International Journal of Educational Research
25 (3) 1996
ISSN 0883-0355

CHAPTER 4

OFFICIAL INDICATOR SYSTEMS IN THE U.K.: EXAMINATIONS AND INSPECTIONS

C. T. FITZ-GIBBON

Curriculum, Evaluation and Management Centre, University of Durham, Durham DH1 1TA, U.K.

Abstract

The reason for developing indicators is to make it more likely that a high quality education is available in each and every school, for every child. Is this purpose being served by present systems in the U.K.? This chapter provides an update on how official and unofficial systems in the U.K. have been set up to ensure quality and how these systems are undergoing changes and adding a new component: measures of "Value Added". Copyright © 1996 Elsevier Science Ltd

Indicator Systems in the U.K.

In the United Kingdom (U.K.) there are two long-established systems which have provided indicators: the examination system and the inspection system. The examination system for 16- and 18-year-olds has undergone some significant changes in recent years, notably in the publication of results in the press, school by school, in tables generally referred to as "League Tables", but officially called School Performance Tables.

A recent major change is that additional examinations, called "Key Stage" assessments, are now almost in place for the ages of 7, 11 and 14 years and the intention is clearly that there should eventually be "value-added" measures made between each two points of assessment. Value added is the term which has come to be applied to the regression residuals which researchers have sometimes called school effectiveness indicators (e.g., Mandeville & Anderson, 1987). Further major changes have recently taken place in the inspection system with its redesign under new legislation as the Office for Standards in Education (OFSTED).

The functioning of each of these three systems — examinations, key stage assessments, and inspection — is illustrated with brief vignettes to highlight important features and give a flavor of the activity. Each vignette is followed by interpretative comments.

Vignette 1. The Summer Exams

Throughout May and June, as the fleeting British summer is finally in full swing, about 95% of 16-year-olds and 30% of 18-year-olds are into a cycle of "swotting"

and taking examinations. Growing numbers of students are now taking vocational courses but external "end tests" are not yet popular on these courses which emphasize modules and project work. The examinations at ages 16 and 18 years are what would be called in the United States "authentic, high stakes, curriculum-embedded tests". They are run by Examination "Boards" which are non-profit organizations originally associated with universities. The cost to the school, per student per subject examined, is of the order of £15 (~$10.00) for each 16-year-old and £25 (~$15) for each 18-year-old. Sixteen-year-olds generally take seven or eight subjects and 18-year-olds two or three subjects. The cost per year in a school of 1,000 students for these two external examinations might well be over £30,000.

Teachers enter their students for the examinations through one of six boards, making the choices on the basis of such considerations as the syllabus, costs, and the tradition of the school. Although examinations are set by different boards, the standards within each subject are expected to be consistent and generally are (e.g., Tymms & Vincent, 1994). This consistency is due to various quality assurance procedures implemented by the boards including collaboration on a regular basis, agreed schemes for the allocation of marks, and statistical moderation.

The examinations are generally completed by students in their own schools and colleges under regulation examination conditions regarding seating and procedures. Scripts are then sent to external markers. These are usually teachers earning extra income (about £2 per script for the age 16 examinations and £3 per script for advanced papers from 18-year-olds) and gaining valuable professional experience in their collaboration with the Examination Boards. Markers work to precise schemes and are supervised by a Chief Examiner. The results are collated, grade-boundaries are assigned for the marks, selected scripts are re-marked as one of several methods of "moderation", and results are published by mid-August. Students and teachers alike await the results with trepidation.

A consistent feature of the results is that some subjects are either more difficult or are more severely graded than other subjects (Fitz-Gibbon & Vincent, 1994). Mathematics, sciences, and foreign languages are consistently taken by more able students and yet the grades achieved are lower. The measured value-added indicators between 16 and 18 years of age are therefore less in these subjects. Since only three subjects are typically studied for the age-18 examinations ("A-levels") this tendency for lower grades in mathematics, sciences and foreign languages provides an incentive for schools to counsel students away from these difficult subjects. The incentive to do so is now particularly strong since (1) the examination results published in School Performance Tables are undifferentiated by subject, and (2) schools and colleges are now in direct competition with each other since 80% of the funds of most Local Education Authorities (LEA) are now devolved to schools on a *per-pupil basis*. This problem arises at a time when the nation is worried by possible shortages in the technical and scientific work force and when entry into Europe means foreign languages are needed.

