Vernon-Wall

The Vernon-Wall Lecture for the annual meeting of
the Education Section of the British Psychological Society
Saturday 4 November 2000

Value Added for those in despair: research methods matter

Carol Taylor Fitz-Gibbon, Professor of Education at the University of Durham and Director of the Curriculum, Evaluation and Management Centre[1]

'SLIDES' refers to a set of powerpoint slides. These will be available with colour and animation on a website and are reproduced in black & white at the end of this booklet. C.T.Fitz-Gibbon@dur.ac.uk

Why should anybody be in despair, particularly with reference to value added? Sixteen reasons immediately spring to mind. These are described below and followed by some antidotes to despair.

Given the 15,000 hours of compulsory treatment meted out to the young in the name of education, it is important that we get things right by examining the problems and approaching solutions scientifically.

Here are some problems, the sixteeen reasons for despair:

1: The blunderbuss approach to school improvement

We are all tempted to believe that we know a good deal about how to improve the education system. Unfortunately inexperienced politicians and their advisers have not generally had the sobering experience of finding out that the world is not so simple that solutions can be found by guessing and easily implemented. Proper research is needed, tedious as that might be.

Instead of carefully researched, piloted and independently evaluated initiatives, teachers in England face a deluge of urgent interventions, with the consequent danger of inducing fatigue and cynicism. The plan, if there is one, is naive: try many well-intentioned, new (or newly named) activities and things will get dramatically better, at least in some places. Then we will know 'what works' and can spread 'best practice'.

In contrast, experience makes researchers and psychologists properly cautious regarding the power of well-intentioned guessing. For example, McCord followed up the careers of over 500 men who had been deemed to be 'at risk' in their teens. Counselling and support had been given over a period of five years. Thirty years later they

[1] The CEM Centre provides value added analyses and a range of additional indicators, affective, social and behavioural, to about one in three schools in England, at their request.

remembered their social workers with affection and many reported that the help had led them away from crime:

'I was put on the right road'

'I think I would have ended up in a life of crime…'

'I probably would be in jail.'

When objective measures were used based on records… such as the number of arrests the seriousness of crimes and recidivism…some results showed no differences between those helped or not helped but other differences were in favour of those *not* provided with help. Thus a well-intentioned programme of help to 'at risk' youngsters heading for delinquent careers had actually made the outcomes worse.

The negative effects, such as increased seriousness of crimes and higher recidivism, would not have been known had there not been proper evaluations using no-treatment control groups. The equivalent non-treated groups enabled us to see what the outcomes might have been without the programmes. (McCord, 1978; McCord,1981; McCord, Tremblay, Vitaro and Desmarais-Gervais, 1994; Dishion, McCord & Poulin, 1999). Not only had public money been wasted but it seemed to have been used to make the outcomes worse.

Such negative findings are immensely valuable, even if unwelcome.

But delinquency may be particularly ill-suited to guessing what will work. Perhaps achievement is different? To 'drive up standards' can we not try lots of good ideas to boost achievement and simply watch for rising test scores, rather than going to the trouble of using carefully designed evaluations? Slavin typified the approach of not using control groups as

"an assault on the very core of social science"

and pointed out that

"In practice, what many developers do is to amass easily collected data from large numbers of schools and only report the ones that made great gains in a particular year. ....It is how makers of miracle cures advertise their products: there's always someone who took their pills and lost weight, grew hair or recovered from cancer (for reasons unconnected to the treatment.' (Slavin, 1999 p. 36)

This is a caution that in England we should remember when we hear the evaluations of hopeful Education Action Zones or the see the label 'Beacon School' applied.

But could any ill come of literacy hours, or more mathematics, or increased homework or breakfast clubs? Possibly. Pressure often has deleterious effects particularly on some vulnerable children. Indeed early childhood interventions in which

"teachers initiated activities and children responded, adhering to a script with academic objectives for the children"

were associated years later with a delinquency rate twice that of other interventions that involved more child-centred and less pressured approaches (Schweinhart and Weikart, 1997). Could the present 'grad-grind' approach to schooling be a reason for the reported increasing levels of drunkenness and violence among young people in England?

2: Lack of evidence-based evaluation

Once it is understood that we may set out to do good and actually do harm there is a clear moral imperative to run 'reforms as experiments' (Campbell, 1979) and to conduct thorough evaluations with rich qualitative descriptions, particularly of implementation. When such proper research methods are used and the innovation or intervention *does* work, the benefit of the controlled design is that the *amount* of benefit can be estimated and we can begin to guide policies by quantifying costs and benefits.

There are currently unparalleled opportunities for evaluation, due to computing power, the framework of tests in England, and widespread, voluntary use in schools of self-evaluation systems, including value added measures. Yet initiatives have largely been launched without adequate evaluations to accompany them. With no systematic design to find out the effect of the initiatives we are left wondering why huge amounts can be spent, and teachers' lives disrupted, without knowledge that the initiative is at least more likely to do good than harm. Where are there estimates of likely effects? Targets yes …but no evidence that the targets are reasonable or reachable.

Of course there is national testing to assess progress, and test scores are rising. Perhaps everything is working? But the tests are newly designed each year and in such a situation the *maintenance* of standards is extremely difficult and to claim *rising* standards is not supportable (Tymms and Fitz-Gibbon, 2001).

In summary, policy makers have ignored the advice of psychologists and researchers and now have only contestable data and the hunches derived from conversations and impressions. … and in the 21st century! Decades after Campbell! This produces tear-out-your-hair despair.

3: Lack of use of value added.

The use of raw outcomes in the 'School Performance Tables' is widely recognised as unfair to schools since about 50 percent of the variation in examination results or Key Stage test[2] outcomes can be predicted from intake measures, as psychologists know. *Schools* in England lead the way internationally in the use of value added so why the slowness to adopt such approaches nationally?

The A Level Information System (ALIS)[3] started in 1983 using what is now called a value added approach and the work was noticed by the Scottish Office seven years later, in 1990. Following a year's consultancy they introduced value added measures from Standard Grade (age 16) to Highers (age 17) and this data has been made available to schools and Education Authorities in the 'standard tables' ever since. Thirteen years after the ALIS project started, we won a tender to undertake a two year contract with QCA (then the School Curriculum and Assessment Authority, SCAA) to design a national system of value added measures that would be 'statistically valid' and 'readily understood'. By this time we had hundreds of schools working on value added indicators with us at every key stage. The final report is available at

www.qca.org.uk/ca/5-14/durham_report.asp

[2] . Key Stages are the modules of education between the ages of 7,11,14, 16 and 18 years. End of Key Stage tests are externally set and graded. At age 16 these examinations are called GCSEs - -General Certificate of Secondary Education - - and at age 18 they are called A-levels - - Advanced Levels.

[3] ALIS = Advanced Level Information System to evaluate examinations at age 18 years.

and contained 32 recommendations.

To the general public it may seem that schools are still judged on raw outcomes since the most prominent information about schools is in the 'School Performance Tables', generally called the 'League Tables'.

The slow adoption of a value added approach is another reason for resigned despair….. but there are worse things…

4: Over-use of value added

Ofsted defines a good school as one in which pupils make better than average progress, i.e. a school with good 'value added'. What if the school also turns out racists? What if the school also has more than their fair share of delinquents? What if pupils are unhappy and their childhood is not as pleasant as it should be? Not all children survive to adulthood. Schools are *not* only a preparation for adulthood: they *are* life for children, and for teachers, and that life should be of high quality, with variety, with enjoyment, with challenges of all sorts, and not just academic.

5: Use of unethical indicators.

Can indicators be 'unethical', immoral, culpable? Certainly they can because *the most important effect of statistical indicators is the impact they have on behaviour.*

Some years ago I was somewhat naïve about the extent to which indicators will drive behaviour. I cannot escape this conclusion when I recall the answer I gave when called into the Department of Trade and Industry following a report showing that A-levels in mathematics, sciences and foreign languages were 'severely graded' (Fitz-Gibbon and Vincent, 1994). Policy people at the Department of Trade and Industry asked

'If some A-levels are easier will not schools and colleges be tempted to push students towards choosing to take the easier subjects?'

They were concerned since they saw mathematics, sciences and foreign languages as important for international competitiveness. I replied that I thought schools and colleges would put the interests of students first and advise with a view to subsequent careers and employment prospects.

I thought institutions would be strong enough to be indifferent to this latest, newly introduced game of publishing indicators….. But they cannot afford this luxury; funding, and even survival, might rest on being responsive to whatever political interference requires. The League Tables may influence enrolment and the data must therefore look as positive as possible…. Hence I was wrong to reassure the DTI .. and they were absolutely right to be concerned.

Now, I would agree with their concerns.

For an example of an unethical indicator consider the percent of students obtaining 5 or more GCSE passes at a grade of C or higher. This provides the major, published criterion against which schools are measured. This arbitrary indicator produces a false dichotomy by introducing a threshold value into a continuous distribution. This therefore focuses attention on the threshold or borderline: the D students.

Incidentally, it is indicative of the extraordinary world in which Whitehall mandarins live that when I mentioned this concentration-on-D-students at a seminar in 1999, a mandarin

said, severely, that ministers would be very annoyed if they thought schools were concentrating on D students. Rather flabbergasted and taken-aback, I protested that Ministers had no right to be annoyed since their policies made this inevitable. Since it would be difficult to find many schools that were *not* making special efforts with D students, this comment showed how out-of-touch the mandarin was, poor soul.

Despair is the order of the day whenever we see indicators reported that represent the percent passing some arbitrary line in a continuum of scores, for this borderline then becomes the focus of attention and distorts practice. Despair also is induced when one realises the lack of contact between some parts of the DfEE and real schools.

6: Poor statistical modelling.

Politicians have the excuse of ignorance. But there is another cause for despair: failure by those who should know better (researchers) *to model the process that produces the data.* The most glaring example arises from the analysis of A levels, but the same points apply at the end of each Key Stage.

The first question to ask when creating a statistical model is 'What process produced the data?' For examinations the answer is that an examining team for each syllabus, led by a Chief Examiner, produces the data. A regression line is therefore *needed for each syllabus.* This provides the only comparisons that are fair to participating schools because it is the Chief Examiner and the procedures followed by the markers for each syllabus that produce the particular relationship between intake and output for that subject that year. SLIDE 1 shows some regression segments for subjects at A-level[4] .

The use of regression *segment*s rather than a regression *line* is an important technique for representing the data visually. We use the term regression segment to indicate the part of the regression line that represents some range of the intake (the x-axis). We usually use a range defined by the mean plus or minus a standard deviation. Thus the regression segment represents visually 68 percent of the intake as well as the general trend relating intake to output scores.

Please observe the regression segments in SLIDE 1. The regression segments are not all in the same place; they are all over the place. They are substantially different. They are not coherent. They do not coincide. They certainly cannot be represented by one single regression line.

Yet counting all A-levels as equivalent underlay a major analysis done for the Department for Education and Employment (Donoghue, Thomas, Goldstein, Knight, 1996) and in the year 2000 the DfEE have conducted a 'value added exercise' that also treats all A-levels as equal. Oi weh! Despair!

7: Heads-in-the-sand regarding standards.

You might think that 'The minister for Lifelong Learning' was a character in a Gilbert and Sullivan comic opera, but it is actually a real 'new labour' title, currently held by Malcolm Wicks. He was reported as saying

[4] A-level: curriculum embedded, high stakes, authentic tests (ie examinations) for 18 year olds.

'It's a scandal that every year people come out of the woodwork to complain about standards at A-level' (The Times Higher Education Supplement, Sep 29th 2000. p.3)

It would be fine for him to call for more data, more evidence, but to just hurl insults is a poor example for someone promoting lifelong learning. To save him the bother of having to learn to practise what he preaches and ask for more evidence when he is in doubt about something, I have some evidence here regarding standards at A-level.

Concentrating first on the notion of the 'difficulty' of a subject (also known as the severity of the grading or what has been called the 'demand' of a subject e.g. SCAA and Ofsted, 1996): two approaches are possible and independent. One approach is to try to evaluate the content - - -what is being taught and is it difficult to learn? Indeed whether or not something is worth learning is surely a very important question. However, important as it is, this kind of judgement suffers from the problems of all judgements - -- the need for clear criteria and the need to demonstrate inter-judge reliability. When applied to *different subjects* judgements of comparative difficulty are almost impossible. Like is not being compared with like. However, within the same subject looking at examinations*over the years* , the judgements might have some value, though they will be more difficult to make the more the syllabus has changed. We could have inspectors read syllabuses and make judgements, as recommended by Goldstein and Cresswell (1996) but rejected as a means of comparing difficulties in a response from Fitz-Gibbon and Vincent, (1997). Reading the content of examinations, whilst important, will be a long and fraught process and inconclusive with regard to the level of 'difficulty'.

The other approach to subject difficulties is within the limits of the '50 percent framework' (Fitz-Gibbon, 1997). This takes the position that if students of the same aptitudes or prior levels of achievement routinely obtain higher grades on subject E than on subject H, then E is considered easier and H is considered harder. The comparisons are made solely within the framework provided by the fact that prior aptitude or achievements predict about 50 percent of subsequent variation in outcomes.

That differences in difficulty, as defined within the 50 % framework, occur and always have, is due to the fact that whilst most subjects attract a full ability range of students there are large differences in the general pattern of intake to different subjects. SLIDE 2 shows, for example, the prior achievement levels of students taking A-level Sociology and those taking A-level Physics. If the sociology grades were awarded on average 'in line' with Physics, there would be a huge failure rate in sociology. SLIDE 3 shows this in a general diagram: *Lower ability intakes must result in more lenient grading if massive failure rates are to be avoided.*

Whenever intakes differ but the failure rates are the same, there is a difference in what is variously referred to as the difficulty of the subject, the severity of grading or the 'demand' in the examination.

