UNIVERSITY OF CAPE TOWN

FACULTY OF EDUCATION

QUALITY ASSURANCE IN HIGH SCHOOLS THROUGH REGRESSION ANALYSIS

A dissertation
presented in fulfilment
of the requirements for the Degree of

MASTER OF EDUCATION

by

J W WATERMEYER

SEPTEMBER 1996

Quality Assurance In High Schools Through Regression Analysis

Abstract

School Effectiveness is a relatively new and poorly defined domain for which a structure is proposed, to facilitate future discussion. Three fields within School Effectiveness are identified, namely School Effectiveness Research (SER), School Improvement (SI) and Quality Assurance (QA). Three divisions are identified within each field on the basis of various criteria. SER has methodological generations, SI is classified by decade, and three themes of QA are described, including performance indicators (PIs). A definition of effectiveness in terms of regression lines is described and the concept of added value or adjusted achievement developed.

This study is concerned with the development of PIs for use within a single school to monitor and promote improvement. The context of the study (a model C senior high school in a predominantly white southern suburb of Cape Town) and the data collected is described before a review is made of some of the analyses which could be used to monitor effectiveness. A technique whereby pupil achievement is adjusted (for prior achievement and other background variables) and the residuals (or adjusted achievement) derived from the regressions investigated with one-way ANOVAs is described and tested using various models and subjects.

With respect to groups, it is proposed that statistical significance of differences between mean residuals could be used as a PI. With respect to individual pupils, educators could set their own criterion for investigating cases where adjusted achievement is very large or very small. Statistical significance requires interpretation, however, and the role of professional judgement in modelling and monitoring adjusted achievement is discussed. The view that techniques such as regression analysis can only indicate when professional investigation and intervention might be necessary is stressed. It would seem unwise to rank teachers or subjects on the basis of adjusted achievement.

Acknowledgements

This study would not have been completed without the support and encouragement of many people, to whom I am extremely grateful. They include the Headmaster and colleagues of Fish Hoek Senior High School, supervisors Mr J. D. Gilmour and Assoc. Prof. T. T. Dunne, and my long-neglected wife and daughters.

Table of Contents

Chapter 1. The Problem

How can school leaders monitor and assess the quality of the education their schools provide? How can parents choose between schools? How can the state demonstrate that the education services contribute towards equal opportunity? These are some of the questions researchers into school effectiveness have tried to address since the sixties when Coleman *et al* (1966) concluded that children left school with unequal levels of achievement and schools were relatively unimportant with respect to long term success. Since then there have been serious efforts to demonstrate that schools do matter, can be improved, and should be accountable for the quality of their "products". Investigation has revealed that schools are not uniformly effective and that effectiveness is not a consistent and monolithic characteristic but "a very fragile construct" (Frechtling, 1987, quoted by Mandeville and Kennedy, 1991). Not only do schools affect children from different backgrounds differently, as the early work of Coleman *et al* (1966) and others found, schools also differentiate in many other ways. Children of different ages and gender but the same ability, for example, may achieve differently (Nuttall *et al*, 1989). School influence also varies with academic subject and level (primary and secondary) (Scheerens, 1992:70). Very little is known about the sizes and causes of these effects. Even the characteristics related to high levels of achievement in school are not clearly identified and apparently only matter in specific circumstances.

There are more questions than answers, and much of the research into school effectiveness has been on such a large scale that the characteristics of the individual pupils have been lost from sight. Nevertheless, it should be possible to apply some of the accumulating wisdom about school effectiveness to individual schools. This study was intended to investigate the application of School Effectiveness Research techniques to the information available in schools. The aim was to provide management with better analyses of information with which to guide school improvement and demonstrate effectiveness.

Motivation

Large scale research has found that school effectiveness is unevenly distributed across pupils and subjects, and unstable. Some schools are apparently more effective in some subjects than others (Luyten, 1994) and more helpful for some pupils than others (Mortimore et al.,1989; Nuttall et al.,1989). These conclusions are

not surprising but the fact that techniques exist for measuring the differences in effectiveness between schools, or subjects, offers interesting opportunities to school leaders.

If, as the large scale research seems to imply, it is possible to distil from a variable that part of achievement which can be explained by school or classroom characteristics, is it possible to apply similar techniques to smaller groups or even individuals within a single school in order to monitor the consistency of their achievements or progress? The question is pertinent in view of the instability of effectiveness over time and the potential for schools to affect pupils unequally.

Facilities which easily and routinely identify possible under- or over-achievers would be valuable to educationists who wish to ensure that achievement is consistently high and that all pupils achieve equally, according to their potential.

The motivation for this study is, 1) that a great deal of information is available in schools, particularly the records of prior achievement, 2) that statistical techniques (such as regression analysis) exist for adjusting achievement for background factors, and 3) that there is a need at school and classroom level for routine monitoring of quality, or consistency, with a view to timeous response to need. The intention was therefore to investigate the application of some of the concepts and techniques of large scale research to small scale situations for use by educational (rather than statistical) professionals.

Much of the large scale research is motivated by the needs of politicians and bureaucrats to demonstrate the effectiveness of the delivery systems and the needs of parents to choose between schools and, having chosen a school, to be assured that quality mechanisms are in place. The league tables in use in the United Kingdom, for example, are extremely crude indicators and probably only reflect the socio-economic status of the pupils a school is able to attract rather than the school's relative effectiveness (Gray and Wilcox, 1994).

It was hoped that this study would also lead to techniques which could help schools to demonstrate their effectiveness, at least against their students' histories, if not by comparison with other schools.

By initially attempting to include to include performance indicators for use in the education market in which parents exercise choice, this study had a very broad purpose. Due to both technical and theoretical issues, however, it eventually focused on indicators which might monitor the achievements of individuals or small groups within a school. The techniques explored may be used to identify situations in which pupil achievements (after controlling for prior achievement) are similar, and those where anomalies or inequalities need to be investigated.

Chapter outline

The following chapter is a review of the theoretical background of the study. As a relatively young domain, School Effectiveness lacks a recognized structure or taxonomy. An analysis which outlines the history of School Effectiveness studies and shows some of the relationships between the main fields is offered. The analysis provides a systematic means of reference and a summary of the findings in School Effectiveness Research. The model used in this research is described and some of the values embedded in it acknowledged.

The school and community context of this research study are described in chapter 3.

After the description of the data collected, chapter 4 deals with the rationale for using regression analysis, and some of the initial findings. Chapter 5 is devoted to a description of the main study, including the regression models and the investigation of their residuals by analysis of variance. Some of the relationships observed in the results are discussed. Techniques for generalizing from the results are also described and applied to examples, then the findings are listed.

The role of the professional judgement of educators when using regression analysis is considered in chapter 6. The difference between educational significance and statistical significance is emphasized. In chapter 7 the conclusions and the possibilities for further research are noted.

Chapter 2. From school effects to added value

Introduction

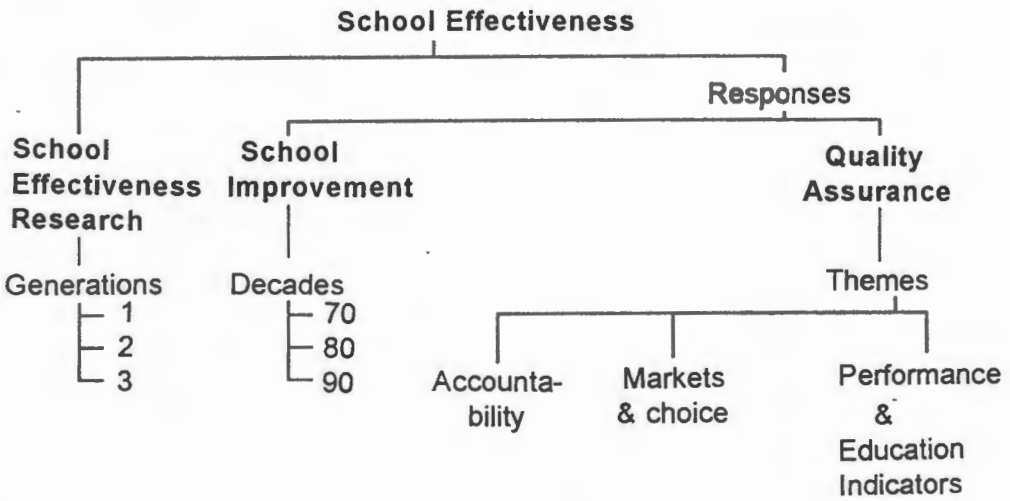
Over the last 30 years a substantial body of literature has been developed in the domain of School Effectiveness. There is an annual International Congress for School Effectiveness and Improvement, which has met since 1988, and a dedicated international journal of School Effectiveness and School Improvement, which first appeared in 1990 (Reynolds *et al*, 1994: 31). Research has been conducted widely in First and Third World countries and cross-nationally. The central concern of the domain, according to one critic, "is to identify techniques and procedures that can be applied directly in any educational or management situation" (Angus, 1993:335). Although its structure is not yet clear, School Effectiveness has its own concepts and a distinctive methodology. Given the nature of school effectiveness, experimental procedures are seldom possible but some recent studies have used control groups (e.g. Harbison and Hanushek, 1992; Reynolds *et al*, 1993). So while School Effectiveness does not fulfil all of the requirements of a form of knowledge such as "distinct and peculiar concepts" (Hirst, cited by Graves and Simons, 1973:28) it does seem to have a claim to a distinctive area of knowledge which is referred to here as a domain. In Cultural Geography "domain" describes the zone which immediately adjoins the core area of a culture and into which the culture spreads (Clark, 1985:149). The domain of School Effectiveness is certainly one in which teaching, economics and politics overlap.

On inspection the domain of School Effectiveness appears to be variously and very idiosyncratically described. Perhaps it is to be expected in such a new area of knowledge. Until recently there has been little consistency in the definitions of terms and their operationalization. Findings seem to be contradictory and initially were reported in forms which made comparison difficult. For example, findings were often given without interpretable units. So, in order to identify what is meant by School Effectiveness the first part of this study consists of a review which shows a structure inherent in the domain. The review concludes with a discussion of the gap between School Effectiveness as an academic interest and as a resource for practical school management.

An analysis of the domain

As noted already, the domain of School Effectiveness is young and largely unsystematic. The following analysis is intended to describe an inherent structure or system. It is based upon the division between School Effectiveness Research and School Improvement described by Reynolds *et al* (1993), and the general agreement that School Effectiveness falls into different periods or generations (Kreft, 1993; Reynolds *et al*, 1994). In the following analysis these divisions are extended to provide a reference system which should facilitate future discussion.

Fig. 2.1 The domain of School Effectiveness



It is convenient to consider School Effectiveness in terms of three aspects, namely 1) research, 2) school improvement and 3) quality assurance. The last two are in part responses to the first (see fig. 2.1), one from the education practitioners and the other from the wider community. These three divisions will be referred to as fields, to borrow a term from Hirst (cited by Graves and Simons, 1973). A field of knowledge is based upon concepts borrowed from several forms of knowledge.

The fields differ in their purposes, methods and conceptions of effectiveness, amongst other things (see table 2.1). School Effectiveness Research (SER) is largely

quantitative and based upon a production-function model. It is primarily concerned with demonstrating how important schools are to society and why they influence achievement, for the practical purpose of informing school improvement. Effective schools should be able to help all pupils by raising their levels of achievement. School Improvement is a field occupied mainly by practitioners concerned with the effects of individual schools. Initially the focus was upon the disadvantaged but now the concern is to improve the quality for all pupils (Teddlie in Reynolds *et al*, 1994). The contributions tend to focus on the correlates of effectiveness and the processes involved in changing schools. Any procedure or tool, such as the one this study is intended to develop, could be applied in this field by school management. The third field, Quality Assurance, derives from the needs of politicians for proof that the education services in general meet their obligations, and the needs of parents to be able to decide whether a particular school will meet their needs. Most of the contributions deal with the information required to improve accountability and parental choice, but provide very little information on how the information should be analyzed.

Before reviewing the fields in detail, note that each one is itself subdivided (see fig. 2. 1). Research has been divided into three generations of a developmental sequence on the basis of the form of the production-function model used in each. The generations are not chronologically discrete since the older techniques are still in use, where appropriate. Different phases are also discernible in School Improvement, which has changed over the years. According to Reynolds *et al* (1993) each decade may be characterized by a different approach (see fig. 2.1). Each of these decades will be briefly described in this review. Quality Assurance is a younger field than the others and differs from them in having developed very little so far. Instead there are a number of related themes or centres of interest. They include accountability, choice in an "education market", and performance indicators.

Now that the structure of the domain has been laid out, each field may be described in some detail.

School Effectiveness Research (SER)

The three generations are described in table 2.2. The purpose, models and findings of each will be discussed here.

Table 2..1 A comparison of the fields in the domain of School Effectiveness

SER	School Improvement	Quality Assurance
Purposes		
Measure the extent of inequality, then demonstrate the import-ance of schools & find correlates of effectiveness.	Formulae for making schools effective, especially for the disadvantaged.	Demonstrate that resources are being effectively and efficiently used to provide education of a high standard.
Focus		
Change in society but found little basis to believe that education could contribute. Later concerned with independ-dent variables such as socio-economic status rather than change strategies .	The characteristics of effective schools and how to change individual schools. The concern tends to be with the journey rather than with the destination.	The relative quality and efficiency of systems or schools to inform account-ability and consumer choice in the education market.
Models & Methods		
Production function model and quantitative analysis of large data bases, supplemented by questionnaires, and later, by interviews.	Five factor model, synthesis and meta -review with rare empirical evaluation. Based upon practitioner knowledge.	Free market, served by the reduction of large data sets to a few easily-interpreted indices, profiles or rankings, i.e. statistical generalization.
Backgrounds of leaders		
Sociologists & economists.	Educational practitioners.	Politicians served by statisticians.
Definitions of effectiveness		
School able to change all its pupils to desired standard.	Empowered staff move the school towards their desired quality.	School able to deliver education services to a specified quality.

With acknowledgments to Reynolds, Hopkins and Stoll,1993.

The purposes of SER

While the present purpose of SER may be typified as a search for tools for pragmatic management (see Angus, 1993, above), the focus of the research has varied across the generations. Certainly the early studies were not much concerned with school level management practices, beyond resourcing. They were interested in the influence of social background on educational outcomes (Coleman *et al*, 1966; DES (i.e. Plowden), 1967) and vice versa, i.e. the impact of education on social inequalities (Jencks *et al*, 1972). Their findings that achievement was dominated by background and not by schooling, and that schooling offered little prospect of breaking the cycle of poverty (Jencks *et al*, 1972) were widely interpreted as "schools don't matter". This conclusion shocked educationists and provoked criticism of both the model and the designs of the research. The reaction was particularly strong in the US where the conclusions contradicted an important tenet of the American dream that anyone from any background could achieve both economically and politically.

The second generation of researchers had a single primary objective. They wanted to show that schools did matter because the schools did make a difference. The difference could be demonstrated either by showing that school effects were larger than found by the first generation, or by finding schools where pupils achieve in spite of their poor, urban minority group backgrounds. When such effective schools were found, the characteristics which correlated with achievement were identified. Some of the problems with the findings of these studies will be considered when the field of school improvement is discussed.

The third generation studies have similar purposes, with the additional goal of establishing which characteristics relate to achievement across a range of different circumstances, e.g. not just in schools serving urban communities, but rural and suburban ones too. In other words the purpose is to identify correlates of achievement which are independent of the context.

The models of SER

The model used in SER is derived from industrial economics (Hough, 1991). Education is viewed as a production function, which in its simplest form relates inputs to outputs (fig. 2.2 A). The production unit itself is simply regarded as a black box which mediates inputs and outputs but is of no direct interest (Kreft, 1993). The

Table 2.2 Three generations of School Effectiveness Research.

2.2.1 The main conceptual differences

1st	2nd	3rd
Purpose		
To demonstrate inequalities.	To demonstrate that schools make a difference & to identify the related correlates.	To show larger influence & more generalizable correlates, i.e. for various contexts. Developing a theoretical basis for SER.
Focus		
Socio-economic differences between pupils and schools. Mainly static factors used.	Measures of variables which might influence school processes such as norms and interaction.	Similar measures but less aggregated to single levels, such as the school.
Version of the production function model		
Input-output.	Input-process-output.	Input-process-output but prior achievement now used as an input.
Models and methods		
Input-output model in cross sectional case studies. Partition of variance & regressions.	Input-process-output models in cross sectional outliers and case studies. Correlations & regressions.	Longitudinal-process-output studies & a major outlier study. Same techniques & multilevel analyses, allowing, inter alia, the use of smaller units of analysis.
Periods		
1966 onwards.	1971 onwards (outliers '71-'79; case studies '79 onwards)	1986 onwards.

2.2.2 Examples

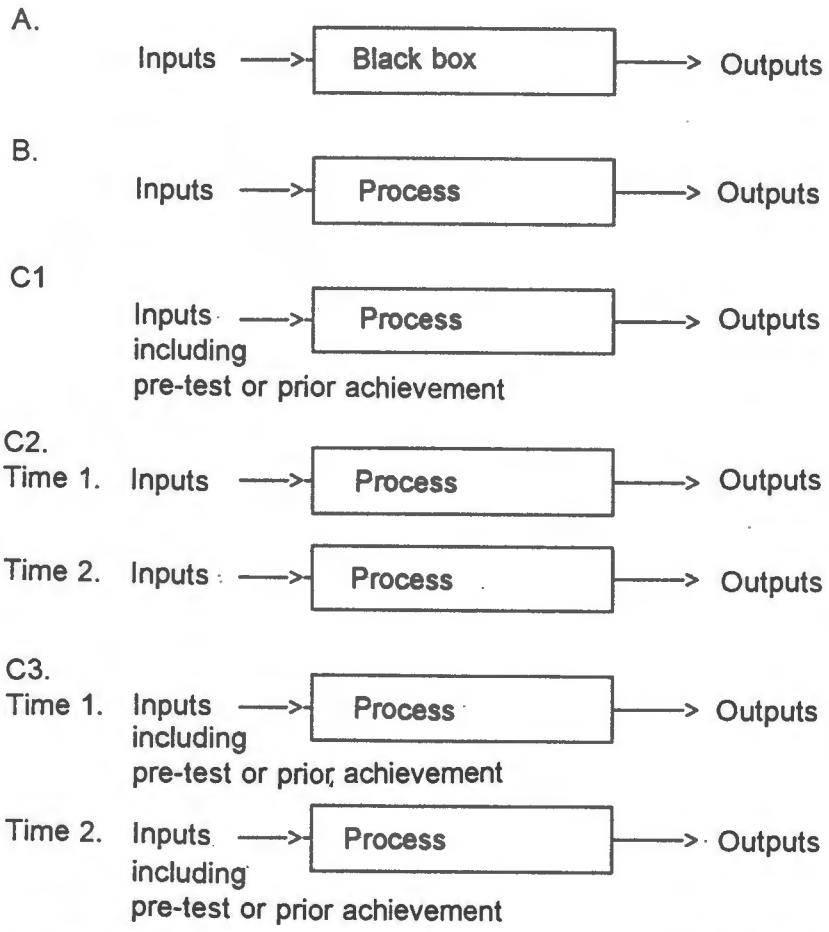
1st	2nd	3rd
Important examples		
Coleman et al,'66; Plowden(DES),'67; Jencks et al.,'72.	Weber,'71; Brookover,et al.'79; Rutter,et al.'79; Coleman, et al.'82.	Mortimore et al.'88; Brandsma & Knuver'89; Gray, Jesson & Sime'90; Teddlie & Stringfield '93; Luyten'94.
Examples of programme evaluations		
	Armor et al.'76.	Harbison & Hanushek '92.
Examples of Third World studies		
Heyneman & Loxley'83.	Lockheed & Longford'89; Lockheed et al.'89.	Riddell'89;
Exemplar reviews		
	Glasman & Biniaminov '81; Purkey & Smith'83; Ralph & Fennessey'83; Good & Brophy'86; Fuller'87.	Scheerens'92; Reynolds & Cuttance '92; Angus'93; Reynolds et al.'94.
Examples of international studies		
IEA's 1st Maths and Science Studies; Six Subject Science Survey.	IEA's 2nd Science and Maths Studies; IEAP Science, Maths and Literacy projects.	Scheerens et al. '89.

assumption was that if one can control for differences in input, then the differences in output reflect the effectiveness of the schools. Stated differently, the value of outputs could be adjusted by deducting the value of the inputs to assess effectiveness. Another perspective was that if one knew how the inputs related to each other and the output, then manipulation of the inputs should have ensured equality of outputs.

When it became clear that schools had less impact on outcomes than was expected, some researchers changed the model by including more variables about the schools themselves and what happens there. The production function thus became an input-process-output model and it is this model which distinguishes the

Fig. 2.2 The production function models

The basic models for SER generations 1, 2 and 3 are A, B and C1 respectively.



first two generations of research from each other (see fig. 2.2 B). Examples of studies using these models are given in table 2.2.2 (p.12).

