



**Curriculum, Evaluation and Management Centre**

PUBLICATION NO.23

EVALUATION, MONITORING AND  
SCHOOL IMPROVEMENT

C. T. FITZ-GIBBON

1993

# EVALUATION, MONITORING AND SCHOOL IMPROVEMENT

Carol Fitz-Gibbon

*School of Education, University of Newcastle Upon Tyne, NE1 7RU*

**Abstract** The advent of TVEI gave a tremendous boost to programme evaluation in the UK and brought teachers and evaluators together in professional meetings to an unprecedented degree. What has been its legacy? Combined with the rapid growth in computing power there has been a growth in the notion of using performance indicators (i.e. monitoring key features of the system at regular intervals.) TVEI has assisted in the growth of monitoring systems, not least as a result of the valued professional networks it established. A National Audit Office report seems designed to prompt moves towards the wider use of monitoring systems. Such moves may change the nature of evaluation.

There are two main reasons for this paper: to reflect, from the perspective of a distance of several years, on the nature of the TVEI evaluations and, secondly, to suggest that evaluation and evaluators must accommodate to the likely growth of 'monitoring' systems. Monitoring systems, by providing regular reports on the status of various 'Performance Indicators', change the information base considerably and this must have a substantial impact on the practice of evaluation.

## **The Impact of TVEI on 'Evaluation'.**

If you look at old textbooks you find that the word 'evaluation' usually referred to the evaluation of *pupils* by educational psychologists, to see if they needed special placement or in the context of individual case conferences (e.g. Thorndike & Hagen, 1969). A change in the meaning of 'evaluation' appears to have taken place in the US during the years when Lyndon Johnson was president. There were funds then, for some reason, not only for the Vietnam war and the space race but also for 'The Great Society'. Johnson, himself a former teacher, wanted to use Federal money to improve education, especially for 'the disadvantaged'. His strategies for getting legislation through congress, documented by McLaughlin (1975), included winning the support of 'Bobby' Kennedy. Kennedy did not trust the educational establishment and therefore wanted some form of accountability, particularly so that parents could have a voice. The education department in each State was consequently required to obtain 'evaluation reports' from all projects aimed at improving the education of disadvantaged children. No one was quite sure what an evaluation report looked like and the early ones were diverse indeed, some being a few pages of basic information about the project, some prolix descriptions and others tables of standardised

pre-test and post-test scores, usually from the project group only; rarely from a comparison group and never from a randomised control group (Alkin, Kosekoff, Fitz-Gibbon & Seligman, 1974).

The distinction between formative and summative evaluation was not widely recognised at that time and whilst there was evidence that the evaluators' activities were largely formative (being helpful friends of the project) the evaluation reports were written as if they were summative (judging the project) although, methodologically, the design of the data collection was not such as to enable strong inferences to be made about the value of the projects. There were, essentially, no adequate control or comparison groups.

These close links between projects and their evaluators were almost inevitable given that the projects were to hire their own evaluators. For formative evaluation this was fine but for summative evaluation this closeness might have been problematic. The closeness with which evaluators and Project Directors worked was illustrated by the fact that in a study of evaluation of bilingual projects (Alkin *et al.*, 1974) as many as 36% of the Project Directors reported assisting the evaluators by actually writing parts of the evaluation reports. Furthermore 90% (38 out of the 42 Project Directors) indicated that evaluation activities helped to guide the project. Clearly the role adopted by evaluators was formative not summative. Evaluators thus tended to join the professional team managing the project. Such closeness made it highly unlikely that any negative reports would be produced, and none were. Yet surely not all projects were flawless successes.

You can imagine that anyone working in that situation in the States and then joining the TVEI evaluation activities in the UK would have a sense of *déjà vu*. Again the situation had the following features:

- projects hired their own evaluators;
- evaluation reports were almost uniformly positive;
- much evaluation was formative but reports were prepared with a summative flavour and problems encountered were rarely mentioned;
- reports often focused on the project alone...comparisons with 'normal classrooms' were implicit, viewed as non-problematic.

