School production modelling to strengthen government monitoring programmes in developing countries

Martin Anders Gustafsson

Thesis presented in partial fulfilment of the requirements for the degree of Magister Artium in the Department of Economics, University of Stellenbosh

Supervisor: Prof Servaas van der Berg

April 2006

Declaration

I, the undersigned, hereby declare that the work contained in this thesis is my own original work and that I have not previously in its entirety or in part submitted it at any university for a degree.

Signature:

Martin Anders Gustafsson

Date: 10 March 2006

Abstract

Education production function analysis is widely recognised as one important area of research that needs to inform education policymaking, specifically policy relating to the mix of funded inputs in a schooling system. Arriving at production functions is a complex task, and is fraught with methodological pitfalls. This thesis sets out to establish a framework for undertaking education production function analysis, and in discussing its various elements, including its pitfalls, recommendations for good practice are arrived at. The material analysed is of four types: texts on econometric theory; existing production function analyses; documentation relating to three data-intensive school monitoring programmes, namely Brazil's SAEB, South Africa's Systemic Evaluation and the international SACMEQ programme; and lastly data, relating mainly to South Africa, from the 2000 run of SACMEQ. The thesis is organised according what can be regarded as seven key analysis steps. These steps include a focus on the importance of a 'mental model', the relative benefits of the one-level regression model and the hierarchical linear model (HLM), and the formulation of actual production functions for South Africa based on the SACMEQ data, using both one-level and HLM models. Key conclusions are, firstly, that the HLM, though still under-developed, offers great analysis potential and, secondly, that production function analyses ought to be translated into budgetary terms in order for them to become fully meaningful to the policymaker.

Keywords

Brazil, economics of education, education planning, education monitoring, production function, SACMEQ, SAEB, school efficiency, school quality, South Africa, Systemic Evaluation

Opsomming

Onderwysproduksiefunksies word wyd erken as ’n navorsingsgebied wat belangrike agtergrondinligting vir onderwysbeleid kan bied, veral rakende die mengsel van befondsde insette in ’n skoolstelsel. Om by produksiefunksies uit te kom is ’n komplekse taak, en daar is heelwat moontlike metodologiese strikvalle. Hierdie verhandeling probeer ’n raamwerk daarstel hoe om onderwysproduksiefunksies aan te pak, en doen voorstelle vir goeie praktyk in die bespreking van die onderskeie elemente (insluitende die moontlike strikvalle) van hierdie benadering. Vier soorte materiaal word ondersoek: teksboeke oor ekonometriese teorie; bestaande produksiefunksie-ontledings; dokumente oor drie data-intensiewe skoolmonitering-stelsels, naamlik Brasilië se SAEDS, Suid-Afrika se Sistemiese Evaluering en die internasionale SACMEQ-program; en ten laaste data, hoofsaaklik vir Suid-Afrika, van die 2000 SACMEQ-opnames. Die verhandeling is rondom sewe belangrike analitiese stappe georganiseer. Hierdie stappe sluit in ’n fokus op die sleutelrol van ’n “denkmodel”, die relatiewe voordele van die enkelvlak-regressie-model teenoor die hiërargiese liniêre model (HLM), en die daarstelling van produksiefunksies vir Suid-Afrika gebaseer op die SACMEQ-data, waarin beide enkelvlak- en HLM-modelle gebruik word. Belangrike gevolgtrekkings is eerstens, dat alhoewel HLM nog onderontwikkeld is, dit groot analitiese potensiaal inhou en tweedens, dat produksiefunksie-ontledings in begrotingsterme vertaal moet word om sinvol vir die beleidmaker te wees.

CONTENTS

Figures and tables

ACRONYMS USED

ABET	Adult basic education and training
CLRM	Classical linear regression model
CNLRM	Classical normal linear regression model
EFA	Education for All
GLS	Generalised least squares
HA	Historically advantaged
HD	Historically disadvantaged
HLM	Hierarchical linear model
HSRC	Human Sciences Research Council
IEA	International Association for the Evaluation of Educational Achievement
IIEP	International Institute for Education Planning
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais (National Institute of Educational Studies and Research)
L1	Level 1
L2	Level 2
LLECE	Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación
LSM	Learning support material
MDG	Millennium Development Goal
NEPA	National Education Policy Act
NGO	Non-government organisation
OECD	Organisation for Economic Co-operation and Development
OLS	Ordinary least squares
PISA	Programme for International Student Assessment
SACMEQ	Southern and Eastern African Consortium for Measuring Educational Quality
SAEB	Sistema Nacional de Avaliação da Educação Básica (National Basic Education Evaluation System)
SES	Socio-economic status
TIMSS	Trends in International Mathematics and Science Studies
UNESCO	United Nations Educational, Scientific and Cultural Organization
UNISA	University of South Africa
WLS	Weighted least squares

SOME POINTS ON TERMINOLOGY

Generally, the South African terms ‘educator’ and ‘learner’ are used, rather than the internationally more common ‘teacher’ and ‘pupil’ (or ‘student’). However, the non-South African terms are used whenever the discussion relates to a particular study or database where the international terms appear.

DETAILS ON COMPUTER SOFTWARE AND DATASET

The SACMEQ II dataset used to test much of the modelling theory is was Version 3 of the dataset. Permission to use this data was obtained from the SACMEQ office in Harare (contact person there was Saul Marimba). Support from that office is hereby acknowledged and much appreciated.

Three software packages were used to analyse the SACMEQ II dataset, as well as dummy data created to test particular theories:

- Microsoft Excel 2003
- Stata (Intercooled for Windows 8.0)
- HLM for Windows Version 6.0.

1 INTRODUCTION

The challenges around the provisioning of basic school education in developing countries are both daunting and exciting. Major advances have been made. For instance, the primary schooling gross enrolment ratio for East Asia, a region accounting for one-third of the developing world's school aged children, increased from 85% to 110% between 1990 and 2000. Sub-Saharan Africa, the region most behind with respect to this indicator, increased from 74% to 82% over the same period. Yet serious challenges remain. One in five girls of primary school going age are not at school in the developing world as a whole. The quality gap between developing and developed countries, measured for instance in standardised tests, continues to be large (UNESCO 2003: 96, 334).

Tackling the basic education challenges implies that governments must tackle the research question of *how education happens in (and around) schools.* What are the inputs that make the greatest difference? What policies and budgets can optimise access to education, and the quality of this education? Is it more textbooks, or more in-service training of teachers, or school lunches, that yield the best returns? Or is it some intervention outside the school, as in Brazil's *bolsa escola* programme that dispenses grants to parents whose children attend school regularly? If combinations of policies and programmes are required, what is a desirable mix?

During the last few decades relatively solid foundations have been laid for understanding the economics of school production. There is a growing awareness of the importance of economic research and data in the education decision-making of developing country governments. In this thesis, what seemed like an interesting combination of texts and data are analysed with a view to responding to a few key questions.

As a point of departure, texts relating to three government monitoring programmes were examined. All three programmes analyse data on school inputs and outputs in a sample of schools. The three programmes are:

- **Brazil's SAEB programme.** This programme, which has been run by the Brazilian government since 1990, was chosen due to parallels between South

Africa and Brazil in general, as well as parallels between SAEB and the Systemic Evaluation. (The author's ability to read Portuguese was another key factor.)

- **South Africa's Systemic Evaluation programme.** This programme, run by the South African government since 2001, was chosen as it arguably represents the South African government's most comprehensive attempt to monitor the relationships between inputs and outputs in schooling.

- **SACMEQ.** This multi-country programme, run by a consortium of Southern and East African governments, in collaboration with UNESCO, since 1995, was chosen partly because of its global importance as a developing country school monitoring programme, and partly because permission was obtained to analyse data from this programme. Statistical outputs and interpretations emerging from this data are included in the thesis.

The fact that the discussion relating to the above programmes deals mainly with Brazil's SAEB and SACMEQ is a result of the greater volume of documentation available for these two programmes. However, many of the points made about SAEB and SACMEQ would be applicable to the Systemic Evaluation.

In addition to texts on the three monitoring programmes, other texts were consulted that could provide, firstly, the economics of education context to school production modelling, secondly, examples of other production analyses and, thirdly, the econometric and statistical theory required for the production modelling.