In the third generation a longitudinal element is added by including measures of the same variable as an input and an output (see fig. 2.2 C1). This modification allows School Effectiveness to be considered in terms of added value. For example, the variable concerned is typically achievement in an area such as mathematics. Each pupil may be tested twice, i.e. before and after a course, and the improvement

claimed as value for which the school is at least in part responsible[1]. The longitudinal element has been accommodated in other ways too. Some studies assess the effect of schools on separate cohorts (see fig. 2.2 C2 and table 2.2, p.11) for examples). Such a procedure may be useful in monitoring the stability of the processing unit, the school, as a whole but is less helpful in assessing added value because it does not measure changes in the same sample of pupils. There is less control for sample differences.

The most powerful way of assessing added value would be a repeated use of pre- and post-tests on a series of cohorts passing through the same system, i.e. the repeated use of model C1 (see fig. 2.2 C3, and table 2.2.2, p.12, for examples). Repeated measurements should give both a mean measure of School Effectiveness and some idea of its variation over time.

The longitudinal production function model has been enhanced in the third generation by the use of multilevel analysis techniques such as the Hierarchical Linear Modelling (HLM) (Bryk and Raudenbush, 1989). The use of these techniques in addition to more conventional procedures is a key characteristic of this generation. They only became available to researchers in 1986, when Aitken and Longford demonstrated the appropriateness of the new technique for modelling educational situations. The hierarchical, nested nature of education systems (e.g. pupils "nested" in classes, which are in turn in schools, themselves in education departments within a national system) had been recognized (Purkey and Smith, 1983:428; Ralph and Fennessey, 1983) as a confounding factor beyond statisticians' control. Either a measure at one level had to be disaggregated to a lower level (and in doing so making unjustified assumptions about the distribution of achievement) or aggregated to a higher one (and losing information in the process). Multilevel analysis has, however, several advantages. It improves estimation of relationships at each level (i.e. within a unit of analysis such as a class or school), makes possible the testing of cross-level hypotheses and the partitioning of variance between different levels (Bryk

1 - The idea of measuring added value apparently stems from Coleman who noted that "it is the increment in achievement that the school provides which should be the measure of the school's quality"(1975, quoted by Rutter *et al*, 1979:5).

and Raudenbush, 1992:5). So it is possible, for example, to estimate how much variance is due to school and classroom factors respectively (Scheerens *et al*, 1989).

With or without the sophistication of multilevel analysis the production-function model is of course problematic in many respects (Angus, 1993). By its use researchers presume to reduce the complex interaction of education to a simple linear relationship. It is patently impossible to capture the range of inputs and outputs through statistical data collected largely by questionnaire. Quite a number of studies have used more output variables than examination or test marks (see Fuller, 1987) but even so it is difficult to measure the full range of immediate outputs such as skills, insights and self esteem, let alone the longer term outcomes. Angus is suspicious of the added value concept, because he fears that "controlling for intake" implies a deficit on the part of the disadvantaged rather than difference (1993:341; also Davies, 1994:210). He feels that researchers do not try to see the wider, more holistic picture and lack curiosity about the mathematical connections they establish between the correlates and achievement.

Before leaving these methodological matters and considering the findings we should note that the second generation of SER is characterized by a number of studies using outlier designs of one kind or another (see table 2.3 overleaf). Outlier designs were used because it was felt that normative analyses (i.e. those based upon average characteristics and achievements) could not "detect the influence of schools on achievement if most schools were equally ineffective" (Edmonds, 1979, quoted in Teddlie and Stringfield, 1993:17). So, for example, researchers began to compare the best with the worst schools (positive and negative outliers) or look for common characteristics amongst the best schools (positive outliers only) (Purkey and Smith, 1983; Stringfield in Reynolds *et al*, 1994:75-6). The procedures for identifying the "best" and "worst" schools vary from simple polling of opinion amongst informed professionals (e.g. Jubber, 1988) to sophisticated regression analyses (e.g. Teddlie and Stringfield, 1993). Procedure aside, concepts such as "best" and "worst" are extremely problematic. Many researchers have been undeterred, however, and their results are starting to show some consistency.

Table 2.3 Types of outlier designs

1. **Positive outliers only** which directs attention to the features being sought, but provide no contrasts.
2. **Positive and negative outliers** which show the strongest contrasts and highlight the desirable features.
3. **Positive outliers and typical examples.** Also known as an outrigger design.
4. **Positive** outliers, **typical** examples **and negative outliers.**

Stringfield (1994: 75-6).

The findings of SER

The general findings of the first generation have not been changed much (see table 2.4, overleaf) although the use of smaller units of analysis has brought awareness that effectiveness is not evenly distributed over all groups in a school.

The first generation of research showed that background factors are more influential than schools in determining achievements on tests or examinations. Coleman *et al* (1966) found that schools explained less than 10% of the variance in pupil achievements and much of that was due to the average socio-economic characteristics of the school (Scheerens, 1992:34). The second generation was generally unable to refute the results of the first (Purkey and Smith, 1983:428) although widely differing claims confused the field. The third generation research confirms that schools contribute of the order of 10% to differences in achievement (Willms, 1987; Mortimore *et al*, 1988, 1989; Brandsma and Knuver, 1989; Gray *et al*, 1990; Mandeville and Kennedy, 1991; Scheerens, 1992; Teddlie and Stringfield, 1993; Luyten, 1994). 10% is considered educationally significant and has been estimated by Jencks *et al* (1972) and Purkey and Smith to be equivalent to as much as one school year by the end of secondary education (1983: 428).

School influence is usually greater for subjects such as Mathematics and Science which are more dependent upon school teaching than language skills such

Table 2.4 The findings of three generations of SER

1	2	3
Pupils' background explained more of the variance in achievement than schools.	A wide range of variables found to correlate with achievement in school but differ from one culture to another.	About 10% of the difference in pupil achievement explained by school factors.
Schools accounted for less than 10% of achievement. Much of this due to the average pupil background.	Curriculum-based exams are more sensitive to school differences than standardized tests.	Prior achievement a good predictor of performance.
Schooling could do little to reduce inequalities in society.	Social factors such as norms & patterns of interaction were found to be important when explaining difference in achievement.	Cognitive outcomes tend to be closely related in primary school.
School seemed to have larger effects in poor countries.	Private, selective schools seem to be more effective than public in the USA and the Third World.	High school subjects show different sizes of school effects.
Physical resources had much larger impact in poor countries.	Great confusion as to whether effectiveness is stable.	Evidence of schools affecting various groups differently.
		Leadership and other correlates vary with socio-economic context.
		The relative size of school and class effects varies across countries.
		Stability of effects (across time and grades) is greater in high than primary schools but not evenly distributed.

as reading which may be acquired "at one's mothers knee" (Fuller, 1987; Mortimore *et al*, 1988, 1989; Raudenbush, 1989; Luyten, 1994).

It would appear that high schools have greater and more stable influences than primary schools (see table 2.5, p.18). The interpretation is that the closer the coefficients are to one, the more stable the effects. However, when smaller units of analysis are used, such as subjects of pupils, then differences in the stability of effectiveness are found. Luyten (1994) has also shown that the influence of Dutch high schools is much less stable or consistent over time for some subjects (e.g. History and Geography) than others (e.g. foreign languages). Fitz-Gibbon has also found that even in the rarefied area of A- levels, effectiveness is unstable (in Reynolds and Cuttance, 1992). High school pupils are taught by a number of subject teachers. It is tempting to suggest that the greater instability of primary schools may be related to the differences between the class teachers to which the pupils are allocated, i.e. a teacher rather than a school influence (Scheerens, 1992:71) but experience in this investigation has shown that class characteristics may play a greater role. See chapter 6.

Table 2.5 Stability estimates of school effects.

	Primary education	Secondary education
Stability over time	0.36 - 0.65	0.70 - 0.95
Stability over grades	0.10 - 0.65	0.25 - 0.90

From Scheerens, 1992. Luyten (1994) identifies them as Pearson's *r* correlation coefficients.

This distinction between school and classroom influences has been investigated in a cross-national study by Scheerens *et al* (1989). It would appear that the portion of school effects which are actually due to classroom variables differs widely. In some countries, e.g. Sweden and New Zealand, all school effects were found to be explained by classroom factors while at the other extreme they explained

very little of the total school effects in Belgium and the Netherlands. These findings serve as a reminder that effectiveness seems to be context dependent.

Given that effects are apparently not very stable across subjects or grades it is not surprising that effectiveness at school level is not very stable either. Only about half the primary schools retained their effectiveness status (as positive or negative outliers) over eight years in a study by Teddlie and Stringfield (1993:218). They point out that it is unreasonable to expect effectiveness to be stable because ineffective schools will be under pressure to improve while changing circumstances may undermine the efforts of the effective ones (pp.45-7). If schools are dynamic organizations it follows that they need to be monitored regularly.

Socio-economic status is almost universally used in SER to control for background differences between pupils because it discriminates between them and explains some of the variability in their achievements. The explanatory power of socio-economic status confirms that schools do not compensate for differences in socio-economic status (Brandsma and Knuver, 1989) but managers would like to ensure that as far as possible children of different backgrounds benefit equally. There is some evidence from the third generation studies (Mortimore *et al*, 1989; Nuttall *et al*, 1989) that schools are differentially effective for ethnic groups, ability groups and genders. All are aspects which may need to be monitored. The techniques for monitoring are not widely known, however.

A definition of School Effectiveness

From the above it is clear that school leaders concerned with quality need to monitor the distribution of the effectiveness of their schools over the pupils. It is appropriate to sum up by establishing what is meant by effectiveness in SER.

In terms of the production function, effectiveness is "the extent to which the desired output is achieved" (Scheerens, 1992:3). Although the model has been discussed above, no description has yet been provided of the input, process and output variables used.

Inputs, independent or explanatory variables, may include measures of individual socio-economic status, family background, neighbourhood and ability. Prior

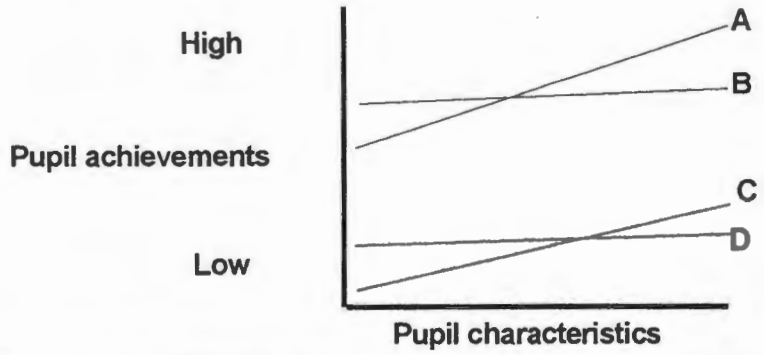
achievement, when used as an input, may capture part of the influence of many such factors. Process variables, or co-variables, may relate to school policies such as streaming or subject choices, as well as teacher and leadership characteristics, such as gender, age, training and style.

Output or dependent variables are usually measures of achievement on tests and examinations but other factors such as self-concept and behaviour (e.g. "contact with the police", Rutter *et al*, 1979) have been used. Long term variables such as admission to tertiary institutions, career choice and success may be termed outcomes to distinguish them from the immediate outputs (Scheerens, 1992: 3).

The definition and examples given here tend to hide the complexity of expressing the production function in practical terms. Most of the variables are problematic. What, for example, are "desired outputs"? Are outputs measured on standardized achievement tests taken by all pupils, or curriculum-based examinations where only one or two subjects may be common to the majority of pupils? Is the output norm-referenced (e.g. percentiles) or criterion-referenced (e.g. the proportion achieving an "A-aggregate")? The measurement of inputs is also problematic. For example, the classification of socio-economic status is a subjective process. Even the selection of a measure of prior achievement may be problematic. Final Std 5 or std 7 marks or std 8 first term marks (aggregate or individual subjects) will all yield different assessments of added value. (The reasons for using these explanatory variables is discussed in chapter 6. See table 6.4, p. 91.)

In spite of these problems, researchers have continued to relate the inputs and outputs (the explanatory and dependent variables) through regression lines. It has been suggested that there are at least two important dimensions to effectiveness - quality and equity (Brandsma and Knuver, 1989; Cuttance, 1992:81; Creemers, 1994: 4). The quality dimension refers to the level of achievement. The equity dimension refers to how this achievement is distributed over pupils of different characteristics. Fig. 2.3 shows these dimensions and illustrates how effectiveness is a relative term. Pupils at schools A and B enjoy higher quality outputs than pupils at schools C and D of similar characteristics and in that sense the schools may be said to be more effective. Within schools B and D pupils of different backgrounds seem to achieve fairly similar results, reflected in the flatness of the regression lines. Where the

Fig. 2.3 Two dimensions of effectiveness illustrated by regression lines



background variable relates to socio-economic status schools associated with flat lines may be described as egalitarian while schools where achievement differs widely between various social groups may be said to be elitist. Schools A and C are examples, indicated by their steeper lines.

Terms such as egalitarian and elitist are value-laden and imply social values. Their use encourages the assumption that all background circumstances, other than those used as explanatory variables, are equal. Experience in this study has shown that such an assumption is unrealistic and invalid. It would be wise to avoid such judgmental terms in all but the most specific circumstances.

A third dimension of effectiveness should be considered too. Research has shown that effectiveness is unstable over time. The changes in a school's position therefore has to be monitored. Performance indicators are used by School Improvers and Quality Assurers to demonstrate the changes.

In conclusion, effectiveness may be considered to have three dimensions: quality, equity and stability. Their expression in terms of regression lines may be appropriate to large scale SER. It is another question whether the concepts in this form are valid or useful for monitoring quality, equity or stability within a school. It will be shown later that the distribution of pupils' achievement above and below the regression lines provided more useful information than the lines themselves.

In this section the purpose, models and findings of SER have been outlined, and a definition of effectiveness described. In the following section the response of education practitioners to the research, namely School Improvement, is reviewed.

School Improvement

Introduction

Since this project is intended to provide a procedure for improved school management it is important to review the field and establish the present position. In much of the recent literature School Effectiveness research and school improvement are dealt with together. The latter is often the justification for the former. Scheerens (1992) is a good example and the title of the journal *School Effectiveness and School Improvement* and the regular international congresses of the same name bear out the point. There is, however, a gap between the two fields. The researchers have a quantitative approach which differs from the social-skills approach of the improvers. It would appear that very little work has been done to bring the techniques of SER to assist school improvement.

The strategies, predominant model and features of promising projects will be outlined here.

The strategies for school improvement

The main difference between SER and Improvement lie in the aims or purposes of the two fields (see table 2.1, p. 9). While the researchers seek to show that schools can be important and to establish the differences between the effective and the ineffective, School Improvers try to find a formula, recipe or procedure which will help schools to change into more effective institutions. Unlike SER, which has changed its aims over the generations, the purpose of improvement has remained constant with minor changes in focus (see table 2.6). However, the means to the end have changed considerably over the decades. According to Reynolds *et al* (1993), in the 1970s attempts were made to introduce innovations from external sources and from the top down. Improvement was sought through schoolwide organizational and curriculum changes without much thought to classroom practices or consultation. Teachers were only involved to the extent that changes (often unwelcome) were made to their tasks. The consequent lack of ownership and commitment to the improvements is considered the main cause of their general failure.

As a consequence, in the 1980s the ideal project was thought to be school-based and owned by the teachers. The strategy was bottom-up in that the

problems identified by the teachers were the ones addressed. Practitioner experience was now central to the strategy, rather than some theoretical research finding. The focus was now on the process of change, e.g. introducing group work skills to arrive at objectives by consensus.

Towards the end of the 1980s, improvers began to recognize that the current approach was reactive and ill-suited to prompting or generating school improvement. Interest then grew in self-evaluation of school processes and outcomes as a necessary and routine element of improvement (e.g. Joyce, reported in Scheerens, 1992:98). A third and more balanced approach may be distinguished in the nineties.

Table 2.6 Three decades of School Improvement.

	1970s	1980s	1990s
Purpose	Improving achievements of the disadvantaged.	Improving achievements of all.	Improving over a wider range of outputs.
Focus	School organization and curriculum.	Improving teacher skills to facilitate the process of change.	Information gathering and consultation including pupils & the wider school community.
Orientation	Top-down.	Bottom-up.	All levels of the school community.
Site	Outside school.	Within school.	School community.
Evaluation	Quantitative.	Qualitative.	Both.

With acknowledgment to Reynolds, Hopkins and Stoll, 1993.

The ideal strategy might now be summed up as an integrated and carefully managed data driven approach involving the whole school community. Change is recognized to be a slow and gradual process which is only considered complete when the alteration has become institutionalized, i.e. routine. The school community is not only the pupils, staff and management but may be all the stakeholders, including parents, bureaucrats, sponsors, higher education institutions and employers.

In conclusion, it should be noted that even such an apparently uncontentious objective as school improvement can be problematic. At one level the issues are similar to those of SER, e.g. how are goals appropriately expressed as outputs, and what constitutes "improvement"? More fundamentally though it has been suggested that the idea that everyone would like highly effective schools is a myth (Davies, 1994). It is argued that if all schools were highly effective the existing social order would be threatened by "qualification inflation .... The last thing a fragile state wants is too many articulate, well-qualified students" (p.206). To ensure that only a limited number of pupils succeed standards are raised or reforms introduced to side-track teachers "into other activities to diminish their efficiency" (p.206). It is claimed that, with the exception of Japan and Tanzania, few governments are honest about the filtering role of education. In Japan school certificates are viewed as paper qualifications unrelated to preparation for work (beyond an "orientation to hard work, competition and loyalty"). In Tanzania state secondary education is "severely rationed" (pp.210-211). However, whether the filtering role of education is a matter of policy or honesty is not for this study to resolve.

The observation that an effective education system might not be a priority for all serves as a reminder to reformers of the need to consider the political dimension of their reforms. Innovators have to "work out what rewards everyone will officially and unofficially extract from a change" or their project will fail (p.209).

The predominant model

One of the early assumptions in School Improvement seems to have been the concept of an ideal school. The logic was that effective schools have many common features which correlate with achievement, and the more all schools could be organized like the effective schools, the better. Although the hunt for characteristics of effective schools, especially those serving poor urban minority groups, was started by

Weber (1971), Edmonds is regarded as the father of the Improvement tradition. His 1979 study of schools in the north eastern USA yielded 55 effective schools, defined as institutions where there was "essentially no relationship between background and achievement" (Teddlie and Stringfield, 1993:17). Although similar to SER of the time, Edmonds' study is considered here because of the use which has been made of the study. What started as a demonstration that there were schools in which all pupils could learn became a significant development movement.

From his own and others' research Edmonds and the improvers advocated a five factor model which became the basis of many development projects. Teddlie suggests that "Edmonds and his colleagues were no longer interested in only studying poor, inner-city schools that worked, but also in creating them" (p.87). The five "malleable correlates of educational achievement" (Creemers, 1994:12-13) were considered to be strong educational leadership, high expectations of achievement, an emphasis on basic skills, a safe and orderly climate and frequent evaluation of progress.

Since it is based upon correlations, the five factor model itself has been criticized on statistical grounds (pp.12-13). Although correlations only measure the extent of relationships between variables, the five factors have been accorded the status of causes of achievement. In the case of high expectations and achievement, it is not clear which variable might be the cause and which the effect (p.12). In addition, Creemers and Scheerens feel that it is tautological to emphasize basic skills and then to use basic skills exclusively as a measure of output (1989, quoted by Creemers, 1994:12). They also question whether the five factors are really independent. The last four might easily be considered functions of the first, namely strong educational leadership.

Besides the statistical weaknesses, the model is probably situation dependent, i.e. only valid in the schools with which Edmonds and his colleagues were concerned. Be that as it may, the five factor model has been widely applied. Reviews such as that of Purkey and Smith (1983) probably encouraged the approach since they noted the correlates of effectiveness identified by many studies. By the 1980s, however, it would appear from Reynolds *et al* (1993) that characteristics of the ideal school organization were being replaced by the concept of empowered staff.

Some of the critical features of improvement projects

A number of principles characterize the 1990s approach to improvement.

1) The school is the centre of change.

2) Change should be a carefully planned and managed process.

3) There is a wide range of educational goals, broader than test achievements and including "the developmental needs of students, the professional development of teachers and the needs of (the) community" (Reynolds *et al*, 1993:42).

4) A multilevel perspective is required, which defines and harnesses the roles of people at different levels in and above the organization.

5) Integrated implementation to ensure that both top and bottom work together.

6) Change is only complete when the alteration has become routine i.e. institutionalized (van Velzen *et al*, 1985, quoted by Reynolds *et al*, 1993:42).

7) Improvement should be data driven, i.e. routine and systematic analysis of information, and a commitment to scrutinize the data and act upon the conclusions (Hopkins *et al*, 1994: 143). This is a key characteristic of 1990s Improvement.