Measurements of these differences are essential in making fair judgements about schools and colleges. Furthermore, employers and admissions officers regard grades as indicative of ability but such inferences are clearly not accurate unless differences in the severity of grading of the various subjects are taken into account. Employers are probably right to be interested in grades as a general indicator of ability since they do predict job performance in a wide range of occupations (Schmidt and Hunter, 1977; 1981; 1993; Pearlman and Schmidt et al 1980).

The regression segments can also be used to develop hypotheses about trends over time, the subject of the outburst by the Minister for Lifelong Learning. As shown in SLIDE 4, the segments for more recent years are moving left (i.e. A-level courses appear to be

enrolling less able groups) and are moving upwards, indicating that higher scores are being awarded. (Of course, for comparisons across different years, the x-axis must stay constant. Fortunately we have been giving the same aptitude test to 18 year olds- - the International Test of Developed Abilities - - since 1988 so we can check the patterns against this constant baseline.)

Does the trend to more lenient grading imply falling 'standards'? I wouldn't put it like that. I would prefer to note that A-levels were being adjusted to a vastly increased staying-on rate. From 4 percent going to universities in the 1950s to 30 percent today, there has to be an adjustment. You can have inclusion or you can have standards, but you cannot have both the same standards and greater and greater inclusion.

Politicians want to claim *rising standards* but that is unlikely unless 'standards' means 'suitability for purpose', i.e. adjusting the difficulty to the intake. The mechanism of this grade inflation is difficult to know. Perhaps schools and colleges are searching for examination syllabuses that yield higher grades in order to improve their institution's standing in the mis-named [5] School Performance Tables. Perhaps examination boards are competing for customers. Perhaps there is covert pressure to award higher grades to justify the 'driving up standards' rhetoric. The origin of the grade inflation may be unknown but the effects are observable….the regression segments are floating up the page, year on year, particularly in the sciences and mathematics.

It is remotely possible to consider that the ever rising regression segments represent higher and higher achievements due to greater effort (e.g. more homework) and better teaching. I am fairly sure that teachers are trying exceptionally hard to get higher grades from their students since so much now rides on grades, for the school and for the teacher personally. However, efforts are not always as effective as we deeply believe. 'Work smarter not harder' as the Americans say.

So do we have evidence of harder work and higher achievement? The amount of homework reported by thousands of students answering independently in hundreds of sixth forms and colleges shows a steady decline in reported time spent on homework.[6] Not much evidence for harder work.

How about higher achievement due to, say, better teaching? Here we need evidence or corroboration. One very measurable subject for competence is mathematics. If students are actually achieving more highly in mathematics, this should be observable when they arrive at the university. A report from the Engineering Council brought together over 60 studies conducted by universities to assess A-level students' competence in mathematics. The conclusion was unequivocal

"There is strong evidence from diagnostic tests of a steady decline over the past decade of fluency in basic mathematical skills and of the level of mathematical preparation of students accepted onto degree courses." (Engineering Council Report, 2000, p. iii)

Furthermore the decline was there among the top 5% and was not just due to the greater range of abilities now taken into universities.

[5] Scores in these tables of raw results reflect not 'school performance' but the intake. The 'School' generally accounts for only 10 to 15% of variance in the outcomes.

[6] Data from the A-level Information System (ALIS), CEM Centre, University of Durham

In summary, the observed changes in the severity of grading in mathematics and science subjects is a reasonable response to changing intakes. Despair is induced only by the refusal to acknowledge parsimonious and reasonable explanations backed by data… Heads are in the sand - - - which is not at all healthy.

8: Spinning

If reality isn't palatable you need a Public Relations person, a spin doctor. A DfEE advertisement was blatant, or one might even say naively honest. (SLIDE 5) Is it not, however, rather insulting and demeaning towards teachers to imply that their views can be manipulated by someone skilled in PR and journalism?

Fortunately schools now often have good data and that will be the antidote to political spinning. **Educational Psychologists could become very important in working with schools on the cool and accurate interpretation of data.**

9: Getting the unit of reporting wrong.

There is only one virtue associated with computing *whole school* indicators: it sells newspapers.

When a single number is calculated to represent the effectiveness of an entire institution (a rather unconvincing quantification) it permits the ranking of the institutions and this simple ranking fascinates politicians and everyone else. It sells newspapers, by providing them with (supposedly) glowing examples and dire disasters.

But the use of *whole school* (*or whole college*) indicators hides the variation *within* institutions which is the variation that provides teachers and pupils with feedback and provides school management with the information they need to monitor. Moreover the 'school-effect' is not nearly as large as the 'class-teacher-pupil group' effect the ratio being probably about 1 to 4. In our work on the national contract to design a value added system Vincent (1997) used multi-level modelling to show the percentage of pupil-level variance associated with classrooms and schools. Classrooms accounted for far more variance than schools. [SLIDE 6]. Indeed this pattern (though not the precise quantities) had been apparent and commented on in early reports to schools in ALIS. And schools could themselves group by classroom or by teacher. Thus practitioners participating in monitoring projects could early know more than most researchers.

Vincent's tables reproduced in SLIDE 6 show the percentage of variance related to the school, the 'class' and the pupil in a multi-level analysis. Each 'class' is of course a confound, a coming together of not only the teacher but also the subject content and the particular group of pupils. The greater apparent impact of this 'classroom-effect' in mathematics (43% of pupil-level variance) as opposed to English (34%) is perhaps not surprising but it does raise the important question of *sensitivity to instruction.* Is student achievement more dependent on the quality of teaching (or the peer group) in mathematics lessons than in English lessons? Is this effect apparent at A-level as well as GCSE? Is the effect the same in the early years of secondary school when there might be changes in teachers from year to year? What are the implications for the evaluation of teachers and, perhaps, their remuneration?

There is one reasonable argument advanced for the use of whole school indicators: since parents and students have to choose an institution, the institution is therefore the unit of choice for analyses. This issue was confronted in the final report of the Value Added

National Project for SCAA/QCA[7]. The recommendation made was that the unit of reporting should be the curriculum-group, such as Maths-Science, Humanities, Performing Arts, Vocational-Business, etc.(Fitz-Gibbon, 1997). This would provide more useful information for those choosing schools than is provided by whole school indicators and these expanded 'league tables' would display the variation *within* the institutions for all to see. Indeed if you were seeking a sixth form, it would certainly be more important to know how good the progress was in the curriculum-areas in which you were interested rather than in irrelevant ones. (It would not be acceptable to go to the other extreme and report value-added at the level of the individual teacher for two important reasons: 1) it would be unreliable due to small sample sizes, and 2) it would essentially involve doing personnel work in public, which is probably unacceptable.)

The over-simplifications of rank-ordered league tables, the lack of general understanding that the differences between positions are trivial in the centre of the distribution and larger in the tails, the distortions of the data due to different curriculum choices (schools and college preparing students for science careers may be unfairly represented) all point to the need to prepare a better presentation of the data, as recommended. There is not a glimmer of a suggestion that the 'League Tables' will be appropriately revised (although finally an average-points score is going to appear as well as the unethical indicator mentioned above.)

10: The slowness of learning at the highest of levels

The DfEE recently ran a Value Added exercise in which all A-levels were counted as equal, and a distinction in Advanced GNVQ was considered equivalent to an A and a B at A-level. The baseline was the Total Points at GCSE which is not a good predictor since the value does not indicate the capabilities of the student so much as the entry policies of the school. Some schools enter students for 11 GCSEs others for only 5. The total points score will be strongly affected whereas the average would not be. The Value Added was a simple difference score, not based on regression.

How long before the DfEE learns to adopt fair comparisons as a criterion that must be met before data are published? How long before the recommendations of the 1995-1997 Value Added National Project are either implemented or rejected with reasons? The one that was immediately adopted was the giving of unique pupil numbers (Recommendation 2.5 [8]). How can the DfEE get value added so wrong?

I'm very fond of slow learners. I like to work with them and explain things till they 'get it'. But I'm not used to their being in control. Despair.

11: Ignoring Einstein.

Einstein said that everything should be as simple as possible, but no simpler. This is surely reasonable, particularly in an applied discipline. Educational research should be a practical, reality-contacting, scientific enlightenment not a mathematical weight-lifting

[7] SCAA = School Curriculum and Assessment Authority now re-named the Qualification and Curriculum

[8] *Recommendation 2.5: "A unique pupil identifier, for use in keeping track of children during the years of compulsory schooling, should be introduced with the use of check-digits or other methods of ensuring accuracy."*

exercise. So if someone wants to convince you that you should move from easy-to-run, readily understood methods of data analysis to purchase a new software package that few people understand, and that gives teachers figures they cannot check nor re-analyse, you need to ask what requires this move away from simplicity.

Specifically, what is the advantage, if any, of moving away from ordinary least squares analyses and using multi-level modelling instead? To answer this we need to answer:

Question A: How different are the results in typical datasets?

Question B: Are there underlying differences in the assumptions driving the analysis?

The differences in the results found in two primary-school datasets analysed by Tymms (1996 and 1997) and three secondary-school datasets analysed by Vincent (1997) were presented in the final report of the national project on Value Added. Various models were used yielding 13 analyses investigating the effects of taking account of compositional effects (the average ability of the group), curves vs. straight lines and allowing slopes to vary. In the primary school data the lowest correlation found between a simple Ordinary Least Square (OLS) analysis and a multi-level modelling (MLM) analysis was 0.93 and at the secondary level the lowest was 0.94. The modal value was 0.99 in both sets! No reason to switch there. As a compromise we recommended OLS for initial feedback to schools followed by a competition among researchers to 'explain' outliers among schools before any data were published, using MLM or other techniques.

Comparisons between OLS and MLM were made in the research paper for the DfEE. (Donoghue et al. 1999). Astonishingly the paper treated all A-level grades as having the same value for indicating progress i.e. it tried to ignore the different regression lines for different subjects as shown in SLIDE 1 and as recognised in ALIS since 1983. It is ironic that such a gross error is made in a paper looking at elaborate statistics. The error means that any discussion of 'differential slopes' was pointless because of the lack of the kind of analysis in SLIDE 1, relating the data to the processes that produced it.

The high correlations between OLS mean residuals and MLM mean residuals reported in 1997 in the Value Added National Project were replicated;

*...**we conclude that for the purpose of calculating overall institutional effects which reflect the performance of average students, the choice of OLS or MLM specification may not be critical**.....but...the confidence intervals associated with OLS estimates are biased and less efficient (i.e. larger) than those of the MLM estimates and will provide less discrimination between institutions.*

Donoghue, Thomas, Goldstein and Knight, 1996 page 19, section 7.9

Setting aside the glee associated with achieving less discrimination between whole-school indicators by use of OLS, the serious point is that confidence intervals are not particularly useful. Who is in favour of 95%? 99%? Do we have a bid for 99.5%? or how about 25%? Confidence intervals are a guessing game.

The usual choice is 95%, equivalent to testing for significance at the $p<.05$ level. Why is the latter common practice? According to Professor Robin Plackett, winner of two gold medals from the Royal Statistical Society, the profound reason for the choice of .05 is nothing but the historical accident of Sir R. Fisher's being unable to get the copyright for other levels. So Fisher published the .05 tables and said he'd found them useful. Yet even today there are people who have taken statistics courses who will tell you that if a

difference is significant at the .05 level it is 'true' difference. Otherwise, it is not. Despair.

The immensely serious point is that this is something that cannot be simplified: when are two institutions substantively different? We need to know what is the cost of a Type 1 or a Type 2 error and we need to know what is alterable and what is not. Just settling for the traditional .05 level and worrying if the test is actually at the .07 or the .10 level is not a defensible basis for practice, certainly not for the high-stakes practice of judging institutions, teachers, pupils…

But in case you worry about the size of confidence intervals, there is a table in the Donaghue, Thomas, Goldstein and Knight (1999) paper showing theoretical ratios of standard deviations from MLM and OLS models. For year groups over 50 there is little difference.[SLIDES 7 & 8] with the ratio of standard deviations being 0.92. For groups of n=200 the ratio is 0.98.

So, the answer to question A above, is that there is little difference in the important results, the value added averages, between MLM and OLS and only slight differences with small samples in the measures of variation, measures which are not much use anyway.

The reason for the differences in variation relates to a difference in underlying theoretical approaches (question B above) . MLM sets out to provide inferential statistics, treating the dataset as samples from a population of datasets. The aim is to estimate the long term likely values of various statistics, such as the mean. When using an early version of MLM, back in the 1980s, this bothered me. I'd worked in the inner city and got, I thought, good results. MLM would have shrunk these good results back towards the mean for similar pupils- - -unfair! The better the results the more the shrinkage and the smaller the group, the greater the shrinkage. Noted US statistician Stephen Raudenbush agreed: the way MLM shrinks extreme scores is a statistical equivalent of biased expectations affecting perceptions (Fitz-Gibbon, 1991, page 79). An alternative approach to the data is to describe it (descriptive rather than inferential statistics). Since the datasets represent 100 percent of each year's data, they can justifiably be regarded as population data rather than sample data.

To summarise, the high correlations between Value Added measures calculated by OLS or MLM, typically 0.99, indicate little is gained by using separate, specially purchased software with analyses not generally accessible, reproducible nor available for re-analysis by ordinary teachers, even mathematics teachers.

