piaget's construct of formal operational thinking. PhD: University of London.

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

Shayer, M., & Adhami, M. (2007). Fostering cognitive development through the context of mathematics: Results of the CAME Project. *Educational Studies in Mathematics*, *64*, 256–291.

Shayer, M., Ginsburg, D., & Coe, R. (2007). 30 Years on – a large anti-'Flynn effect'? The Piagetian test Volume & Heaviness norms 1975–2003. *British Journal of Educational Psychology*, *77*, 25–41.

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.

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*, *948*, 62–70.

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. (2007). Secular declines in cognitive test scores: A reversal of the Flynn Effect. *Intelligence*, *36*(2), 121–126.

Received 11 August 2008; revised version received 15 October 2008

Appendix 1. Descriptive (concrete) and interpretative (formal) thinking

It is not widely realized how little is communicated by simple cause-and-effect models, although Hume did point out that all they say is that *this* event was concatenated in time with *that* event. But the distinction can be shown by contrasting Hooke's Law with Newton's Laws. When it is said that the extension of a spring is proportion to the weight applied it does indeed enable us to predict what will happen if the weight is doubled; but the usefulness of this descriptive prediction may conceal from us that nothing has been said on *why* the prediction is so.

By contrast consider this dialogue resulting from a use of Newton's Laws with a gifted Y8 Class. Teacher puts a book on the wooden table where all can see it:

Teacher: 'What forces are acting on this book?'
Pupil 1: 'There are no forces: the book is at rest' (descriptive thinking)
Teacher: 'Oh! – what about gravity?'
Pupil 2: 'Well, gravity is pulling down both the book and the table'

Teacher: 'Then if gravity is pulling the book down why doesn't it continue through the table on to the floor?'

Pupil 2: 'the table pushes up on the book and stops it'

Teacher: ' . . . and the book pushes down on the table because of gravity. But how do the two forces compare?' *(long wait)*

Pupil 3: 'I suppose that unless the forces are the same the book would move.' (interpretative thinking).

In this approach to Newton's 3rd law there is an explanatory model in use where concepts intervene between the concatenation of events.

Teacher: ' I want you all to remember this when you do the investigation on the rate of falling of cup-cakes' – groups of five pupils each with stop-watches timing a 2 metre drop of paper cake cups at successive 50 cm. intervals through the air during which the cups achieve terminal velocity. Some at least may realize that this happens when the forces acting on them are equal. (Newton's 1st Law interpreting constant velocity).

(Such teaching may be considered by many teachers as unhelpful for increasing SATs results!)

Equilibrium in the Balance likewise involves formal modelling, not simple cause-and-effect. Just changing the weight on one end gives the possibility of a descriptive model:' the more the weight is increased the more it will go down at that end.' But to make more precise predictions pupil has to set up a mathematical relation which relates the two variables of distance and the two variables of weight in such a way that predictions can be made from it. Because of the mental work involved in working his/her model pupil is now using interpretative thinking to make his/her predictions, which are also explanations in terms of relative forces.