Reynolds *et al* (1993) described three examples of nineties projects, namely the Cardiff Change Agent Study, the Halton Board of Education's Effective Schools Project and the Improving Quality of Education for All project (IQEA) (Reynolds *et al*, 1989, Stoll and Fink, 1992, and Ainscow and Hopkins, 1992, respectively cited in Reynolds *et al*, 1993). Key aspects of the Cardiff Study seem to have been the equipping of senior staff with group work skills, in order to facilitate schoolwide consultation, and to increase awareness of pupils' problems and the ways in which the schools impact on their pupils (p.49; Scheerens, 1992:104). The Halton Project in Canada is notable for a highly structured but flexible process which permitted local adaptation. The four stages of the process are assessment, planning, implementation and evaluation. The initial assessment involves pupils, parents and staff. The whole process is expected to take at least three years (Reynolds *et al*,

1993:48). Flexibility of process within a focused strategy is common to all three of these examples, not least the IQEA (Hopkins *et al*, 1994:102). The main approaches of the IQEA are also to encourage schoolwide collaboration and to help schools improve their internal conditions. Collaboration is used particularly for adapting externally imposed requirements into school priorities and in evaluation. The internal conditions have to be developed so that they will both manage and support the change (Reynolds *et al*, 1993:46). Rather like the Halton Project, the process is seen to involve a number of facets, namely staff development, inquiry and reflection, leadership, co-ordination and planning. But unlike Halton, these procedures are not seen as stages in a linear process. They tend to "coalesce" (p.46).

Conclusions

The common elements of these models include staff development and training, wide consultation, assessment of situation prior to the project, evaluation of outputs or outcomes, co-ordination and support for the changes at all levels, and flexibility to accommodate local conditions and wishes.

The evaluation of outputs is the concern of this study. The key characteristic of nineties School Improvement is that it should be data driven. The importance of an information system is widely recognized in the Improvement literature (e.g. Jenkins, 1991:61; Joyce, 1991, cited in Scheerens, 1992:98; Scheerens, 1992:102; Potter and Powell, 1992; Fitz-Gibbon, 1992; Riley and Nuttall, 1994; Hopkins *et al*, 1994:143). Few, however, go beyond pointing out the power of modern computers (e.g. Jenkins, 1991) and with the exception of Fitz-Gibbon, and MacBeath (1994), very little advice is available on how the data might be analyzed at school level. This shortcoming may reflect the gap between the quantitatively oriented researchers and the social-skills orientation of the improvers. This project is intended to find ways of bridging this gap.

So far School Effectiveness Research and the response of educationists have been outlined. The political response is considered in the following section.

Quality Assurance

School Effectiveness Research provoked the Improvement movement amongst practitioners. The response of the parents and politicians to findings such as "schools don't matter" and to international league tables which placed the USA and United Kingdom far down the rankings, for example, in Science (Reynolds *et al*, 1994: 226), was to call for greater accountability and proof that the education services were delivering. For example, in the UK the Treasury was concerned about value for money (Riley, 1994: 87). Kogan, however, notes that the reason for seeking greater control might be paradoxical (Kogan,1986:21). Is control intended to increase efficiency and reduce expenses, or is it (as Coleman argued) about the input of resources? In the latter case control may lead to greater expense. Despite this confusion a search began for performance indicators which would inform public auditing and parental choice. Quality Assurance is used here as a general term for accountability and the means for serving it, although the term has a more specific meaning in the popular concept of Total Quality Management (Murgatroyd & Morgan, 1993). In this section the concepts of accountability, education market and education performance indicators will be discussed in order to recognize the political dimension of SER and School Improvement.

Accountability is said to exist when "role holders are liable to review and the application of sanctions if their actions fail to satisfy those to whom they are in an accountability relationship" (Kogan, 1986:25). Kogan outlines several ideal models of accountability. See table 2.7. An important difference between them is the identity of the group whose interests are best served by each model, be they the politicians and bureaucrats, the professional staff or the consumers. The identity of the role holders depends on the level of the system under review, starting at the top with the politicians and ending at the bottom with the classroom teachers. The nature of the sanctions referred to in the definition are very limited in the professional accountability model but may be extreme in the free market model, namely loss of post. The nature of the review required by each model also differs from model to model. In the professional accountability and self-reporting models, review is essentially private while in the free market situation information is required so that consumers may make informed choices. It has been suggested that parents will discriminate upon inappropriate

criteria (such as sports results ?) if there is no better information available on which to assess the effectiveness of schools (Walsh, 1994:60). Since it is the aim of this project to improve the quality of the information available for review, whether internal or external in a free market, it is appropriate to note some of the problematic aspects of the education market.

The education market

In the education market approval or disapproval are signalled through choice. Having to compete for trade is supposed to improve the quality of schools' service. It has been argued that the logic is faulty and that competition does not automatically lead to better service. Apart from the fact that there is often only one school available and hence no choice, for example in rural areas, loss of public support and dwindling enrolments may force a school to reduce its staff and the subjects it offers. If the process forms a vicious cycle of decline and the unpopular school goes to the wall the remaining schools can relax somewhat because there are plenty of pupils while the consumer now has less choice than before - fewer schools with fewer subjects (Ball, 1993:8). Another concern is that schools will find ways of making the results look better without actually improving performance. One way would be to refuse to accept weak children. If efficiency is an issue (the best possible outcome for the least inputs) then children with expensive learning needs may also be excluded.

In the market model schools are considered to be service providers motivated by the same interests as other businesses in the market. The analogy is of limited value since generally businesses are run for the benefits of either their owners (higher profits) or the staff (higher salaries). In contrast, schools are supposed to serve the pupils and their communities. Experience suggests that privatization increases the costs of services (Walsh, 1994:60) but service quality does not necessarily improve (Ball, 1993:7).

Given the above problems with the market model, the inefficiency of bureaucratic control and the self-interest of professional domination (Scheerens, 1992) there may not be a great deal to choose between the various models. There seems to be considerable support for the free market model in the United Kingdom and the USA, however.

Table 2.7 Models of accountability.

1. **Public or state control**

   Managerial or bureaucratic hierarchy to whom teachers are accountable for externally imposed standards. The sanctions available to superiors are circumscribed by the tenured position of staff.

2. **Professional accountability**

   Protection of schools from "product-oriented outcomes" although accountability is to parents and/or the pupils so as to improve school "responsiveness to clients".

3. **Professional accountability & self-reporting**

   An extension of (2) with little external validation. Favoured by naturalistic evaluators (e.g.Simons) who feel that the developmental potential of evaluation is spoiled by publicity.

4. **Consumerist control - partnership**

   Parents and school are equals and "accountable to each other for their contributions to a shared task". Partnership requires consensus on objectives, exchange of information (e.g. on methods, their limitations and implementation) and dialogue to evaluate what has been done.

5. **Consumer control - free market**

   Parental power of choice through free schooling through subsidies or vouchers. A "negative relationship" between the school and the clients, but no political or bureaucratic control.

From Kogan (1986:25)

Performance Indicators (PIs) and Education Indicators (EIs)

Whichever model of accountability is adopted, information is required. For most purposes, the specifications for this information are probably the same as those for politicians (Ruby, 1994) : simple, comparable and timely. To which could be added inexpensive and based on incorruptible data (Nuttall, 1994). How these ideals are to be achieved is a dilemma which has prompted investigation of performance indicators.

Performance indicators (PIs) are routinely recorded quantitative and comparable statistics used as a set. Each may be compared with others in its set and with earlier indices. They are commonly used to permit generalization about large scale activities such as trade, crime and sport. The Olympic medals table is an example. Nuttall and Gray distinguish between performance indicators and performance criteria (1994:76). The former are quantitative and relate to organizational input and output characteristics. Criteria are qualitative descriptions of the processes which "are assumed to mediate input and outputs". They may also be known as education indicators (Cuttance, 1994: 105).

In Britain, PIs were first used in the public domain to audit the National Health Service. Experience there has shown them to be susceptible to manipulation and inflation. The number of indicators grew from 70 to 450 in less than a decade (Walsh, 1994). Pressure for accountability in education grew until the publication of examination results referenced on some external standard or norm became mandatory and the league tables appeared. Measuring achievement against a benchmark is known as the standards model of School Effectiveness (Cuttance, 1992:76; Jesson, 1992a).

In the USA, indicators seem to be widely used. For example, state by state comparisons are used by a federal agency, the National Assessment of Educational Progress (Seldon, 1994:44). In some areas indicator systems report variables such as achievement test scores, pass rates, attendance and pupil-teacher ratios regularly at state, district and school levels (p.44). Some states publish this data in the form of a school report, much as schools issue to parents. In addition to diagnosis and planning, indicators are also used in some states as the basis for incentive rewards

and sanctions (p.44: Mandeville and Anderson, 1987; Mandeville, 1988). The use of performance criteria or education indicators in the US is apparently recent and limited.

PIs in education typically appear in the form of tables of indices, such as examination results, post-sixteen destinations, attendance, and comparison with national averages. Descriptive statistics, such as means and rates of change, and graphs may be used to guide interpretation, (e.g. Cuttance, 1994). In the UK, according to Gray and Wilcox (1994), such statistics have caused a debate as to what they mean. Do results tell one anything about the schools or just about the type of pupil they attract? Efforts have been made with regression analyses, first, to place examination results of schools into their socio-economic context and, second, to assess the value added by each school. These efforts have been made in spite of the conclusions that rankings, of schools or systems, based upon regressions are unstable (Mandeville and Anderson, 1987; Anderson, 1988; Woodhouse and Goldstein, 1988). The dilemma is that the better the regression equation fits the data the smaller the residuals and the less the discrimination between institutions (Woodhouse and Goldstein, 1988: 314). Small changes in the input variables may then result in considerable changes in the ranks of schools based upon either regression coefficients or residual analyses.

Attention has also turned to performance criteria or education indicators (EIs) since the PIs alone seem to some to be an inadequate basis for judging systems. Gray and Wilcox suggest that the use of critical lists of education indicators in the UK is due to the need to make inspection more credible. The lists stem from three sources: the correlates of effectiveness beloved of the early school improvers, "generally agreed notions of so-called good practice", and the practices of the inspectorate. Little is known about how these lists are or should be used. A promising use of the EIs is to combine them with rating scales, as was done in the evaluation system used by the former Cape Education Department in South Africa.

It is unrealistic to regard indicators and the features they measure as objective and unproblematic (Angus, 1993; Gray and Wilcox in Riley and Nuttall, 1994). Their selection and presentation is inevitably value-laden (see Simons, 1984:51). In addition, indicators may result in a distorted evaluation and priority given to inappropriate aspects of education (Helsby and Saunders, 1993:66-7). It is argued

that indicators need to be "finely differentiated in terms of their purpose, their audience, their use and their source" (p.67). Such differentiation requires a clear understanding of the procedures by which the indicators are generated. Some of the techniques are discussed below.

Some techniques for Quality Assurance.

There are two aspects to assuring quality, namely information and management's response to it. Management strategies such as the Total Quality Management (TQM) package of Murgatroyd and Morgan (1993) are attempts to systematize the 1990s approach to school improvement in a form appropriate to the education market. Potter and Powell (1992) in their management guide do mention some simple PIs. But while the key characteristic of the 1990s School Improvement is that it should be data driven, very little guidance is given by the management experts on the derivation or analysis of PIs. For this information it is necessary to turn to other sources.

Four techniques used by quality assurers to inform educational leaders are briefly outlined here.

1. Regressions have been generally used, from the first generation of SER (see Glasman and Biniaminov, 1981) in efforts to assess added value by adjusting the apparent response or output for background variables. They seem to be the basis for any ranking procedures (e.g. Willms,1987: Woodhouse and Goldstein, 1988) and more sophisticated than the standards model (Jesson, 1992 a,b; Cuttance in Reynolds and Cuttance, 1992; Birnbaum, 1993). Background variables such as ability, prior achievement and socio-economic status may be used to explain differences in the examination results. The coefficients of regression lines may be found for different groups, e.g. boys and girls, or pupils (or whatever the unit of analysis being used) may be ranked according to the size of their deviation from this regression line. An assessment of School Effectiveness is obtained when prior achievement is used as the main explanatory variable. In terms of the production function, it reflects the value added by the school. When using other explanatory variables, such as socio-economic status, Cuttance refers to the resulting residuals as intake-adjusted estimates (in Reynolds and Cuttance, 1992: 78).

2. The regression approach may be enhanced by the use of multilevel analysis (Aitken and Longford, 1986; Mandeville, 1988; Woodward and Goldstein, 1988; Bryk and Raudenbush, 1992; Kreft, 1993). Not only does it show stronger associations within levels of analysis, but makes it possible to allocate influence to different levels.

   The purpose of using background or explanatory variables, as in regression analysis, is to ensure that comparisons as far as possible are made on the same basis by controlling for some of the differences, so that like is compared with like. Unfortunately it is impossible to take all differences into account and sometimes schools, and pupils, are very different from each other. A technique known as Data Envelopment Analysis has been proposed to assist in the comparison of schools or systems with similar goals or circumstances (Jesson *et al*, 1987; Mayston and Jesson, 1988). It is intended to accommodate "the multi-dimensional, and inter-dependent, nature of educational outcomes" (p.332), using a multiple linear model to establish an "achievement possibility frontier" (p.325) against which a school, for argument's sake, might be compared. This frontier estimates what the "best" schools with those circumstances and goals have achieved, i.e. what is possible. The technique allows comparisons to be made with the top achievers, not with the average as is the usual practice with average residuals or adjusted achievements. Despite these advantages, there has not been much reference to Data Envelopment Analysis in recent literature.

3. A number of researchers have investigated the effectiveness of schools on different ability groups by dividing the sample into sub-sets such as ability bands (Rutter *et al*, 1979; Nuttall *et al*, 1989; Blakey and Heath, in Reynolds and Cuttance, 1992). Glogg and Fidler (1986) report a similar approach to monitoring what they considered to be the added value measured by examination results in an English secondary school. They used the average scale points and average IQs of different ability groups over a five year period. The average scale points and IQ for the whole cohort were also calculated. All the indices could be compared against long term averages. Improvement was defined as the improvement in average scale points scored per pupils divided by proportional improvement in ability at entry of each cohort. A similar

technique might be used for other groups, e.g. boys and girls, if long term averages are available.

4. An ipsative model which is not based upon the production function, which makes none of the assumptions underlying regression analysis and does not require records of background data, has been described by Birnbaum (1994). It can only be used to analyze the distribution of marks achieved within the sample, i.e. there is no longitudinal element. Each pupil's marks are converted to a deviation from his/her own average. The deviations are then totalled in the groups being compared, say, English, Mathematics or Art. This produces a profile or ranking which shows which subjects get most marks for pupils. Within each subject it would be possible to sub-divide the total deviation between groups of interest, e.g. gender, ability or prior achievement groups. Although this technique is described as providing "a reasonably pure record of value added" this claim seems to be a misconception as none of the pupils' history is used and no account is taken of previous achievement or differences between pupils. It is therefore not a value-added model.

From the above description of accountability and indicators it is apparent that Quality Assurance could be a minefield. As will be shown in chapter 6, safe navigation requires clarity on the use which is to be made of the indicators, the prevailing model of accountability and the techniques by which they are generated. The uses could include the monitoring of such diverse objectives as material efficiency or affirmative action within or across schools. The model of accountability will be important since it will indicate who is accountable to whom, while the techniques used and the operational details may profoundly influence the resulting indicators. Even the achievement in examinations may be measured in various ways, such as marks, scale points or standard scores of marks. Given the variations possible, there is a danger that when indicators are produced, the targeted recipients will not be sure of their meaning.

Conclusions

In this chapter an analysis has been made of the domain of School Effectiveness, from the early estimates of school effects by Coleman *et al* (1966) to the recent expressions of added value and intake-adjusted achievement. Each field (SER, SI and Quality Assurance) has different contributions and limitations as far as a school manager is concerned. Despite the focus of SER upon between - school comparisons, as opposed to within - school issues of performance, some of the findings of SER are important for individual schools. Apart from the encouraging interpretation that School Effectiveness is significant, the findings that effectiveness is inconsistent and unstable have implications at school level. However, the extent to which the techniques of the large scale analyses can be applied to small samples is unclear. For many, the findings of SER are also inaccessible because of the statistical techniques used. Quality Assurance also requires the application of similar techniques for public use and is hampered by the same technical problems, quite apart from the questions raised by the basic models of production function and accountability. Nevertheless, on the basis of SER and the inequalities in education, School Improvers have long seen the need for change. Despite an appreciation that improvement should be informed by performance indicators, very little attention has been given to developing information systems for use at school level. In response to this situation, the aim of this study has been the investigation of analyses of data already available in a single senior high school which could generate measures of added value, or adjusted achievement.

The school and the context of the study are described in the next chapter.

Chapter 3. The context of the study

Introduction

In the previous chapter it was argued that the impact of School Effectiveness on individual schools has been hampered by technical problems. In effect, there is a gap between the theory and the practice of School Effectiveness. The gap exists mainly because the concept of effectiveness and the statistical tools of SER have not been applied within a single school. The School Improvers, who acknowledge that change should be informed by indicators, have not addressed the problem of how these should be derived. This study is intended to try to develop such an indicator or indicators.

In this chapter the context of the study will be generally described, i.e. the nature of the community served by the school and its schooling infrastructure. Attention will also be drawn to the instability of class characteristics which arises from the subject choice available, the characteristics of the cohort on whom the study is based, and the performance indicators in use at Fish Hoek Senior High School (FHSHS).

The community and the schools

When the 1994 matric cohort used in the study entered standard 8 at FHSHS, the community was relatively homogeneous. According to the census of 1991 (CSS, 1992), Fish Hoek and Sun Valley were almost exclusively white, 60% of the population had incomes of less than R30 000 p.a. (per person, not per household), 60% were not economically active (mainly retired and home managers), 24% were 65 years of age or older, and 15% were under the age of 15. The single largest economically active group worked in 'clerical and sales' jobs, and, while half the population had std 10 certificates, very few were graduates. Several implications for educational achievement may be expected to follow from these conditions. As we have seen, socio-economic status (SES) may influence educational outcomes, so one might expect the pupils not to achieve as well as those from schools serving higher SES groups, but better than those from poorer areas where the general level of education is lower. Another implication is that there is probably very little difference between the SES of FHSHS pupils, on the grounds that the community is fairly

homogeneous. Whether the fact that a large proportion of the population is not economically active has an influence on pupils' ambition or work habits is unknown.

The schools

The schooling infrastructure in Fish Hoek is unusual, with a Preparatory School (up to std 1), four Primary Schools (including Kommetjie) being the main feeder schools for the Middle School (stds 5-7), and a Senior High School (std 8-10). Thus it is not uncommon for children to attend four schools in their career, each for three years, in addition to any pre-primary schooling. Perhaps this unusual number of changes of school makes the teachers particularly sensitive to community opinion, but there is a strongly perceived need on the part of the schools to gain and hold the confidence of the community. Very specifically, the FHSHS needs to be able to demonstrate publicly that it is effective in terms of the final examination results and post-school careers of its pupils, and it would assist the Fish Hoek Middle School (FHMS) if the Senior High School could demonstrate that the products of the FHMS do better than, or at least as well as, pupils who enter the FHSHS from other schools. Over and above the general principle that the efforts of the schools will be more fruitful if the community has faith in them, the needs of the schools arise from the wish to encourage enrolments in the developing education market in which most model C schools (and their successors) will find themselves. The urgency of the situation is increasing in the face of the current process of rationalization in Western Cape education. Pupil numbers, after all, equate to jobs for teachers and wider subject choices.

The pupils

The general context has been outlined. More specifically, the cohort of 1994 std 10 pupils consisted of 153 pupils, although another 53 had moved in and out of the group during the three years at the Senior High. Of the final group, 142 would probably have been considered white. There were slightly more boys (52%) than girls. The majority had come directly from the Middle School (72%). Most still had two parents at home (79%). Many lived in Fish Hoek or Sun Valley (66%). Nearly 40% had mothers who were not economically active. The majority thus had stable backgrounds.

Table 3.1 IQ distributions of std 10 classes

Stanine	1989		1990		1991		1992	
	% cl.	8&9	% cl.	8&9	% cl.	8&9	% cl.	8&9
9	11.7		15.2		8.1		9.3	
8	13.3	25.0	12.6	27.8	16.7	24.8	18.5	27.8
7	26.1		25.2		21.9		23.5	
6	21.7	<6	20.5	<6	28.3	<6	25.4	<6
5	18.3	27.1	19.2	26.5	15.0	24.8	12.1	23.6
4	7.7		6.6		8.6		7.3	
3	1.1		0.7		1.2		4.2	

Stanine	1993		1994		1995	
	% cl.	8&9	% cl	8&9	% cl	8&9
9	9.8		4.8		8.2	
8	15.6	25.4	10.9	15.7	16.5	24.7
7	19.7		15.0		19.8	
6	21.3	<6	30.6	<6	27.2	<6
5	21.3	33.3	21.7	38.7	18.1	28.0
4	9.0		12.9		7.4	
3	3.3		4.1		0.8	
2					1.7	

With respect to their academic ability, however, the class of 1994 was not an able group. On the basis of the stanine distribution (table 3.1, overleaf) it was apparently the least able group for some time, with a smaller proportion of the class

with stanines above 7 and a larger proportion below stanine 6. The variability between cohorts displayed in table 3.1 may in part be due to the fact that, as a community school, no child who might benefit from the school's teaching is turned away, i.e. there is a non-selective admission policy.

The practice of generating and analyzing the stanine distribution began several years previously when the std 10 results were felt to be disappointing and some basis on which to evaluate the perception was sought. When the perception was confirmed - a very able cohort had produced very mediocre results - a new policy was adopted in order to improve examination results.