The major difference between the US and UK reports lay in the greater amount of sociological discussion in the UK reports.

In Newcastle, whilst the majority of reports to the projects were positive, the 'hard data' yielded publications which were an exception. In other words, the papers from Newcastle (Fitz-Gibbon, Hazelwood, Tymms & McCabe, 1988; Tymms, Fitz-Gibbon, Hazelwood & McCabe, 1990) reported negative, quantitative findings, albeit couched about with caveats, and rightly so.

It is of interest to ask how this occurred. We never, for a moment, wanted to write negative reports. 'Appreciative evaluation' was our model and the vast majority of our reports to the projects were descriptive and complimentary. We liked our friends and colleagues who were working hard, enthusiastically and creatively. We liked TVEI and the way it brought teachers together to an unprecedented extent (murmuring occasionally about the expense of the hotels and meals, such a change from the desperately inadequate funds available in the



normal course of schooling.) There was genuine excitement for the introduction of residential courses, work experience, computing and other technical subjects, and 'new' styles of teaching and learning (e.g. experience-based and student centred learning.) And it must have been very good for the hotel industry. (It is said that Blackpool had a second season: after *The Lights* there was TVEI.)

Why, then, did Newcastle nevertheless publish negative findings, showing TVEI pupils at age 16 obtaining fewer qualifications than similar pupils on whom vast sums had not been spent? (And the sums were truly vast in the pilot phase of TVEI, as, for example, £400,000 to be spent per year for four years on cohorts of 250 pupils.)

The negative evaluation event can be traced to the methodology: we had designed the evaluation to collect good comparative data: Performance Indicators and measures of value-added, although those terms were not being used in education at that time. Random assignment had not been implemented but we did our best to locate and measure pupils similar to those in TVEI. Having this 'hard' data, we had to publish. We tried every statistical control we could apply but the finding would not disappear. This was not the arrogance of opinion but the evidence of carefully collected data.

Once the data had been thoroughly analysed, written up, submitted to a journal and accepted, following anonymous peer review, we made the manuscript available to Peter Wilby, a highly regarded education correspondent for *The Independent*. We felt some urgency in that the government was urging TVEI-type 'enterprise' projects on universities and preparing to extend the TVEI 'pilot' to all schools, claiming unmitigated success. Yet our data sounded a warning. There were parts of TVEI not working. These findings needed to be 'taken on board'. Here was a nice 'problem solving exercise' for TVEI personnel. They were very keen on having students take on problem solving exercises which were as close to real life, as 'relevant' as possible. They seemed less keen on being presented with a real life problem of their own. Surely problem solving was not only for TVEI pupils? We had provided information to assist in the improvement of the projects; information relevant not only to the stated aims of the initiative but also to the concerns of parents and pupils. Examination results were, and are, important. As I have pointed out at greater length elsewhere, the data could have been used to improve the programme (Fitz-Gibbon, 1990).

I would like to emphasise the fact that we were genuinely surprised as the examination results came back to us and showed the negative associations with TVEI, despite our best efforts to remove the effects with fair statistical adjustments (e.g. looking at boys only or taking account of the larger number of TVEI pupils who intended to leave at age 16). Our data, provided in the middle of the pilot phase, should have been useful in adjusting the TVEI courses. Every project needs formative evaluation and time to be developed and adjusted as necessary in the light of experience informed by evidence. Alas, that was not the response received.

We had regarded the data collection as simply good practice. The first stated objective of TVEI was that pupils should leave school with better qualifications

so we had collected data relevant to that objective. When we adopted the approach of collecting quantitative data from both TVEI and non-TVEI pupils we had not envisaged the surprising findings that emerged. Not only was the approach we adopted standard evaluation methodology but the existence of good comparison groups might have enabled us to show gains which would otherwise have been undetectable. In planning the TVEI evaluations we had worried about the 'crunch statistics' (Fitz-Gibbon, 1986). I had thought, rather naively it seems, that subsequent funding might be dependent upon having evidence of an effective project. Trying to guess which data would count as evidence was an early concern. Was this pure naïveté ... to expect rational, caring-about-evidence behaviour? Perhaps the idea or expectation was just a few years early. A 1991 report from the National Audit Office appears to support moves towards the kind of data-based evaluation which we conducted. In their summary and conclusions, for example, we find:

It was not until 1988 that the Department (of Employment) began to formulate national performance indicators for monitoring the achievement and progress of the Initiative. In 1990 they issued a set of 27 standard national indicators against which education authorities were required to report...the considerable discretion given to authorities over the formulation of other indicators may render the information inconsistent and therefore difficult to analyse and aggregate nationally. (NAO, 1991: 2)

The report notes the extension of TVEI:

In 1987 the Department decided to extend the Initiative nationally, to allow all 14–18 year olds to become involved and to cover the whole curriculum.

There is no statement about the basis on which this extraordinarily wide-ranging decision was taken but later in the document we find that there was a review which has not been released:

Although they (the Department) completed an interim assessment of the pilot projects in 1988, this was used only for internal policy considerations relating to the further development of the TVEI extension phase and was not published or disseminated.

Nor, it seems, is much use made as yet of the performance indicators recommended by the NAO:

In the view of the National Audit Office, there is considerable scope for extending the use of the information provided by the national performance indicators. For example, it could be used to help direct the evaluation strategy towards areas where problems are being encountered or where targets are not being met.

Quite so. It is encouraging, I believe, to see a nationally influential organisation advising people to take Performance Indicators seriously; to use them; to learn from data. Looking into the indicators for assistance in understanding what is

happening is rational and advisable behaviour. Instead, when our findings were published, the Director of MSC's Education Programme suggested that the less able had not been able to cope with the absences from the classroom required by work experience and enterprise projects. This speculation was not substantiated by looking at the indicators. In fact further analyses of the data suggested almost an exactly opposite conclusion: it was the *more* able whose examination results were lower in TVEI, not the *less* able. The simple suggestion that more able pupils should be entered for more examinations might have reduced or eliminated the difference in examination outcomes between TVEI and non-TVEI students. (Fitz-Gibbon, 1990). The indicators were there; strategies could have been advised by the indicators but instead of using the data the administrators responded with speculation and defensiveness. We need to set up systems which promote rational responses to data rather than defensiveness. Such systems must probably incorporate at least two features.

- (1) Sufficient time for formative evaluation, time in which jobs are not on-the-line. A commitment to developing quality requires taking risks in the interest of advancing to higher levels of performance. Risks may involve temporarily reduced outcome levels and people will not feel able to take risks or face up to problems if they are under threat.
- (2) Sufficient feedback, widely disseminated, to enable everyone to make adjustments as needed. The monitoring and feedback must be relevant to the smallest units of management, such as the department in a school.

However, the NAO report recommends setting national targets. This does not seem advisable and is perhaps either an indication of a certain level of inexperience with Performance Indicators or an acquiescence to current political pressures. Not knowing how to interpret indicators people invent targets against which to judge them. What is needed instead is a basis for fair comparisons, such as looking at the value added by various groups compared with that added by groups which are working in sufficiently similar situations.

Did the TVEI evaluations influence the conduct of evaluation in general? The enduring *positive* effect has lain in the contacts between people in TVEI: the professional networks which TVEI brought into being. Networks are important and TVEI strengthened these at almost every level of the education system. Alas it now seems that, among universities and schools alike, the valuable professional co-operation, and the sharing of expertise that was so well promoted by TVEI are at risk. Schools are tempted to abandon collaboration as the market economy puts them into competition with each other. It is apparent from the NAO report that the sharing of information on practice was seen as a central feature of TVEI. The report includes four tables of instances of 'good ideas' and it is stated:

The National Audit Office consider it important that all evaluation studies should have a clearly-defined plan for disseminating findings.



That the climate of competition between schools will now jeopardise this development does not seem to have been taken into account. Lack of sharing of 'good practice' runs against the understood aims of TVEI.