The cost of insisting on multi-level modelling is to remove the data from the overview of the practitioner. The cost is to provide data that that can be used to judge teachers but cannot be checked by teachers. The cost is obfuscation and an unscientific approach to data analysis, somewhat like applying relativity to analysis of the impact of a moving train: theoretically correct in a sense but ridiculous when orders of magnitude are taken into account.

This is not to detract at all from the beautifully produced multi-level modelling software developed in England by a team led by Goldstein (Goldstein, 1985; Goldstein, H., J. Rasbash, Plewis, I.Draper, D., Browne, W., Yang, M., and G. Woodhouse, G. and Healy, M. (1998) and another led by Aitkin & Longford (Longford, 1985; Aitkin and Longford, 1986; Longford, 1988). Participants in the discipline of Education are impressed and pleased that these professors have provided tools now widely used in many disciplines.

But you have to keep your mind focused on the purpose of analysing data for schools to know when an analysis can be regarded as totally flawed (e.g. Donoghue et al, 1999, because of ignoring differences between A-level subjects) and when simple techniques are best.

12: Cookery book significance testing.

Whilst we are considering the lamentably poor modelling adopted by certain statisticians, let's stay on their case a little longer.

Let's apply statistical reasoning to a piece of metal moving in a magnetic field:

Move a metal through magnetic lines of force and you get electric currents (p<.001). Electric currents cause heating (p<.000). Heating causes melting (p<.001) Therefore, since cars are made of metal and move through the earth's magnetic lines of force, they will melt when driving east-west [SLIDE 9].

Fortunately, physicists have the concept of the *magnitude* of an effect and at last the Effect Size is introducing this concept into social science (Glass, McGaw and Smith, 1981; Hedges and Olkin, 1985) [9] Testing for statistical significance is not sufficiently clarified and often not taught in terms of the costs and benefits of various decisions, which is the only way to make sense of type 1 and type 2 errors.

Professor Sir David Cox, in lectures given as president of the Royal Statistical Society, often made the distinction between passive observational data and experimental data. It is an important distinction yet insufficiently emphasised. [SLIDE 10]

The good news is that many of the most useful statistical procedures obey Einstein and keep things as simple as possible. But are teachers in training introduced to these vital concepts? Are they given practice in running experiments to check the efficacy of teaching strategies or other interventions? Rarely it seems. Despair.

Could educational psychologists provide up-dating training for teachers to help usher in the new era of evidence-based practice?

13: Using any available data, adequate to the job or not (e.g. FSM[10])

Schools are judged by Ofsted on the basis of inadequate gross comparisons currently known as PANDAs[11].

In October 1993, Sammons et al were to undertake a six month research project for Ofsted. The introduction to their final report (Sammons, Thomas, Mortimore, Owen and Pennell, 1994) records that

'The intention was to enable comparisons of GCSE performance to be placed in better context so that schools could be compared in "like with like terms".

---

[9] for an easy introduction see Fitz-Gibbon, 1984 and Fitz-Gibbon and Morris, 1987, chapter 7

[10] FSM = Free School Meals. The percentage of pupils eligible for FSMs is used as an indicator of poverty

[11] PANDA = Performance AND Assessment, an acronym from Ofsted

The project specification explicitly recognised, that,

ideally, a 'value added' approach ..... which employed baseline measures of students' prior attainment would provide the most appropriate basis for evaluating school performance. ..... The project was intended to investigate the usefulness of developing other less sophisticated ways of contextualising performance in the interim. It was intended that any grouping method developed would be of use in the short- to mid-term and would be superseded by value added methodology when prior attainment databases were instituted'.

In the event, tables were developed which Mike Tomlinson, currently the Acting HMCI, represented to the Select Committee of the House of Commons as the 'ULI method', but in fact, in the event, became only a grouping based on free school meals data, not on the variety of indicators which the Sammons team developed for this 'interim measure'.

It is, unfortunately, on the basis of this interim effort that schools have been compared - - -and with no sense of tentativeness. Ofsted spokespersons, even well educated ones, will show you with delight the relationship between FSM (the percentage of pupils on free school meals) and the average achievement of a school or college. The fact that a strong relationship ('explaining' about 60 percent of the variance) arises at this aggregate level (a means on means analysis) whereas the underlying correlation between any measure of social class you choose to use and a pupil's achievement is only about 0.3 and therefore explains only 9 per cent of the variation, does not seem to disturb Ofsted.

The strong relationship in the aggregate, means on means data, is due to the fact that many schools are segregated, to a greater or lesser degree, by social class. The segregation results in a strong relationship at the aggregate level which will vary from LEA to LEA according to the amount of segregation. [SLIDES 11-13]. This variation in itself is unfair to schools, since it results in some school performance measures being adjusted more than others.

The use of Free School Meals data is a source of despair for three reasons. One is that the datasets are often open to question. For example, observant Headteachers have noted that recorded eligibility for FSM often varies by sex, reflecting differences in reporting rates rather than real differences and thus casting doubt on the entire exercise. Secondly, Free School Meal eligibility is a very crude criterion, with little differentiation between depths of deprivation.

Thirdly, the use of poverty as an excuse is not necessarily excusable. It is particularly unacceptable now it has finally been admitted that after using the essential, major, appropriate and acceptable predictor of subsequent achievement, namely a measure of *prior achievement or developed abilities,* the amount of additional variance in achievement outcomes explained by adding in a socio-economic measure is generally zero (Goldstein, 1998) [SLIDE 14].

It might not have been so, and life might have been more complicated, but it is so: about 50 per cent is predictable from prior achievement or ability measures, and about 50 per cent or more is not predictable… and nothing much is gained by adding other *intake* measures. (What we need to know is the impact of *process* measures such as, for example, how and by whom the pupils are taught.)

When Ofsted let a contract to find some way of making adjustments for school results, it was explicitly stated that this was to be an interim measure until value added became available. Mike Tomlinson from Ofsted told a Select Committee of the House of Commons, 'We would love to use value added, but it simply isn't available'. They then

seemed to forget that they would love to use value added and asserted that free school meals percentages were good enough benchmarks of 'similar schools'.

I am hopeful that under the new regime[12] Ofsted will take more care about accuracy and fairness. Despair is on hold…..

14: The real need to 'drive up standards' - - - in inspection

Research methods matter very much when schools are to be judged on the basis of the resulting data.

From the inception of Ofsted, inspectors were required to judge raw achievement against national standards (which was easy) but also the progress of pupils i.e. the value added. The mysterious means by which they would judge progress in the absence of data was never revealed. They seemed to think that they could sit in classrooms and judge whether the pupils were making appropriate progress but they offered no evidence that this was true. One high-up Ofsted official told me that 'Five minutes of conversation with a pupil and I can judge that pupil's ability'.

Such arrogant belief in their own powers of insight is simply indicative of their totally unscientific approach to the whole issue of inspection of schools. They have done no studies of the adequacy of their sampling procedures. They spend 70 per cent of their time in school sitting in lessons. Why? The time in school is a pre-announced visit, which results in a week of charades and special 'Ofsted lessons'.

Ofsted have never produced an adequate study of the reliability of different inspectors in field situations. Their one published study was based on 17 per cent who volunteered, and who knew they were being studied and who had often worked together, and even so could not agree on which lessons were failing. It came up with an over-all correlation between these self-selected inspectors on ratings of quality of lessons on a seven point scale of 0.82. This cannot be generalised to ordinary field situations in which no studies have been conducted.

As for validity, SLIDE 15 shows average progress made over several years in a number of schools. Ofsted defines a good school as one in which pupils make better than average progress (thereby, at a stroke, ruling out half the schools who will be below average,) By their own definition many of the schools shown are very good schools --- schools in which pupils make better than average progress. Yet every one of the schools in SLIDE 15 was declared to be failing! I have published this data (Fitz-Gibbon, 1998) but Ofsted has kept *significantly* quiet about it.

Ofsted is a depressing topic. It is incomprehensible, except perhaps when seen as having been instituted by a government in which ministers would state, off the record, that all schools would soon be private schools. Ofsted certainly has demoralised teachers and must be partly held responsible for crises in recruitment. The damage they have done to schools called 'failing', to the children in those schools, and the parents of those children, needs careful study. Would that the inspectors could be prosecuted for administering cruel and unusual punishment ….to innocent teachers and inaccurately!

The major reason for the despair that must surround Ofsted is that it was ever allowed to operate without having any evidence of quality.

[12] Chris Woodhead resigned Nov. 2000

The major lesson to be learned is that research methods do matter; bad methods produce hurt, mislead the profession, detract from reality testing, hinder progress, waste public money. Ofsted's value for money is almost certainly negative.

Another lesson to be learned is that we must teach research methods in schools to avoid the current widespread ignorance of standards in research.

15: Unqualified Chief Inspectors

Young people are constantly exhorted to obtain qualifications so that they will be eligible for good positions in the labour market. The idea that people need to know various facts and be skilled in various procedures before they can be assigned to tasks is a very reasonable one. Skills and knowledge are essential to the adequate performance of many jobs. This rule, however, seems to break down in the higher echelons.

Civil servants move from department to department, perhaps carrying with them some aura of ineffable generic skills. But this practice may be a source of the quite amazing levels of mismanagement in the highest levels of government. Reference has already been made to ill-designed value added measures proposed by the Department for Education and Employment.

One of the most damaging instances of lack of qualifications must surely be the first two Chief Inspectors appointed to Ofsted. Stewart Sutherland was invited to set up Ofsted on a part-time basis, whilst he was also the Vice Chancellor of London University. Tasked with designing a system for evaluating schools and even individual teachers, one might have hoped to have somebody with some experience of education, evaluation, and statistics. He had no qualifications nor experience in any of those fields[13]. He was a professor of religion, with books to his name such as *Faith and Ambiguity*. Little wonder then that the design of Ofsted inspections was amateurish and anachronistic.

Furthermore, anyone seriously setting up a high-stakes system of evaluation should run pilot evaluations and evaluate their impact and their adequacy. There were no studies of whether a pre-announced visit yielded valid information as opposed to an unannounced visit to a school. There were no studies of the number of days that needed to be spent in a school to get good data, or the length of time inspectors needed to be in classrooms to get an adequate measure of a teacher. There was no justification for the accuracy of the judgements made.

It would not have been impossible for a person ignorant of basic scientific procedures to produce a reasonable inspection system had he or she had the humility to consult experts, or the wisdom to consider, at least tentatively, some general principles of design. Those who wrote the American Constitution, for example, had in mind some very simple rules about checks and balances. They were wary of power being one-sided. Lord Acton might not have been experienced in evaluation, but his observation that 'Power tends to corrupt, and absolute power tends to corrupt absolutely' is a well-known warning to those designing systems. The legislation that set up Ofsted ensured that the judgement of inspectors could not be challenged. Inspections can only be challenged in court if they have failed to follow procedures. Consequently, when Breeze Hill Comprehensive School in Oldham was first found to be satisfactory and then with a visit by HMI a few

[13] When asked about this he claimed ‘some’ experience of education but did not elaborate. Maybe Sunday school?

months later declared to be failing, everyone should have fallen about laughing at this demonstration of inconsistency. But the matter was too serious. The school wanted to challenge the judgement in court and the Local Authority was keen to do so. But lawyers advised that the *judgements* could not be challenged, and Ofsted had followed their own procedural regulations.

The second Chief Inspector of Ofsted was Chris Woodhead, an English teacher who claimed no understanding of statistics. His response to ever-increasing complaints was to appoint an Ombudsman (failure in the system to separate powers, an Ombudsman appointed by Chris Woodhead might not take the same positions as an Ombudsman appointed independently by, say, a Select Committee of the House of Commons), and to offer re-inspections to schools who might be dissatisfied with their first inspection. The last thing schools want is another highly disruptive, distressing inspection.

Schools want to look after pupils and even when they were unjustly described as failing, most did not want to fight the judgement for fear of prolonging the press attention and the strain on children and teachers.

Despair arises from the apparently routine idea in government that a system of evaluation can be managed by somebody with no experience of evaluation. This needs to be severely challenged. Management and leadership demand knowledge of the underpinning processes. Wise decisions are based on an evaluation of the evidence. Neither Sir Stewart Sutherland nor Chris Woodhead had the knowledge or experience to evaluate the adequacy of Ofsted's procedures. Neither had the humility of decent scientists who have tested their own opinions against carefully collected evidence and realised that the world is not so easily assessed as the naïve or opinionated believe.

The results of Ofsted are widely believed to be a demoralised and angry profession. The anger should not be focused on any pair of individuals but on the inadequacies of our systems of creating public policy.

16. Over-concentration on unalterable variables: perpetuating stereotyping.

When statisticians say there is a significant difference between the achievement of boys and girls, the lay audience, alas, does not say 'How big is the difference?'; 'What is the overlap between boys and girls?'; 'Can you tell by the fact that someone is a boy that they are going to be better at maths than someone else who is a girl?' And if they say ‘Do we need to teach boys and girls differently?’ there are no useful answers except ‘probably not’ or ‘which boy? Which girl?’. What use, then, are these stereotyping research ‘findings’ of mean differences between boys and girls, ethnic minorities, socio-economic groups? The answer is that they provide rather easy topics for academic papers for the research assessment exercise.

But they delay the search for what Bloom called ‘alterable variables’ (Bloom, 1979). What can make a difference for all students? What can be changed?

If there is sexism or racism, let it be proved by observation directly, not simply arise as a statistical accusation with a lack of any supporting evidence. And let’s find the interventions that get rid of prejudice and labelling people by group-membership.