The aim of improving the results has so far been expressed as various objectives, such as :

1) to increase the proportion of the cohort achieving "A" and "B" aggregates[1] ;

2) to reduce the number of failures to zero;

3) to improve the averages in each subject in relation to the provincial medians.

As recognized by the approach of the 1990s to School Improvement, raising academic standards of a school is a slow process, requiring sustained effort. It has been likened to changing the course of a fully-laden ship. The wheel goes over a long time before the ship is brought round. To some extent, this study of effectiveness in general and indicators in particular is part of a prolonged effort to raise academic standards at FHSHS.

The subject choice and timetable at Fish Hoek are important elements of the context of this study since they prescribe the groups which are compared and hence those within which individuals operated.

A very wide subject choice was available. Since several subjects were offered in more than one timetable group, a pupil could take almost any combination of four subjects selected from more than a dozen, many differentiated into Higher and Standard Grades, in addition to the two official languages. Despite deliberate streaming on prior achievement, this flexibility resulted, inter alia, in subject classes (or sets) which varied in size (from less than 10 to greater than 30 - see appendix 4.1), prior achievement, ability, gender mix and background. In addtion, the composition of subject sets was not stable from standard to standard as the cohort progressed.

In the circumstances, an analysis which could take different set sizes into account would be preferable to one which could not. Further, the instability of the composition of sets requires the analyst to be cautious in comparing achievements over more than one year since changes in group composition might themselves result in apparent variations in group achievements.

1 A represents an aggregate 1679 marks, or 80%. B represents a mark 1469 - 1679, or 70 - 80%.

The indicators

The stanine distribution is used as a guide to the interpretation of the performance of each cohort. IQs are mainly used in general terms such as stanines or distributions. The IQ distributions, in conjunction with the standard average subject marks, have provided part of the justification for deliberate changes in teaching style and the introduction of additional support structures in the past. Intervention has taken place either when a large tail of weak pupils or particularly strong head of high achievers has been identified.

Annually the indicators used to monitor performance in the final external examinations at FHSHS at present include:

1) the failure rate, as a percentage of the candidates;

2) the number of matric exemption candidates and the number of matric exemption passes, both expressed as a percentage of the whole standard;

3) the numbers of A, B and C aggregates achieved (for some reason not considered as percentages of the candidates);

4) the symbol distribution in each subject and the total number of subject As;

5) subject means in the final examinations in relation to provincial medians and the September trial examinations;

6) the number of subject failures.

On a quarterly basis similar indicators are used, i.e. proportions of failures, matric exemptions, A and B aggregates and subject means. However, the means of only a few subjects are monitored closely: English, Afrikaans, Mathematics (Higher and Standard Grades) and Physical Science (again both grades).

The external examination indicators are used to identify strengths and weaknesses in the results of the latest std 10 class. Such post mortems often provide the basis for a subject department's goals for the coming year, and possibly even for the staff as a whole, bearing the qualities of the incoming std 10 class in mind. The

quarterly indicators are employed in the same way, to review achievements in recent internal examinations and make adjustments at the subject level where necessary or possible.

Most of these indicators deal with the cohort as a whole, or large subsets. The progress of individuals is reviewed quarterly on the basis of their previous marks. Changes in aggregate are noted. This review is often informal, however, and little account is taken of the distribution of the marks or overall changes amongst the pupils. For the std 10s particularly, aggregate failure or a drop of 5% or more from term to term may result in compulsory revision classes. Although positions in class and standard on the basis of aggregate are available, little use is made of them, apart from recognition given to the top ten in the standard. The top ten improvers, on the basis of aggregates, are also recognized.

The major shortcoming of these indicators is the absence of any control for background differences between cohorts, classes or individuals. As with the league tables of schools in Britain, it is not possible to establish whether a subject which has a higher average and more subject As is more effective than one which had a low mean and no As, or whether there are other explanations, such as differences in prior achievement, ability or background of the pupils. If achievements could be adjusted for such variables (Cuttance, 1992:78), then more informative comparisons could be made between data subsets of interest, e.g. between school subjects. Such a technique would also make possible the useful monitoring and comparison of other groups which have so far been ignored, such as: the pupils from FHMS as compared with those who came from other schools; girls and boys; pupils who take different curricula, such as Technika or more commercial courses; ability groups; racial groups; socio-economic groups; class sets. Marked inequalities between groups such as these could be investigated further and adjustments or interventions made where appropriate.

Some of these analyses would be for public consumption and school promotion, but the majority would primarily signal further professional investigation and response.

Conclusion

The literature of school effectiveness indicates that adjusting apparent achievement for prior achievement and other background factors in single school analyses should be possible and practical. This study was undertaken to establish if that indication is correct, and whether it might be possible to improve the indicators used at FHSHS. In developing a school level indicator, the study should assist in bridging the gap between School Effectiveness theory and school leadership.

The collection and initial exploration of the data and various models used for refining data towards fair comparisons are set out in chapter 4.

Chapter 4. Data and Analytical Techniques

Introduction

The purpose of the preliminary exploration was to consider the procedures by which the data could be analyzed. The main focus was on regression analysis since much of the literature defines school effectiveness in terms of regression lines. Furthermore, regressions have been used to derive residuals on the basis of which groups such as schools have been compared. Alternatives to regression analysis, both simpler and more complex, were also briefly considered.

The data collection is described here rather than in the previous chapter because the procedures followed influenced the findings. Data was assembled in a Lotus 2.4 spreadsheet and analyzed with the aid of the Statgraphics package, version 6.

Data collection

Information was collected from three sources:

1) The official record card, which should follow all pupils through their school careers in the state system, and on which the final marks of each year up to std 9, as well as some family background information, the schools attended, and the child's IQ scores are recorded. Unfortunately, the marks do not include any means (apart from the standard average aggregate) or standard deviations. Marks obtained at different schools are therefore not comparable.

2) The enrolment form completed by parents when the child arrived at the Senior High School. This form includes the child's date of birth, the parental occupations, home address, family size and the child's rank in the family, all at the time of enrolment. Although information such as addresses and parental occupation may have changed later, enrolment was taken as the synoptic moment and more up-to-date data was not sought.

3) The school's record of quarterly marks, the so-called "green book", which is certified by the headmaster and superintendent of education at the end of each year.

Eventually, over 180 variables were included in the database. They included information about pupils' background and middle-schooling, i.e. final std 5, 6 and 7 marks in selected subjects, and aggregates. The vast majority were quarterly aggregates and marks in the subjects included in the study. The subjects were English First Language Higher Grade (HG) and English First Language Standard Grade (SG) (in std 10 only), Afrikaans First Language HG and Afrikaans Second Language HG, Mathematics HG and SG, Physical Science HG and SG, Geography HG and SG, and Business Economics SG (although, languages apart, all subjects are only offered on the Standard Grade in the first term of std 8). The languages, Mathematics and Science were included because they are regarded as key subjects at Fish Hoek on the grounds that they are taken by almost all of the pupils who do well academically. Business Economics is offered as the alternative to Mathematics and these pupils tend to be less successful in examinations. The inclusion of Business Economics was intended to make the sample more representative of the pupils. Geography was included because it is the researcher's own subject. Apart from the possibility of bringing special insights to the analyses of one's own subject, it was thought to be wise to be seen to submit one's own results to the same scrutiny as colleagues'.

Aggregates were the sum of each pupil's marks from their best six subjects.

No pupil took all the subjects in the study since a wide choice was available over and above those listed. While all pupils had language marks and aggregates, Mathematics and Business Economics were mutually exclusive. There was some overlap between Mathematics, Science and Geography. The numbers of pupils taking each subject in std 10 varied from over 150 in the languages down to around 20 in Business Economics.

The many other subjects offered at Fish Hoek were excluded to ensure that the study did not become too big and unwieldy simply through the size of the database and the number of calculations.

Although Higher and Standard Grade marks were recorded separately, it was decided to treat them as if they came from the same equal-interval scale and merge the marks into one variable. The same decision was made with respect to First and Second languages. Such assumptions are in line with the Education Department

procedures by which Higher Grade marks are converted to Standard Grade, and marks from different Grades are used to calculate the aggregates. The main reason for making the assumption was to increase the size of each subject sample since, for example, the Geography Higher and Standard Grade pupils could be treated as a group. Merging the grades also reflected the composition of some classes, where both grades were taught simultaneously. Merging was also consistent with the use of the std 8 first term marks as the main explanatory variables (as is described later) because all subjects other than the languages were only offered on the Standard Grade. The assumption of an equal-interval scale seemed more acceptable than the scale point system (e.g. HG A = 8 points, HG B = 7, HG C = 6, SG A = 6, etc.) but turned out to be invalid for the small Afrikaans First Language HG and English First Language SG sets.

Race was not included since more than 90% of the cohort would probably have been considered white in apartheid terms. It was felt that race as an explanatory variable would not discriminate adequately between pupils.

The cohort which finished school in 1994 consisted eventually of 152 pupils in std 10, although only 132 of these had done all of standards 8, 9 and 10 at Fish Hoek. None of the 10s were repeating the standard although some had spent two years in earlier standards. Appendix 4.1 contains the descriptive statistics of the cohort.

While most of the information was already numeric, variables such as parental occupations and residential address involved considerable subjectivity. The same classes of occupation were used as were applied by the Central Statistical Services in the census of 1991 (CSS, 1992) but uncertainty arose when parents reported their occupation in general terms such as "Navy", "self-employed" or "builder". Equally subjective decisions had to be made when classifying or grouping residential areas. Should Simons Town and Glencairn have been combined, or Kalk Bay and Mitchell's Plain? The fact that these variables later proved to be of limited use in explaining the variability between pupil's achievements may in part be due to the subjectivity of the coding despite efforts to ensure that it was internally and educationally defensible.

Possible analyses

Once the data had been collected and the descriptive statistics calculated (see Appendix 4.1) various techniques for analyzing the data were considered. Some possibilities are outlined here, working from the simple to the sophisticated.

First, however, it will be useful to define some of the terms and the notation used in this account. Variables will usually be described as dependent or explanatory. The output or dependent variable (DV) is the set of marks being explained, analyzed or investigated. They are "dependent" upon other variables, such as prior achievement, ability and background, which are known as the explanatory variables (EVs). To distinguish one mark variable from another in each subject and aggregate, each variable is labelled according to the standard and term in which the marks were achieved, such as standard 8, term 1. These labels have been abbreviated in the text to 8'1 through to 10'4 for the results of final examinations in std 10. In some cases, particularly in the appendices, the labels have been abbreviated still further, to 81 through to 14. Also please note that "subjects" includes the aggregate variables too, unless otherwise specified.

1. Comparing raw scores

Recent achievements (dependent variables, DVs) may be directly compared with previous marks (the explanatory variables, EVs), either recent or old. For example, std 10 final marks (designated as 10'4 marks) could be compared by subtraction with the previous term's outcomes (i.e. the 10'3 marks) or with the final marks of stds 9 or 8 (i.e. the 9'4 or 8'4 marks). One of the limitations with this approach is that the dependent variable may only be compared with one explanatory variable at a time, and the background variables cannot be used as explanatory variables. The main stumbling block, however, is that tests differ in their means and distribution of marks. So a hard test in one term creates the impression that all pupils are underachieving and the following term's results may look good in that everyone has improved since the hard test. For example, in table 4.1 (p. 50) all three pupils seem to have done well in 9'4 when compared to 9'2. Perhaps the English papers were easy or everyone worked harder for the finals. But when the 9'4 marks are used as a basis for assessing the 10'2 marks, all three appear to have under-achieved.

Table 4.1 Examples of simple indicators for individuals in English

A : Difference in marks B : Difference in Z scores C : Residuals

		High Achiever			Average Achiever			Low Achiever		
DV	EV	A	B	C	A	B	C	A	B	C
8'2	8'1	13	.17	7.3	-42	-.92	-39.1	1	-.07	-3.7
8'4	8'1	6	-.31	-9.4	0	-.22	-6.0	2	-.38	-12.3
9'2	8'1	5	.0	-1.0	-13	-.22	-2.8	-1	-.11	-5.1
9'4	8'1	21	-.16	-4.6	2	-.38	-7.6	0	-.61	-23.6
10'2	8'1	15	-.09	-5.4	0	-.24	4.8	-42	-1.33	-59.3
10'4	8'1	62	1.27	49.8	-20	-.36	1.1	-51	-1.37	-59.2
8'4	8'2	-7	-.48	-21.5	42	.70	32.4	1	-.30	-13.6
9'2	8'4	-1	.31	12.1	-13	-.01	5.7	-3	.26	10.6
9'4	9'2	16	-.16	-5.3	15	-.16	-2.0	1	-.50	-20.0
10'2	9'4	-6	.07	1.0	-2	.14	12.0	-42	-.72	-36.0
10'4	10'2	47	1.36	51.6	-20	-.12	1.8	-9	-.74	-14.9

In regression terms, in the technique of using differences in comparing raw scores a constant slope coefficient of 1 is assumed. As we shall see later, the assumption is invalid.

2. Comparing standard scores

A sophisticated approach would be to convert all the raw marks to standard scores, with the same average and standard deviation, before subtracting the EV from the DV. This adjustment gives a more accurate indication of the change which has taken place. For example, in table 4.1, although all the 9'4 English marks were larger than the 9'2 marks (resulting in a positive difference), the subtraction of the standard scores shows that the differences are all negative. This negative sign

suggests that although the marks were higher in 9'4, the three pupils represented in the table actually did less well relative to the rest of the standard than they had in 9'2. Using standard scores in this way, however, still has two weaknesses: 1) only one EV can be used at a time, and 2) a constant slope coefficient of one (for the standardized scores) is assumed as before.

3.1 Comparing simple regression lines

Simple regressions (ordinary least square) assess the changes in slope on which the definition of effectiveness is based (see chap. 2). As the name "simple" indicates, only one EV is used for each DV. For each pair of explanatory and response variables, the best fitting line may be described by its slope coefficient and intercept or constant. A steep regression line will occur when pupils with very similar prior achievements obtain widely differing outputs. In some circumstances a steep regression line could be a cause of concern, a sign that a subject (or school) discriminates too much between pupils. One would not, for example, expect two pupils who had 8'4 marks of 55% and 60% respectively to achieve 9'4 marks of 30% and 55%. A flat line, especially at a low level of achievement, might also signal the need for investigation. One would not expect pupils of widely differing prior achievements to obtain almost the same outputs. In practice, what constitutes a "good" or "acceptable" distribution of achievement appears to be highly subjective. Where the distribution of the EV is consistent from class to class and year to year, norms may be established. Since the composition of classes (or sets) is unstable at Fish Hoek Senior High School, it was felt that interpretation of differing slopes was not likely to be helpful to teachers whose experience of performance indicators has been limited to crude measures such as pass rates and the proportions of a cohort achieving matriculation exemption and distinctions.

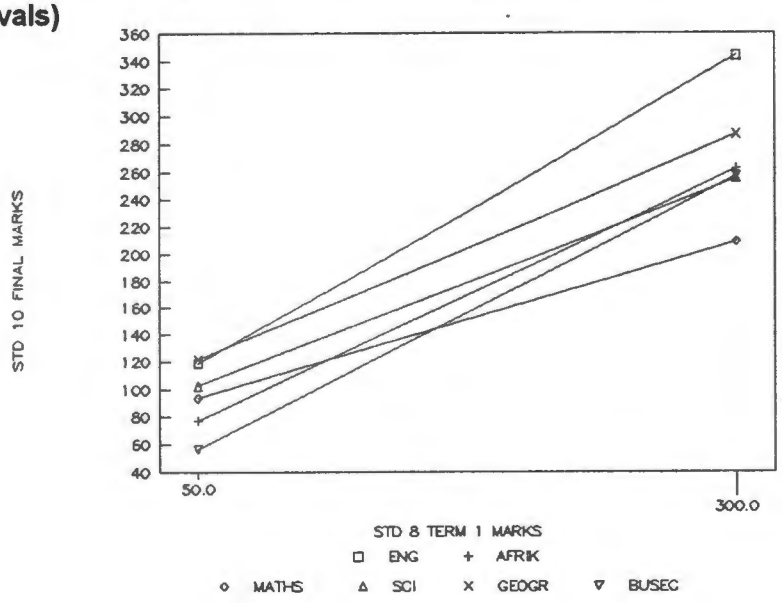
Consider the examples set out in table 4.2 and fig. 4.1 (both on p.52). While all the slopes have coefficients of less than 1 (45°), they vary considerably. The criteria for labelling any of the regressions as too steep or too flat would have to be defined in terms of historical record and the objectives of the school. For example, is it the policy to extend the more able pupils (in which case a steep slope might be required) or is it to ensure that all pupils achieve mastery (for which a flat slope would be expeced)?

Table 4.2 Details of some simple regressions

DV	EV	Subject	Adj. R^2%	Intercept Estimate	t	Slope Estimate	t
10'4	8'1	English	46.8	73.8	3.1	0.90	6.1
10'4	8'1	Afrik.	69.8	40.3	5.7	0.74	17.0
10'4	8'1	Maths	29.6	70.7	5.0	0.46	6.4
10'4	8'1	Science	36.9	72.0	4.4	0.61	6.4
10'4	8'1	Geogr.	49.6	88.8	4.9	0.66	6.6
10'4	8'1	Bus.Ec.[1]	41.9	16.7	0.4	0.80	2.7
10'4	8'1	Agg.[1]	63.0	260.0	4.5	0.77	14.5

[1] All subject EVs had a maximum of 300 except English, which was 400. Aggregates had a maximum of 2 400.

Fig. 4.1 Simple regression lines for subjects (extrapolated to common intervals)



should not extrapolate beyond the observations. Slopes for non-intersecting intervals (unlike those in fig. 4.1) could cloud the interpretation further, particularly if the number of observations varies considerably. In this study, for example, there were 126 pupils for whom there were both 10'4 and 8'1 marks in English but only 11 pupils for whom the same records were available in Business Economics.

Even when separate regressions were fitted to observations for approximately equal size sub-divisions, there was no basis for establishing whether or not the differences were educationally significant. An example of the Aggregates of boys and girls is shown in fig. 4.2 (p. 54).

When comparing the achievements of different ability groups (based on IQ stanines) in Mathematics (see fig. 4.3, p. 54), there could be considerable debate as to whether the mix of slopes or uniformity of slope might be more desirable. In some circumstances, for example, where marks of a subject discriminate strongly between pupils of different abilities, a mix may be more acceptable than parallel slopes, and where a mix of slopes is consistent with objectives, the norms for that mix are nonetheless likely to be uncertain.

Plotting the intercepts and slopes of regressions did not provide more easily interpretable graphs either. See fig. 4.4 (p. 55) for examples. While it was recognized that such graphs could be made more informative, for example, by plotting the t-values of each estimate, it was felt that the problems of interpreting the distribution outweighed the possible insights into the relative effectiveness within each ofof the subjects.

3.2 Residuals - comparing the observed with the expected

Residuals may be used instead of regression lines. A residual is the difference between the observed and expected scores. The residual so calculated may be investigated at the levels of either the individual or groups. Since the units used for the residuals are marks, it is easy to interpret the information relating to a single pupil. For example, in Table 4.1 (p. 50) , the negative residual of 59 marks out of 400 indicates a mark nearly 15% below what might have been expected for this student after the first term in std 8.