The focus in this thesis is less on *how school production occurs*, than on *how school production can be modelled,* though both of these areas receive considerable attention. There is thus a slight bias in the thesis towards understanding the methodology, as opposed to understanding school production. Moreover, three key questions narrow the focus down further:

- What general and econometric analysis lessons from Brazil's SAEB programme could inform South Africa's education monitoring programmes, and those of developing countries in general?

- More specifically, how useful are hierarchical linear models (one of the models used to analyse SAEB data) for rendering knowledge about school production?

- On a broad level, what can be considered an overall and adequate set of data analysis procedures for a government monitoring programme generating knowledge on school production?

Structurally, the thesis begins with an account of the economics and political backgrounds. It then moves through seven steps involved in producing knowledge on school production. These steps are:

- Understanding the data
- Building a mental model
- Selection of a statistical model
- Variable selection and manipulation
- Iterative modelling
- Translation of the model into policy information
- Recommendations for future data collections

These seven steps, and their corresponding seven sections, constitute the main body of the thesis. The weight attached to each one is less dependent on the topic's general importance, than on the three key questions identified above.

2 THE BACKGROUND

2.1 The economics of production background

We live in a world where individuals, firms, and other institutions such as schools, industries and countries depend on each other economically. That dependence takes the form of a continual exchange in goods and services, and some less tangible things such as legitimacy and accreditation. Exchange is driven partly by the demand for goods and services, which in turn depends largely on consumption patterns. And exchange is partly driven by the supply of goods and services, which in turn depends on production patterns. Production thus underpins supply in the overall economic picture, and is clearly one of the main pillars of any model embracing a whole economy.

The first ever model of production is attributed to Fransçois Quesnay (1694-1774), a French surgeon and economist who put together what we can consider a paper-based spreadsheet describing aggregate agricultural and industrial production in France. Adam Smith (1723-1790) can be said to have initiated the theory of the firm and its production processes. In Smith's theory, grounded far more on general observation than on the analysis of quantitative data, the profit-motive of the capitalist brings about efficient production, which at an aggregate level is good for the country as it increases the average well-being of citizens. In Smith's model, which was very much about the welfare of Britain, colonial expansion played an important role. Karl Marx (1818-1883), like Smith, put the productive processes of the firm at the centre of his economic vision. For Marx, however, the mode of production of the firm, in particular the factory, carried the seeds of capitalism's demise because this mode of production inevitably led to ever increasing levels of worker exploitation, and hence class conflict.

Classical economics upheld production as the centrepiece of economic theory. This changed, however, with the advent of neoclassical economics during the late nineteenth century. Though still an important concern, the theory of production had to make way for the theory of exchange as the central concern of economic study. The neoclassical economists were moreover responsible for an increasing use of mathematics and eventually econometric models as the acceptable language in which to express economic realities. Production thus became increasingly a subject of mathematical and statistical models, often of extraordinary levels of complexity, as opposed to a topic of less structured discussion, as in the texts of, for instance, Smith (Marx's writings are from a late enough period to be significantly influenced by a more mathematical style of argumentation) (Skinner, 2002; Mandel, 2002; Smith, 1991; Marx, 1991).

Production models are typically concerned with either the firm, in which case they are referred to as microeconomic models, or with an entire country, or an entire industry within a country, in which case they are referred to as macroeconomic models. However, they can also span other units, such a group of firms in a region, or a particular industry globally. Whatever the level of the model, there are features which are more or less similarly applicable to all production models. Key concepts are the following:

- Production models typically describe the conversion, through some **process**, of several **inputs** into one **output**. They may describe more than one output, and ways in which the outputs influence the inputs, but such models would be considered to be members of a more advanced class of the general group of production models.

- The inputs in a production model are typically grouped into **flows** of recurrent goods and **stocks** of capital equipment or infrastructure. Very often models are differentiated by whether they allow for a change in the stock of capital infrastructure, or whether this is fixed. The former are referred to as **long run** models, and the latter as **short run** models.

- All production models are **time-bound** in that they describe inputs inserted into the production process and outputs emerging from the production process over a set period of time.

- In production, **technical efficiency** is said to prevail when it is impossible to raise the level of output with the basket of inputs that is available.

- **Allocative efficiency** is said to prevail when it is impossible to raise the level of outputs within the budget needed to purchase the basket of inputs. In other words, the concept of allocative efficiency takes into account the possibility of exchanging the present basket of inputs for a better basket of inputs within the same budget. Strictly speaking, allocative efficiency should also take into account the influence of the demand for inputs on the prices of all possible inputs. This level of complexity is, however, often avoided in production models.

- Production models must have a way of attaching **monetary value** to inputs, and possibly outputs, in particular if allocative efficiency is to be considered. Market prices may suffice, but they may not, in particular if there is not a market for particular inputs. In this regard, the **opportunity cost** of not utilising inputs for alternative productive processes is something that we may want to take into account in a model. (As an example, the rent that could be earned from a school building standing empty during a school holiday should perhaps to be considered a part of the value of the school building as an input in the schooling process.)

- A production model may incorporate the **price elasticity of substitution**, or the degree to which the productive entity is able to substitute between different kinds of inputs in response to price changes.

- Production models often consider as inputs less tangible factors, in particular the **technology** by which workers and machines undertake the production process. It should be kept in mind that the broad meaning generally attached to 'technology' in economics embraces not only technical procedures, but 'inputs' such as management styles (or classroom practice in the case of education). **Positive externalities**, roughly meaning non-purchased inputs coming from beyond the productive unit, such as clean air or the absence of civil strife (or a national culture of reading in the case of education), may also be considered inputs in certain models. **Negative externalities** would include factors such as labour relations instability in the country as a whole, which could increase time spent on reaching agreements, or **transaction costs** in economics terms, within individual firms.

- Production models may incorporate as *outputs* certain externalities, or unintended outcomes of the production process. Pollution (a negative externality) or the technological prestige of the country (a positive externality) would be examples.

- Production models are frequently concerned with the relationship between the processes and the size of the productive entity. **Increasing returns to scale** refer to the increased efficiency of, say, a firm, as that firm grows in size – this is Smith's first topic in *The wealth of nations.* However, **diminishing returns to scale**, or increasing inefficiency, are said to set in once a firm has reached an unwieldy size. This second phenomenon only began receiving serious attention in neoclassical economics.

- All first-year economics students are required to grapple with the concepts and graphs relating to the **average cost** of all inputs and the **marginal cost** of additional inputs. Because of increasing and decreasing returns to scale, average cost is generally not equal to marginal cost. The point at which they are equal is said to indicate the optimal size of the productive entity, for instance the firm.

- More sophisticated models of production might model the important matter of **technical change**. No productive entity is static, they all change in terms of their technology, and this change may be linkable to particular stages of development.

- Inputs have complex inter-relationships, and they must often be utilised in particular configurations. Moreover, individual inputs often come in fixed bundles, or they may be indivisible. In the context of the school, one need only think of the complex relationships between educators, their specialisations, physical classrooms and normal school hours that must be optimised within the school timetable. **Activity analysis** and **linear programming** are established techniques for attempting to deal with these problems in production models.

A brief look at the Cobb-Douglas function, an early and well-known neoclassical model of production, illustrates some of the concepts referred to above.

$$y = cx_1^{\alpha} x_2^{\beta} \tag{1}$$

This model has two inputs, x_1 and x_2, and one output, y. The model has three fixed parameter values, c, α, and β. A typical rendition of the model has x_1 representing capital inputs x_2 representing labour inputs, not necessarily in monetary terms. Technical efficiency is assumed, in other words it is assumed that with a given set of inputs, output has been maximised. Allocative efficiency can be gauged if we represent both x_1 and x_2 in monetary terms, where capital inputs are amortised as a flow. If moving expenditure from x_1 to x_2 or vice versa increases output y, then the original situation was not allocatively efficient. If α and β add up to 1, then constant returns to scale prevail. In other words, doubling the quantities of x_1 and x_2 would double y. However, if the sum of α and β is greater than 1, then decreasing returns to scale prevail, and if the sum of α and β is less than 1, then we have increasing returns to scale. One weakness of the model is that it cannot represent the typical situation of increasing returns up to a particular size, and decreasing returns thereafter. Another weakness is that the price elasticity of substitution is always 1. If the price of one input increases by, say, 50%, maximum output is achieved by maintaining total expenditure on that input constant, and just halving the quantity of the input used. There is thus no reason to substitute between inputs.