Differences between groups can arise for all kinds of reasons that are beyond the control of the classroom teacher, and if there's one thing this world needs to do, it is to stop thinking in terms of groups and respect each person as an individual. If you want to know

if a group of students can do maths, you have to test them on maths, not judge them as members of a group.

Antidotes to Despair

"It is better to light a candle than to curse the darkness.

And if you are a little grannie from the north, you'll need to light a flaming bonfire"

Old Chinese proverb (improved)

A: Dreaming

Day dreaming is comforting and important. Furthermore, imagining a better future may help to bring it about.

My daydream is that the pensions of Chief Education Officers will be tied to the long-term outcomes (adjusted for suitable co-variates) of students in their care for 15,000 hours. Some former student is in prison? Then the pension decreases by £100. Former student who had been 'at risk' is fully employed? The pension goes up £200. Then Chief Education Officers would be asking how to create really effective schooling; they would ask for evidence; the long term consequences of education would become important….

I am even fairly confident that a fair system could be devised that would attract people because of the challenge of finding out what works. If their pensions depended upon it, they would ask very searching questions, demand validated answers from scientists and have creative solutions that they would test out over the 30 or 40 years of their stewardship in order to create good outcomes and contribute to the body of knowledge about how to create a better society through education.

This proposal may seem desperate and unrealistic but one must follow the admonition of Bertrand Russell: 'See the world as it is, not as you would like it to be'. And what we see is that people do respond to being held accountable. However, the responses in the short term are often damaging to a concern for the long-term. Hence the pension payoffs.

However, perhaps the spirit of science is a better antidote to short-termism. We don't know. **We need to conduct experiments with local government systems to find designs that work**.

B: Distributed Research: Working with those who do the job

Deming (1986) stressed that those who do the job are best placed to suggest improvements. Over the last two decades the experience of working directly with schools and LEAs in England[14] has been inspiring and also sustaining, a real antidote to the despairs described above.

The key role played by listening to and involving practitioners is illustrated with a few examples.

[14] Particularly in England with both the state and the independent sector but also with New Zealand, Scotland, and scattered schools in many other countries

The ALIS project was started to help one particular school to solve a problem. It was a practical question about a mathematics department drawn to my attention by a school governor who was a mathematics lecturer at the university. Didn't I think there were 'too many Ds' in A-level mathematics results? This question was put in 1982, before there was any talk of League Tables of examination results, years before Local Management of Schools[15], a decade before Ofsted and when the typical School Effectiveness Research project collected data only on small samples, often used means-on-means analyses rather than pupil-level data and did not give data back to schools. The Headteacher was concerned that the results were poorer than in English but the head of the mathematics department retorted that mathematics did not get the best students and also that mathematics was severely graded.

The school had, indeed, a very charismatic English department that was probably attracting some of the best students away from mathematics. As to whether mathematics was severely graded or not, only more data could cast light on the question. In fact only quantitative data could resolve this conflict. So started ALIS, originally as an unfunded research project with 12 schools (Fitz-Gibbon, 1985). Once the dataset was available and analysed it turned out that both the Head of the school and the Head of Mathematics were right. Mathematics consistently yielded lower grades than English for students of the same prior achievement but, even after taking into account this severity of grading, and taking account of the prior achievement levels of the students, the residuals (value added) were still strongly negative at the school causing concern. Because the Mathematics department could see how the calculations were made they then accepted what they had rejected for several years: that there was a problem. The school and the mathematics department then pulled together to change the results.

I originally saw the data collection as a six year school effectiveness study that would provide a useful answer and some research data. But I called it my reprehensible research. Why reprehensible? Because it was not experimental and therefore not contributing to a solid body of evidence that established cause and effect.[16]

As the project expanded to more schools, several of England's most well-known professors told me to stop letting the projects grow because, they said, the projects were: consuming too much time; research associates needed too much attention; I had enough data now; it was too much effort to measure attitudes and processes; a 25 per cent sample would be enough; etc, etc. Working with 100 per cent samples, pupil by pupil, syllabus by syllabus was indeed a considerable effort but it had gradually dawned on me that providing the data to schools was the most important outcome of the effort, far more important than writing research papers.

The provision of data to practitioners meant that they participated in the research. Indeed they were the *only* ones who knew the surrounding circumstances for their classrooms, their department, each pupil, each family, etc.. They were the major players: the ones who could interpret and learn from the detailed data. Likewise with research associates: far from being a burden they can be self-organising and contribute vital computing skills when they are given excellent facilities and time. Their motivation also is enhanced by contact with practitioners at conferences and by phones and email. The sense and usefulness of what we are doing induces creative thoughtfulness.

[15] 'Site based management' with 80% of LEA funds devolved to schools

[16] for a critique of the inadequacies of much school effectiveness research see Coe and Fitz-Gibbon, 1998

Just as mainframe computers with all the power at the centre, have given way to distributed computing, with power in the desktops at the periphery, so we now have, in England, distributed research with schools. **Schools are the laboratories for education and the teachers are in the laboratories. They are major players in this new era of 'Distributed Research'.** Given their own data, teachers are not misled by outliers; they conduct sensitivity analyses in intelligent ways without using the terminology. People learn quickly about data in which they have an interest. They analyse their own data well, particularly when it is presented graphically. They also see clearly after two or three years of data that the sampling variation is inherent, it will happen no matter what they do, and it must not be over-interpreted. By setting the average residuals in a statistical process control chart, ([SLIDES 15 & 16] , a procedure developed by mathematician Shewhart and adopted by W Edwards Deming, there is good guidance, quickly understood, regarding the interpretation of the data and taking cognisance of inherent variation and its relationship to sample size.

The key roles of those who were closest to the job continued throughout the next decade. As the projects grew it was more often Head teachers who took decisions to join the projects rather than LEA personnel[17]. Furthermore Heads and other teachers regularly participate as presenters in the eight conferences a year that we now run. It is not in every country that a deputy head who teaches sociology gives a detailed explanation of his use Statistical Process Control charts, a co-ordinator offers workshops on data use and analysis, and a teacher of Religious Education participates in research on under-aspiring pupils and reports this to rapt audiences.

The inspection system too was tackled by Distributed Research and a professional pressure group. Given the lack of standards in inspection, it was necessary to create Ofstin, the Office for Standards in Inspection. An outstanding group of retired teachers and headteachers organised a conference in Oxford in June 1996 with accounts collected into a book: *A Better System of Inspection* carefully edited by former Head Michael Duffy and amusingly illustrated by cartoonist Bill Scott. Ofstin then applied to the Joseph Rowntree CharitableTrust. This Trust does not deal with education, but it funded research into Ofsted under its *democracy* remit, perhaps because of the apparent violation of teachers' rights to what the Americans would call 'due process'. A team led by Maurice Kogan issued a report substantiating the high costs and stress associated with Ofsted inspections, and the lack of evidence that they produce improvement [Kogan et al, 1999].

The strength of Distributed Research.

Democracies are stronger than totalitarian states because they draw on the strength of their people and have self-correcting mechanisms, choice and diversity. **Monolithic systems are dangerously inefficient and liable to collapse and that concept almost certainly applies to centralised research, driven by national forums and effectiveness units, and judged by a centralised inspection system.**

[17] although LEAs in the northeast (Newcastle, North Tyneside, Durham and Northumberland) had responded positively when I requested they fund a post and were followed by Staffordshire and then many others eventually.

C: More data not less

There have been suggestions that just a few key indicators should be reported. Researchers Gray and Wilcox (1995) for example suggested three. On the contrary, if indicators are to be beneficial in their impact then we need more data not less.

It is increasingly recognised that a few key indicators are a formula for distorting practice. In the Health Service, for example, a focus on hospital waiting lists leads to these being manipulated but to little benefit. Reference was made earlier to the percent of students obtaining 5 A* to C grades… the unethical indicator that has focussed attention on the D students, those on the borderline of this arbitrary dichotomy. Use of just a few indicators also leads to the ranking and labelling schools.

In contrast there is ever growing interest …from schools…in a wealth of good data.

Again it was an observation of current practice that led to an important development, the Chances graphs. Early in the ALIS project some schools and colleges were reporting that only students with a C or better in the mathematics examination at age 16 were allowed to take A-level mathematics. It seemed important to look at the chances students had of passing from various starting points (Fitz-Gibbon, 1985). In all projects we now present 'Chances graphs' showing the chance (probability) of obtaining each grade from a given level of prior achievement. The empirical 'Chances' data pointed to a new way to present the data, a way particularly useful in working with students: the weak can be encouraged to try ('Look! Others have made it from your starting point) and the high-achieving can be encouraged not to relax their efforts ('From similar heights some students still failed').

With the Chances graphs we make visible the unpredictable 50 per cent of the 50 % framework. (Examples can be seen on the website www.cem.dur.ac.uk).

The Chances graphs were particularly important in helping to remove the worry that tests would be used to stereotype children. In the early years we did not give the scores on baseline tests back to schools, fearing the possible consequences of setting up negative expectations for some students. We were also alarmed at statements in the literature on testing referring to 'innate' abilities or 'potential', as though these could be measured by a current test. We studiously adopted the term '*developed* abilities'.

Fortunately individual human beings are not accurately predictable. I think it was the psychologist William James who said, 'individual biographies will never be written in advance'.

A very interesting example of the malleability of the mind has arisen from our work with a national survey of deaf students. In a sample of over 116,000 students, taking the MidYIS[18] test, 117 were identified as deaf. . On a test of Perceptual Speed and Accuracy[19] the deaf pupils scored a whole standard

[18] MidYIS: Middle Years Information System which contains newly developed baseline tests for ages 11, 12 and 13 years (the first three years of secondary schooling in England)

[19] A scale on the MidYIS test for 11 year olds

deviation above an average. This finding challenges the facile concept of the mind-as-a-computer with a built-in speed of processing. Furthermore it provides another example of the knowledge among practitioners: we are told by deaf colleagues that deaf persons were often employed by printers as proof readers. Thus the practitioners knew of these *developed* abilities. Indeed proof-reading is another innovative component in the MidYIS test and, sure enough the practitioners were right: the deaf scored better than average on that section also. [SLIDE 17].

Another example that confirms the concept of *developed* abilities is the article that has recently appeared reporting that the speed with which blind people can comprehend sentences is about twice that of non-blind people (Roder, 2000) . Indeed David Blunkett, the blind Secretary of State for Education is said to listen to audio tapes on a fast rather than normal speed.

As we move to more computer-delivered testing the possibility for widening the baseline measures are considerably enhanced. And by providing personal indicators on the '*current aptitudes*' or '*developed abilities'* of students we shall expect to find ever-richer insights from teachers, thus guiding developments.

D: The dawning (perhaps) of Evidence-based policies.

Another antidote to despair is provided by noting that at least the words regarding experiments are now being used: 'evidence-based'. Unfortunately there is backsliding from parts of the DfEE who hedge their options with the term 'evidence informed' meaning there is data somewhere. But data/numbers do not alone make science. Astrology uses numbers. *Designs* are needed to establish effects and enable cost-benefit analyses to be conducted and it is to good experimental designs that the term 'evidence-based' refers.

Deming (1986) also stressed the need for what he quaintly called 'profound knowledge' but which we might more simply call science. If Taguchi can conduct thousands of experiments before a production line is started, perhaps we can one day take the same care with designing the production process of schooling. We may need to make some changes. As David Hargreaves has pointed out:

"**Schools are still modelled on a curious mix of the factory, the asylum and the prison.**"

(1994, The Mosaic of Learning p. 43)

Recommendations

Learning from data can be speeded up by collaboration among schools but this research, takes time to organise. Hence there is a need for 'teacher-researcher' posts on the Senior Management Team with a brief to develop research that is useful. Given time in the timetable, thousands of teachers could become active researchers providing RFTs (randomised field trials[20]) of many innovations. This role would

[20] See Boruch (1997) for a book on RFTs

professionalise the work of teaching and would be likely to improve the retention of insightful and creative teachers in the profession.

There should be a career path that leads to becoming a 'teacher-researcher', equivalent to a senior management post.

Educational psychologists have a very important role to play in this new climate of distributed research. Almost alone among social scientists, psychologists are usually well-trained in methodology. They know that correlation is not causation; they know how to set up and evaluate an experiment; they can work with schools to help them find out what works. Thousands of experiments should now be designed, reported and summarised in meta-analyses, providing the kind of evidence recognised in medical research as the soundest basis for practice. [Figure 1 shows the Hierarchy of Evidence adopted in medical reviews of the literature].

I wish educational psychologists growing joy, and declining despair, as they face the challenge of making the best possible use of the 15,000 hours of compulsory treatment in schools.

Figure 1

Hierarchies of evidence

(from Medical groups)

I-1	**Systematic review and meta-analysis of two or more double-blind randomised control trials**
I-2	**One or more large double-blind randomised control trials**
II-1	**One or more well-conducted cohort studies**
II-2	**One or more well -conducted case-control studies**
II-3	**A dramatic uncontrolled experiment**
III	**Expert committee sitting in review: peer leader opinion**
IV	**Personal experience**

REFERENCES

Aitkin, M. and N. Longford (1986). "Statistical Modelling Issues In School Effectiveness Studies." Journal of the Royal Statistical Society A(149): 1-43.

Bloom, B.S. (1979) Alterable variables: the new direction in educational research. Edinburgh: Scottish Council for Research.

Boruch, R. (1997). Randomised Experimentation for Planning and Evaluation: a practical guide. London, Sage.

Campbell, D.T. (1979) Assessing the Impact of Planned Social Change. Evaluation and Program Planning, 2: 67-90.