Fig. 4.2 Extrapolated regressions Aggregate 10'4 - 8'1 Male / Female

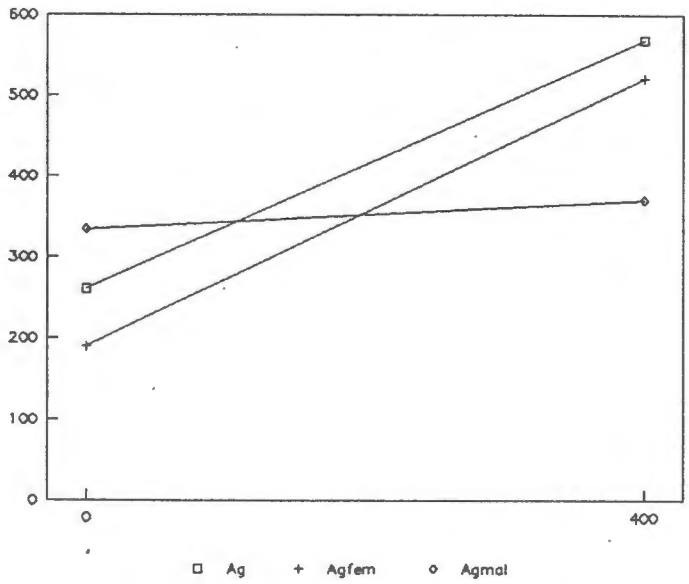
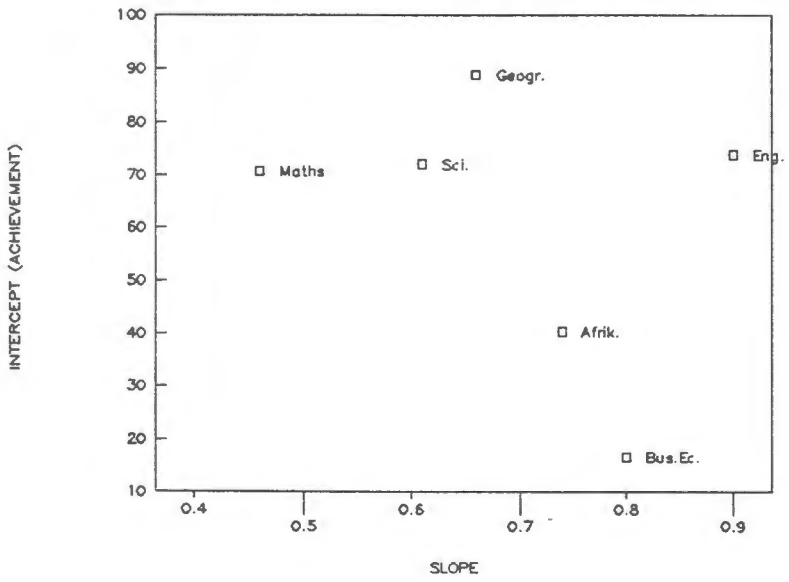


Fig. 4.3 Intercept versus Slope of simple regressions



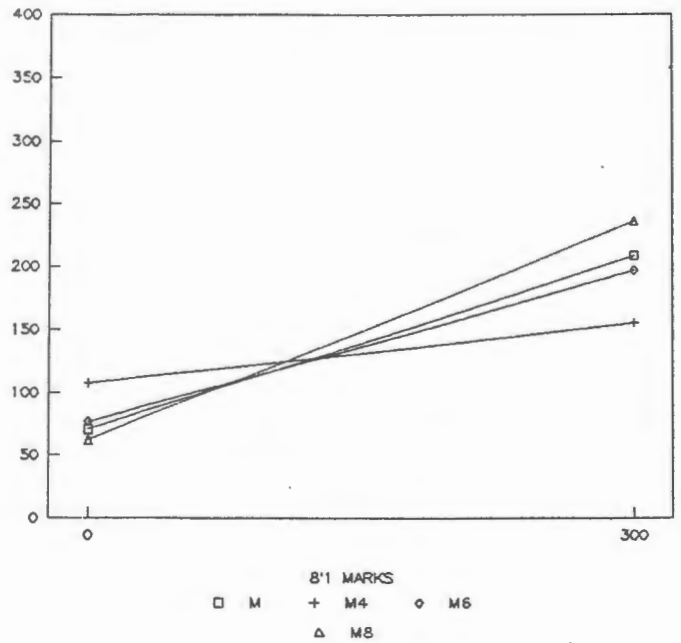
The means of sets of pupils such as gender groups may be compared too. The mean residuals for groups will be non-zero when the groups are performing unequally, after adjustment for prior achievement.

Comparison of individuals or groups on the basis of residuals will only be meaningful in relation to a common regression line.

3.3 Comparing mean residuals

It was found that the residuals may be used in a one-way analysis of variance (ANOVA), which gives a measure of the likelihood of the observed differences between mean residuals occurring by chance, i.e. the statistical significance level of the observed mean residual differences between the groups. For an example, see table 4.3 (p. 56). Note that since the significance level in this example is smaller than 0.0500 the three groups may be said to be statistically significantly different at the 5% level.

Fig. 4.4 Extrapolated Maths 10'4 - 8'1 for three ability groups



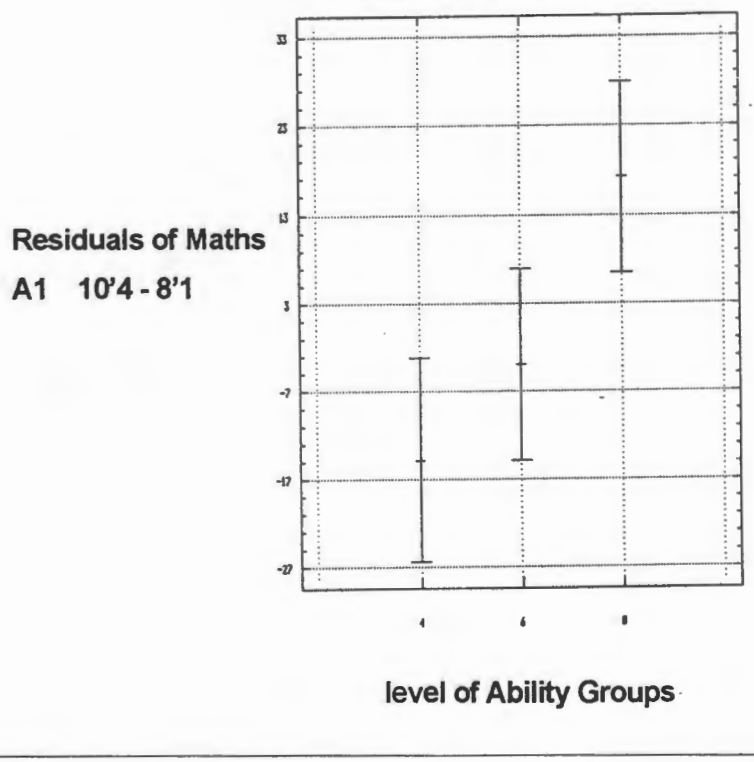
Key : M All Mathematics pupils; M4 pupil with ability less than stanine 6; M6 pupils with ability of stanine 6; M8 pupils with ability greater than stanine 6.

3.4 Comparing the explanatory power of regressions

For each regression calculated one may find the (unadjusted) index of the regression. This index is an estimate, inter alia, of how well the regression line fits the points which result if one plots the DV against the EV. (Yamane, 1967:399). The R^2 index may be adjusted to take the number of EVs being used into account and it is this more conservative index which is used in this study.

Expressed as a percentage, the index may be interpreted as an estimate of the portion of variability in the DV explained by the EV. It was found that by using more than one EV in a multiple regression the R^2 index could be improved in some cases. So, much of the preliminary exploration related to the choice of explanatory variables, i.e. the models which could be used to derive residuals for use in one-way ANOVAs. The modelling was done with the aid of an automatic stepwise regression procedure, with the criterion for selection being F=2.0.

Fig. 4.5 Interval plots for mean residuals of Maths model A1 3 ability groups



After the ANOVA a multiple range test may be used to describe how the various groups relate to each other (table 4.4, p. 57) and the distribution of the residuals may be graphed (e.g. fig. 4.5, p. 58, and fig. 4.6, p. 59). Note that the fact that two of the graphs in fig. 4.5 do not overlap indicates that those groups are statistically significantly different from each other.

The technique outlined above, i.e. analyzing the residuals from simple regressions by analysis of variance, has at least two advantages. First, the (statistical) significance of differences between groups is indicated and this facility may be exploited as a flag for further investigation (see chapter 6). Second, ANOVAs

Table 4.3 Results of an ANOVA of the residuals from a simple regression using Mathematics marks (DV 10'4 marks, EV 8'1 marks) and the three ability groups referred to in fig. 4.4. One-way Analysis of Variance

Data: Residuals from Maths model A1 extrapolated 10'4 - 8'1

Level codes: Three ability

Means plot: LSD Confidence level: 95 Range test : LSD

Analysis of variance

Source of variation	Sum of Squares	d.f.	Mean sqaure	F-ratio	Sig. level
Between groups	16922.52	2	8461.2585	4.263	.0169
Within groups	184585.33	93	1984.7885		
Total (corrected)	291507.85	95			

110 missing value(s) have been excluded.

The preliminary exploration was restricted to a limited number of models and mainly one DV, the final marks in each subject, the 10'4 marks. The findings suggested the following generalizations:

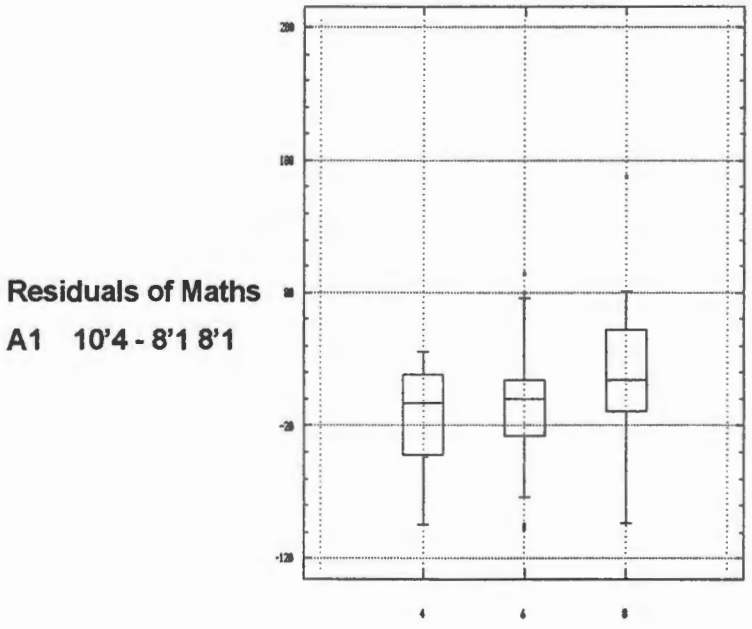
1) Prior achievement, in the form of marks from the same school, is the most powerful EV.

2) Recent marks have more explanatory power than older or more distant marks.

3) The selection of background co-variables is unstable, liable to change both within a subject when the main EV changes, and from subject to subject.

4) effectiveness, i.e. achievement adjusted for previous achievement through the process of finding the residuals, could appear stable if the same main EV was used for a series of DVs. However, if the EV changes regularly the

Fig. 4.6 Box and whisker plots of sets of residuals for 3 Maths Ability Groups



effectiveness will appear unstable, for example, when using lag 1, i.e. the marks of the immediately previous term, as EV.

5) Various groups were found to have statistically significantly different (SSD) mean residuals, but few clear patterns to the differences were discerned. For example, the girls sometimes have a positive mean residual and sometimes negative.

6) Subject set is a key variable. The set is the group of which the pupil was a member and identified by its teacher.

7) Statistically significant differences are not necessarily educationally significant differences. A little investigation may reveal several possible explanations for differences in effectiveness between groups such as subject sets. Professional judgement would be required to decide whether under-achieving pupils or groups need some form of intervention. (Indeed, it may be that the absence of statistically significant differences between some groups, e.g. ability groups, might be a stronger signal for intervention than their presence).

4. Comparing large educational organisations

There are more sophisticated techniques available. In the literature on School Effectiveness Research (SER) hierarchical linear modelling is pre-eminent. However, since this study was restricted to a single cohort within a single school the situation was not considered truly hierarchical. Hierarchical linear modelling is more appropriate for comparing schools from several school districts and different provincial departments within one national system.

Conclusions

After considering these techniques with the equipment and software available at FHSHS and likely to be accessible and used at other schools, it was concluded that regression analysis permits at least two approaches. One is to fit regressions to sub-divisions of the sample, such as genders, and compare the slopes of these lines. An alternative to this multiple slope approach is to fit a single line to the data and to analyze the residuals from this common line.

The common slope approach appeared to have two advantages. One, that the (statistical) significance levels yielded by analyzing the variance of the residuals from a common slope could be used as a flag for professional educational evaluation of sets such as gender groups. Two, the common slope approach permitted the monitoring of individual pupils for anomalous performances.

Since the interpretation of multiple slopes appeared extremely complex (pp.51 - 53) and less promising, it was decided to investigate the common slope approach further.

In the above discussion of regressions, the assumption was made that the relationship between the DVs and EVs was linear. The validity of this assumption could be tested in a larger study. The purpose of the subsequent investigation was, however, to apply the explanatory variables in various models and confirm the conclusions of the smaller preliminary exploration. From the process described in chapter 5 it was also possible to identify some of the patterns in the relationships between explanatory and dependent variables.

Chapter 5. The analysis of adjusted achievement

Introduction

The object of the main investigation was to use a family of regression models to test more thoroughly the conclusions suggested by the preliminary exploration .

After a brief recapitulation of the analysis used, this chapter therefore contains a description of the models and the process by which variables were selected for use in the models to adjust achievement. The models are then compared on the basis of their adjusted R^2 indices and their residuals (i.e. the adjusted achievement) analyzed by ANOVA.

In the second half of the chapter the differences between various groups of the std 10 cohort of 1994 identified by the analysis are described in general terms. The chapter concludes with a reassessment of the conclusions of the preliminary exploration in the light of the wider analysis, and a brief discussion of the trends in effectiveness at FHSHS.

In outline, the analysis involved the following steps:

1) a number of regression models were identified;

2) explanatory variables (EVs) were selected for use in the models;

3) each model was applied (where appropriate) to the six subjects and aggregate, and the residuals saved;

4) each set of residuals was used in 6 one-way ANOVAs employing monitoring variables such as gender and subject set;

5) where the differences between the groups of pupils, such as boys and girls, or the various English sets, were found to be statistically significant at the 95% level a multiple range test was used to analyze the residuals further;

6) summaries were then made of the results in order to compare the models and generalize about the differences identified by the monitoring variables.

The models

The family of models considered in the analysis are summarized in table 5.1 (overleaf) and illustrated in fig. 5.1 (p. 66). From fig. 5.1 it will be appreciated that each subject model could involve from 9 to 12 regressions, if the models were applicable.

In general, the group B models are group A models repeated with additional explanatory variables, usually background variables, but occasionally a second mark variable from the same subjects. The use of marks from other subjects was considered in the initial investigation but found to reduce the sample size dramatically since relatively few pupils took, for example, Mathematics and Science.

The 8'1 marks were used most widely in the modelling because they were the first marks for all the pupils from the same school. About 25% of the pupils came to the FHSHS from schools other than the FHMS -see appendix 4.1. It was hoped that the 8'1 marks would provide a common benchmark measure of achievement at the start of the pupils' careers at the FHSHS. The model groups 1, 5 and 6 (i.e. both A and B) were all intended for use in value-added analyses, i.e. to assess the change that occurred while pupils had been at the FHSHS. Model groups 3 and 4 would assist value-added assessments of changes which occurred while pupils had been in a certain standard. In contrast to the other models, models in group 2 were intended to measure changes in achievement from term to term. Model B2d is unique amongst all the models in using both the latest and the benchmark variables to explain achievement (and as such perhaps deserved a separate classification, such as B7).

Model C1 was intended to derive a benchmark assessment of pupil potential, rather than prior achievement, from the background variables.

The background variables

It was expected that for each subject a different set of background explanatory variables (EVs) would be required. The following steps were used to identify these sets of EVs:

Table 5.1 The models investigated

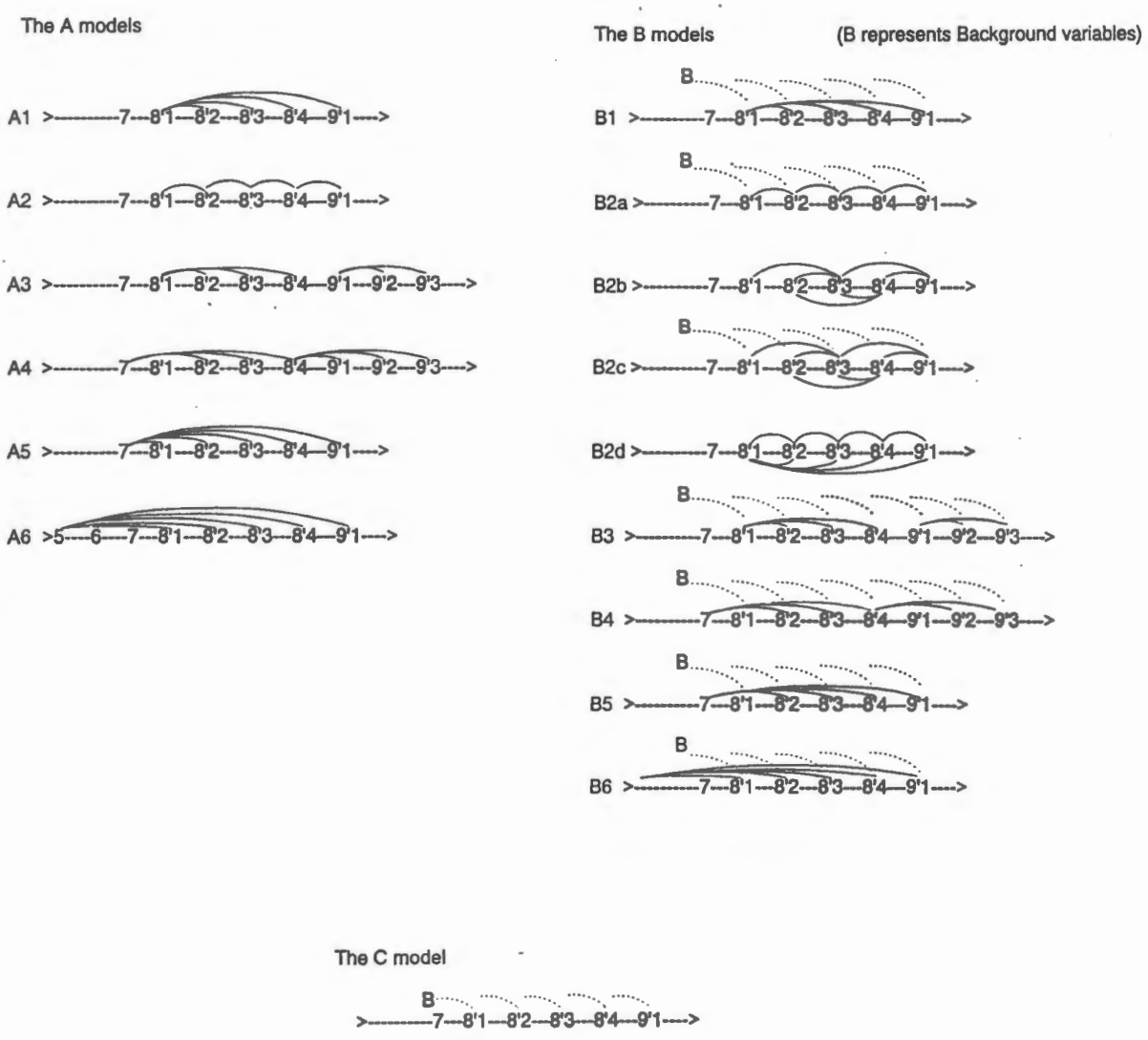
Bak = Background variables

Simple regressions			**Multiple Regressions**		
Group A models			**Group B models**		**The C model**
Model No.	EV	DV	Model No.	EVs	EVs
A1	8'1	8'2-10'4	B1	A1 + Bak	C1 Background only, excluding the monitoring variables.
A2	Lag 1[1]	8'2-10'4	B2a	A2 + Bak	
		8'3-10'4	B2b	A2 + Lag 2[2]	
		8'3-10'4	B2c	A2 + Lag 2 + Bak	
		8'3-10'4	B2d	A2 + 8'1	
A3	Term 1 of same std	8'2- 8'4 9'2- 9'4 10'2-10'4	B3	A3 + Bak	
A4	Term 4 of std before	8'1- 8'4 9'1- 9'4 10'1-10'4	B4	A4 + Bak	
A5	Std 7	8'1-10'4	B5	Std 7 + Bak	
A6	Std 5	8'1-10'4	B6	Std 5 + Bak	

[1] Lag 1 : The mark variable of the term immediately previous to the EV.

[2] Lag 2 : The mark variable two terms previous to the EV.

Fig. 5.1 The models represented on time lines



1) A standard list of EVs was drawn up. It included all the background variables available except the six to be used as monitoring variables. (Erroneously, the std 5, 6 and 7 mark variables were included in the list too, but excluded from the final selection because nearly 30% of the pupils in the data set had not been to the FHMS. This error may have slightly influenced the final selection, but probably not the generalized results. While such an error may be defensible in an exploratory exercise, it would not be admissible in application.)

2) Twelve automatic stepwise regressions were then run for each subject (criterion for selection, F=2.0). The models are described in table 5.2. Only the marks of pupils from the FHMS were used for models T, U and V. See appendix 5.1 for the detailed results.

Table 5.2 Models used to select background variables.

Model	DV	Main EV	Model	DV	Main EV
L	10'4	9'4	S	10'4	5
M	9'4	8'4	T	9'4	5
N	8'4	8'1	U	8'4	5
O	8'4	none	V	7	5
P	10'4	7	W	6	5
Q	9'4	7			
R	8'4	7			

From the selection made by these models, two sets of background variables were identified: 1) those for use when the main EV was a std 8, 9 or 10 mark variable, and 2) those for use when the EV was a std 7 mark variable. See table 5.3 (overleaf). The first set contained all the variables selected by any of the models L, M, N and O, and the second all the variables selected by any of the models P, Q and R.

Table 5.3 Final selections of background variables

5.3.1 Variables for model groups B1 to B4

	Father's occup.	Mother's occup.	Married or single	Suburb	Size of family	Position in family	No. of Schools
Eng.	X			X		X	
Afrik	X	X	X	X		X	
Maths	X	X	X				X
Sci.			X	X	X	X	
Geog.		X	X	X	X		X
Bus.Ec.		X			X		X
Agg.		X	X			X	

5.3.2 Final selection of background variables for model B5.

Eng.			X	X		X	X
Afrik	X	X			X		X
Maths		X					
Sci.			X		X		X
Geog.	X		X	X		X	X
Bus.Ec.	not applicable - no std 7 marks available						
Agg.	X					X	

Two models were not considered after this stage. Very few background variables were selected when the std 5 marks were used as the main EV (models S to W). As a result A5 and B6 were almost identical. B6 was therefore dropped. When trying to select EVs for model C1, the background variables were found to have so little explanatory power that they were not selected without the monitoring variables being available. Only in Afrikaans were any (three) background variables selected. So model C was dropped.