The above concepts indicate the range of possibilities, and the complexities, in the field of production modelling (Glahe and Lee, 1989: 198-278; Fuss, 2002; Koopmans, 1965: 33-7; Samuelson, 1983: 65; Brown, 2002). Some will be referred to explicitly in the rest of the thesis. Others will not, though they may lurk in the background as important provisos or gaps in the discussion of school production models.

2.2 The economics of education background

What is known as the economics of education has its roots in human capital theory, which was elaborated in the 1960s. This theory explains a systematic relationship between investments in education, at the individual and country level, and lifetime earnings of individuals, which in aggregate terms would be national income. Human capital is considered very much like physical capital. Growth in the stock of human capital, where this stock is the sum of human skills and knowledge in society, is translated into greater overall efficiency in the economy, and hence greater overall welfare. Investments in education comprise not only the direct costs of, for instance, educational materials and teaching time, but also the opportunity cost of time spent not working. According to the theory, households invest in the education of their young on the basis of perceived rates of return to investments in education. These rates are calculated largely as they would be calculated in the case of investments in physical capital. Furthermore, the theory sees government's role as one of investing public resources in the education of poorer households, which are subject to two forms of constraints. Firstly, they experience financial constraints – they lack funds to achieve optimal rates of return on educational investment. Secondly, they experience household-level human capital constraints – poorly educated parents are unable to give their children the educational start in life that better educated parents are capable of offering (Rosen, 2002).

The private cost-benefit relationship with respect to university education (to take an example) is captured by the following:

$$\sum_{t=1}^{40} \frac{(W_u - W_s)_t}{(1+r)^t} = \sum_{t=1}^{5} (W_s + C_u)(1+r)^t \qquad (2)$$

Here the household considers the benefits of additional income accrued over (in this case) the 40 years that the university graduate can earn an income, where this is additional to the income the individual would have earned during the 40 years as a secondary school graduate. This is the left hand side of the equation. Simultaneously, the household considers the direct private costs of five years of university education, C_u, plus the cost of five years of forfeited income as a secondary school leaver. Single variable optimisation yields the value of *r*, the private rate of return to university education (Psacharopoulos, 1995). If this rate of return is higher than the rate of return in equities or a new tractor on the farm or any other investment, the household chooses to send someone to university. The theory is elegant and useful as point of departure, but the caveats are numerous. One constraint is that the model assumes households know the ability of the potential university student to cope with the studies. Another is obviously that future earnings cannot be known. Even if good data are available on the earnings of current university graduates, this data may not be applicable in future labour market scenarios. The informational difficulties of the household are more or less replicated at the level of government, where decisions must be taken regarding optimal levels of public expenditure on education, and the degree of poverty targeting to be pursued in education expenditure. Moreover, government must take into consideration two separate sets of rates of return. On the one hand, government must be aware of the private rate of return, or the rate of return perceived at the level of the individual household. This awareness assists in predicting the response of households to policy changes. On the other hand, government must take into consideration the social rate of return, or the rate of return obtained after including the public cost of education and tax revenue paid by income earners.

Two kinds of policy information emerging from human capital models stand out. Firstly, rates of return for education that are below other rates of return in the economy can point towards inefficiencies in the education system, or, alternatively, an over-supply of a particular kind of education. Secondly, comparing rates of return for the primary, secondary and tertiary education levels can point towards the need for public expenditure shifts between these levels. Much of the empirical research has revolved around the World Bank and George Psacharopolous, and has involved the estimation of actual rates of return. Psacharopoulos (1995) provides rates of return for the three levels of education for 29 countries. Brazil's private rates of return, for example, are 37%, 5% and 28% for primary, secondary and tertiary education respectively. The social rates of return are 36%, 5% and 21% – private rates are necessarily greater than or equal to the social rates. In the case of Brazil, the low rate of return for secondary schooling should be cause for concern amongst policymakers. This statistic could be indicative of poor efficiency in this sector. The rates of return for the primary and tertiary levels, on the other hand, are clearly comparable to or better than those for other types of investment. As in the case of most developing countries, the Brazil statistics indicate that public expenditure on primary education is a better national investment than investment in the other two levels. In fact, rates of return studies have been instrumental in a very strong, some would say excessive, emphasis, partly on the part of donor countries, on primary school public spending in developing countries. These studies have also been used to oppose the notion that investment in vocational education is good for economic growth – modelling has indicated that the rates of return for general education are in fact better than those for vocational education (Jimenez and Patrinos, 2003). South Africa is not included in Psacharopoulos's list. Case and Yogo (1999) provide a rare example of an education rate of return analysis dealing with South Africa. Their analysis focuses largely on the degree to which the learner/educator ratio makes a difference to the rate of return of one additional year of schooling.

Human capital theory assumes that education is the process whereby human capital is generated. Spence has put forward an alternative hypothesis, known as screening or signalling theory, which opposes this fundamental assumption. In the more extreme form of the screening model, the education system works as an elaborate mechanism for categorising individuals by level of human capital. The human capital each individual possesses is innate, or a product of family background. It is not *produced* by the education system. All the education system does is that it categorises each person's innate human capital, through performance scores and attainment levels. Rosen (2002: 686) sees the screening model and the classical human capital model as complementary, rather than as opposites. It is not difficult to understand this. To some degree progression through schools and universities produce important signals to the labour market relating to the innate talent of individuals. On the other hand, the education system certainly does produce skills and knowledge in the population, which clearly assist people to be more productive in their work environments.

Proof of the productive capacity of education systems can be established through production functions dealing with the education system. This topic takes us back to the concerns of this thesis. In its simplest form, the education production function is represented as follows:

$$Y_i = F\left(X_{1i} \ldots X_{ni}\right) \tag{3}$$

Y_i is the educational output, or level of skills and knowledge, or human capital, of a group of learners (in particular the learners of one grade in one school), or one individual learner. X_{1i} is an input in the education process, for example contact time between educators and learners in one year. X_{2i}, X_{3i} and so on, up to X_{ni}, are other inputs, or 'explanatory variables'. The function F in equation (3) is some function that describes the education production process in the population of learners or schools as well as possible. The details of this constitute a large part of the thesis. In education production functions, explanatory variables typically include factors relating to the home background of learners, and school and classroom management processes followed in utilising the various physical and human inputs. This makes education production functions rather distinct from production functions describing, for instance, production in a manufacturing plant. This frequently leads to doubts as to whether education production functions can in fact be classified as production functions (Hanushek, 1979: 352). In this thesis, whilst the term 'production function' is adopted, the X variables are mostly referred to as explanatory variables, and not as input variables. The word 'input' on its own, in the thesis, refers mostly to the things that make schooling possible, including purchased and hired items such as textbooks and teaching time, but even efficiency items such as good school management, or good classroom management.

If the production function can establish that the educational outputs of one group of learners are higher those of another, *similar* group of learners, because more inputs are provided, then this can be taken as proof that education inputs do have some productive role with respect to human capital. Growth in human capital can be more directly gauged if one of the X_i explanatory variables is the level of human capital at an earlier point in time. Eric Hanushek stands out in the education production function field, with respect to both developed and developing country schooling systems. Harbison and Hanushek's (1992) modelling of input and output data from schools in the north west of Brazil is regarded as a milestone study in this area.

Production functions do more than prove that education produces human capital. They can provide information on the efficiency of education services in a far more detailed way than do typical human capital models relating years of education to future earnings. In particular, they can demonstrate that increases in one input, relative to increases in another input, will produce a more substantial improvement in outputs, for example performance scores. Production functions can thus inform policymakers regarding matters such as the spread of public expenditure across different types of inputs, for instance learning support materials and educators, and the way in which the various inputs are combined and managed. Production functions can also provide crucial information relating to pro-poor expenditure targeting. The following policy questions stand out:

1. **Could output be improved with the current basket of inputs?** This question deals with the technical efficiency of the schooling system. If certain schools are found to produce better results than other schools which have the same inputs, the suggestion is that improved management in the other schools could improve their results.