### **The Impact of Monitoring Systems**

Most TVEI evaluation reports were focused on the TVEI projects themselves. This has always been a deficiency in evaluation: large amounts of money and effort have been focused on the small part of the system in which there is a new project. The assumption is that what is happening in the system as a whole is well known. The new project can therefore be described in detail and the comparison with 'normal classrooms' is implicit. Yet there is no adequate consensus about what 'normal' education looks like, perhaps because there is considerable variability in 'normal' classrooms.

Another problem with evaluation, which is a general one, not specific to TVEI, is the expectation that evaluators will locate a dream project (cf. the 'dream team'), write it up, disseminate it and education will be improved. This expectation is rather similar to the hope of researchers that they will 'discover' processes which can be disseminated and copied by others. There are real problems with this hope, not least that it has not proved to be a particularly effective strategy to date. A major difficulty is in the likelihood of very strong influences from the context in which a project takes place and from the nature of the (perhaps specially hired and enthusiastic) staff taken on for the project. There are all the problems of getting people to take on an innovation from 'outside' and the faithfulness of the implementation. We need to look at other methods of improving education because the dream project may not emerge.

Monitoring with feedback is a method which, thanks to computers, is now feasible. How would TVEI evaluations have been different if adequate monitoring systems had been in place when TVEI was introduced? For one thing, there would have been a body of professionals looking at their own data year after year. They would know a reasonable target from a fiction. They would be able to identify changes larger than usual. Thus not only a small cadre of evaluators would have been involved in the evaluation but also an entire body of concerned professionals, i.e. every teacher.

Given adequate monitoring there would have been more than just descriptive, quantitative indicators. There would have been evidence for every pupil of the 'value added' and this data would have presumably been available to schools and evaluators. The value-added for pupils in TVEI could have been readily compared with value-added for non-TVEI pupils, not just in the 20 odd schools in four LEAs for which we had set up a monitoring system, but for the whole country, school by school, LEA by LEA.

Furthermore specific questions could have been asked of the data regarding, for example, the value-added for ethnic minorities, for girls and for special needs pupils and any other groups of concern. If, for example, the Local Education Authorities (LEAs) had been monitoring achievement in schools there would have been no possibility that the surprising finding that Newcastle uncovered

could have failed to have been noted, or admitted, nationally. Perhaps the thought that the effectiveness of their interventions in education will be displayed by Performance Indicators could deter politicians from pressing for the use of indicators. However, the NAO, at least, wants audits of effectiveness.

However, I am increasingly coming to wonder whether indicators should be immediately in the public domain. If airline pilots had to give their names when they reported 'near-misses' they might not report as readily nor as accurately. People need room to face up to and recover from mistakes. Furthermore, the mistakes may be popularly attributed to the persons involved although the reality may be, more likely, that the problems arise because of *features of the system rather than because of the actions of particular personnel*. This notion is clearly spelled out by Deming, the guru of Total Quality Management who makes the distinction between system effects which he calls 'common cause' and 'specific cause' which might be attributable to an individual (see, for example, Neave, 1990: 264).

The difficulty of pinning down cause is, of course, at the heart of research methodology. Given this difficulty, one not recognised by those who simply trust their intuitions and opinions in the pre-scientific culture in which they seem content to live, two important lessons will need to be learned.

- (1) Innovations need a period of confidential, formative evaluation during which time careful measures are developed and collected but the findings are only used for confidential feedback *to those responsible for the projects*, not for public judgements. As Campbell (1969) noted, people need to be allowed to fail or they will not take the risks needed to make progress.
- (2) Interpretation demands comparative data and the better the comparisons the fairer and more secure are the interpretations which can be made of the performance indicators. Fair Performance Indicators are not only fair but necessary if cost-benefit analyses are ever to be feasible. The benefits accrued must be compared with benefits which might fairly have been expected from other expenditures of the funds. The best comparisons are based on the kinds of designs used in clinical trials in medicine. These 'true experiments' will eventually be used in a rational society concerned to provide the maximum benefits to its people within its resources.