The monitoring variables

Six categorical variables were used in the study. They are :

1) Gender, with two levels, female and male;

2) Ability group, with three levels, namely above stanine 6, stanine 6, and below stanine 6;

3) Age group, with three levels, namely youngest third, middle third, and oldest third;

4) Middle schooling, with two levels, namely those who moved from the FHMS directly to the FHSHS, and the others;

5) Curriculum group, with four levels, namely academic, commercial, Technika and other (see appendix 4.1);

6) Subject set per standard, with varying numbers of levels depending on the number of classes formed.

It was hoped that these variables would be representative of the range of factors which might be used to monitor the effectiveness of a school. Race could be used as a monitoring variable too. In practice it would be preferable to use fewer monitoring variables and make the rest available as background EVs. The selection, however, would depend upon the analyst's interests.

Linear or other functions?

So far, the assumption had been made that the relationship between the EVs and DVs could best be summarized by a straight line, that is a linear relationship or function. When all the subject model simple regressions (the A models) were calculated, the opportunity was taken to test this assumption. The R^2 indices of each regression was calculated for linear (y=a+bx), multiplicative (y=ax^b) and exponential (y=exp(a+bx)) functions. For more than 70% of the regressions, the linear function had the largest R^2 index and thus the best "fit". See table 5.4 (overleaf).

Table 5.4 The number of multiplicative and exponential regressions which fit better than the linear function on the basis of R^2 indices of the group A models.

Total no. of regressions	Multiplicative (1)	Exponential (2)	Either (1) or (2)
371	60 (16,2%)	68 (18,3%)	97 (26,1%)

The models evaluated

The models will be compared on the basis of the R^2 indices of the regressions and in terms of the number of subject model-monitoring variable pairs which proved to be statistically significant. The R^2 indices are a measure of how well the EVs explained the variability in the achievement of pupils. The phrase "subject model-monitoring variable pair" refers to the ANOVAs by which the residuals were analyzed. For example, the residuals from English model A1-8'2 analyzed by ANOVA for gender is a pair. It may be that in each pair there are statistically significant differences between the levels of the monitoring variable - in this pair between the boys and the girls. The frequency with which these pairs appear to be statistically significantly different (SSD pairs) was analyzed.

In the following discussion "subjects" will be used to refer to the subject and aggregate variables where the procedure applies, for instance there are Aggregate model - gender pairs but no Aggregate model - set pairs. All R^2 indices are adjusted where appropriate (for the B models) and expressed as %.

The average sizes of the R^2 indices of all the regressions derived from each model are set out in table 5.5 (overleaf). The average for model A1 of 57.5%, for example, is the average of all the A1 subject models (i.e. 8'2 8'1, 8'3 8'1 .....10'4 8'1 regressions for English, Afrikaans .... Aggregate). See appendices 5.2 and 5.3 for details.

Table 5.5 The average R^2 indices by models

Mode	Avg. R^2%	Model	Avg. adjusted R^2%	*n* for A and B model where applicable
A1	57,5	B1	56,3	77
A2	71,9	B2a	71,7	77
		B2b	74,6	70
		B2c	74,4	70
		B2d	73,1	70
A3	66,2	B3	66,1	63
A4	62,5	B4	62,2	80
A5	50,1	B5	51,4	72
A6	44,2			72

Since multiple regressions (the B models) use more explanatory variables one would expect them to have more explanatory power and to have larger R^2 indices than the A models. Improvement in the explanatory power is after all the reason for using the B models. The R^2 indices are, however, adjusted for the number of EVs used and R^2 indices for the group B models are penalized or adjusted downwards for the inclusion of many background variables -see the account of the selection of the EVs earlier in this chapter.

While the R^2 indices of both the A and B group models are very similar it will be shown later that their residuals yield different results when analyzed by ANOVA (see tables 5.7 and 5.15 on pp. 73 and 81).

As fig. 5.1 (p.66) shows, the interval (in time or terms) between DVs and EVs differs by model. The most extreme difference is found between A2 (which uses the previous term as EV) and A6 (std 5 marks). In table 5.6 (overleaf) the models are ranked from those with the oldest EVs to the most recent.

The table shows that the shorter the lapse of time between the DV and the main EV, the larger the R^2 index, i.e. recent EVs have more explanatory power than old EVs. The table also shows the average frequencies of statistically significantly different pairs (SSD pairs) at the 5% level by model. It will be recalled that for each subject model there were a number of regressions - see *n* in table 5.5. The residuals

from these regressions were used in six ANOVAs, one for each monitoring variable. The number of ANOVAs which identified statistically significant differences between the levels concerned (e.g. genders) were counted. They are expressed in table 5.6 as percentages of *n*, the total number of all the ANOVAs done per model, i.e. the number of regressions per model times the number of subjects times the number of ANOVAs. See appendices 5.2 and 5.4 for details.

Table 5.6 Models ranked on age of the explanatory variables, from the oldest to the most recent.

Model	Avg. R^2%	SSD Pairs Avg. % frequency (5% level)	Model	Avg. R^2%	SSD Pairs Avg. % frequency (5% level)	*n* for A & B models where applicable
A6	44,2	34,9	-			348
A5	50,1	35,4	B5	51,4	34,0	348
A1	57,5	36,8	B1	56,3	35,1	451
A4	62,5	25,7	B4	62,2	25,3	444
A3	66,2	33,3	B3	66,1	31,1	366
A2	71,9	21,6	B2a	71,7	20,5	448
			B2d	73,1	16,1	408
			B2c	74,4	17,1	408
			B2b	74,6	17,5	408

The general relationship between the size of the R^2 indices and the frequency of SSD pairs is negative. As the index gets bigger, the frequency of SSD pairs drops. This relationship may suggest that the fewer SSD pairs yielded by a model the more likely those SSD pairs are to be important and educationally significant. The converse would be that a model with a low R^2 index will yield a high frequency of SSD pairs. However, much of the variability was unexplained by the EV so that many of the differences identified by SSD pairs might be explained by variables other than the monitoring variable. On the criteria of large R^2 index and low frequency of SSD pairs, the B2 models seem to be most useful, particularly B2d.

Table 5.7 Subjects ranked by sample size, with the average frequency of SSD pairs by model groups A and B.

	No. of pupils term 1,4	Average frequency of SSD pairs as % Model As	Model Bs	Total
English	152	53,5	36,1	44,9
Aggregates	152	43,6	28,5	36,1
Afrikaans	152	39,6	30,1	34,9
Maths	120	26,7	26,2	26,5
Science	92	24,1	9,7	21,9
Geography	71	26,5	16,1	21,4
Bus. Econ.	26	5,3	7,4	6,5
n		2116	2778	4894

Whether compared on the basis of their R^2 indices or SSD pairs, there is little difference between the group A and B models, except with respect to the group 2 models. The use of additional EVs, either background or mark variables, in conjunction with the latest marks yields appreciably lower frequencies of SSD pairs than model A2. Model B2d achieves the lowest frequency of SSD pairs, 25% less than A2, without the use of background variables.

In addition to the size of the R^2 index, the frequency of SSD pairs is also influenced by the sample size, i.e. the number of pupils taking each subject. In table 5.7 the subjects are ranked from largest subjects to smallest (and from highest frequency to lowest amongst the three subjects of equal size). The smaller the group the lower the frequency of SSD pairs. The number of pupils and the total frequency of SSD pairs by subject model are strongly correlated. This relationship may suggest that the differences found within larger subjects may exist in the smaller subjects too, but they are not identified by the analysis. The principle illustrated in table 5.7 has important implications for analysts who might wish to compare subjects, or any units of analysis of different sizes using the frequency of SSD pairs. Comparisons could only be considered between units of equal or large size. With regard to this study, consider the apparent anomalies in table 5.6. On the basis of the average R^2 indices model groups 1 and 3 have frequencies of SSD pairs which are too high, or model

groups 4, 5 and 6 are too low. Groups 1 and 3 were based upon the pupils who had attended FHSHS from std 8, while groups 4, 5 and 6 only used pupils who had attended the FHMS, a smaller number. If the numbers of pupils had been of equal size the frequencies of SSD pairs for model groups 4, 5 and 6 might have been larger.

Table 5.7 also shows much larger differences between the subject model groups A and subject model groups B than appear in table 5.6. (p.72). For details see appendix 5.4. The sizes of the differences suggest that using background variables in the subject models improves the power of the analysis to identify subject model-monitoring variable pairs which should be investigated by teachers.

While comparing models, it should be reported that the group 2 models create the impression that achievement is very unstable. When ranking of sets and other groups is discussed below, it will be shown that a set may go from first to last and back in the space of three terms. This apparent variation is because a new EV is used each term. If a group or individual did well one term, then in relation to that good term it must appear as if there had been underachievement the next term. Even a sustained high level of achievement is unrecognized.

Generalizing from the ANOVAs

The ease with which the results of the ANOVAs may be interpreted depends upon the number of levels in the categorical variable being used as the monitoring variable. Some monitoring variables only have two levels, e.g. gender and middle schooling. Some have three levels (age groups and ability groups). Curriculum group has four and subject sets are often more numerous than four. The results of the ANOVAs will be reviewed in that order, from simple to complex situations.

The approach was to establish the general pattern or distribution of residuals. Results could then be compared against the general pattern and unusual cases identified.

Two examples will illustrate the findings with regard to gender. The aggregate model-gender pairs showed the most consistent pattern up to the second term of std 9. As table 5.8 (overleaf) and appendix 5.5 show, the boys tended to have

positive mean residuals and the girls negative. After 9'2 the differences between boys and girls cease to be statistically significant.

Table 5.8 Mean residuals of Aggregate model-gender SSD pairs all at the 5% level

Model	DV	Main EV	Girls' avg.	Boys avg.	Diff.	Model	DV	Main EV	Girls' avg.	Boys' avg.	Diff.
A1	8'3	8'1	-29.4	27.1	56.5	B1	8'3	8'1	-29.7	24.8	54.5
	8'4	8'1	-23.3	22.7	46.0		8'4	8'1	-21.8	21.5	53.5
A4	9'3	8'4	22.5	-20.6	43.1	B2b	8'3	8'2	-16.9	15.6	32.5
	1'2	8'4	17.1	-15.6	32.7		1'2	1'1	13.4	-12.3	25.7
A5	8'2	7	-47.2	38.2	85.4	B2d	1'3	1'2	-14.5	15.5	30.0
	8'3	7	-55.4	44.8	100.2	B3	8'3	8'1	-26.7	24.8	51.5
	8'4	7	-53.6	44.6	98.2		8'4	8'1	-21.8	21.5	43.3
	9'1	7	-46.3	38.2	84.5	B4	8'2	7	-46.1	38.3	84.4
	9'2	7	-40.6	33.5	74.1		8'3	7	-56.1	46.6	02.7
A6	8'2	5	-36.5	28.2	64.7		9'3	8'4	21.7	-20.0	41.7
	8'3	5	-45.6	35.2	80.8		1'2	9'4	16.9	-15.8	32.7
	8'4	5	-40.9	33.0	73.9	B5	8'2	7	-42.4	34.2	76.6
	9'1	5	-41.4	33.5	74.9		8'3	7	-50.5	40.7	91.2
							8'4	7	-48.8	40.5	89.3
							9'1	7	-41.5	34.2	75.7
							9'2	7	-37.2	30.6	67.8

The English model-gender pairs reflected the opposite pattern. All the model group 1 pairs showed statistically significant differences between the girls and boys but here the girls had positive mean residuals. See table 5.9 (p. 76).

The other English models yielded far fewer SSD pairs. After English, Afrikaans had the most SSD pairs (see appendix 5.2), but the advantage was shared equally between the genders. Afrikaans also differed from English in that none of the group 1 model-gender pairs were SSD, whereas all the English pairs were.

Middle schooling was also a two-level monitoring variable. Relatively few of the subject model-middle schooling pairs were SSD. Where the SSD pairs occurred, the

Table 5.9 Mean residuals of English A1 and B1 model-gender SSD pairs, all all the 5% or the 1% level (*)

	Model A1 (EV : 8'1)			Model B1 (EVs : 8'1 & background)		
DV	Girls	Boys	Diff.	Girls	Boys	Diff.
8'2	8.0	-7.3	15.3*	7.5	-6.7	14.2*
8'3	4.8	-4.3	9.1*	4.7	-4.2	8.9*
8'4	4.4	-4.2	8.6*	4.1	-3.8	7.8*
9'1	5.1	-5.1	10.2	5.8	-5.6	11.4*
9'2	8.1	-8.0	16.1*	8.6	-8.3	16.9*
9'3	6.7	-6.4	13.1*	6.4	-6.1	12.5*
9'4	8.4	-8.0	16.4*	8.6	-8.0	16.6*
10'1	6.4	-6.5	12.9	6.2	-6.2	12.4
10'2	8.6	-8.9	17.5*	8.8	-9.0	17.8*
10'3	4.8	-5.0	9.8	5.0	-5.1	10.1
10'4	10.4	-10.9	21.3*	10.9	-11.3	22.2*

pupils who had attended FHMS usually had positive mean residuals. See table 5.10 (p. 77) for an example, and appendices 5.2 and 5.6. (It should be noted that about 70% of the pupils were from the FHMS).

Age groups and ability groups both had three levels and both had clear patterns. The youngest group tended to have a positive mean residual and the oldest a negative one, with the middle age group close to zero. See table 5.11 (p. 77) and appendix 5.7. Similarly, the most able group had positive means and the less able group negative means, again with the middle group near zero. See table 5.12 (p. 78) and appendix 5.8.

Since the residuals represent what remains of achievement after adjustment for prior achievement, the pattern illustrated in table 5.12 apparently implies that the longer pupils are at school the wider the gap in achievement becomes between the most and least able groups. Schooling does not seem to compensate for initial inequalities amongst pupils of different abilities. One should, however, remember that model A1 uses 8'1 marks as the main EV, which are fairly distant from the DVs in std 10 (see table 5.6 (p.72) and the accompanying discussion for the consequences of using 'old' EVs). The group 2 models yielded much lower frequencies of SSD subject

model-ability group pairs than the model 1 groups. See table 5.13 (p. 79). Value-added models using a distant benchmark variable may thus exaggerate the differences between the most and least able groups. A similar caution should be applied when inter]preting results such as those in table 5.11, i.e. with regard to age groups.

Table 5.10 Mean residuals of Mathematics model A1 - middle schooling pairs, all significant at the 5% or the 1% level (*)

	Model A1 (EV : 8'1)	
DV	FHMS	Other
8'3*	5.1	-30.8
8'4	3.9	-21.7
9'3	4.1	-12.1
10'3	3.6	-24.2

Table 5.11 Mean residuals of Aggregate model A1 - age group pairs, all significant at the 5% or the 1% level (*)

	Model A1 (EV : 8'1)		
DV	Youngest	Middle	Oldest
8'2	21.0	11.2	-38.4
8'4	24.6	3.6	-40.4
9'2*	36.4	-3.1	-52.4
9'3*	40.9	-11.5	-48.5
10'2	40.6	-16.6	-47.7
10'3	39.7	-16.4	-44.8
10'4*	42.0	-16.6	-50.5

Table 5.12 Mean residuals of Aggregate model A1 - ability group pairs, all significant at the 5% or the 1% levels (*)

Model A1 (EV : 8'1)

DV	Stanine groups >6	6	<6
8'2 *	64.4	-6.5	-33.1
8'3 *	66.9	-13.1	-21.2
8'4 *	76.9	-11.3	-34.3
9'1	54.1	9.1	-31.1
9'2 *	63.5	-5.7	-31.5
9'3	50.4	-16.4	-13.7
9'4 *	65.2	-9.4	-28.7
10'1	48.1	6.1	-34.6
10'2 *	64.8	-1.0	-47.0
10'3 *	75.1	2.1	-56.2
10'4	73.4	-9.8	-46.3

Curriculum groups and subject sets had more than three levels, so ranking (on mean residuals) and multiple range tests were used to describe the distribution of achievement across the various sub-sets, e.g. curricular groups. Four curricular groups were identified. On the basis of distant or old EVs, the ranking and homogeneity of groups was stable. See table 5.14 (p. 80) for an example, and appendix 5.9. Group 3 was always in the first rank and groups 2 and 0 in the last rank.

In contrast, on the basis of model B2a there were far fewer SSD pairs and the ranking was very unstable. In the SSD English and Afrikaans model-curriculum group pairs the rankings differed considerably from the pattern for Aggregate model A1 (in table 5.14). For example, in Afrikaans model group 1 SSD pairs curriculum groups 1 and 2 (Academic and Business) were always a homogeneous group with larger mean residuals than 3 and 0 (which were also consistent members of a homogeneous group, see appendix 5.9). In the English model group 5 SSD pairs, the ranking was usually 1, 0, 2 and 3.

Table 5.13 Frequencies of SSD subject model-age group and ability group pairs for two model groups (n = 77 for each model)

Subject model-age group pairs

Model	5%	1%	Model	5%	1%
A1	13	20	B1	14	14
A2	10	2	B2a	8	4

Subject model-ability group pairs

Model	5%	1%	Model	5%	1%
A1	16	27	B1	19	18
A2	8	12	B2a	6	12

The ranking therefore changed with the model and the subject. Within subject models, however, sets which might require further professional investigation can be identified.

Finally, the interpretation of the analyses using subject set as the monitoring variable must be discussed. Generalizations about the subject model-subject set pairs are complicated by two characteristics of the sets. First, the number of levels. English in std 10 had as many as seven sets. Second, the pupil composition of the sets and their teachers could change from standard to standard. The response to the first problem was to identify the outlier subject sets rather than to monitor the rankings. The second circumstance meant that patterns could be sought mainly within standards.

The concept of homogeneous sets is illustrated in table 5.14 (p. 80) For example, in the last four pairs of model A1 (i.e. DVs 9'3, 10'1, 10'2 and 10'3), groups 1, 2 and 0 are members of a homogeneous group. If all the model A1 pairs in the

5.14 Ranking and homogeneous groups of SSD Aggregate model-curriculum group pairs, all significant at the 5% or 1% level (*).

DV	Model A1 (EV : 8'1)	Model B2a (EVs : previous term & background)
8'2	3 1 0 2* X X X X	3 1 0 2* X X X X
8'3	3 1 0 2* X X X X	
8'4	3 1 0 2* X X X X	
9'1	3 1 0 2* X X X X	
9'2	3 1 0 2* X X X X X X	
9'3	3 1 2 0 X X X X	
10'1	3 1 2 0* X X X X	3 2 1 0* X X X X
10'2	3 1 2 0 X X X X	
10'3	3 1 0 2 X X X X	1 0 3 2* X X X X

Note: Ranking from largest (on the left) to smallest mean residual.
Key: Curriculum groups 3 = Technika, 2 = Business 1 = Academic, 0 = Other.

Table 5.15 Frequency of all model A and B pairs significant at 5% level by monitoring variables, as %

Model group	Gender	Ability Group	Age Group	Middle School	Curriculum Group	Subject Set
A	12.0	51.9	29.7	13.9	24.0	61.4
B	15.0	31.8	22.0	14.4	17.0	55.5

table are considered, only 0 and 2 are consistent or common members of the lowest ranking set. They could be described as the lower outliers, while set 3 would be the upper outlier.

The frequency of SSD subject model-subject set pairs is high, especially in the languages, with many significant at the 1% level (see table 5.15 (p. 81) and appendix 5.10). The pattern of outlier subject sets is illustrated in tables 5.16 and 5.17 (pp 82 and 83).

Note that the rankings yielded from the A2 models is too unstable over four terms to identify consistent outlier sets. Also note that Afrikaans set 7 was a First Language HG class. The set may have been identified as the upper outlier because their marks were out of 400 rather than 300, or because they had been selected on the basis of their high achievement. Coincidentally, the std 10 English set 7 was also a small group, but of First Language SG pupils. They too had selected themselves on the basis of past achievement, or more precisely, poor past achievement. Their marks were out of 300 whereas all the other marks were out of 400. Set 7 is a lower outlier perhaps as an artifact of their selection, or because of the difference between the maximum marks, or there may be some other reason.

The educational significance of the differences between these sets will be discussed in the next chapter.

Table 5.16 Consistent members of homogeneous outlier groups of SSD English model A1 - English sets, by standards

Model		Std 8	Std 9	Std 10
A1	Upper	9	3	3
	Lower	1,4,10	4	4,11
A2	Upper	6	6	X
	Lower	1,4,10	X	7
A3	Upper	9	6	3,6
	Lower	1,4,10	1,2,5	1,4,11
A4	Upper	6	1,3	1,11
	Lower	1,2,4	4	7
A5	Upper	6	6	6,11
	Lower	1,4	4	4,8
A6	Upper	6	6	6
	Lower	1,2,4,10	4	7

X No consistent or common member of the homogeneous outlier-group over one year.