2. **Could output be improved by changing the mix of inputs within the current budget?** This question, and the next three, deal with matters of allocative

efficiency. They obviously require the use of monetary values in the model (the previous technical efficiency question can be answered without the use of financial data).

3. **On what should new education funds resulting from budgetary increases be spent?** This question would be particularly important in a developing country where real expenditure per learner can be expected to improve substantially over time.

4. **What education inputs should be cut following budgetary cuts?** This is the opposite of the previous question.

5. **How should education inputs be rearranged following a change in the relative prices of those inputs?**

6. **How much additional education funding is required by poor communities in order to achieve greater equality in the outputs?** This is a key equity question. The variables describing the learner's socio-economic status found in a typical education production function make it clear that SES advantages are strongly associated with better outputs. It is possible to use the production function to model what level of additional school inputs are required to at least partly offset the learner's SES disadvantage. This can form the basis of a pro-poor education financing approach.

7. **What are education outputs likely to be in the future?** Even in a situation where the government has minimal leeway in actively rearranging the production process in schools, it is important for the government to predict future trends in education outputs, even if those trends depend only on endogenous factors such as an ageing educator corps. Production functions, whilst by no means forecasting models, can nevertheless assist in the forecasting process.

Scepticism around the ability of education production functions to provide answers to all or some of the above questions is common. Monk (1990: 338) notes that production function findings are often contradictory, and hence not very useful for the decision-making processes within government. However, Monk does see potential if more detailed data, in particular time-series data, can be collected and analysed. He emphasises such monitoring programmes are costly. This has implications for, above all, developing country governments. These cost limitations also explain why the focus of this thesis is on the analysis of cross-sectional data which, whilst subject to some serious analysis constraints, is relatively inexpensive to collect, and is a more likely basis currently for production function analysis in developing countries. It could be argued that policy analysis using this less than ideal data is a necessary step before a demand for better but more costly time-series data is created in the education planning offices of the government.

Pradhan (1996: 75), in providing an analysis framework for developing country public expenditure reviews (PERs), supports the use of production functions in determining expenditure optimality. He explains how the more limited levels of public expenditure in developing countries, added to the fact of overall human capital under-development, brings about higher levels of impact when marginal increases in education inputs are implemented. Production function analysis in developed countries often suggests very low impacts on output of small changes in the input basket. In developing countries, on the other hand, such marginal shifts have been shown to have an impact, in particular when inputs such as learning support materials are increased. More discussion on our current stock of knowledge on the education production function in developing countries is provided in section 4.3.

The effective schools research approach to determining what constitutes an optimal bundle of inputs in the education process is described by Monk (1990: 413) as an inductive approach that is often taken as standing in opposition to the education production function approach. Instead of focussing on the production process in a fairly large group of schools in a very structured and statistical manner, as is done in the production function approach, the effective schools approach advocates the identification of a few well performing schools, and the in-depth analysis of how educational outputs are produced in those schools, using not just statistical models, but more qualitative and intuitive methods. The effective schools approach can be said to be inductive insofar as it begins looking at outputs, and then moves backwards into the inputs that produce the outputs. Monk makes the sensible point that the two approaches are complementary, and should both receive attention by education researchers.

2.3 The development economics background

Bell (2002) describes development economics as the study of how one group of countries, the 'latecomers', do or should catch up to another group of countries, the 'pioneers'. To be studied are thus both the pioneers, or the developed countries, particularly how they got to be pioneers, and the latecomers, or developing countries. The ahistorical framework of Bell is fairly typical – no significance is attached to the fact that nearly all the latecomers were at some time colonies of a few key pioneer nations. A typical refinement of the developing-developed country distinction is the categorisation of developing countries into low and middle income countries. The World Bank, for example, has a classification system that puts around 15%, 45% and 40% of the world's population into, respectively, the high, middle and low income country groups. Through its focus on South Africa, and to some extent Brazil, this thesis concerns itself more with middle income country issues and systems than those in low income countries.

Three key development economics debates stand out. One is the degree to which the developing country state should be involved in steering the economic development of the country. Stiglitz (2002) and others have argued that ideology and the interests of lobby groups in developed countries have over-emphasised the ability of market forces to drive development, and under-emphasised ways in which the developing country state can and should promote development. A second debate is whether developing countries should simply adopt the technologies currently being utilised in developed countries, or whether there should be more investment into the development of technologies that are more appropriate for developing country contexts. In this regard, the degree of emphasis to place on capital intensive technologies is a key concern. Thirdly, and very importantly, there is an extensive debate around what economic development should aim to achieve. Amartya Sen (2001) underlines the need to look beyond the classical economic growth concern, to concerns around the quality of life, specifically economic, social and political freedoms enjoyed over a long and healthy life. As an illustration, South Africa and Brazil are more developed than Sri Lanka in terms of GDP per capita, yet Sri Lanka enjoys less inequality, and a longer life expectancy, and is thus in a sense the most developed of the three (Sen, 2001: 6). Sen does not go as far as critiquing the environmental implications of the typical economic growth paradigm. Global warming and other environmental considerations pose immense challenges to how we conceptualise development, and emphasise the need for the re-development of *developed* country economies in the interests of environmental sustainability (Gupta, 2001).

These development economics debates resonate within the education system. The role of the developing country state in funding and managing the education system is a critical concern. Whilst there is general acceptance that the state should fund increased access to education amongst the poor, funding the non-poor, and the public-private mix and national-local mix in the funding and management of the system is widely debated. The source document for the technology of an education system is the curriculum, and whilst the establishment of national curricula gains wide support, how prescriptive the curriculum should be, in particular with respect to the inputs required, is much debated. In South Africa, the new post-apartheid curriculum has been described, by the Minister of Education, as a methodology not sufficiently sensitive to local, South African, needs (Chisholm, 2003: 4). There is considerable debate around the optimal mix of primary, secondary and tertiary education at the various stages of a country's development (Gillis, Perkins, Roemer and Snodgrass, 1983: 250). With regard to outputs, the debates around the importance of average GDP per capita in economic development, as opposed to, for instance, the equality of income earned, finds a parallel in the education debates relating to average test scores and the equality of the distribution of those scores.

Formal models have been constructed to deal with the dynamics of an education system of a developing country. Mingat and Jee-Peng Tan (1998) provide a simple but useful model dealing with inputs and access (but not quality) at the macro level. An adaptation of their model, so that non-personnel expenditure could be included as a consideration, looks as follows:

$$\frac{G_e}{GNP} = \frac{P_e}{P_{sa}} \cdot \frac{P_{sa}}{P_t} \cdot \frac{TS}{GNP/P_t} \cdot \frac{1}{P_e/T} \cdot \frac{G_e}{TS \cdot T} \qquad (4)$$

Essentially, the model explains the relationship between, on the one hand, education expenditure over GNP, $\frac{G_e}{GNP}$, and, on the other hand, enrolled learners as a proportion of the school-age population, $\frac{P_e}{P_{sa}}$, the school-age population as a proportion of the total population, $\frac{P_{sa}}{P_t}$, the average teacher salary as a proportion of GNP per capita, $\frac{TS}{GNP/P_t}$, the inverse of the pupil/teacher ratio, $\frac{1}{P_e/T}$, and a ratio that indicates expenditure on non-personnel items in the education system, $\frac{G_e}{TS \cdot T}$ (for example, a ratio of 1.1 would indicate that the expenditure on non-personnel items came to 10% of the personnel cost).

In terms of policy challenges, Mingat and Jee-Peng Tan emphasise two things that developing country governments must guard against, and both have to do with teacher policy. Firstly, governments should guard against reducing the pupil/teacher ratio, or P_e/T, at too early a point in the development trajectory, particularly if the enrolment ratio, or $\frac{P_e}{P_{sa}}$, is still below 1.0. This warning is linked to the lack of clear evidence that lowering the pupil/teacher ratio improves performance. Secondly, governments should guard against a relative teacher salary, $\frac{TS}{GNP/P_t}$, that is too high for the level of development of the country. In this respect, the point is made that the chief factor that allows developed countries to spend more on each pupil, is the fact that teacher salaries, relative to GNP per capita, are lower in rich countries. In other words, a developing country government should plan for a progressive lowering of teacher salaries, not in absolute terms, but relative to what others in society are earning.