Coe, R. and C. T. Fitz-Gibbon (1998). "School Effectiveness Research: criticisms and recommendations." Oxford Review of Education **24**(4): 421-438.

Deming, W.E. (1986) Out of the Crisis: Quality, Productivity and Competitive Position. Cambridge: Cambridge University Press.

Dishion, T.J., McCord, J., Poulin, F. (1999) When Interventions Harm. American Psychologist, 54(9): 755-764.

Donoghue, M., Thomas, S., Goldstein, H. and Knight, T. (1996) DfEE Study of Value Added for 16-18 Year Olds in England. London: DfEE.

Duffy, M. (1997). A Better System of Inspection? Ofstin (the Office for Standards in Inspection) 9 Quatra Bras, Hexham, NE46 3 JY

Fitz-Gibbon, C. (1998). "Ofsted: time to go?" Managing Schools Today **7**(6): 22-25.

Fitz-Gibbon, C. T. (1984). "Meta-analysis: an explication." British Educational Research Journal 10(2): 135-144.

Fitz-Gibbon, C.T. (1985) A-Level Results in Comprehensive Schools: The Combse Project Year 1. Oxford Review of Education, 11(1): 43-58.

Fitz-Gibbon, C.T. (1991) Multilevel Modelling in an Indicator System. In S.W. Raudenbush and J.D. Willms (Eds) Schools, Classrooms and Pupils: International Studies of Schooling from a Multilevel Perspective. London: Academic Press, 67-83.

Fitz-Gibbon, C.T. (1997) Listening to students and the 50 per cent framework. In A.D. Edwards, C.T. Fitz-Gibbon, F. Hardman, R. Haywood and N. Meagher (Eds) Separate but equal? A levels and GNVQs. London: Routledge.

Fitz-Gibbon, C.T. (1997) The Value Added National Project: Final Report: Feasibility studies for a national system of Value Added indicators. London: School Curriculum and Assessment Authority.

Fitz-Gibbon, C.T. and Morris, L. (1987) How to analyze data. Beverly Hills CA: Sage Publications.

Fitz-Gibbon, C.T. and Vincent, L. (1994) Candidates' Performance in Public Examinations in Mathematics and Science. London: School Curriculum and Assessment Authority (SCAA).

Fitz-Gibbon, C. T. and L. Vincent (1997). "Difficulties Regarding Subject Difficulties: developing reasonable explanations for observable data." Oxford Review of Education **23**(3): 291-298.

Glass, G.V., McGaw, B. and Smith, M.L. (1981) Meta-analysis in social research. Beverly Hills, CA: Sage Publications.

Goldstein, H. (1985). Multi-level mixed linear model analysis using iterative generalised least squares. . Biometrika.

Goldstein, H. and Cresswell, M. (1996) The Comparability of Different Subjects in Public Examinations: a theoretical and practical critique. Oxford Review of Education, 22(4): 435-442.

Goldstein, H., J. Rasbash, Plewis, I.Draper, D., Browne, W., Yang, M., and G. Woodhouse, G. and Healy, M. (1998). MLwiN, Multilevel Models Project. Institute of Education, London

Gray, J. and Wilcox, B. (1995) Good School, Bad School: Evaluating Performance and Encouraging Improvement. Open University Press.

Hedges, L.V. and Olkin, I. (1985) Statistical methods for meta-analysis. New York: Academic Press.

Kogan, M. (1999) The Ofsted System of School Inspection: An independent evaluation. A reportof the study by the Centre for the Evaluation of Public Policy and Practice and Helix Consulting Group, CEPPP, Brunel University.

Longford, N. T. (1988). "Variance component analysis: manual" Princeton Educational Testing Service.

Lightfoot, S.L. (1983) The Good High School: Portraits of Character and Culture, New York: Basic Books.

McCord, J. (1978) A thirty-year follow-up of treatment effects. American Psychologist, 33: 284-289.

McCord, J. (1981) Considerations of Some Effects of a Counseling Program. In S.E. Martin, L.B. Sechrest and R. Redner (Eds) New Directions in the Rehabilitation of Criminal Offenders. Washington DC: National Academy Press, 393-405.

McCord, J., Tremblay, R.E., Vitaro, F. and Desmarais-Gervais, L. (1994) Boys' disruptive behavior, school adjustment, and delinquency: The Montreal Prevention Experiment. International Journal of Behavioral Development, 17: 739-752.

Pearlman, K., Schmidt, F.L. and Hunter, J.E. (1980) Validity Generalization Results for Tests Used to Predict Job Proficiency and Training Success in Clerical Occupations. Journal of Applied Psychology, 65(4): 373-406.

R⇗der, 2000 Blind people process language faster than sighted people Neuropsychologica v.38 p 1482 (Reported in New Scientist, 14 Oct, p.18)

Sammons,P, Thomas, S., Mortimore, P., Owen, C. & H.Pennell (1994) Assessing School Effectiveness: Developing Measures to put School Performance in Context. London: Institute of Education report for Ofsted.

SCAA and Ofsted (1996a) Standards in Public Examinations 1975 to 1995 Hayes Middlesex : School Curriculum and Assessment Authority Publications:

Schmidt, F.L. and Hunter, J.E. (1977) Development of a General Solution to the Problem of Validity Generalization. Journal of Applied Psychology, 62(5): 529-540.

Schmidt, F.L. and Hunter, J.E. (1981) Employment Testing: Old Theories and New Research Findings. American Psychologist, 36(10): 1128-1137.

Schmidt, F.L. and Hunter, J.E. (1993) Tacit knowledge, practical intelligence, and job knowledge. Current Directions in Psychological Science, 2: 8-9.

Schweinhart, L.J. and Weikart, D.P. (1997) Lasting Differences: The High/Scope Preschool Curriculum Comparison Study Through Age 23. Ypsilanti MI: High/Scope Press.

Slavin, R.E. (1999) Rejoinder: Yes, Control Groups are Essential in Program Evaluation: A Response to Pogrow. Educational Researcher, 28(3): 36-38.

Somekh, B., Convery, A. Delaney, J. Fisher, R., Gray J., Gunn, S., Henworth, and A. Powell, L. (1999) *Improving College Effectiveness: raising quality and achievement*, Further Education Development Agency: London.

Tymms, P. B. (1996). The Value Added National Project Second Primary Technical Report: An analysis of the 1991 Key Stage 1 data linked to the 1995 KS2 data provided by Avon LEA. London, The School Curriculum and Assessment Authority.

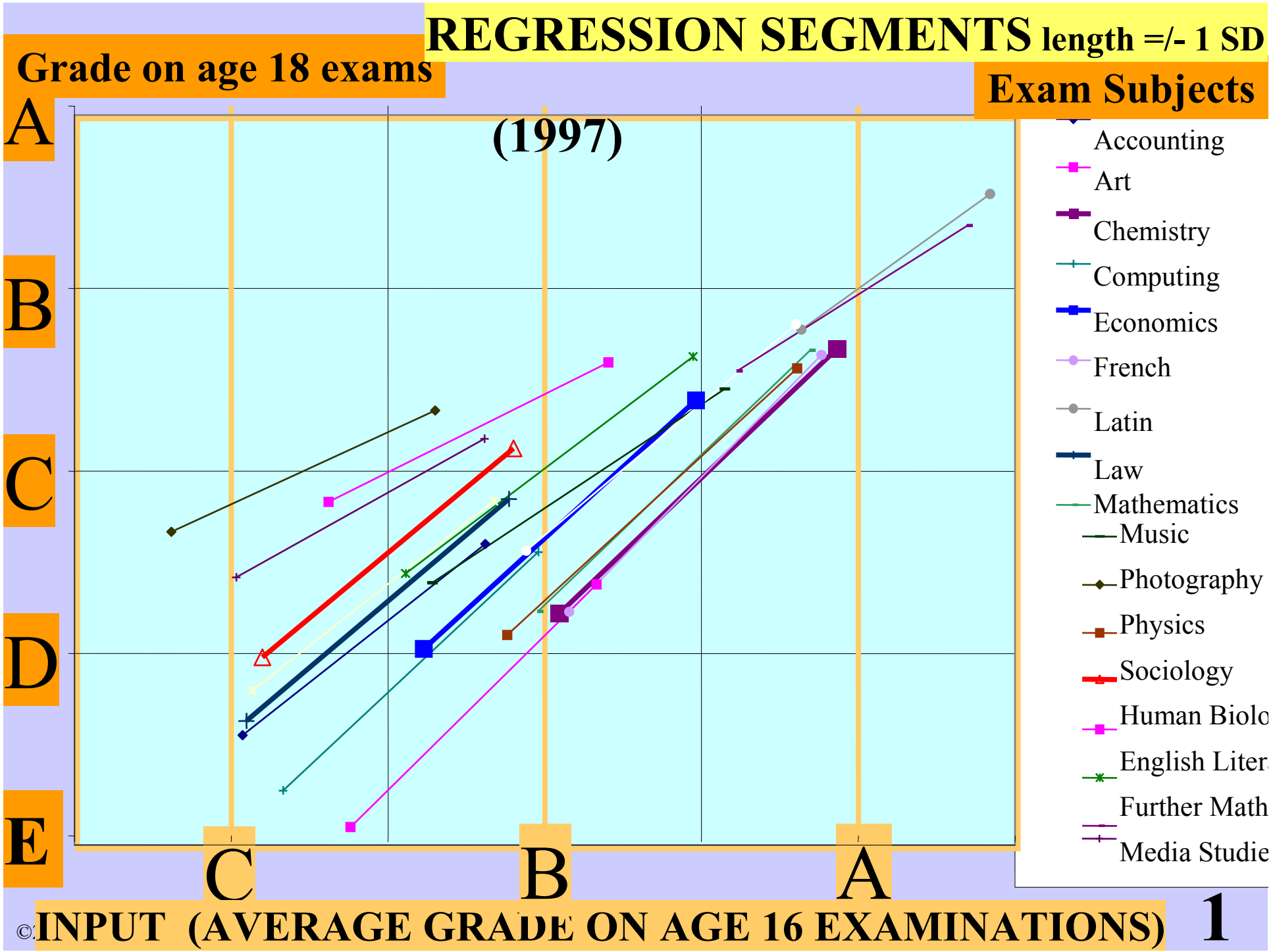
Tymms, P.B. (1997). The Value Added National Project Technical Report: Primary 4. London, School Curriculum and Assessment Authority.

Tymms, P (1999) Baseline Assessment and Monitoring in Primary Schools: Achievements, Attitudes and Value-added Indicators. 1999, London: David Fulton Publishers. 104.

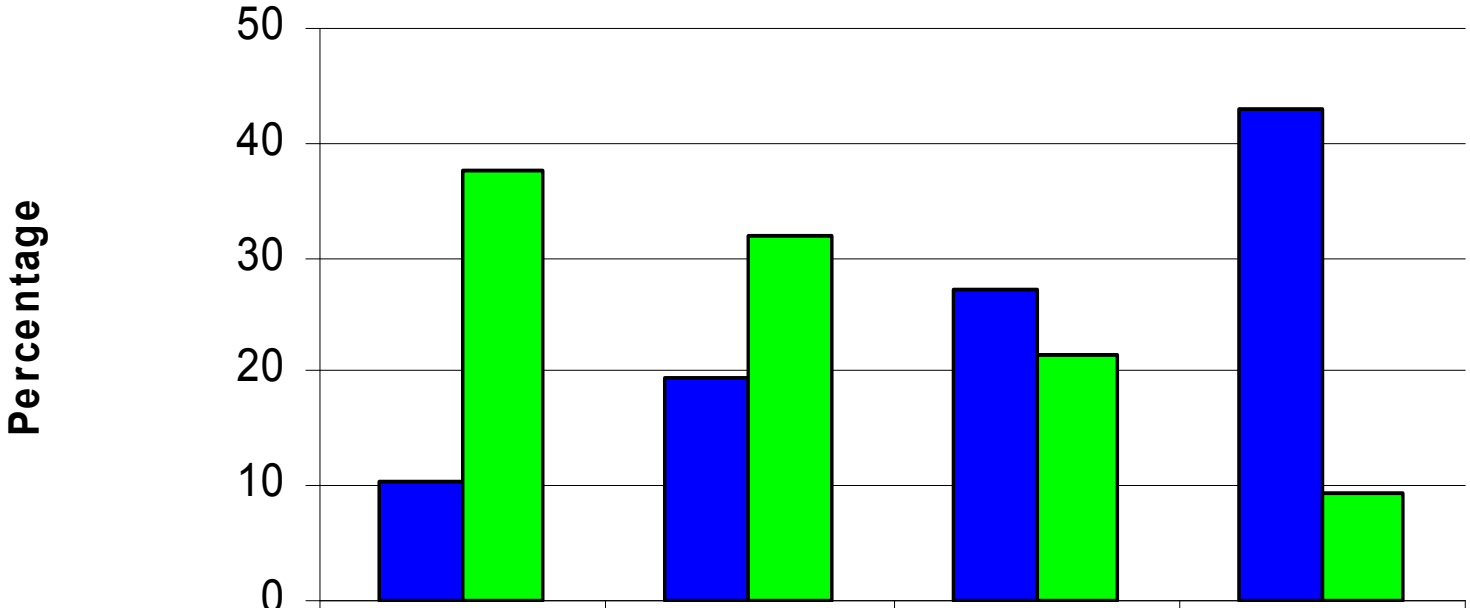
Tymms, P.B. and Fitz-Gibbon, C.T. (2001) Standards, Achievement and Educational Performance, 1976-2001: A Cause for Celebration? In R. Phillips and J. Furlong (Eds) Education, Reform and the State: Politics, Policy and Practice, 1976-2001, London: Routledge.