Table 5.17 Consistent members of homogeneous outlier groups of SSD Afrikaans model A1 - Afrikaans sets, by standards

Model		Std 8	Std 9	Std 10
A1	Upper	7	3,7	3
	Lower	3,4,5,6	1,2,6	2,5,6
A2	Upper	X	X	X
	Lower	X	X	X
A3	Upper	7	~	6
	Lower	3,4,5,6	~	1,5
A4	Upper	7	1,3,7	2,3
	Lower	3,4,5,	2	X
A5	Upper	7	7	X
	Lower	3,4,5,6	2	X
A6	Upper	7	7	7
	Lower	3,6	1,2,6	6

X No consistent or common member of the homogeneous outlier group over one year.

~ No SSD pairs in any of the terms of that year.

The findings of the preliminary exploration evaluated

The following findings of the preliminary exploration have been confirmed:

1) Prior achievement is the most powerful explanatory variable;

2) Recent mark variables have more explanatory power than older or more distant mark variables. In addition, the use of distant EVs may exaggerate the differences between groups within a variable;

3) The selection of background variables does change with subjects and models;

4) Effectiveness, or adjusted achievement, appears less stable in some models than in others, depending upon the frequency with which the main EV is changed;

5) Subject set is a key variable in that it differentiates between collective effectiveness of groups of pupils very frequently.

The finding that there are "few clear patterns" in the SSD pairs should be revised. With many more SSD pairs to generalize from, it has been easier to identify some of the standard patterns in the distribution of adjusted achievement, i.e. effectiveness.

Additional findings:

1) Great care should be exercised when considering the merging of Higher and Standard Grade marks in a single school context. Against the disadvantages of statistically evaluating sets incorrectly must be weighed the statistical advantages arising from large samples;

2) Without the monitoring variables (such as age and ability) and middle school marks, no useful set of background EVs could be found. As a result, regressions and ANOVAs were not practical using background variables alone;

3) Linear regressions usually fit the data better than multiplicative or exponential regressions.

4) The larger the R^2 index, the fewer SSD pairs are identified;

5) The use of background variables slightly reduces the frequency of SSD pairs identified;

6) The larger the number of pupils taking a subject, the more likely the analyses of variance are to find statistically significant differences;

7) The greater the number of levels in the monitoring variable, the more difficult it becomes to generalize from the results of the ANOVAs.

The limitations of this study

These findings and the conclusions drawn later need to be qualified. This study was based upon a single, perhaps atypical cohort of pupils (see the distribution of ability, table 3.1, p.40) of a co-educational and predominantly white 'Model C' school, i.e. where fees are required, attended by pupils for stds 8, 9 and 10 only. Middle school marks were from schools other than FHSHS, predominantly the FHMS. The socio-economic background of the pupils is relatively homogeneous. Such a limited sample is an inadequate basis from which to extrapolate about school effectiveness. This study was never intended to contribute directly to the findings of the large-scale School Effectiveness Research (SER), but only to adapt the techniques of SER to school leadership at the local level. The intention has been rather to establish procedures which could alert professional educators to phenomena which might otherwise go undetected, rather than to apply them exhaustively in a school-wide audit.

Even as an investigation into FHSHS, the study is limited. For example, only a few of the subjects offered by the school were included, and while more than 150 pupils took the languages, there were far fewer in some of the other subjects. There was only limited overlap in the pupil composition of the smaller subject groups.

The limited overlap between subjects influenced the modelling in that it was not practical in this study to use other subjects' mark variables as EVs. For example, Mathematics and Science marks could not be used as EVs of apparent achievement in Geography because very few pupils took all three subjects.

Fish Hoek trends in relation to SER findings

Although testing the findings of SER was not the purpose of this study, there are some interesting comparisons to be drawn between the patterns noted here and the findings of SER. In addition to the limitations noted above, it should be born in mind that the studies quoted here are all multi-school studies, and deal with various phases of education.

Generally, effectiveness within subjects at FHSHS is unstable from term to term. This observation agrees with the finding of Luyten (1994) that school effectiveness is neither monolithic nor uniform across subjects. There is also evidence that at least at a statistical level FHSHS was not equally effective for all groups of pupils (e.g. in languages). This finding is consistent with the findings of Nuttall *et al* (1989) that schools discriminate between various groups. In this study prior achievement proved to be the most powerful explanatory variable of current achievement, as it was found to be by Mortimore *et al* (1989). These authors also found that achievement (in primary schools) was positively related to age, whereas in this FHSHS cohort (without removing the influence of the repeating pupils) the relationship between age and achievement was negative.

Conclusion

The main study has been described in this chapter. The emphasis has been upon the techniques used and the patterns or relationships observed. The relationships identified in the study have augmented and confirmed most of the findings of the preliminary exploration. The most important aspect of the whole project has still to be discussed, namely the relevance of statistical significance in educational decisions. The question of how educational practitioners might interpret or employ the information produced by regression analysis is the subject of chapter 6.

Chapter 6. Professional judgement

Introduction

An information system employing tools such as regressions and analyses of variance requires the school leadership to exercise professional judgement. Decisions must be made both while designing the system and, once the indicators are available, judgement must be exercised in interpreting them. The analysis cannot replace professional decisions. In this chapter an algorithm of the procedure followed when applying regressions and ANOVAs summarizes the analysis, and the key choices in the modelling process are identified. Most of the chapter is, however, devoted to the interpretation of the results of the analysis. It will be argued that statistical significance is best regarded as an indicator to educators for professional response in the form of debate and understanding rather than a threshold beyond which praise or blame, rewards or intervention should necessarily follow. Examples of curriculum groups and language sets will be used to demonstrate the role of professional judgement in interpretation

Modelling

The procedures for quality assurance through regression analysis (or QATRA) are set out in table 6.1 (overleaf). In applying the algorithm professional judgement will be required in the choice of models and monitoring variables. These choices will be dictated by the purpose for which the educational indicators are required.

With respect to modelling, if the purpose is to detect changes in achievement as soon as possible then the previous term's marks should be the main explanatory variable (EV, as in model group 2). If the intention is to assess added value over longer terms then other model groups would apply. See table 6.2.

The purpose of the system will also determine which factors are to be monitored. Categorical variables such as subject set, gender and race could be considered. Set would be important in large schools to check that all classes progress equally. Gender and race would be important to monitor the equity of the school's effectiveness. If continuous variables such as age or ability (measured by IQ) are to

Table 6.1 QATRA Algorithm

1. Clearly express the purpose of the analysis.
   See **Interpretation** (p. 91) and **Conclusions** to this chapter for restrictions..

2. Determine the models to be used, i.e. decide which DVs, monitoring variables and group of EVs will be used.

3. Collect the data and calculate the descriptive statistics, such as means, minima and maxima. These statistics give some insights into the data and a check on the accuracy of the collection process.

4. Decide whether Higher and Standard Grade variables are to be merged, or treated separately. Where merging is used, calculate the descriptive statistics again.

5. Establish which subject and background variables are to be used as EVs with the aid of a standard list of background variables and stepwise regression procedures. The DV for the process could be the first variables to be analyzed, e.g. 8'1, or the final mark variables for each standard if they are available.

6. Apply the models to the DV in each subject and:
   - 6.1 use the resulting adjusted achievement (residuals) with ANOVA/s employing the monitoring variables;
   - 6.2 where the monitoring variable has more than two levels, use a multiple range test to investigate the distribution of adjusted achievement;
   - 6.3 identify all individuals with adjusted achievements larger or smaller than some criterion level, for example, 5% of the maximum possible DV mark.

7. Investigate the results.
   - 7.1 The SSD subject model - monitoring variable pairs, to see if they are educationally important. Take action if necessary and practical.
   - 7.2 The individuals who have the very large or very small adjusted achievements. Praise or intervene where necessary and practical.

be used then important decisions have to be made about the number and size of the statistical classes to be used when recoding.

Another area calling for professional judgement relates to the selection of background EVs. Which variables should be considered and how should they be selected? For example, in the procedure followed in this study (see chapter 5) choice was exercised in the selection of the models used in the stepwise regressions. The analyst may also control the order in which variable are offered for selection.

In coding some of the background categorical variables, such as suburb or occupation, more choices are required of the analyst. For example, care needs to be taken to ensure that suburbs which are grouped together are in fact similar in terms of the characteristic educational backgrounds of the pupils who live in them. In other words the classifications used must be defensible.

One further restriction on the modelling process should be observed to avoid the technical problem of multicollinearity. When selecting explanatory variables (EVs) one should avoid using as an EV any aggregate with *all* its components. Dunne[1] comments upon the use of an aggregate and its components as follows: "This usage will create a rank deficiency in the matrix of observed EVs, and while the resulting residuals from a regression analysis may be examined as previously, the slope coefficients are essentially unspecified (not exactly determined). A host of possible solutions for the coefficients will exist and any chosen set of slope coefficients is essentially arbitrary."

The use of a common intercept term along with several intercept terms for subgroups or sets constitutes a similar collinearity issue. Only the differences between set intercepts are meaningful and not the estimated coefficients themselves. The estimated coefficients are meaningful only when the common intercept term is dropped from the model.

The discussion so far has related to exact collinearity. Near collinearities can also give rise to unreliable coefficients. When several measured variables (from

1 Assoc. Prof. T. T. Dunne, Department of Statistical Sciences, UCT, personal communication.

absolute scales) are used the analysis should be done with care. However, it is unlikely in the context of marking, where values may range between 0 and 400 at most and 0 and 100 at least, restricted to integer values, that near collinearities will arise in the somewhat large data sets envisaged by this type of study.

Once the data is ready and the system begins to generate the indicators, the more difficult task of distinguishing between statistical and educational significance begins.

Interpretation

Before the examples of SSD pairs are considered, two general warnings should be noted. First, explanatory power (reflected in the R^2 indices) does not necessarily indicate causality. A powerful explanatory variable is one in which change is closely related to change in the dependent variable. For example, high marks in one term may be good grounds to expect good marks in the following term, and similarly, low marks would explain low marks next term. The first term's marks are not, however, the cause of the following results unless the skills are cumulative. One cannot do advanced trigonometry without a knowledge of the basic ratios, for example. However, where new work is not directly based on previous learning, then the prior achievement does not cause the subsequent achievement. Carefully controlled experiments are required to show causality (Clegg, 1982). Secondly, comparisons between subjects or cohorts (such as, English has more SSD subject sets than Afrikaans, or English discriminates more between boys and girls than Afrikaans) may only be made if each has similar numbers of pupils and equally well-fitting regressions. These restrictions arise from the observation (in chapter 5) that the frequency with which subject model-monitoring variable pairs yield statistically significant differences is an artifact of the size of the R^2 index and the size of the sample. The index is a measure of the extent to which a regression line fits the data, and is seldom the same for any two pairs of variables. The size of the sample is simply the number of pupils for whom EVs and DVs are available. It was noted that the larger the R^2 index, the smaller the percentage of SSD pairs, and the smaller the sample, the less the likelihood of SSD pairs. Comparisons should therefore only be considered if the subjects have similar numbers of pupils and the regressions fit equally well in both subjects. Since the second of these conditions will seldom be

Table 6.2 Value-added models

Model group *	Main EV	Assess change over:
6	Std 5	High schooling
5	Std 7	Senior secondary phase
4	Last term of previous year	A year

* As defined in this study. See fig. 5.1 and table 5.1.

Table 6.3 Characteristics and achievements of two outlier curriculum groups (n = 22)

	Technika	Business
Actual mean std 7 aggregates %	54.4	56.1
Mean residuals model A1 when DV was 8'4 aggregates	125.9	-39.4
Average age in relation to the standard average	+3 mths 15 days	+4 mths 27 days
% pupils > stanine 6	31.8	13.0
% female	9.1	91.7
Mean residuals model A1 when DV was 10'3 aggregate	83.9	-62.5
Difference between mean residuals	146 marks or >7%	
Actual means 10'3 aggregate	999.3	958.33
Difference between actual means	41 marks or 2%	

met, comparisons of the effectiveness of any two units, such as subjects, may not be validly made.

There will be a great temptation to attribute the success of positive outliers and the weakness of the negative outliers to the associated teachers. Similarly, subjects or courses whose pupils have large, positive adjusted achievements will wish to claim some of the responsibility. The following examples suggest that even in the rare event of comparisons being admissable, it is unlikely that solid grounds will be found for ranking subjects or rewarding some teachers while pressing others to improve. While the explanatory variables take some of the background differences into account, a little investigation will often reveal other possible explanations for differences between groups. Tables 6.3, 6.4 and 6.5 show how much some outlier groups differ from each

Table 6.4 Characteristics and achievements of outlier language sets (n > 17)

Subject	English		Afrikaans	
Set	3	4	3	5
Actual mean std 7 marks, %	62.3	55.5	64.2	49.6
Mean residuals model A1 when DV was 8'4 mark	13.7	-4.2	8.8	0.3
Average age of set in relation to std average	-3 mths 26 days	+2 mths 23 days	-5 mths 22 days	+3 mths 28 days
% of pupils > stanine 6	36.0	0.0	66.7	3.3
% female	67.9	36.4	76.0	40.1
% academic	40.7	16.6	66.7	7.4
% Technika	7.4	33.3	4.6	22.2
Mean residual model A1 when DV was 10'4 mark	23.8	-4.9	11.9	-4.4
Difference between mean residuals	28.7 or 7%		16.3 or 4.1%	
Actual mean 10'4 mark	226.0	175.2	189.7	146.1
Differences between means	50.8 or 12.7%		43.6 or 10.9%	

Table 6.5 Characteristics and achievement of two Afrikaans sets with same teacher

Set	3	1
Actual std 7 means, %	64.2	48.8
Average age of the set in relation to std average	-5 mths 22 days	-1 mth 30 days
% pupils > stanine 6	66.7	4.2
% female	76.0	40.0
% Academic	66.7	22.7
% Technika	4.6	18.2
Mean residuals model A1 when DV was 10'4 mark	11.9	-3.3
Difference between mean residuals (signif. at 5% level)	15.2 marks or 3.8%	
Actual mean 10'4 mark	189.7	139.3
Difference between actual means (signif. at 5% level)	50.4 marks or 12.6%	

other. Some interesting contrasts may be identified. For example, in the case of two of the curriculum groups (see table 6.3, p.92) the difference between the mean residuals (when the DV was the 10'3 marks) is greater than the difference between the actual means (i.e. the mean calculated from raw marks). Teachers comparing the actual means of these groups would probably have decided that the difference between their mean marks (41 or 2% of aggregate) was not important. The difference between the mean residuals is much bigger (7% of aggregate) and further investigation would probably be justified. In this instance, then, the statistical significance of the difference (at 5%) would flag the need for such an investigation. Nevertheless, the analysis of differences between group means is an inadequate

basis for arguing that the Technika Department should be rewarded or the Business subject departments censured. The main deduction which may be drawn from this analysis is that the pupils taking a Business curriculum needed assistance in some form.

The inquiry would reveal that the Technika group had more than twice as many able pupils (IQ > stanine 6) and much larger adjusted achievement by the end of std 8. The proportion of girls in the Business group is also very much higher, but no grounds were found to suggest that this characteristic may be relevant.

Although adjusted achievement was not found to differ significantly (i.e. statistically) between boys and girls in the Technika or Business curriculum groups (see appendix 5.2) differences between the genders were found to be statistically significant in English (table 5.9, p.76). The contrast between the percentages of females in the curriculum outliers thus does probably not explain the difference between the achievements of the Technika and Business groups. The proportion of girls in the language sets, especially English, is much more important, however (see table 6.4). In the examples of the outlier language sets the differences between the sets is more obvious than the differences between the curriculum groups because the gap between the actual means is much larger than the 2% between Technika and Business. The difference in the mean residuals or adjusted means is smaller than in the previous example (but still significant at the 5% level). For both languages these differences might be deemed educationally significant. Investigation of the characteristics of the sets would reveal that the upper or positive outliers had younger, more able pupils, many of them female and many taking Academic subjects. The investigator would be justified in concluding that the differences between the sets could be explained as much in terms of the set characteristics as the subject teacher or methodology. There do not seem to be grounds here for assuming that the differences are due to teacher rather than classroom effects.

The argument that differences between adjusted achievement (i.e. mean residuals) should not be wholly attributed to the teachers but to the set characteristics is supported by the example in table 6.5, where two classes with the same teacher but different backgrounds achieve outcomes which are significantly different statistically. These classes are also of interest as an example of a pair of close

residual or adjusted means where the difference would probably not be considered educationally important despite the statistical flag and the very large difference in actual means.

Professional judgements will be required about groups and individuals. In both cases, a standard criterion might be applied to residuals. For example, a difference of 5% or 10% on the mark scale may be the flag for further investigation. In the examples discussed above the differences were expressed in terms of the maximum number of marks available in the subject concerned. In the case of groups, the difference would be between divisions such as boys and girls, or subject sets, while in the case of individuals, the change between terms would be of interest. The terms might be consecutive, or in comparison to a benchmark term such as the end of the previous standard. The criterion chosen could depend upon the circumstances. For example, 5% between outlier sets of different characteristics may not be educationally significant, but 5% between the first and second ranking sets might be. Similarly, at the level of the individual, a 5% change in adjusted Aggregate is much more significant than a change of that size in a subject mark. From another perspective, a 5% change in the adjusted mark of a weak pupil, or someone aiming for a scholarship, may be educationally important, but not for a pupil comfortably in the middle, or recovering from personal dislocation.

Nevertheless, the automatic flagging of individuals whose adjusted achievement differs by a specific criterion, say 5%, would be a useful function of a school information system. Investigations and responses could follow. If the differences between present and prior achievements of individuals are found to be educationally significant, the responses could range from recognition of quality (for example, prizes and Colours) through sanctions (such as homework classes) to assistance, whether revision lessons or pastoral care. In the event of educationally significant differences between groups intervention might range from revision lessons through changing the teaching style, pace and classroom management to reallocation of resources such as staff and facilities.

Conclusion

It has been argued in this chapter that professional judgement would be frequently required in the creation and use of a school information system which monitored the achievements of groups or individuals. Where educationally significant differences are found analysts should be careful of ascribing the differences to the associated variable. Analysts should note that even when prior achievement and background differences are taken into consideration the remaining differences between groups of pupils cannot be causally related to the monitoring variable. The information yielded by QATRA should nevertheless be useful and enable teachers to make more appropriate and timely responses to fluctuations in achievement. Nor may comparisons be drawn by applying ANOVAs to the residuals of groups if the residuals are derived from different regressions. And since causality cannot be ascribed, QATRA should not be used for bureaucratic assessments of effectiveness. The information yielded by QATRA should nevertheless be useful in monitoring the consistency of the performances of pupils or groups over time. Where fluctuations are detected, professionals may be able to make more timely and appropriate responses than is possible with the existing performance indicators.

Analysts may be interested in the distribution of the residuals of different sets or groups beyond the sums of the residuals used in this study. It would be wise to consult a statistician before drawing any conclusions from the observed distributionss.

In the process of reporting this study, a number of problems and unanswered questions have been identified. These are outlined in chapter 7, where the relationship between the study and the domain of SER are also considered.

More general conclusions to the study are also considered in chapter 7.

Chapter 7. Summary and conclusions

Introduction

In asking how quality may be assured in education, one must also decide what constitutes effectiveness. This investigation therefore involved a review of the domain of school effectiveness before the specific problem of the analyses which serve quality assurance could be addressed. Some of the key features of the domain of school effectiveness will be revisited in this chapter. After some aspects of the way forward have been considered, the chapter will conclude with an evaluation of the study in relation to its aims.

The domain of School Effectiveness

In most studies of School Effectiveness there are three elements or concerns, namely research, improvement and accountability. It was however found convenient to partition the work into three fields, depending upon the primary purpose. The School Effectiveness Research (SER) was mainly the preserve of the scientists who wished to establish what proportion of the differences in the outcomes of schooling could be attributed to differences in schools. Improvement and Quality Assurance represent the responses of the educationists and politicians to the findings of SER. The former wished to develop the correlates of the effective schools in as many schools as possible. The latter wished to demonstrate that state money was being effectively used. In the review it was shown that while the Research and Improvement fields have progressed through several generations, Quality Assurance (expressed through the mechanism of the education market) is a relatively recent idea. The name of the field is used to embrace accountability and the mechanisms by which it may be served.

SER has found that school effects seem relatively small, i.e. only about 10% of the variability in pupil achievement can be ascribed to school or classroom effects. Ten percent, however, is interpreted as educationally significant - supposedly equivalent to about a difference of a year's education by the end of twelve years of schooling (Jencks *et al*, 1972; Purkey and Smith, 1983).

Of greater importance than the size of school effects for this study is the conclusion of SER that effectiveness is neither uniformly distributed over all subjects and groups of pupils, nor stable over time.