Because the model in equation (4) does not deal with the qualitative output of the education system, it would relate only obliquely to the matter of production modelling. Nevertheless, the model presents within an interesting structure three key inputs that must be optimally balanced for efficient school production to be realised: teacher salary, class size and the proportion of the budget spent on non-personnel items.

2.4 The politics of education background

The previous section identified some of the key points of contention relating to the education systems of developing countries. This section looks briefly at how the agendas of international organisations and national governments deal with, or evade, these points of contention. We shall see that the agenda has become increasingly globalised.

The political imperative of better education for more people has, roughly speaking, occurred within two fairly different ideological streams during the past fifty or so years. We shall refer to the Marxist and the social democratic streams here.

The Marxist stream has been strongly linked to socialist and anti-colonialist struggles, and is influenced by intellectuals such as Frantz Fanon and Paulo Freire. In terms of the debates of the previous section, both of these men placed great emphasis on expunging the curriculum of the coloniser from the education system of the developing country, and replacing it with an indigenous, democratic and socialist-oriented one. The emphasis on the ideological dimensions of the curriculum could, one might argue, be interpreted as an under-emphasis on the more technical competencies of learners. Yet it is interesting to note that the country that over the last half century has most steadfastly maintained a Marxist agenda in its education system, namely Cuba, has achieved learner performance levels well above its non-Marxist neighbours in standardised international tests. Out of 12 Latin American countries participating in the 1996 Laboratorio tests of LLECE, Cuba obtained an average mathematics score for primary school learners of 357 against an inter-country average of 257 (the second-highest average score, that of Argentina, was 265) (UNESCO, 2001: 53).

Unlike the Marxist stream, the social democratic stream is geopolitically non-committal, and is often associated with the UN and its structures, as well as the donor funding agencies of developed countries. Here the emphasis is largely on increasing access to schooling in developing countries, and enhancing the quality of this schooling, where quality is understood largely in terms of the competencies of individual learners. The education agenda in this social democratic stream is but one part of a larger development agenda whose flagship is arguably the eight Millenium

Development Goals (MDGs). The second MDG deals specifically with education. It is to ‘achieve universal primary education’ (United Nations, 2005).

In some senses, there has been a political convergence towards the middle, represented by the social democratic stream, with ex-Marxists (for instance in South Africa) subscribing to the social democratic agenda, and advocates of the private market becoming more adamant about the importance of public funding in basic education. With regard to the last point, a significant development was the about turn in the World Bank’s policy on school fees. More or less in 2003 the World Bank shifted from being strongly in favour of school fees, on the basis of its research finding that low school fees had little influence on school attendance, to an explicit position against school fees, based on research that school fees *did* influence attendance (Stiglitz, 2002: 76; August 2003 ‘issue brief’ on user fees on the World Bank website).

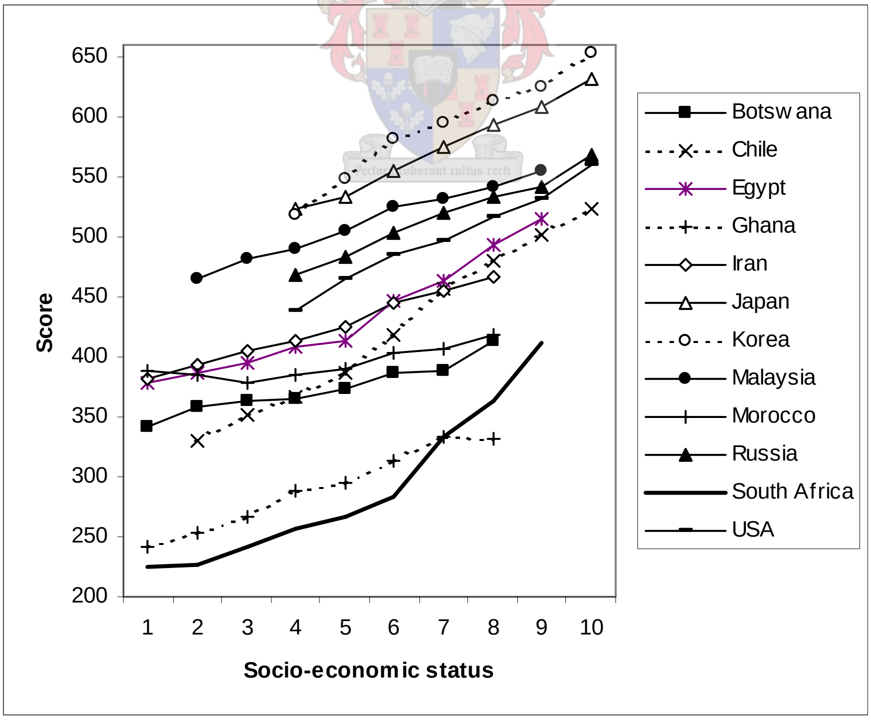
The dynamic between the imperatives of access and quality has been an important one in the social democratic stream. The 2005 report of the Education for All programme, the centrepiece in UNESCO’s education reporting system, is sub-titled *The quality imperative*, and attempts to raise the profile of the quality imperative, relative to the access imperative. It also provides a fairly in-depth overview of what quality is said to mean and which resources and practices in schooling systems are key to promoting quality in education (this is of course a central concern in this thesis, and UNESCO’s position is analysed in some depth in section 4.3 below). Insofar as quality schooling is linked to adequate resourcing, there is clearly a tension between access and quality. A developing country may have to compromise on average quality when there is a large and sudden rise in access to schooling, especially as these rises tend to involve the incorporation of children from the poorest households in society, and since performance is negatively correlated with socio-economic status. As an example, the drop in average scores in Brazil, reflected in the results of the SAEB monitoring programme in the years up to 2001, is clearly attributable to an expansion in access to schooling following the advent of democracy in that country in 1985 (INEP, 2004). At the same time, as stated in the education Millennium Development Goal (MDG), access to schooling means access to ‘high-quality education’. Implicitly, access to poor quality schooling is not access to schooling.

Two concrete challenges stand out for developing country governments wishing to improve access and quality. The first is obtaining adequate education budgets. The second is establishing sufficiently robust performance monitoring systems. We can generally assume that such performance monitoring systems should be sample-based, given the high cost of regular and standardised monitoring of the performance of all learners in the system. We can furthermore assume that such sample-based performance monitoring systems should gather data not just on outputs, or performance, but also on the school, the home background, and teaching practices. Not gathering this full range of data makes it impossible to gauge what causes better or worse performance in the system, and hence makes the planning of education expenditure and school interventions difficult.

On the expenditure side, a loose target of 6% for education expenditure over GNP has been promoted by UNESCO (2005: 142). Brazil's ten year education plan, promulgated in 2000, sets the target of 7% for 2010 with respect to education expenditure over GNP. South Africa's education expenditure over GNP is on a clear and continuous downward trajectory, from around 6.5% at the advent of democracy in 1994 to a projected 5.5% in 2007. The government has argued, on the basis of international comparisons, that the education expenditure over GNP level has been too high in the past. Clearly, pressure to deal with poverty in non-education areas, in particular within the growing social grants system, explains much of education's loss. It is moreover important to note that in real terms, education expenditure has generally been increasing since 1994. There is no clear position, however, on what might be an optimal level of expenditure on education relative to GNP, and on whether the current decline in this regard should continue beyond 2007 (National Treasury, 2005: 11).

With respect to the gathering of performance data, there has been an increase in developing country participation in international sample-based monitoring programmes. International, as opposed to national programmes have at least two advantages. Firstly, they cut research and instrument development cost down in an area where skills are in short supply in most developing countries. Secondly, they provide a basis for the international comparison of results, which can be regarded as important given the elusiveness of the educational quality concept. This is illustrated if we look at the TIMSS 2003 mathematics results for Grade 8. It would be easy to conclude from the emphasis placed on quality improvements in historically disadvantaged schools in South Africa that historically advantaged schools do not have a quality problem. However, as figure 1 below indicates, even learners with a high socio-economic status (SES) in South Africa obtain mathematics scores that are considerably lower than those of their SES counterparts in other developing and developed countries. Clearly, the challenge in South Africa is both to reduce inequalities, and hence reduce the gradient in South Africa's slope, and to raise mathematics performance across the entire SES range, and hence move South Africa's slope vertically upwards. (The TIMSS data and accompanying metadata were obtained off the TIMSS website, and SES index values that were comparable across countries were calculated using the TIMSS background questionnaire data, in much the same way as the SACMEQ SES values described in section 6 below were calculated. Specifically, six pupil variables relating to parent education and characteristics of the home were used.)