Another perverse incentive arising from the publication of performance tables is that, because one indicator in the tables is the percentage of students obtaining A, B or C grades, schools are tempted to concentrate efforts on students likely to obtain a D, thereby violating the ethic of valuing each student equally.

Vignette 2. The Key Stage Assessments

In the summer of 1993, the tests which were to be given to all 14-year-olds in England were almost completely boycotted by teachers. A national newspaper reported:

> At 1pm last Monday, after months of trying to head off a teachers' boycott of the first national tests for 14-year-olds in English, mathematics and science, John Patten, the beleaguered education secretary, was finally forced to admit defeat. Out of 600,000 pupils only a few thousand pupils in a handful of schools put pen to paper. The tests were a flop . . . .
>
> How could it be that such a simple reform — one of the "big ideas" of the Thatcher era — was so comprehensively undermined? If successive Tory governments were able to defeat the miners, privatise public utilities and curb the excesses of left-wing local authorities, why had they failed to ensure children are taught to a national standard and tested on whether they can read and write? (*Sunday Times*, 13–6–93, p. 11).

How indeed? And how does this supposed opposition of the teaching profession to accountability square with our own quite contrary experience? At the time that the key stage tests were boycotted, schools were choosing to spend some of their limited income on obtaining detailed analyses of examination data, including valued-added measures pupil by pupil and subject by subject. Far from opposing accountability data, the teaching profession was buying into information systems which provided good evidence for effectiveness or lack of it.

Indeed, the A-level Information System (ALIS) had already been running for several years, having started in 1983. A wide range of outcomes is assessed by a questionnaire administered to students under examination-style conditions. Attitudes of the students towards the school and to each subject are assessed on scales. Aspirations are related to aptitudes to give residuals and comparative data. All data are fed back to schools under code names and in confidence with separate reports for each subject. Teaching and learning processes as perceived by the students are also analyzed (Fitz-Gibbon, 1985; 1991; 1996a). We have now developed other systems, including one for primary schools (Tymms & Gallacher, 1995). We are working with about 2,000 schools and colleges and many LEAs, and all our projects are actively supported by teachers' associations/unions. Other university-based researchers have worked largely with individual local authorities (Jesson, Jones, & Gray, 1986; Woodhouse & Goldstein, 1988; Sammons, 1995).

So what went wrong with the key stage assessments? The main expressed problems centered on the workload for teachers and the content of the tests. (The English teachers were implacably opposed to the test content and the test was even leaked one year so that it was rendered useless.) There were grave doubts concerning the quality of the data (Hutchison & Schagen, 1994). The primary school tests were called the Standard Assessment Tasks (SATs until the U.S. SATs were noticed when the term had to be dropped) but they were far from "standard". Much weight was placed on "teacher assessments", based on teachers' judgments even though such procedures could hardly be used for accountability.

There was little comment on the extraordinary notion, which was apparently widely accepted, that the tests would be administered and marked by teachers and the results published in an era of massive competition between schools with jobs at stake. Perhaps we are too English to say, "What would prevent cheating?" but we should note Deming's maxim: "Where there is fear there will be wrong figures" (1986, p. 266).

The troubles were partly due to the allure of criterion-referenced testing and "alternative

assessment". Debates on these topics have been ongoing and will no doubt continue (Glass, 1978; Wolf, 1991; Sanders & Horn, 1995).

Perhaps the motivation to have teachers conduct assessment was purely financial. However, good relationships between teachers and pupils are not promoted by having teachers change their role from coach to umpire. Whatever the merits and demerits, the key stage assessments are now to have an externally marked component. A thorough account by Black (1994) conveys the trauma and false starts which have ensued from the attempt to set up testing systems for ages 7, 11 and 14 years.