At this point some readers are thinking 'naive again' (except those in the US and the Netherlands where the response might be 'Does this need saying? Of course we know experiments are needed.'). The argument advanced in the UK against experiments is usually that they are impossible. Would it have been impossible to implement TVEI in such a way as to permit strong inferences to be drawn about its impact, that is, as a true experiment? Not in the least. The pilot phase was restricted to small numbers of schools and LEAs; fewer than 4% of pupils. Rather than selecting these solely on the basis of criteria (unknown in general) applied to submissions made by LEAs to the Department of Employment, some, at least could each year have been based on matching up schools and randomly assigning them to TVEI or non-TVEI. (The 'non-TVEI' schools could simply have been delayed-entry rather than total withholding of funds.)



Within TVEI schools, participation could still have been voluntary. This situation is encountered all the time in clinical trials: not all persons accept the proffered treatment. Not every one swallows their pills. The analysis has to take account of this problem of implementation.

There is not the space, and this is not the place, to elaborate on the multi-factorial design which could have been set up, nor to deal adequately with the complexities introduced by the lack of definition of exactly what TVEI was meant to embrace. Suffice it to say that an equally complex intervention was adequately assessed in the US in a series of true experiments (Lazar & Darlington, 1982) thanks to which US politicians seem to be aware that there are surprising long-term benefits to be derived from the provision of cognitively-oriented, nursery education for severely deprived and at-risk children in US inner-city areas. The positive cost-benefit analyses which proper experimentation made possible are probably the reason why such programmes are generally protected in the budget, even in very difficult times.

In contrast, lest people imagine that every effort by social scientists has a positive impact, there was strong evidence that the provision of 'big brothers' to assist teenagers at-risk, and their families, in fact resulted in more law-breaking, recidivism and dependency than was found in equivalent families which had been left without 'big brother' visits (McCord, 1979).

In short, experiments have been done and can be done again. It is unethical to spend public money and interfere in children's lives in an unsystematic fashion when, with the simple introduction of some systematic decisions in the design of the innovation, we could find out how to improve the innovation and the extent to which it does good.

The advent of monitoring systems will considerably facilitate the conduct of experiments, as Willms argues (Willms, 1992). In considering how TVEI might have been different had there been monitoring systems already in existence, another point must be made: such a quantitative system could have left evaluators who are adept in qualitative methods free to try to analyse the 'Why' questions instead of spending large amounts of time trying to find out 'What' was happening. Qualitative descriptive reports might have been extraordinarily important if they could have been compared and contrasted across sites *from which there was good evidence of various kinds of effectiveness*. This evidence of effectiveness would not, in a good monitoring system, have been confined to examination results. At Newcastle our monitoring work has always included measures of attitudes and satisfaction levels among pupils, in addition to value-added measures. Such measures seem to be called 'Qualitative Indicators' by the authors of the NAO report.

However, useful as monitoring systems will be for evaluators of the future, the reason for setting up monitoring systems is not to aid one-off evaluations of programmes, whether these programmes are based on the latest ministerial flash of inspiration or on evidence accumulated from experiments and pilot studies. The major reason for implementing a monitoring system is to provide valid feedback to professionals responsible for the complex task of educating children.

### Summary

If monitoring systems had been in place when TVEI was introduced:

- Data would have been available nationally, not hidden.
- The qualitative descriptive reports could have been compared and contrasted across sites yielding insight rather than the uncertainties of opinion.
- Evaluators would have been free to concentrate on the *why* questions.
- The repeated measurements from the monitoring would have made at least a quasi-experimental design highly feasible: non-equivalent control groups would have been immediately available.
- Data would have been useful in adjusting the TVEI courses to improve outcomes, as urged by the recent NAO report.

The very complexity of the educational process is a reason why the search for the nugget of truth in a research programme or the search for the dream project both seem equally unlikely to yield beneficial changes as fast as they are needed. The vital reason for setting up monitoring systems is to feed back into the system the information needed for every unit to self-monitor its outcomes. Only carefully developed feedback of outcomes adjusted for intake differences can provide the information needed for quality assurance (equal opportunities) and for learning on the job (the reflective practitioner needs data).