Figure 1: Socio-economic status and mathematics performance in Grade 8 in 2003



Source: Authors own calculations on TIMSS, 2003.

The following are key international learner performance monitoring programmes. All focus on mathematics performance, but the focus on other competencies varies. All use representative samples of learners.

- **Trends in International Mathematics and Science Studies (TIMSS).** This international programme is the longest-standing one of its kind, and traces its origins to data collections conducted in developed countries in 1959. TIMSS is run by the International Association for the Evaluation of Educational Achievement (IEA), an association governed by 62 'institutional members' which include, for example, the HSRC of South Africa and INEP of Brazil, both publicly funded research organs. The head office work of TIMSS is financed through donor funding, and participating country governments are generally expected to fund the in-country data collection. The 2003 run of TIMSS incorporated 48 countries at one or both of the Grades 4 and 8 levels (Martin, 2005). A cursory analysis of the participation trends reveals that there has been a marked drop in developed country participation in TIMSS over the years, and a marked increase in developing country participation.

- **Programme for International Student Assessment (PISA).** This OECD programme was started in 2000. PISA concentrates on the performance of fifteen year old learners. The 2003 run of PISA involved 43 participating countries, ten of which were not OECD members. Brazil, for example, participated in this run. Presumably, the emergence of PISA explains the decline in developed country participation in TIMSS.

- **Southern and Eastern African Consortium for Measuring Educational Quality (SACMEQ).** This programme, also linked to UNESCO, has collected data twice, once in 1995 and once in 2000. In 2000, 14 Southern and Eastern African countries participated, including South Africa. (See Appendix A for more details.)

- **Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación (LLECE)** or **Latin American Laboratory for the Evaluation of Educational Quality.** LLECE was constituted in 1994 as a UNESCO organ. In 1996 it collected learner performance data from 13 Latin American countries, at the Grades 3 and 4 levels.

At the national level, monetary cost, scarcities in the necessary skills and, to some extent, political sensitivities around what actually occurs in schools, hinder the development of national equivalents of TIMSS-like programmes within developing countries. Yet there are good reasons for national programmes of this nature. National programmes may focus on specific national issues missed in the international programmes, for example curriculum specificities and, in the case of South Africa, racial equity issues. Moreover, national programmes can focus on revealing important differences between regions within a country.

Both South Africa and Brazil run national sample-based data collections to gauge learner performance. Brazil's SAEB programme was started in 1990, whilst the first Systemic Evaluation run in South Africa occurred in 2001. Brazil's programme has developed a considerable advocacy component, and in 2005 the Brazilian government announced that SAEB would form the nucleus for a broader learner performance system that would assess all Grades 4 and 8 learners in the country every year. SAEB would be retained, under a different name, as a sample-based nucleus programme, presumably providing more in-depth information than the broader, universal programme. (Details on SAEB and the Systemic Evaluation are provided in Appendix A.)

3 UNDERSTANDING THE DATA

The seven steps for producing knowledge on school production used in this thesis are based loosely on the standard steps for econometric analysis put forward by Gujarati (2003: 3). In those standard steps, the analyst only becomes concerned with actual data after the formulation of a hypothesis, and the specification of a model. The assumption is that the analyst determines what data is collected.

In the world of government monitoring programmes, however, the analyst is typically faced with the *fait acompli* of an already finalised dataset that must be subjected to econometric analysis. For this reason, we begin here with the step of understanding the data. In this step, the analyst gains an overall picture of what economic issues are covered in the data, and of how reliable and robust the data are for producing policy recommendations. The fact that we begin with the data here should not detract from the importance of also formulating a hypothesis or mental model early on in the analysis process. For that reason, the mental model receives attention in the second step, covered in section 4.

It should be borne in mind that beyond the scope of the seven steps put forward in the thesis, and beyond the typical scope of econometrics, lies a vital set of steps relating to the management of the data collection process. Clearly, proper management in this regard is a prerequisite for a good dataset and credible data analysis. Ross *et al* (2004: 3) provide a detailed and useful set of steps for managing the data collection process for a school monitoring programme.

The topics covered in this section are: (1) research and questionnaire design; (2) sampling technique and record weighting; (3) collection methodology; (4) data accuracy and reliability; and (5) typical distribution of values. The approach will be to cover some theory, bring in discussion of specific features of the three monitoring programmes described in Appendix A, refer to school production analysis texts not relating to the three programmes, and report on a limited analysis of values occurring in the SACMEQ database. This is also the approach taken in sections 4 to 9 dealing with the remaining six analysis steps.

What all three of the monitoring programmes in Appendix A have in common is that, firstly, they aim to gauge the academic performance of learners and, secondly, that they aim to find explanations behind the performance trends that can be translated into policy improvements. All three programmes use performance tests and questionnaires as data gathering instruments, and all the programmes gather data from a sample of the schooling system, where the aim is for that sample to be representative of the entire schooling system at the level of the school grade, or grades, in question. All the programmes produce cross-sectional data. It is important to understand this aspect of the data as this has implications for what can and cannot be done with the data.

Figure 2: Types of datasets

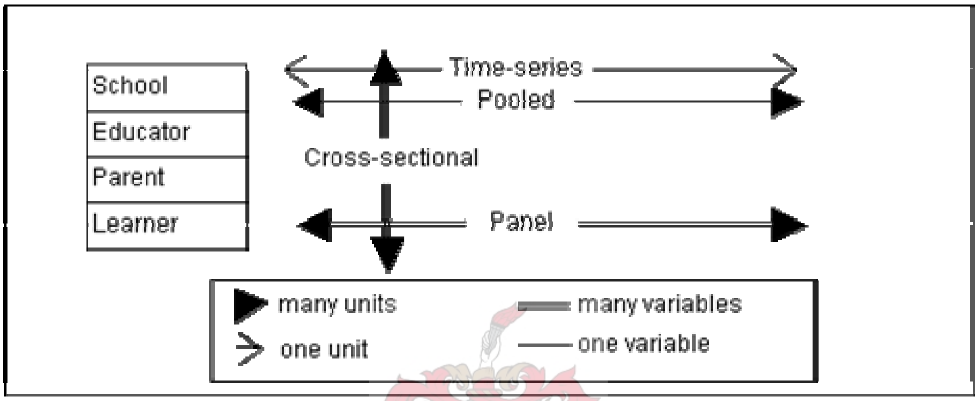


Figure 2 compares cross-sectional data to other types of data. The cross-sectional data that we have in the three programmes cover many variables across four types of units: for instance the mathematics score of the learner, the level of education of the parent, the years of pre-service training of the educator and the age of the school (only one programme includes the unit parent, however). Cross-sectional data describe the units of analysis in one point in time (hence the cross-sectional dataset is represented by a vertical line in figure 2) and in terms of many variables (hence the double line). This must be distinguished from the data types that include data from more than one point in time, and are hence represented by horizontal lines in Figure 2. Time-series data may, for instance, track the country's average mathematics score across many years (the country is the unit here). Pooled data may, for instance, track the average score of many schools (school is the unit) across many years. Finally, panel data would collect data on the *same units at the lowest level* across many years. So, for instance, panel data could comprise the mathematics scores of the same learners across many years (Gujarati, 2003: 27).

Whilst each single run of the three monitoring programmes produces cross-sectional data, over time each of the programmes produces pooled data (and, by implication, time-series data), as each programme has the potential to produce variables such as average mathematics performance in a state (SAEB), province (Systemic Evaluation) or country (SACMEQ) over time. However, none of the programme is designed to produce panel data, as none of the programmes targets the same set of learners, educators, schools or parents from one run to the next. There are good reasons for this. Tracking, for instance, the same schools would bring in serious selection problems as it is likely that education authorities would place special emphasis on the selected schools, leading to a non-representative sample of schools. The approach does limit the scope for analysis, however. In particular, the value-added model of school production explained in section 5.3 cannot be implemented.