Vignette 3. The OFSTED Inspection
(An English phenomenon, not the same in Wales, Scotland or Northern Ireland)

> In November 1993, an inner city comprehensive school was given a good report following a week's full-scale OFSTED visit by 13 inspectors, who were provided with 20 kilograms of documentation. In September 1994, a smaller team of 5 inspectors returned and declared the school to be "failing". One comment was that the school had done nothing about the underachievement of boys on the General Certificate of Secondary Education (GCSE). The inspectors were probably ignorant of the fact that attempts to find significant heterogeneity in regression slopes by sex have generally failed, thus suggesting that the gender gap in achievement is a consistent feature in all schools and colleges, and is not, therefore, something for which individual institutions should be blamed (Tymms, 1992).
>
> The LEA and the school tried to stop the publication of the second report which they hotly contested but the court upheld OFSTED's right to publish. The LEA and the school wished to press a legal suit but were advised by legal experts that as long as OFSTED had followed their published procedures their judgments could not be challenged.

The Office for Standards in Education was set up to inspect schools and, where necessary, label them as "failing". The previous system of inspection had employed full-time professionals, "Her Majesty's Inspectors" (HMI), who were selected for excellence in their professional experience and underwent a year of training. They were primarily responsible for reporting on the general state of schools to the central government. In contrast OFSTED is contracted-out, part-time work. Teams consist of persons with about one week's training. Reports on schools are written within a week or two of the visit and made publicly available. The method of inspection is spelled out in a thick document called the "Framework" which is currently under revision. The school is informed of the forthcoming inspection several months in advance and is requested to produce large amounts of documentation.

OFSTED accepted a poisoned chalice: the legislation creating it required that inspectors attempt to assess by inspection that which examinations assess better. A prime focus of the inspection is on standards of achievement — absolute and relative. But how can a visit assay an answer to such a question? If you wish to know relative standards you need to know pupil "capabilities". Can looking at or talking to a few pupils or sitting in on over-prepared lessons check on standards in the school?

The newspaper coverage of inspection reports makes the OFSTED experience a modern equivalent of being placed in the public stocks. The effect on school staff, parents and pupils of being labeled "failing" can be devastating. Public humiliation would be poor personnel practice

even if the judgments were correct, but is even less excusable in the absence of any study of the reliability or the validity of inspectors' judgments. An inspection should be particularly reliable and valid since its outcome can be a determinant of the careers of teachers and principals.

In terms of effective management and value for money, the quality system represented by OFSTED is in grave trouble. Furthermore, in terms of social science the lack of a justified or adequate methodology is a denigration of carefully developed procedures to collect and analyze judgments fairly and credibly. OFSTED's very existence signals a major failure in the education system: a failure to create politicians who understand the standards of social science methodology.

Recent Changes in the Examination Framework in the U.K.

Probably in response to the alienation of the teaching profession which had led to the near-collapse of the key stage testing system, Sir Ron Dearing, Chairman of the School Curriculum and Assessment Authority (SCAA), was invited to review the National Curriculum and testing. He recommended a period of stability, a reduction in the amount of content, and the investigation of value-added measures. A working party on Value Added was set up by the SCAA, the government agency with responsibility for syllabuses, examinations, key stage assessments, quality control, statistics and all matters indicated by its title. The working party made a number of recommendations, the first of which was:

> . . . the use, *wherever there is sufficient statistical validity and reliability* [italics added], of simple models for value-added performance indicators using data aggregated to the school level (SCAA, 1994, p. 49).

(A common phenomenon is illustrated in this quotation: worries are routinely expressed about the adequacy of *numerical* indicators without any similar concern about the accuracy of *verbal* judgments, such as those made in OFSTED inspections.)