The 1990s will be a time of growth and development of monitoring systems. The educational professionals (researchers, teachers, LEAs) would be wise to design fair and effective systems which work for the benefit of the students, staff, and society as a whole. Such systems could have been of great benefit in the evaluation of TVEI.

One of TVEI's most important legacies might be the networks it created which lead to the climate of collaboration and professionalism in which monitoring systems can be welcomed.

### References

- Alkin, M.C., Kosecoff, J., Fitz-Gibbon, C.T., and Seligman, R. (1974) Evaluation and decision-making: The Title VII experience. *CSE Monograph Series in Evaluation* No. 4. Los Angeles: Centre for the Study of Evaluation.
- Campbell, D.T. (1969) Reforms as experiments. *American Psychologist* 24, 409–29.
- Fitz-Gibbon, C.T. (1986) The roles of the TVEI local evaluator. In D. Hopkins (ed.) *Evaluating TVEI: Some Methodological Issues*. Cambridge: Cambridge Institute of Education.
- (1990) Learning from unwelcome data. In D. Hopkins (ed.) *TVEI at the Change of Life* (pp. 92–101). Clevedon: Multilingual Matters.
- (1992) Performance Indicators, value-added and quality assurance. Paper given at the annual meeting of the British Educational Management Association, Bristol.
- Fitz-Gibbon, C.T., Hazelwood, R.D., Tymms, P.B. and McCabe, J.J.C. (1988) Performance indicators and the TVEI pilot. *Evaluation and Research in Education* 2 (3), 49–60.
- Lazar, I. and Darlington, R. (1982) *Lasting Effects of Early Education* (Monographs of the Society for Research in Child Development). Chicago: University of Chicago Press.

- McCord, J. (1979) A thirty-year follow-up of treatment effects. In I.J. Lehman and W.A. Mehrens (eds) *Educational Research: Readings in Focus* (2nd edn) (pp. 282-92). New York: Holt Rinehart and Winston.
- McLaughlin, M.W. (1975) *Evaluation and Reform. The Elementary and Secondary Education Act of 1965, Title I*. Cambridge, MA: Ballinger Publishing Company.
- National Audit Office (1991) *The Implementation and Development of the Technical and Vocational Education Initiative*. London: HMSO.
- Neave, H.R. (1990) *The Deming Dimension*. Knoxville, TN: SPC Press, Inc.
- Thorndike, R.L. and Hagen, E. (1969) *Measurement and Evaluation in Psychology and Education* (3rd edn). New York: John Wiley and Sons, Inc.
- Tymms, P.B., Fitz-Gibbon, C.T., Hazelwood, R. and McCabe, C. (1990) The effect of a national initiative on examination results. *School Effectiveness and School Improvement* 1 (4), 281-98.
- Willms, J.D. (1992) *Monitoring School Performance: A Guide for Educators*. London: The Falmer Press.

## IMMIGRANT LANGUAGES IN EUROPE

*Edited by* Guus Extra and Ludo Verhoeven

The focus of this volume is on immigrant groups and immigrant languages with a recent or earlier background of migration to industrialized countries in Western and Northern Europe. After presenting some basic concepts in the area of language and immigration, the book focuses on demographic statistics of immigrant groups in European Community countries and Scandinavia, and on research in the field of immigrant language varieties.

### Contents

Introduction: Immigrant Groups and Immigrant Languages in Europe  
 Part 1: Use of Immigrant Language Varieties in Europe  
 Part 2: First Language Acquisition in a Second Language Context  
 Part 3: Code-switching  
 Part 4: Language Maintenance and Language Loss

Hbk ISBN1-85359-179-3

336 pages

1992

£19.95 (US\$42.00)



**MULTILINGUAL MATTERS LTD**  
 Frankfurt Lodge, Clevedon Hall, Victoria Rd.,  
 Clevedon, Avon, England, BS21 7SJ