The next table, focussing on the use of questionnaires (Q) and performance tests (T) for the different unit types, reveals some key differences between the programmes.

Table 1: Data collection units

	Brazil SAEB	*SE*	*SACMEQ*
School principal	Q	Q	Q
Educator	Q	Q	Q T
Parent		Q	
Learner	Q T	Q T	Q T

Brazil's SAEB can be regarded as incorporating the bare minimum for a programme of this nature. Questionnaires gather data from the school principal, the educator, and the learner, and learners are tested. South Africa's Systemic Evaluation goes further and incorporates a parent questionnaire. This allows for important verification of the home background data collected through the learner questionnaire, though there would be serious cost implications. For example, it would be impossible to conduct a school visit in just one day as fieldworkers would need to return on a second day to collect the parent questionnaire from learners. SACMEQ incorporates a teacher test. Given the importance of the knowledge of educators as an explanatory variable (Harbison and Hanushek, 1992), this feature appears important. Unfortunately, SACMEQ teacher tests were not run in South Africa or Mauritius due largely to teacher union opposition.

The following table focuses only on the SACMEQ data obtained through the three questionnaires. Variables in the dataset are classified by their unit type and by the type of data, which relates to the type of question in the questionnaire. The number of units of each type covered in the sample is also indicated, under *n*.

Table 2: SACMEQ questionnaire variables by type and level

	n	*Ratio*	*Interval*	*Ordinal*	*Nominal*	*Total*
Learner	3,162	17	1	37	19	74
Educator	326	26	0	105	7	138
School principal	169	74	1	89	5	169
Total		117	2	231	31	381

Ratio data is data where we can say, for example, that the value for variable *X* in the case of school A is twice that for school B. An example would be the number of visits by the education authorities to the school during a year. Interval data allow us to say that the difference between school A and school B is twice the difference between school B and school C with respect to variable *X*, but do not allow us to say that the value for school A is twice that for school B, basically because zero is not statistically meaningful. Year in which the school was established would be one of the rare examples of interval data we would find in this kind of dataset. Ordinal data allows us to say the value for school A is greater than the value for school B, but in a loose way that does not allow us to say, for example, that school A's value is twice that of school B. Most of the SACMEQ questionnaire data is ordinal data. An example would be the teacher's grading of in-service training received along a scale spanning the levels not effective, reasonably effective, effective and very effective. Many of the ordinal variables in the data are binary, focussing on the presence or absence of something, for instance a teacher table in the classroom. Nominal data does not allow for any ranking of, for instance, school A against school B because we are dealing with a non-rankable difference, for instance the difference between a male and a female school principal. Clearly, the breakdown summarised in the above table is not watertight. In particular, the difference between nominal and ordinal binary variables may be debatable. However, the picture provided by the table seems important, as it has implications for how we analyse our data.

As suggested by the table, we have relatively good information on the characteristics of the school. Our information gaps relate largely to the learner and the educator. In particular, we have a paucity of information about the ‘black box’ of what occurs within the classroom (discussed in the section on our mental model below). Data on the socio-economic background of learners and educators is on the whole good, but a richer set of questions aimed at learners and educators focussing on classroom activities would have allowed for a more informative production model.

The prevalence of ordinal data in the dataset has implications for our variable selection and manipulation step. As we shall see in section 6, there is a need to collapse groups of ordinal variables into ratio variables, partly to reduce the number of variables, and partly because raw ordinal data is often not suitable for regression models.

Ross *et al* (2004: 20) say the following about optimum sample design:

> The “best” sample design for a particular project is one that provides levels of sampling accuracy that are acceptable in terms of the main aims of the project, while simultaneously limiting cost, logistic, and procedural demands to manageable levels.

There is thus a trade-off between methodology and cost. This applies to both the sample size and the sample type. The final choices made can be well or not so well informed, but there is never only one solution. Sample design will be discussed briefly here, to provide a sense of some of the options relating to school monitoring programmes, and to provide a clearer picture of the SACMEQ dataset and its record weights, given that we shall analyse this data.

Three of the basic sampling approaches explained by Blalock (1979: 553) have relevance for our discussion: simple random sampling, stratified sampling and cluster sampling. The first is a useful point of departure for considering more complex approaches. In simple random sampling, we build a sample in one stage, and each learner has an exactly equal probability of being selected. Applied in the education context, the approach would mean, for instance, testing learners in a completely random way, regardless of their schools. We might then end up testing one learner in a school. This is clearly a very costly approach. In stratified sampling, we identify strata, for instance school districts, and then select an equal proportion of learners from each and every district. This too is likely to result in the testing of one learner in some schools. In cluster sampling, we identify clusters, which are smaller than strata. Typically, schools would be clusters. We would include only some schools in the sample as a first stage in the sampling process. Then, in a second stage, we would select only some learners from each of the selected schools. This is clearly more economical. We shall see that actual approaches are more complex than this, and involve a mix of the stratified and cluster approaches.

In SAEB, five dimensions are used to group all of Brazil's Grades 4, 8 and 11 learners into 389 strata. The dimensions are grade, state, school ownership, rurality, and status with respect to school shifts (it is common in Brazil to have separate morning and afternoon shifts involving different learners in the same school). Within each strata, classes are considered clusters, and a number of these are selected according to an algorithm. Within each selected cluster, all learners are tested. Despite the complexity of the approach, there is just one sampling stage, the one in which classes are selected.

SACMEQ, which considers all Grade 6 learners to be the population, has two sampling stages. Regions within countries are used as strata. Within each strata, a number of schools are selected according to an algorithm that ensures that within each selected school, only twenty learners need to be tested. The random learner selection is then the second sampling stage.

Both SAEB and SACMEQ rely on the school census data for the respective countries being adequately reliable. More details on the sampling approaches are provided in Appendix A.

In both datasets, weighting of records would be necessary due to the sampling approach (in SACMEQ) and due to realities observed on the day of the test (this would apply to both programmes). In the SACMEQ data, record weights are proportional to the reciprocal of the probability of including a learner in the sample (Ross *et al,* 2004: 31). In other words, the probability that a learner with a weight of 0.2 was included in the sample is twice the probability that a learner with a weight of 0.4 was included in the sample. Two weights per record are provided in the SACMEQ data: *pweight1* and *pweight2.* The second weight takes into account two important bits of reality not considered in the first weight. Firstly, enrolment on the day of the test would not be the same as the enrolment figures that were used when the sample was constructed (the previous year's school census data would have been used). Secondly, some learners were absent on the day of the test (it was assumed that the absence was random). The sum of *pweight1* and the sum of *pweight2* are each equal to the number of records in the dataset. Wherever data is weighted in the analysis contained in this thesis, *pweight2* is used.

In the cases of both SAEB and SACMEQ, fieldworkers who are not part of the school community managed the completion of questionnaires by respondents in the school, and the writing of the tests. This clearly contributes immensely to the reliability of the data, and hence the policy conclusions.

In discussing the accuracy of a dataset, Gujarati (2003: 29) refers to problems of non-response (the questionnaire was not filled in at all), rounding off (the precision of ratio values was somehow reduced), omission (individual questions were not answered) and commission (respondents deliberately provided incorrect responses to individual questions). In school monitoring programmes, the last two problems are likely to be most prevalent.

Missing or incorrect data can be the result of both the fieldwork, specifically the questionnaire completion process, and the post-fieldwork of capturing data onto computers. Controls and follow-ups can be put in place to improve the situation, but at the point when the data capturers are sent home and ‘normalisation’ or ‘cleaning’ of the data begins, we will inevitably be left with residual problems of three types: missing values, impossible values and unlikely values. What is vital to the analyst, is that the normalisation process and the eventual state of the normalised data should be well documented, as this can have serious implications for how the data is interpreted.

The SAEB 1999 data normalisation reports are relatively comprehensive (see Barbosa *et al*, 2000). Special non-zero values were inserted to indicate where values were missing. Different non-zero values were used to indicate where impossible values had been removed. An example of an impossible value would be number of books in the school library if other responses clearly indicate that the school does not have a library. Variables with unlikely values, for instance extremely high numbers of computers within one school, were transformed so that no values exceeded what was regarded as a reasonable level. Not having these transformations would result in unrealistic grand totals. Critically, these transformations are documented for the use of the analyst.