The Curriculum, Evaluation and Management Centre won the resulting contract and has undertaken initial trials (Fitz-Gibbon, 1996a; Tymms & Henderson, 1996; Trower & Vincent, 1996). As we work on the design of a national value-added system it is clear that numerous technical and practical issues need to be confronted:

1. Is the school the desirable unit of analysis?
2. Is there meaningful heterogeneity of the slopes of the regression lines (in which case any single indicator would be inadequate)?
3. If there is statistically significant slope heterogeneity, what is the extent of the *substantive* significance, (i.e., the magnitude)? Further, if slope heterogeneity exists, is it stable from year to year and/or associated with any observable policies or practices in schools?
4. Is multilevel modeling needed or does it yield essentially the same results as ordinary least squares applied to pupil-level data?
5. Is Bayesian shrinkage appropriate when it might be unlikely to be acceptable or accessible to staff in schools (Fitz-Gibbon, 1991, p. 79)?
6. Are the "school averages" (i.e., means on means analyses) an acceptable way forward for the sake of efficiency?
7. Do curves fit better than a straight line and would other techniques (such as smoothing) yield different results?
8. Should different regression equations be developed across different regions of the intake data?

9. Could modeling be replaced by simple retrospective comparisons of like with like (e.g., with resampling methods used to estimate uncertainty)?
10. Should compositional effects (e.g., the average achievement of a school) be included in models?
11. Should sex be included as an "explanatory" variable? Likewise for socioeconomic status — or is it like an "excuse"? See Willms (1992) on Type A and Type B effects.
12. Are some areas of the curriculum particularly sensitive to instructional and school effects? Should there therefore be differential accountability?
13. How do we move from the arbitrariness of significance levels to a sense of what sizes of differences matter?

An approach suggested in our first report was built on a recognition that, in working with assessment systems:

> We are not dealing with a natural phenomenon which can be modeled, once and for all time, if only its laws are understood . . . . Given this *inherent* uncertainty as to the nature of the assessment data which might arise from year to year, one reasonable system might be a two stage approach:
>
> FIRST STAGE: Statistically representative samples of entry and exit measures are analyzed using rapid and readily understandable procedures. The resulting models are made available to schools for their internal use.
>
> SECOND STAGE: National datasets are analyzed using models at the level of complexity demanded by the data that year. Particular attention is paid to unusual schools to ensure that explanations are explored before judgments might be made. Public reporting would follow the completion of the second stage and may be based on three year averages. (Fitz-Gibbon, 1996b, p. 11).

In other words, the strategy is to provide clear and simple methods for schools' own investigations followed each year by sophisticated analyses which will precede any publication of results.

Scotland and Northern Ireland

It has already been noted that OFSTED is an English phenomenon. Scotland, Wales, and Northern Ireland have different independent inspectorates and slightly different examination and testing systems. The Welsh system is much like the English but there are important differences in Scotland and Northern Ireland.

Scotland

Scotland has led the world in indicator systems. Not only is the Centre for Educational Sociology the source of much research on school effectiveness (Willms, 1986; McPherson & Willms, 1987; Willms & Kerckhoff, 1995; McPherson, 1992) but Scotland must have been the first nation to provide indicators for *every department* in every school and for three consecutive years. How did they do this?

In Scotland, almost all students take "Standard Grade" at about age 16 and many take "Highers" a year later. Moreover, these external examinations are administered by a single Board, the Scottish Examinations Board. Yet, because of the problems in matching the dat

student by student, value-added measures were not produced. Instead the Scottish Examinations Board created an alternative procedure for giving schools feedback on the performance of their departments: relative ratings.

Relative ratings require more explanation than regression approaches because the technique has not been widely used. Following a paper by Kelly (1976) which contained an appendix by Lawley, a method was developed of re-scaling the examination results to obtain "correction factors" which take account of the general level of achievement of those students who take a particular subject *and* the "difficulty" of the subject as evidenced by the average grade. This requires solutions of matrix equations. The same procedure is then used within each school to compare the results achieved in the various subjects. The relative rating is the difference between the two correction terms. Relative ratings in a school will sum to zero — for every winner there will be a losing department — relatively. Thus relative ratings are different from valued-added indicators; however, they tend to correlate about 0.70.