School improvement in the present decade has moved away from the strong emphasis on the five factor model (with its focus upon strong leadership, high expectations, basic skills, orderly climate and frequent testing) to an open-ended approach which permits school communities to set their own objectives. Improvement is now considered to be a slow process requiring wide consultation and careful planning based upon relevant and accessible information. The state of the youngest of the three fields of School Effectiveness, Quality Assurance, is less clear and there is little agreement about the model. The free market model of accountability upon which it is largely premised may be inappropriate where education is provided by the state. Even private schools are rarely businesses. But there are other models of accountability, and some degree of internal and external quality assurance is required in all of them by some role-player or other, depending upon the model. The mechanisms for monitoring and reporting quality are varied and some of the associated practices such as ranking of schools and rewarding apparent levels of quality should be used with care.

It was concluded from the review that there is a gap between the techniques and concerns of the researchers into school effectiveness and the needs of individual schools. Similarly, there is a gap between the prescriptions of school improvers and the detail needed by school leadership wishing to follow their advice. This study has shown that analysis of the stream of information now available in personal computers could be used to inform school improvement, as the basis for professional intervention where pupils need help, and as a source of indicators with which to inform parents and community. This study was intended to help bridge the gap by investigating whether the techniques used by SER could be applied at the level of the individual school.

The way forward

Future studies could follow three directions, firstly attending to the unanswered questions raised by this study in a single school, secondly, research into the applications of the techniques of SER multiple school studies, and thirdly, promoting the techniques explored in this study available.

Some unanswered questions from this single school, single cohort study

A number of questions which would warrant further attention have been identified during this investigation. They include the following issues:

1. The use of the slope co-efficients of the linear regressions.

While the use of these co-efficients was avoided in this study because of the difficulties of evaluating the differences between slopes, the unstable composition of subject sets and the small size of many sets, they may nevertheless prove useful.

The residuals used in this study were derived from a common regression, i.e. one line fitted to the marks for one subject. Where there are enough observations to allow the division of pupils into two or more units of analysis (on the basis of one of the monitoring variables, such as gender) it would be useful to compare the distribution of achievement within each sub-division. For example,it may be found that while boys and girls have similar levels of achievement and near-zero mean residuals, the distribution is very different, with (say) previously strong girls apparently under-achieving and previously weak boys apparently over-achieving. In such a situation the girls' slope would be very flat, even negative, while the boys' slope would be very steep. Given such different slopes, the educational analyst would investigate further statistically and consult the teachers involved. The analysis would consider the distribution of the DV along the slope and the identification of influential outliers, for example, while the discussions could provide background and explanations for the observed trends.

It would also be useful to establish whether comparisons of the slopes of a sub-division (e.g., gender) could be valid across several years, or whether such comparisons should be confined to a single year. The interpretation of the latter

comparisons would not be clouded by the effects of changes in factors such as set composition and size.

2. The use of race as a monitoring variable.

In the present period of transformation in South Africa race should be used to monitor the distribution of achievement over the pupils in a school or system. If achievement, especially adjusted achievement, is found to be unrelated to race it will be important to be able to demonstrate this equality. Where achievement is unevenly distributed intervention may be necessary. For example, remedial teaching or pre-schooling might be encouraged.

3. The use of non-linear regressions when these provide a better fit.

Only linear regressions were used in the study. The use of residuals from simple monotonic non-linear regressions and multiple regressions should lower the frequency of statistical flags, i.e. reduce the number of model-monitoring variable pairs which the professional has to investigate.

4. Standard scores.

The use of standard scores in place of raw marks should eliminate notional differences in standards between subjects. There may be applications for such procedures in a single school analysis.

5. Using as EVs some monitoring variables of this study.

Two aspects need attention here. First, the development of model C (pp 62 - 66). In this study, it was not possible to find variables which could explain achievement without the use of prior achievement or some of the monitoring variables such as ability. Reducing the number of monitoring variables (and thus making them available for use as EVs) may make it possible to adjust achievement according to potential rather than prior achievement, i.e. to use model C. Second, investigating the effects of making all variables, i.e. ability, age, gender, curriculum, middle schooling and subject set, available for selection as EVs, except for the monitoring variable. The inclusion of these variables in the model group B should increase the size of the R^2 and reduce the frequency of SSD pairs.

6. A more controlled study of the relationship between age and achievement.

In the FHSHS 1994 cohort it was found that achievement before and after adjustment was negatively related to age. The relationship may have been influenced by pupils who had repeated a standard at some stage. It would be interesting to repeat the analysis excluding those pupils who had repeated.

None of these matters should be confined to studies of a single cohort or school. They would be better addressed in studies involving a series of cohorts in several schools.

Multiple school studies

On the larger scale, the obvious follow-up to this study would be repeated trials with other cohorts at FHSHS and a study across a number of schools, or even school board districts or circuits. At this scale of research, hierarchical linear modelling would be preferable to regression analysis since it is designed to analyze multilevel or nested data. (See chapter 2, p.15.)

Improving access to the techniques

Many schools have information systems to handle the attendance register, marks and other administrative needs. Amongst other things, these packages generate mark schedules and reports to parents on the progress of their children. The value of these facilities would be enhanced if the databases and programmes could be expanded to include the additional information and statistical functions required for QATRA. The monitoring of quality would be facilitated if the flags signalling the need for professional investigation were routinely and automatically available. The techniques investigated in this study should therefore be promoted to the designers of packages such as *Saspak* and *School Administrator*. At the same time the QATRA concept could be opened to debate by publication and demonstration.

Conclusion

The aim of this study was to bridge the gap between School Effectiveness and practical school management. The intention was to explore the application of the research analytical techniques to information available in a school. The motives for

seeking to analyze the information were twofold. First, to improve understanding of apparent achievement and to inform intervention in order to improve the school's effectiveness; and second, to assist schools in demonstrating the quality and stability of their effectiveness and the equitable distribution of achievement between important groups of pupils.

The study has shown that regression analysis may be used to monitor the various dimensions of effectiveness, although the residuals of the fitted regression lines were used in the analysis rather than the slope coefficients and intercepts of the lines themselves. The residuals provided convenient measures of achievement adjusted for prior achievement and background which could be more easily analyzed and interpreted than the coefficients and intercepts. The advantage of residuals is that they mark individual effects while the slope and intercept relate to group effects.

The technique investigated in the study uses statistical significance as a performance indicator, to signal situations where professional investigation is required because achievement is apparently very different from what might have been expected. The difference in achievement may be either an improvement or a decline in relation to the prior achievement of individuals, or between the mean adjusted achievements of groups.

One of the criteria for performance indicators is that they should be accessible, which means that the audience should be able to interpret them and that the information and analysis should be inexpensive. The comprehension of the audience should not be a problem provided that the notion is accepted that statistical significance is only an indicator or flag to investigate further. The danger is that too much will be constructed upon statistical signals. With regard to the second requirement, the costs of the information system and the analyses should be relatively low since recent versions of at least two popular software packages (*Excel* and *Quatro Pro*) have some multiple regression as well as one-way analyses of variance functions. Probably the most difficult resource to find will be teachers' time. Ideally, however, the indicators would be routinely and automatically available, as suggested above.

The main purpose of the study has been achieved. It has been demonstrated that a technique familiar to School Effectiveness Researchers may be easily used to

assist quality assurance at school level. The models on which the technique is based are value-laden, however, and the education market context of the production function model must be acknowledged. Similarly, care should be taken when interpreting the statistical results. It will be tempting to read too much into statistical significance, whereas the focus of analysts should be upon educational rather than statistical differences.

The second aim, of demonstrating school effectiveness, has been addressed inversely, in the sense that within school subjects one may identify situations where *NO* evidence of (statistically) significant differences is found, for example between genders. Such inverse examination to detect possible concerns of equity would apply to variables such as race or parental social class.

While the exploration of regression analysis could have been more exhaustive, it is concluded that regressions and analysis of variance could be useful tools for school leadership to track and promote achievement, and demonstrate effectiveness.

Reference list

Aitken, M. and Longford, N.(1986) 'Statistical issues in school effectiveness', *Journal of Royal Statistical* Society, Series A,149:1-42.

Angus, L.(1993) 'Review essay : the sociology of school effectiveness', *British Journal of Sociology of Education*, 14(3):333-345.

Armor, D., Conry-Oseguera, P., Cox, M., King, N., McDonnell, L., Pascal, A., Pauly, E. and Zellman, G. (1976), *Analysis of the school preferred reading program in selected Los Angeles minority schools,* Santa Monica, CA:Rand Corporation.

Ball, S.J.(1993) 'Education markets, choice and social class : the market as a class strategy in the UK and the USA', *British Journal of Sociology of Education*, 14(1):3-20.

Birnbaum, I.(1993) 'Digest: value added variations',*Education*, 12 Nov.:i-iv.

Blakey, L.S. & Heath, A.F. (1992) 'Differences between comprehensive schools: some preliminary findings' in Reynolds, D. & Cuttance, P. (eds) *School Effectiveness: Research, Policy and Practice.* London, Cassell.

Bosker, R.J. and Scheerens, J.(1989) 'Issues in the interpretation of the results of school effectiveness research', *International Journal of Educational Research*, 13(10):741-751.

Brandsma, H.P. and Knuver, J.W.M.(1989) 'Effects of school and classroom characteristics on pupil progress in language and arithmetic', *International Journal of Educational Research,* 13(10):777-788.

Brookover, W.B., Beady, C., Flood, P., Schweitzer, J. and Wisenbaker, J. (1979) *School social systems and student achievement: schools can make a difference*. New York, Praeger.

Bryk, A.S. and Raudenbush, S.W.(1992) *Hierarchical Linear Models : applications and data analysis methods.* Newbury Park CA, Sage.

Clark, A.N.(1985) *Longman Dictionary of Geography: human and physical.* London, Longman.

Clegg, F. (1982) *Simple statistics : a course book for social sciences.* New York, Cambridge University Press.

Coleman, J.S., Campbell, E.Q., Hobson, C.F., McPartland, J., Mood, A.M., Weifeld, F.D. and York, R.L. (1966), *Equality of Educational Opportunity.* Washington D.C., US Office of Education, US Government Printer.

Coleman, J.S., Hoffer, T. and Kilgore, S.(1982) 'Cognitive outcomes in public and private schools', *Sociology of Education*, 55:65-76.

Creemers, B.P.M.(1994) 'Background of school effects research', in Reynolds, D., Creemer, B.P.M., Nesselrodt, P., Schaffer, E.C., Stringfield, S. and Teddlie, C., (eds) *Advances in School Effectiveness Research and Practice.* Oxford, Pergamon Elsevier.

CSS (1992) *Population Census 1991: selected statistical region - Cape Peninsula: Part II No. 03-01-12 (1991),* Pretoria, Central Statistical Service.

Cuttance, P. (1992) 'Evaluating the effectiveness of schools' in Reynolds, D. & Cuttance, P. (eds) *School Effectiveness: Research, Policy and Practice.* London, Cassell.

Cuttance, Peter (1994) 'Monitoring education quality through performance indicators for school practice', *School Effectiveness and School Improvements,* 5(5):101-126.

Davies, L. (1994), 'The Management and Mismanagement of School Effects', *Compare,* 24(3):205-216.

DES (Plowden) (1967) Central Advisory Council for Education (England) *Children and their Primary Schools.* DES London - HMSO, 1:The Report.

Digest (1992) 'The value-added tasks', *Education,* 31 July,: i-iv.

Farrell,P. and Oliveira, J.B.(1993) *Teaching in Developing Countries : improving effectiveness and managing costs.* Washington, World Bank.

Fitz-Gibbon, C.T. (1992) 'School effects at A level - genesis of an information system?' in Reynolds. D. & Cuttance, P. (eds) *School Effectiveness: Research, Policy and Practice.* London, Cassell.

Fuller, B.(1987) 'What school factors raise achievements in the Third World?', *Review of Educational Research,* 57(3):255-292.

Glasman, N.S. and Biniaminov, I.(1981) 'Input-output analyses of schools', *Review of Educational Research,* 51:509-39.

Glogg, M. and Fidler, B.(1990) 'Using examination results as performance indicators in secondary schools', *Educational Management and Administration,* 18(4):38-48.

Good, T.L. and Brophy, J.E. (1986), School Effects, in Wittrock, M. (ed.) *Third handbook of research in teaching,* 570-602. New York, Macmillan.

Graves, N.J. and Simons, N.(1973) 'Geography in philosophy', in Bale, J., Graves, N.J. and Walford, R. (eds) *Perspectives in geographical education.* London, Oliver and Boyd.

Gray J., Jesson D. and Sime, N.(1990) 'Estimating differences in the examination performances of secondary schools in six LEAs: a multi-level approach to school effectiveness', *Oxford Review of Education*, 16(2):137- 158.

Harbison, R.W. and Hanushek, E.A.(1992) *Educational performance of the poor: lessons from rural northeast Brazil.* Oxford, OUP/World Bank.

Helsby, G. and Saunders, M. (1993) 'Taylorism, Tylerism and performance indicators: defending the indefensible?', *Educational Studies,* 19(1):55-77.

Heyneman, S. and Loxley, W. (1983) 'The effect of primary school quality on academic achievement across twenty-nine high- and low-income countries', *American Journal of Sociology,* 88:1162-1194.

Hopkins, D., Ainscow, M. and West, M.(1994) *School Improvement in an Era of Change.* London, Cassell.

Hough, J. (1991) 'An economist looks at education', *Educational Management and Administration, 19*(4):218-232.

Jencks, C., Smith, M., Ailand, H., Bane, M., Cohen, D., Gintis, H., Heyns, B. and Michelson, S.(1972) *Inequality: a reassessment of the effect of family and schooling in America.* New York, Basic Books.

Jenkins, H.O.(1991) *Getting it Right: a handbook for successful school leadership.* Oxford, Blackwell.

Jesson, D.(1992a) 'Digest - Performance indicators : beyond the league tables', *Education,* 28 Febr.:179-80.

Jesson, D.(1992b) 'Digest - valuable addition', *Education,* 24July:i-iv.

Jesson, D., Mayston, D., and Smith, P.(1987) 'Performance assessment in the education sector : educational and economic perspectives', *Oxford Review of Education,* 13(3):249-266.

Jimenez, E., Lockheed, M., Luna, E., & Paqueo, V.(1991) 'School effects and costs for private and public schools in the Dominican Republic', *International Journal of Education Research,* 15(5):393-410.

Jubber, K.(1988) *The home and family environment and school performance: a study of 267 pupils from 15 Cape Town schools.* Unpublished report, University of Cape Town.

Kogan, Maurice (1986) *Education Accountability.* London, Hutchinson.

Lockheed, M.E., Fuller, B. and Nyirongo, R.(1989) 'Family effects on student's achievements in Thailand and Malawi', *Sociology of Education*, 62:239-256.

Lockheed, M.E. and Longford, N.T.(1989) *A multilevel model of school effectiveness in a developing country.* Washington, World Bank.

Luyten, H. (1994) 'Stability of school effects in Dutch secondary education: the impact of variance across subjects and years', *International Journal of Educational Research*, 21(2):197-216.

MacBeath, J. (1994) 'A role for parents, students and teachers in school self-evaluation and development planning' in Riley, K.A. and Nuttall, D.L. (eds) *Measuring and quality: education indicators - United Kingdom and International perspectives.* London, Falmer.

Madeus, G.F., Kellaghan, T., Rakow, E.A. and King, D.(1979) 'The sensitivity of measures of school effectiveness', *Harvard Educational Review*, 49:207-230.

Mandeville, G.K. and Anderson, W.A. (1987) 'The stability of school effectiveness indices across grade levels and subject areas', *Journal of Educational Measurement* ,24(3):203-216.

Mandeville, G.K. (1988) 'School effectiveness indices revisited: cross-year stability', *Journal of Educational Measurement*, 25(4):349-356.

Mandeville, G.K. and Kennedy, E.(1991) 'The relationship of effective schools indicators and change in the social distribution of achievement', *School Effectiveness and School Improvement*, 2(1):14-33.

Mayston, D. and Jesson, D. (1988) 'Developing models of educational accountability', *Oxford Review of Education*, 14(3):321-339.

Miller, M.D. and Moore, W.P.,'Private-public school differences in the United States: findings from the Second International Mathematics Study', *International Journal of Educational Research*, 21(15):433-444.

Mortimore, P., Sammons, P., Ecob, R., Lewis, D. and Stoll, L. (1988) *School Matters : the Junior Years* Wells: Open Books.

Mortimore, P., Sammons, P., Stoll, L., Lewis, D. and Ecob, R.(1989) 'A study of effective junior schools', *International Journal of Educational Research,* 13(10):753-768.

Murgatroyd, S. and Morgan, D.(1992) *Total Quality Management and the School.* Buckingham, Open University Press.

Nuttall, D.L. (1994) 'Choosing indicators' in Riley, K.A. & Nuttall, D.L. (eds) *Measuring and quality: education indicators - United Kingdom and International perspectives.* London, Falmer.

Nuttall, D., Goldstein, H., Prosser, R. and Rashash, J.(1989) 'Differential school effectiveness', *International Journal of Education Research*, 13(10):769-776.

Potter, D .and Powell, G. (1992) *Managing a Better School.* Oxford, Heinemann Educational.

Purkey, S.C. and Smith, M.S.(1983) 'Effective schools: a review', *Elementary School Journal,* 83:427-452.

Ralph, J.H. and Fennessey, J.(1983) 'Science or reform: some questions about the effective schools model', *Phi Delta Kappan*, 64:689-694.

Reynolds, D., Creemers, B.P.M., Nesselrodt, P.S., Schaffer, E.C., Stringfield, S. and Teddlie, C.(1994) *Advances in School Effectiveness Research and Practice.* Oxford, Pergamon (Elsevier Science).

Reynolds, D. and Cuttance, P.(1992) *School Effectiveness: Research, Policy and Practice*. London, Cassell.

Reynolds, D., Hopkins, D. and Stoll, L.(1993) 'Linking school effectiveness knowledge and school improvement practice : towards a synergy', *School Effectiveness and School Improvement*, 4(1):37-58.

Riddell, A. R.(1989) 'An alternative approach to the study of school effectiveness in third world countries', *Comparative Education Review*, 33(4):481-504.

Riley, K.A. (1994) 'Designing a system at the local level' in Riley, K.A. & Nuttall, D.L. (eds) *Measuring and quality: education indicators - United Kingdom and International perspectives.* London, Falmer.

Riley, K. A. and Nuttall, D. (1994) *Measuring quality : education indicators - United Kingdom and International perspectives.* London, Falmer.

Ruby, A. (1994) 'Education indicators, officials, ministers and the demand for information' in Riley, K.A. & Nuttall, D.L. (eds) *Measuring and quality: education indicators -United Kingdom and International perspectives.* London, Falmer.

Rutter, M., Maughan, B., Mortimore, P. and Ouston, J.(1979) *Fifteen Thousand Hours: secondary schools and their effects on children.* London, Open Books.

Scheerens, J.(1992) *Effective schooling: research, theory and practice.* London, Cassell.

Scheerens, J., Vermeulen, C.J.A.J. and Pelgrum, W.J.,(1989) 'Generalizability of instructional and school effectiveness indicators across nations', *International Journal of Education Research,* 13(10):789-799.

Seldon, R. (1994) 'How indicators have been used in the USA' in Riley, K.A. & Nuttall, D.L. (eds) *Measuring and quality: education indicators - United Kingdom and International perspectives.* London, Falmer.

Simons, H. (1984) 'Issues in curriculum evaluation at the local level', in Skilbeck, M.(ed.) *Evaluating the Curriculum in the Eighties. London*, Hodder & Stoughton.

Simons, H. (1987) *Getting to know schools in a democracy: the politics and process of evaluation.* London, Falmer Press.

Skilbeck, M (ed.) (1984) *Evaluating the Curriculum in the Eighties.* London, Hodder and Stoughton.

Stringfield, S. (1994) 'Outlier studies of school effectiveness' in Reynolds, D., Creemers, B.P.M., Nesselrodt, P.S., Schaffer, E.C., Stringfield, S. & Teddlie, C. *(eds) Advances in School Effectiveness Research and Practice.* Oxford, Pergamon (Elsevier Science).

Teddlie, C. and Stringfield, S. (1993) *Schools make a difference : Lessons learned from a 10-year study of school effects.* NY, Teachers College Press.

Teddlie, C. (1994) 'The study of context in school effects research : history, methods, results and theoretical implications' in Reynolds, D., Creemers, B.P.M., Nesselrodt, P.S., Schaffer, E.C., Stringfield, S. and Teddlie, C. (eds) *Advances in School Effectiveness Research and Practice.* Oxford, Pergamon (Elsevier Science).

Walsh, K. (1994) 'Quality surveillance and performance measurement' in Riley, K.A. & Nuttall, D.L. (eds) *Measuring and quality: education indicators - United Kingdom and International perspectives.* London, Falmer.

Weber, G. (1971) *Inner city children can be taught to read :four successful schools.* Washington D.C. : Council for Basic Education.

Willms, J.D.(1987) 'Differences between Scottish education authorities in their examination attainment', *Oxford Review of Education*, 13:211-232.

Woodhouse, G. and Goldstein, H.(1988) 'Educational performance indicators and LEA league tables', *Oxford Review of Education*, 14(3):301-320.

Yamane, T. (1967) *Statistics : An introductory analysis* . New York, Harper Row.

Zar, J. (1984) *Biostatistical Analysis , 2nd ed. Englewood Cliffs, NJ, Prentice-Hall.*