Vincent, L. (1997/6) The Value Added National Project: Analysis of Key Stage 3 Data for 1994 Matched to GCSE Data for 1996. London: School Curriculum and Assessment Authority.

REGRESSION SEGMENTS length =/- 1 SD



Percent of students in each quarter of GCSE results
Physics and Sociology A levels: 1999

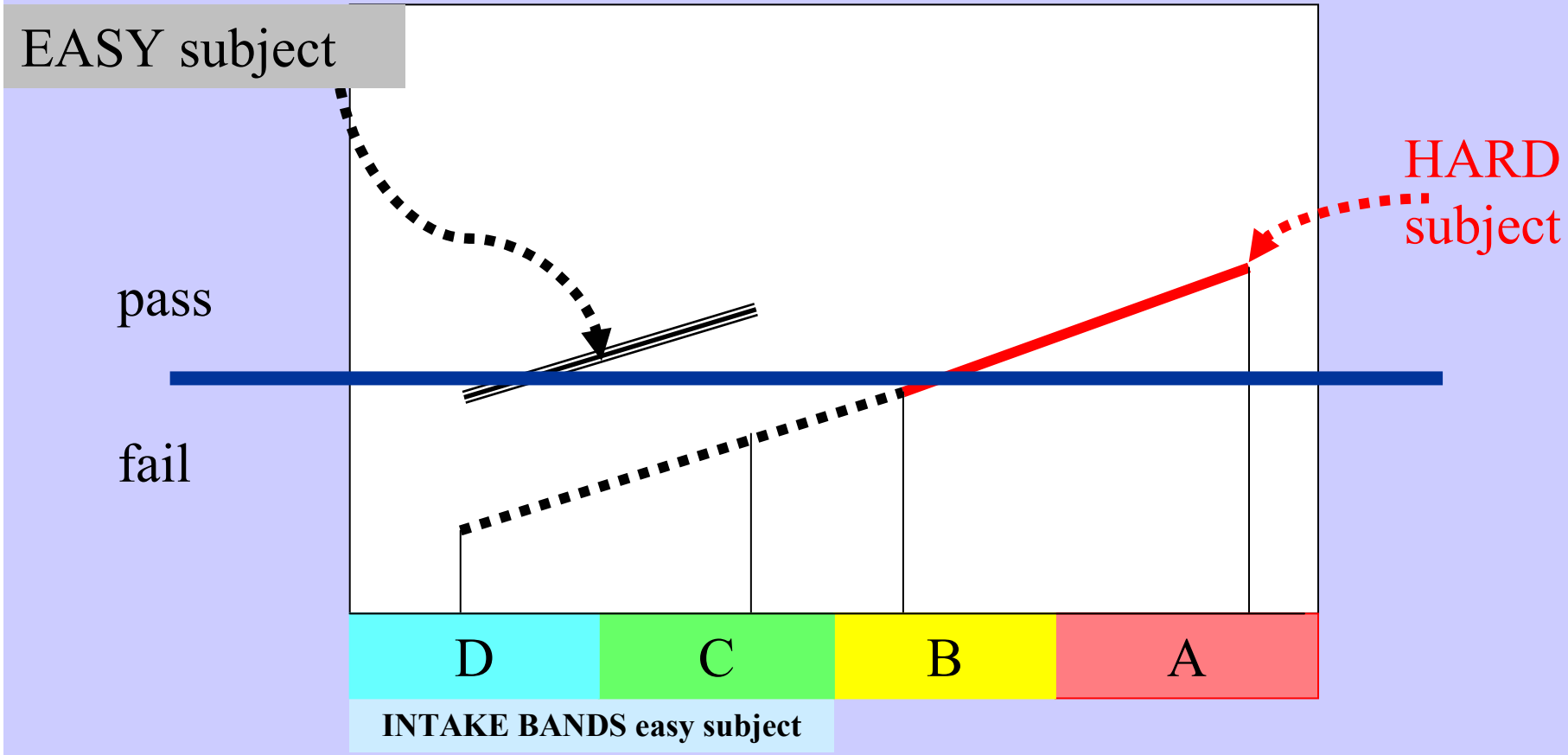


	Bottom quarter	Lower-mid quarter	Higher-mid quarter	Top quarter
Physics	11	19	27	43
Sociology	38	32	21	9

Average GCSE quarters of all students completing A levels

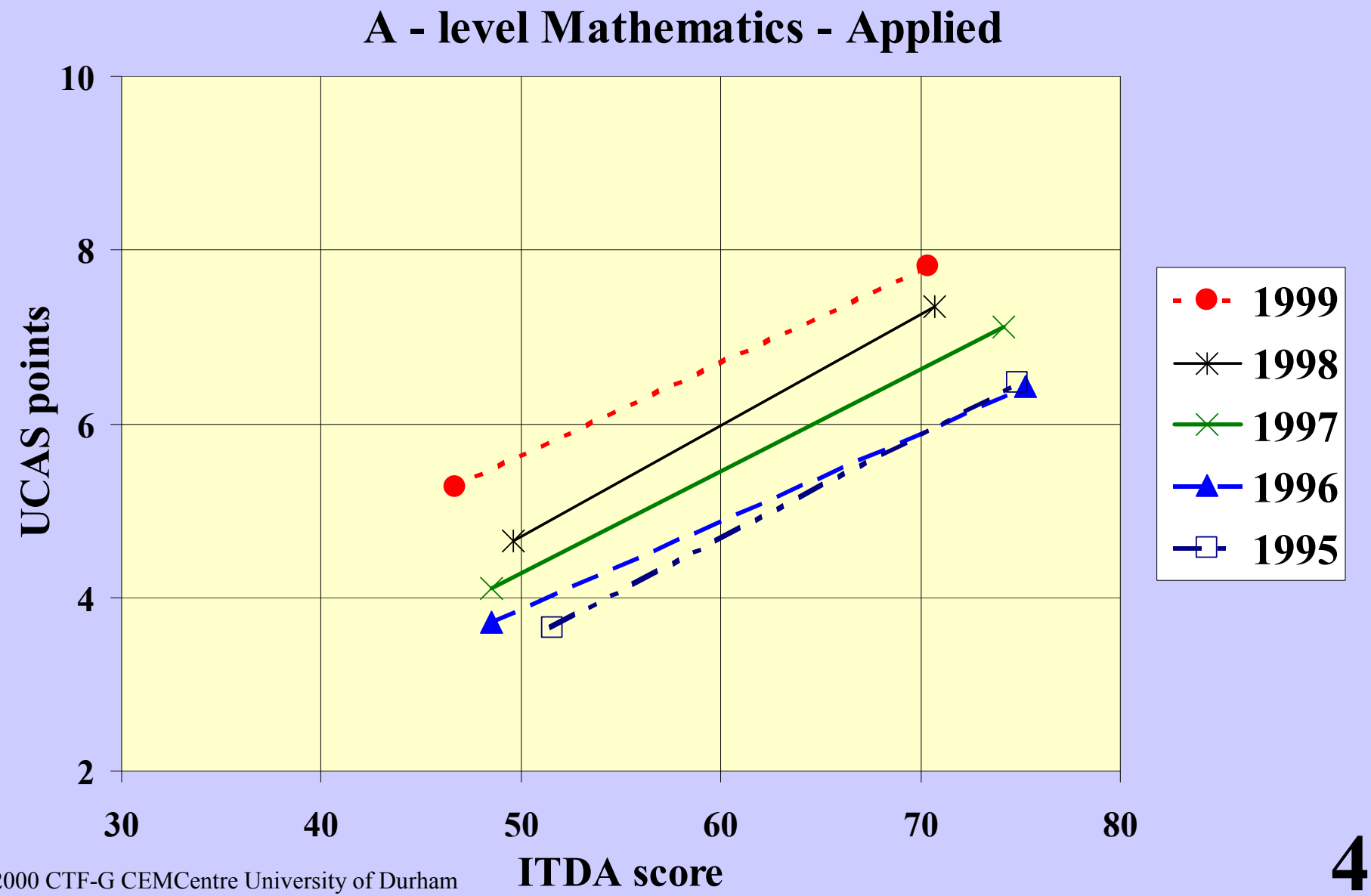
Source: ALIS Project.

Two Trend Segments not in line- - - different ‘standards’?



Similar fail rate even though intakes were different - - - otherwise there would be high failure in the subjects enrolling students that had Cs and Ds at age 16

REGRESSION SEGMENTS
CAN ALSO SHOW TRENDS OVER TIME



Advertisement in the Times Educational Supplement Nov. 1999

D*f*EE
Department for Education and Employment

COMMUNICATIONS STRATEGIST

Teachers – our future in their hands

Imagine you're at a Christmas party. A 45 year-old teacher finds out you work for the DfEE. Will you spend the rest of the evening arguing the case for pay reform?
Or will she tell you
teaching's great, she's getting a £2,000 pay rise
and would you like a top-up?
Well that depends on how good you are.

ie SPIN DOCTOR WANTED

What relates most to maths and English achievement scores?

Using a model that recognised the three level hierarchy of *pupils* taught within *classes* within *schools* suggested most of the variation was at the class level for Mathematics but not for English.

FROM VINCENT 1997

Proportion of variance at each level: Mathematics

Level	proportion of variance 'accounted for'
School	15%
Class	43%
Pupil	42%

Proportion of variance at each level: English

Level	proportion of variance 'accounted for'
School	12%
Class	34%
Pupil	54%

Multi-Level Modelling (mlm):

(special software; advanced statistics)

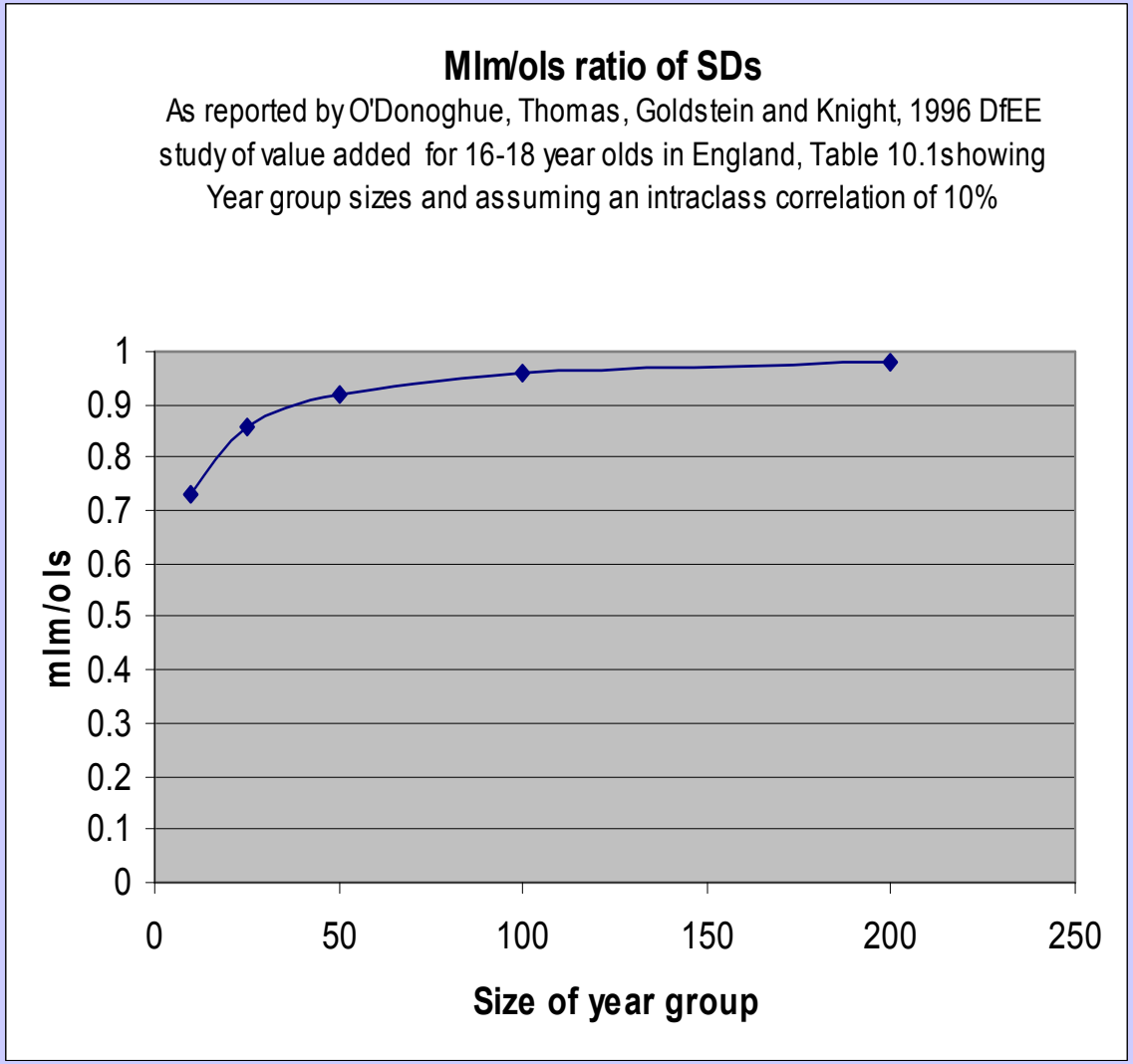
or Ordinary Least Squares (OLS)

(repeatable in Excel; GCSE-level maths)

EFFECT OF SAMPLE SIZE on error terms in mlm and ols analyses

UNIT SIZE (n)	Mlm/ols ratio of SDs
10	0.73
25	0.86
50	0.92
100	0.96
200	0.98

BUT … these differences will generally be less than the errors in guessing what level to call statistically significant……...