In short, whenever there are no good prior predictors which would allow a value-added approach to be used, relative ratings can be particularly valuable. The impressive achievement of the Scottish Office Education Department was to institute these procedures very economically, working with the Scottish Examinations Board, and then to provide the data for three years in "Standard Tables" for use throughout Scotland, years before anyone else had officially monitored *schools*, let alone *departments within schools*.

Northern Ireland

The existence of the examination system in Northern Ireland has provided researchers with excellent outcome indicators for schooling, a feature well exploited by a series of important studies (Kellaghan & Madaus, 1979; Daly, 1991, 1995). As in Scotland there is a single examination board although schools are free to use mainland boards if they wish. The indicators published by the Northern Ireland Department of Education are essentially of the league table type. They differ from the English tables in using the year-in-school group rather than age group as the unit of reporting.

Conclusions: Issues to be Addressed by National Indicator Systems

By way of summary the general, not technical, issues which any national indicator system has to address are listed below. For each issue the current solution in the U.K. is summarized.

1. *The content to be taught to each age group.* There is now a National Curriculum managed by SCAA and a rationalization of vocational qualifications managed by the National Council for Vocational Qualifications (NCVQ).
2. *The nature of the assessment with particular concern for the impact of the assessment on teaching and learning (the backwash effect) and costs.* Examination Boards provide authentic, high stakes, curriculum-embedded tests taken for decades by school leavers. Recently testing has been extended in coverage to key stage assessments (now recovering from early problems).
3. *The uses to which the assessment results are put.* A major issue is that of confidentiality versus open information systems. At present in England we have School Performance

Tables in the national and local press every year and Value Added coming on stream. Meanwhile unofficial systems provide schools with much more than examination analyses, and are keenly supported by the teaching profession.

4. *The impact of indicators on behavior*. The school performance tables have a powerful impact because they are official and published. Their positive impact may be a re-focusing on achievement but examples were given of negative impacts. These could be simply remediated by use of more appropriate indicators.
5. *Non-cognitive outcomes*. Official indicator systems contain only a few non-cognitive measures, largely derived from management information on attendance, enrollments and destinations. As already mentioned, some unofficial indicator systems contain hundreds of indicators of other outcomes of schooling, and also *process* indicators to generate ideas for improvement. The unofficial indicator systems tend to be developed in universities, use specially collected data, maintain confidentiality, and are led by the needs of schools. Their impact needs to be studied.
6. *Accountability — How is accountability most effectively "enforced"?* By the LEAs? Inspection? The media? The market? But to whom are these organizations accountable? How are they evaluated? Is parliament enough? Does methodology matter? Is research sufficiently reliable? Can monitoring-with-feedback ensure improvement (Fitz-Gibbon, 1996a)?

The academic, human, professional, and scientific values associated with indicator systems need to be kept under constant review and one critical necessity would seem to be the use of adequate methodology.

References

Black, P. J. (1994). Performance assessment and accountability: The experience in England and Wales. *Educational Evaluation and Policy Analysis, 16,* 191–203.

Daly, P. G. (1991). How large are secondary school effects in Northern Ireland? *School Effectiveness and School Improvement, 2,* 305–323.

Daly, P. (1995). Science course participation and science achievement in single sex and co-educational schools. *Evaluation and Research in Education, 9,* 91–98.

Deming, W. E. (1986). *Out of the crisis: Quality productivity and competitive position*. Cambridge: Cambridge University Press.

Fitz-Gibbon, C. T. (1985). A-level results in comprehensive schools: The Combse project, year 1. *Oxford Review of Education, 11,* 43–58.

Fitz-Gibbon, C. T. (1991). Multilevel modelling in an indicator system. In S. W. Raudenbush & J. D. Willms (Eds), *Schools, pupils and classrooms: International studies of schooling from a multilevel perspective* (pp. 67–83). London: Academic Press.