AND what is *substantively* (as opposed to statistically) significant……??????

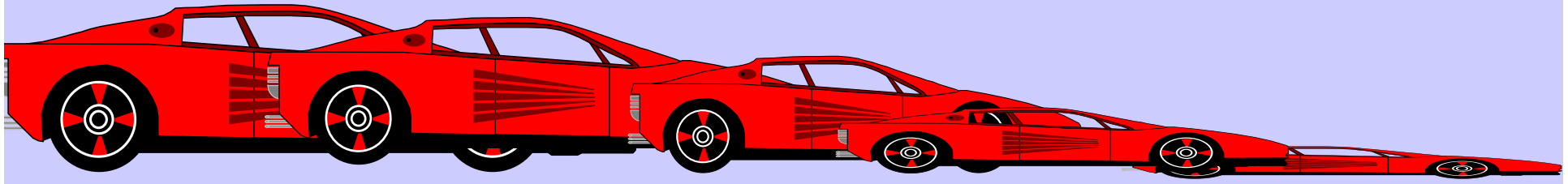
ONLY

experience

and experiments

can settle that $60,000 question

Cars driving from west to east.....



…will melt !!!!

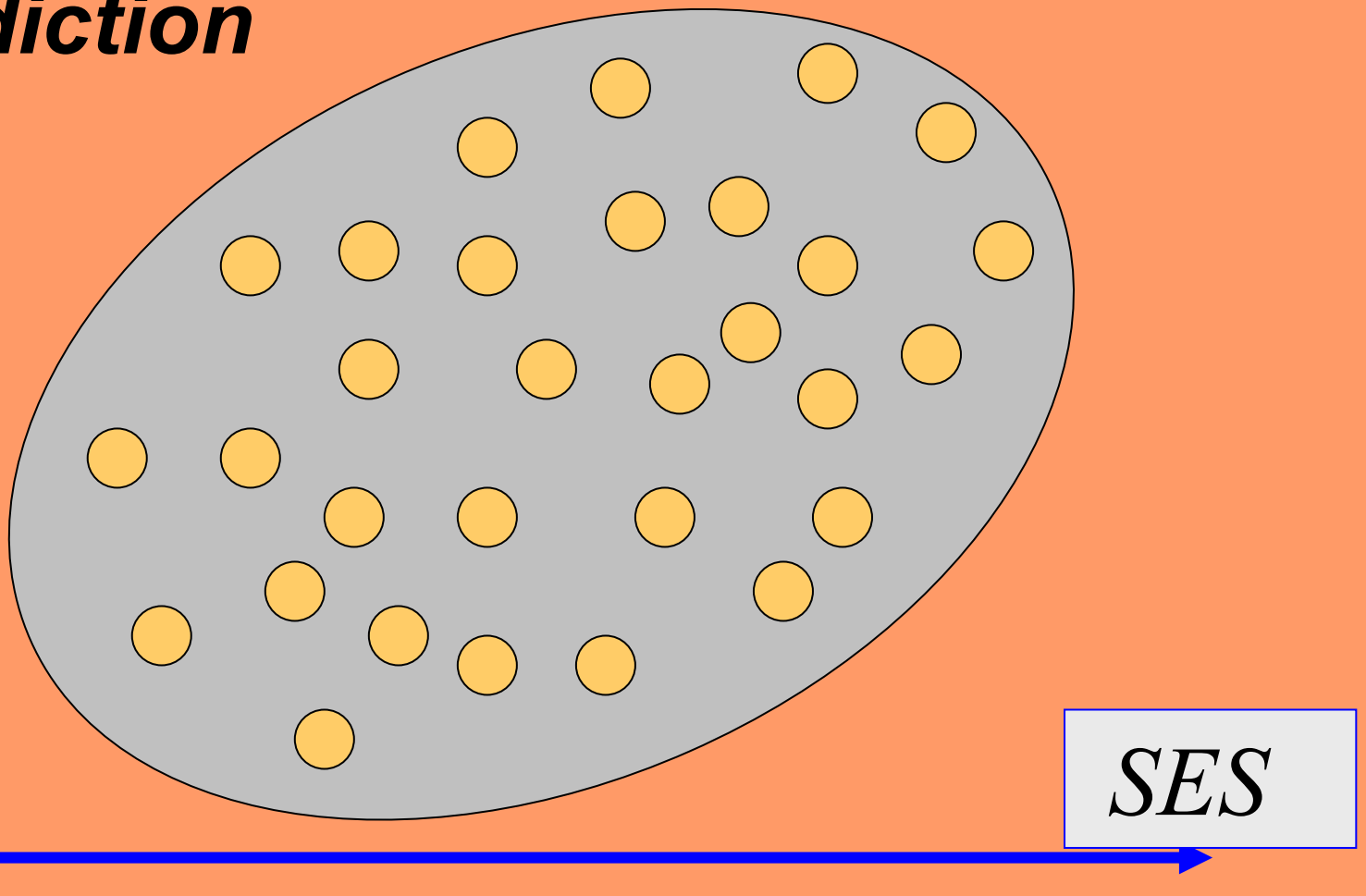
quantification:

Effect Sizes needed, as well as 'significance' levels

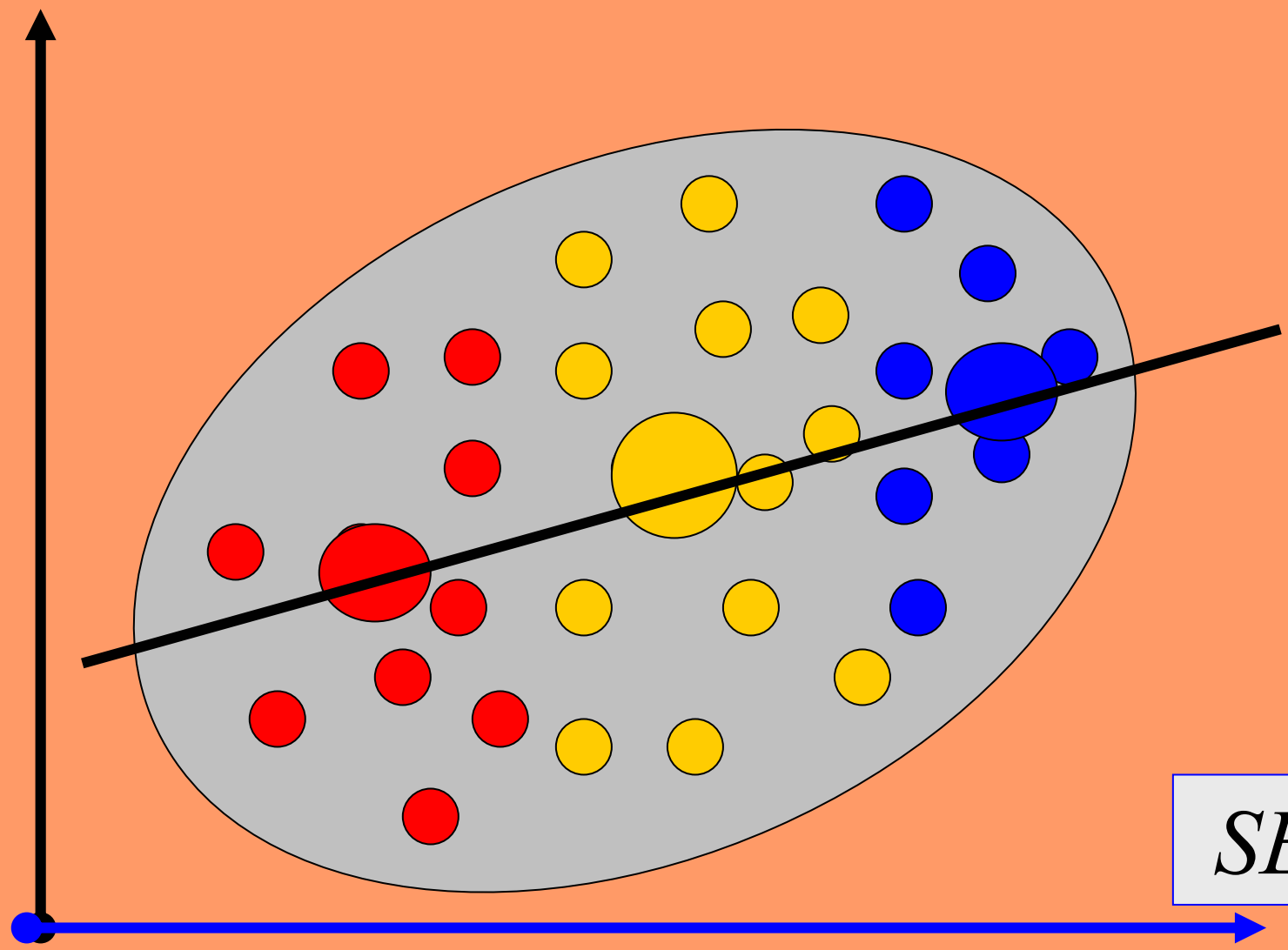
Two kinds of data

Indicators	Experiments
Value added	***Target getting***
Passive observations	Active interventions
EPIDEMIOLOGY	CLINICAL TRIALS

Weak correlation gives poor prediction

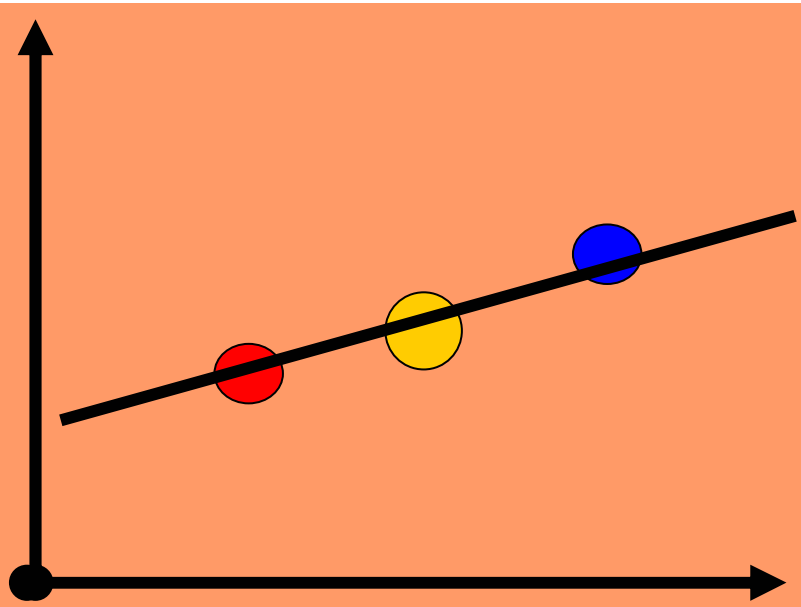


Weak correlation r = 0.3, say



Weak correlation at pupil level - - -

but large dots = school means…..



based on *aggregated* data ('means on means')

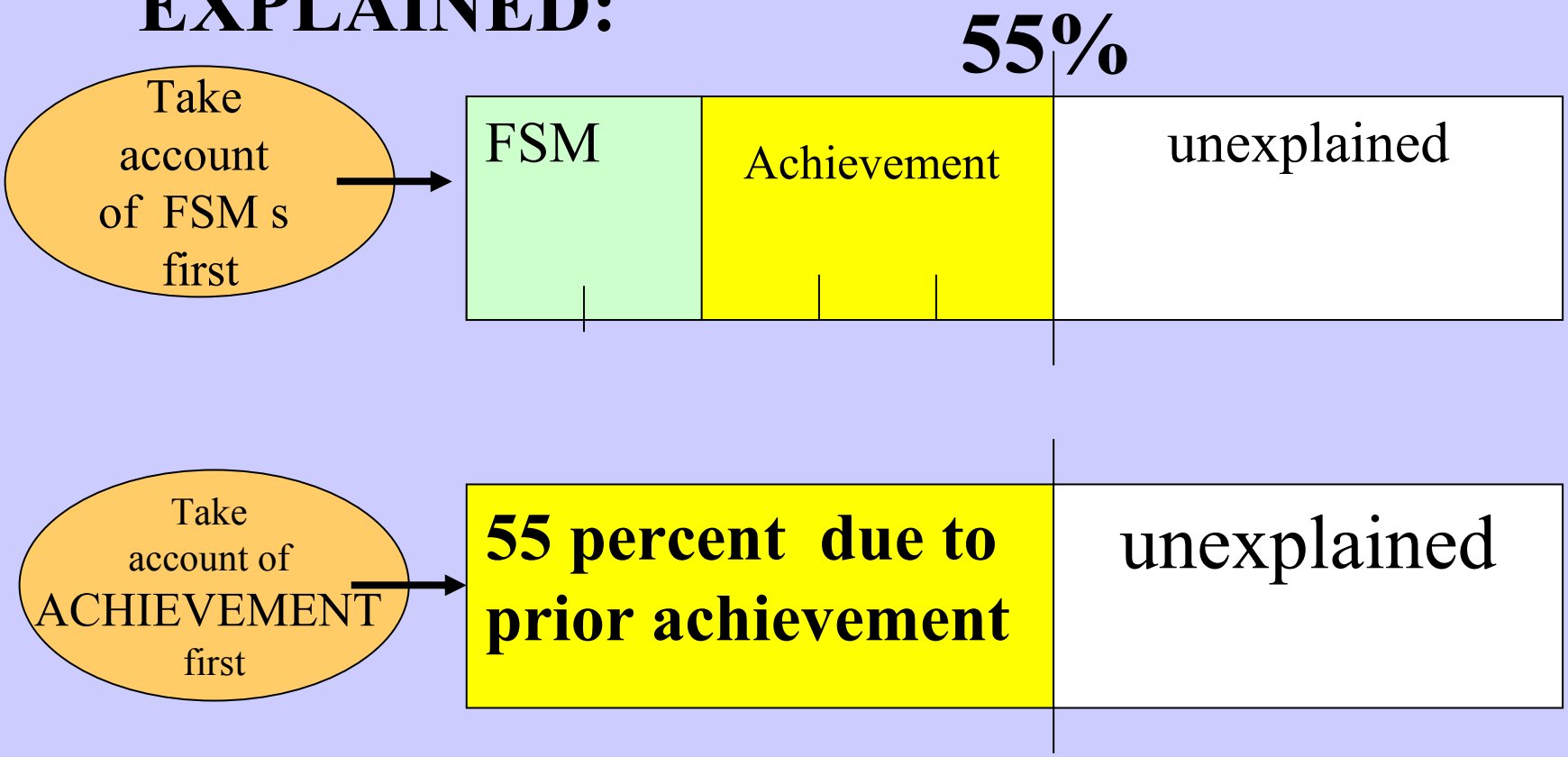
The correlation is strong!