Fitz-Gibbon, C. T. (1996a). *Monitoring education indicators, quality and effectiveness*. London: Cassell.

Fitz-Gibbon, C. T. (1996b). *Issues to be considered in the design of a national value added system*. A CEM Centre report for the School Curriculum and Assessment Authority, London.

Fitz-Gibbon, C. T., & Vincent, L. S. (1994). *Candidates' performance in science and mathematics at A-level*. Report for the School Curriculum and Assessment Authority (SCAA). London: SCAA.

Glass, G. V. (1978). Standards and criteria. *Journal of Educational Measurement, 15,* 239–261.

Hutchison, D., & Schagen, I. (Eds) (1994). *How reliable is national curriculum assessment?* Slough: National Foundation for Educational Research.

Jesson, D., Jones, B., & Gray, J. D. (1986). The search for a fairer way of comparing schools' exam results. *Research Papers in Education, 1,* 91–122.

Kellaghan, T., & Madaus, G. F. (1979). Within school variance in achievement: School effect or error? *Studies in Educational Evaluation, 5,* 101–107.

Kelly, A. (1976). A study of the comparability of external examinations in different subjects. *Research in Education, 16,* 50–63.

Mandeville, G. K., & Anderson, L. W. (1987). The stability of school effectiveness indices across grade levels and subject areas. *Journal of Educational Measurement, 24,* 203–216.
McPherson, A. F. (1992). *Measuring added value in schools.* NCE Briefing no. 1. London: National Commission on Education.
McPherson, A. F., & Willms, J. D. (1987). Equalisation and improvement: Some effects of comprehensive reorganisation in Scotland. *Sociology, 21.* 509–539.
Sammons, P. (1995). Gender, ethnic and socio-economic differences in attainment and progress: A longitudinal analysis of student achievement over 9 years. *British Educational Research Journal, 21,* 465–486.
Sanders, W. L., & Horn, S. P. (1995). *An overview of the Tennessee value-added assessment system (TVAAS).* Knoxville, TN: The University of Tennessee.
School Curriculum and Assessment Authority. (1994). *Value-added performance indicators for schools.* London: Author.
Trower, P., & Vincent, L. S. (1996). *Secondary technical report: Valued added national project.* A CEM Centre report for the School Curriculum and Assessment Authority, London.
Tymms, P. B. (1992). Accountability — Can it be fair? *Oxford Review of Education, 19,* 291–299.
Tymms, P., & Gallacher, S. (1995). Primary science: An exploration of differential classroom success. *Research in Science and Technological Education, 13,* 155–162.
Tymms, P. B., & Henderson, B. (1996). *Primary technical report: Value added national project.* A CEM Centre report for the School Curriculum and Assessment Authority, London.
Tymms, P. B., & Vincent, L. S. (1994). *The comparability of examination boards at A-level: A report for the GCE boards.* CEM Centre, University of Newcastle Upon Tyne.
Willms, J. D. (1986). Social class segregation and its relation to students' examination results in Scotland. *American Sociological Review, 6,* 289–306.
Willms, J. D. (1992). *Monitoring school performance: A guide for educators.* Lewes, Sussex: The Falmer Press.
Willms, J. D., & Kerckhoff, A. C. (1995). The challenge of developing new educational indicators. *Educational Evaluation and Policy Analysis, 17,* 113–131.
Wolf, A. (1991). Assessing core skills: Wisdom or wild goose chase? *Cambridge Journal of Education, 21,* 189–201.
Woodhouse, G., & Goldstein, H. (1988). Educational performance indicators and LEA league tables. *Oxford Review of Education, 14,* 301–320.

Biography

Carol Taylor Fitz-Gibbon is Professor of Education at the University of Durham and Director of the Curriculum, Evaluation and Management Centre which has the contract to pilot official Value Added systems for England. She created ALIS, a comprehensive, unofficial monitoring system in 1983 which now analyzes approximately a third of A-level examinations in England, Wales, and Northern Ireland.