BUT the correlation will be less when there is less segregation by social class e.g. in Scotland

[See chapter 16 in *Monitoring Education*]

Moreover, the danger of excuses must be recognised

PROPORTIONS OF SCHOOL- LEVEL VARIATION EXPLAINED:

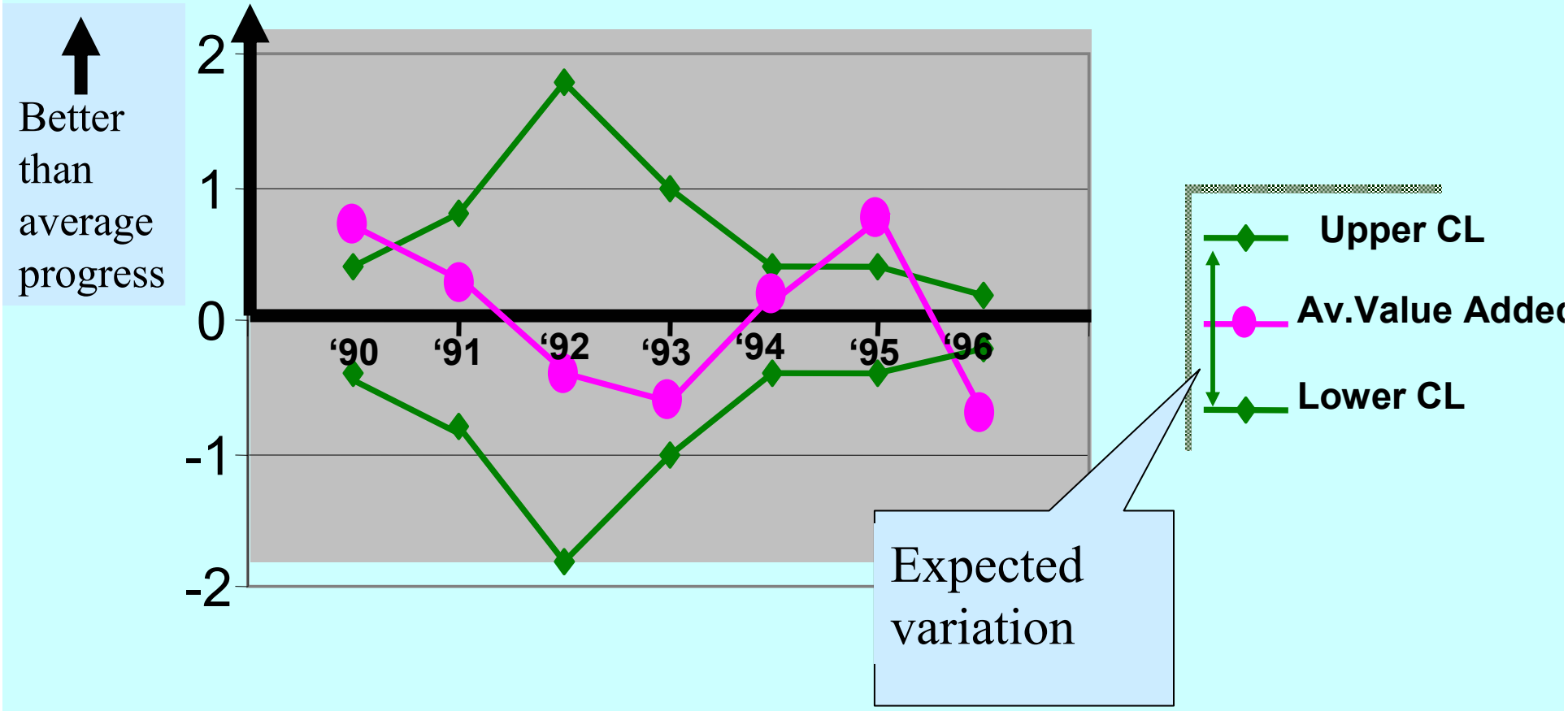


Drawn from data quoted in Goldstein, 1998 Oxford Review of Education, 24(4)

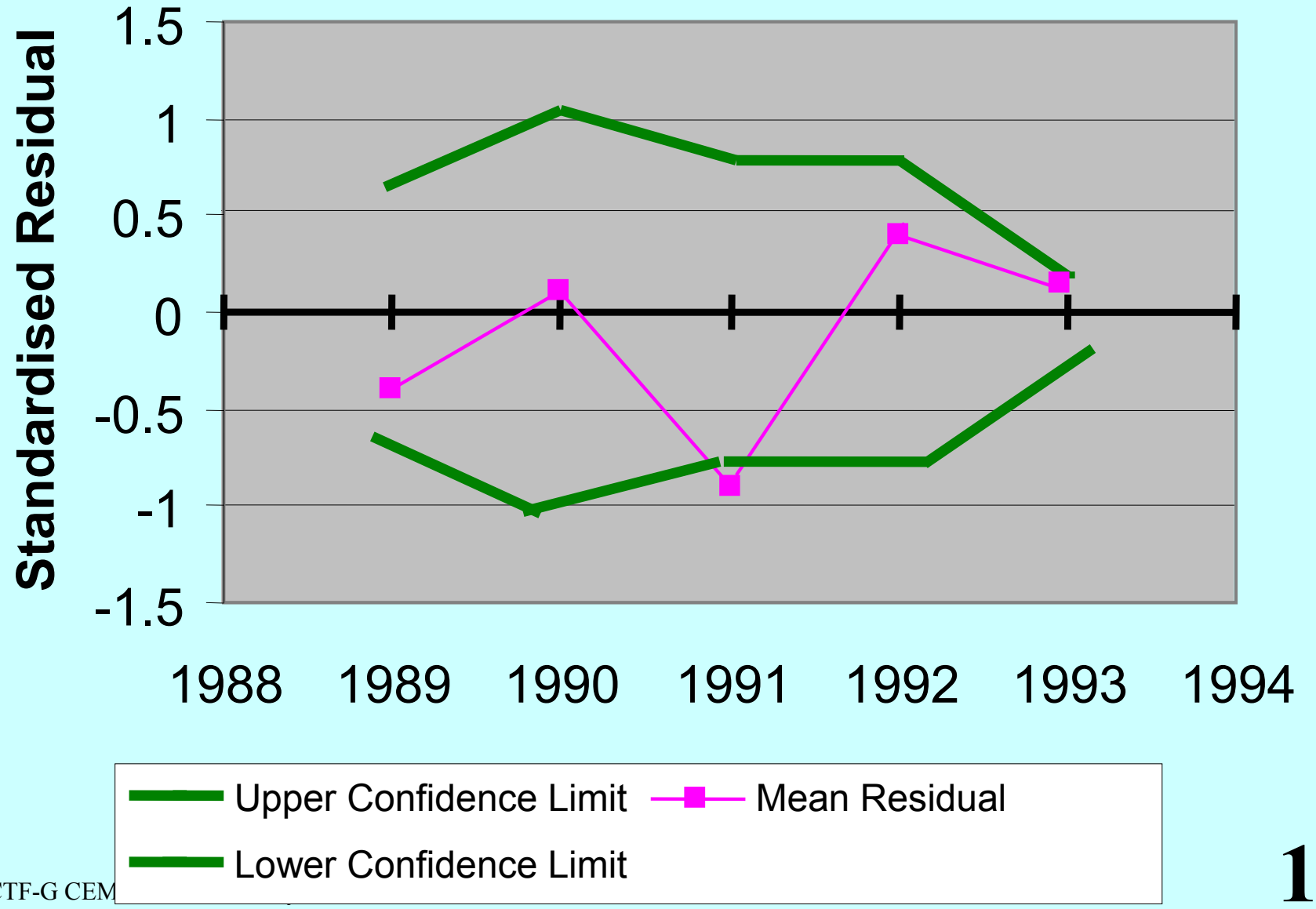
See Coe & F-G in same journal.

Statistical Process Control Charts for each subject

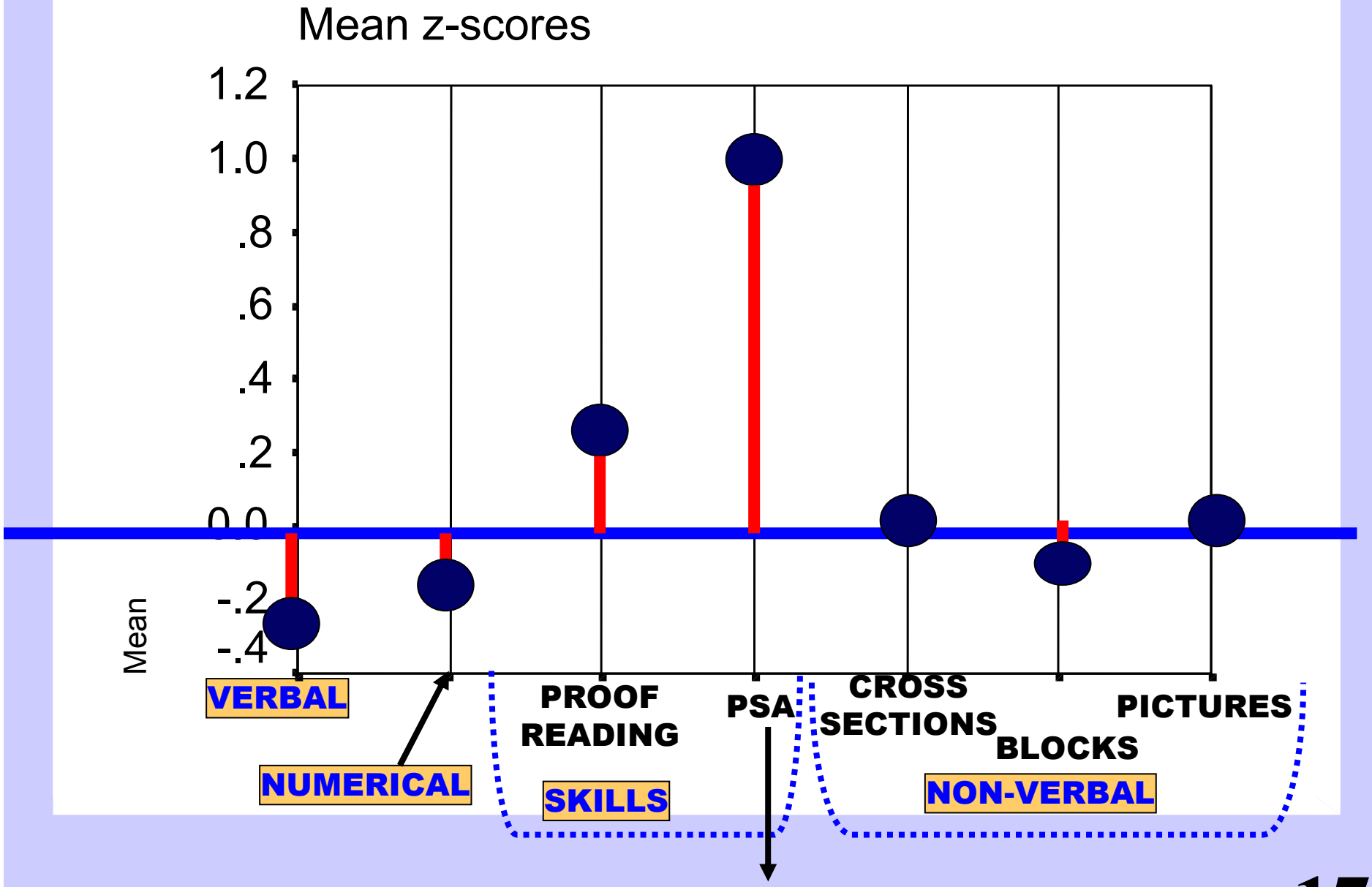
SPC chart: SUBJECT X, school J



CHEMISTRY A-level (real data)



Deaf students from yr7 2000 MIDYIS



Education

15,000 hours of compulsory treatment…

Let's evaluate it carefully … lest we do harm

LET'S CREATE

② DISTRIBUTED RESEARCH
② TEACHER-RESEARCHERS
② ED.PSYCHS. GIVING INSET