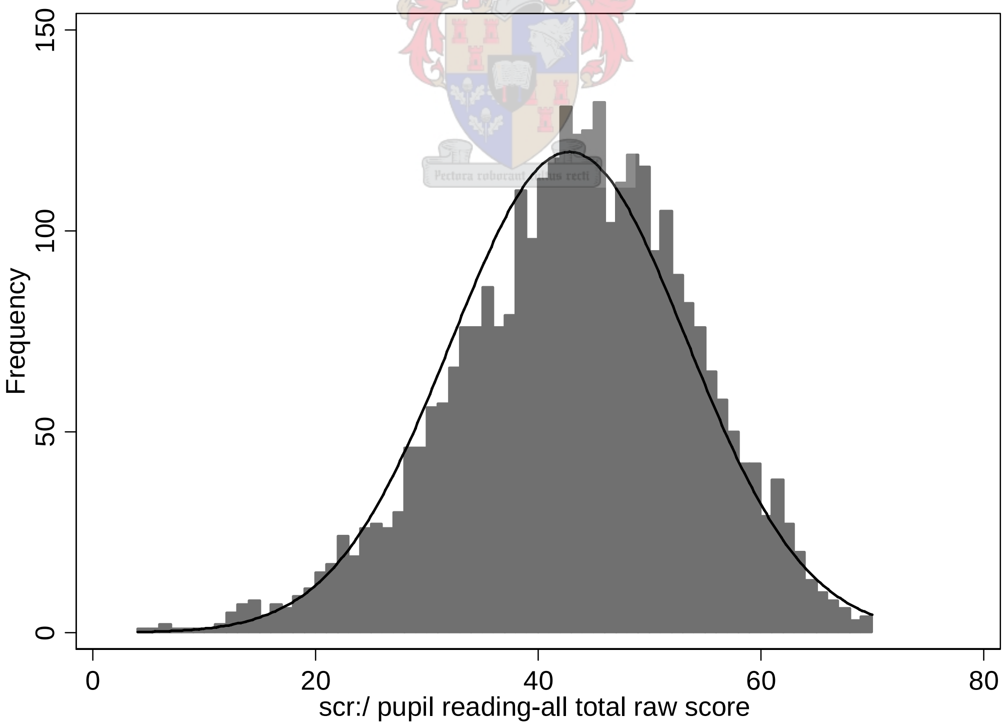
The SACMEQ data also comes with a large volume of technical documentation. This documentation explains how the bulk of the 381 questionnaire variables broken down in table 2 were transformed into new variables with simplified coding. (The 381 original questionnaire variables, plus the recoded versions of these, plus the test results render a total of 1,274 variables in the final dataset.) It is also explained how missing values were, in some instances, imputed from available data, according to a set of rules. The number of imputed values in the data was not covered in the original documentation, but was obtained on request (the pre-imputed dataset, which would make it clear which records had missing values, was not available, however). This information revealed that at least for South Africa, imputation of new values to replace missing values was fairly limited. The number of variables for which 50 or more values were imputed (this cut-off corresponds to about 1.5% of all records) was 2, 15, 7 and 5 for the learner, school principal, reading teacher and mathematics teacher questionnaires respectively. Imputation did not entirely eliminate missing values. One of the principals did not respond at all. In three schools both the reading teacher and the mathematics teacher did not provide any responses, in two schools just the reading teacher data was missing, and in two schools just the mathematics teacher data was missing. A more detailed profile of missing data is provided in Appendix C. Overall, the extent of missing values in the SACMEQ data is low relative to that in Brazil's SAEB. About 20% of the learners did not have any questionnaire data in SAEB. It is relatively easy to establish that the distribution of missing values in a typical school monitoring dataset is not random. Historically disadvantaged learners and schools have a disproportionately high number of missing values. (This was found in the Brazil PISA data, for instance.)

In the selection and manipulation of SACMEQ variables, described in section 6, the recoded SACMEQ variables were avoided, and the original variables were used instead, in the interests of accuracy and given considerations specific to South Africa (the recoding occurred on the basis of assumptions relating to the region as a whole). Moreover, the variables with high levels of imputation were avoided.

Whilst getting to know the data, the analyst stands to gain from an examination of the distribution of values within key variables. A fundamental concept in this regard is the normal distribution of values. In econometrics, gauging normality becomes especially important when we look at error terms (see section 5.2 below). But the techniques are equally applicable to the raw values. An examination of the normality of the SACMEQ performance scores will be shown to be instructive.

The distribution of reading scores in the case of Mozambique serves as a useful point of departure. (Importantly, the score values used here and elsewhere in the thesis are the raw scores obtained from the learner tests. The SACMEQ dataset also includes several recoded score variables, where each variable uses the curriculum of one SACMEQ country as a base, and furthermore adjusts values so that the SACMEQ mean equals 500. For example, the variable named *zr_mal2* carries a reading score for every SACMEQ learner using only the reading questions supported by the Malawi reading curriculum, and scaled in such a way that the all-SACMEQ mean is 500. These recoded values are not used in this thesis at all, though they are commonly used in other analyses of the SACMEQ data. The choice of the variable makes virtually no difference to the findings.)

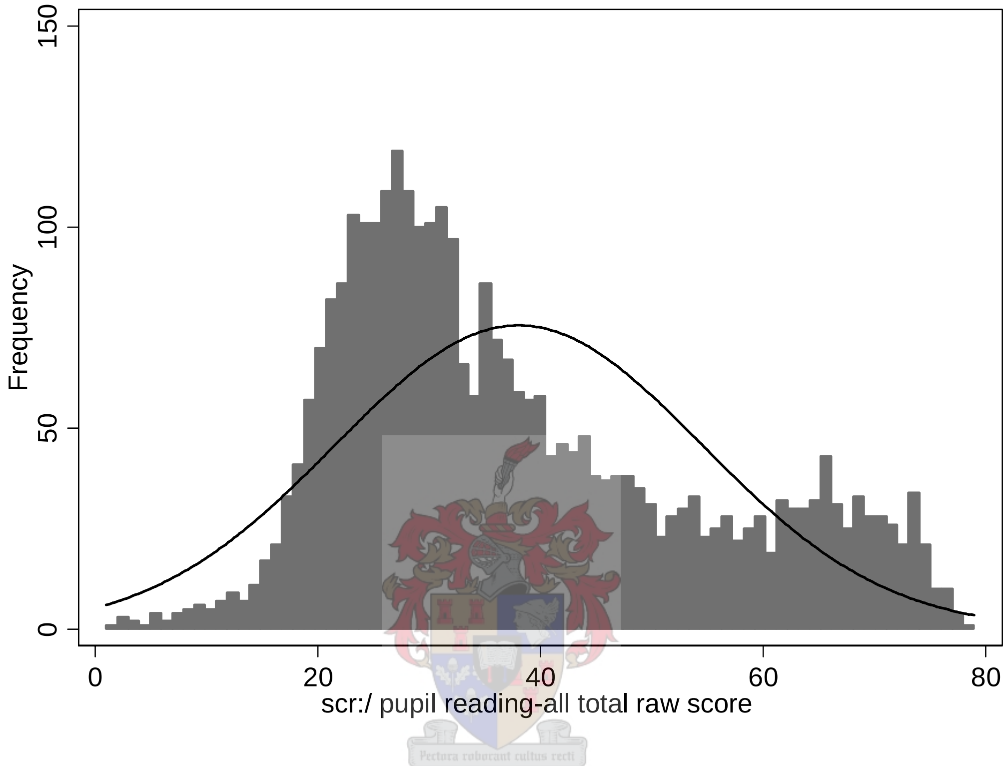
Figure 3: Histogram of learner reading score in Mozambique in 2000



Source: SACMEQ, 2000.

Stata's superimposed normal curve indicates that the actual distribution of reading scores in Mozambique is highly normal. Similar results are noted for Brazil in the SAEB technical documentation. The pattern in South Africa, however, is very different.

Figure 4: Histogram of learner reading score in South Africa in 2000



Source: SACMEQ, 2000.

In the case of South Africa, we find a larger peak for historically disadvantaged learners on the left, and a smaller peak for historically advantaged learners on the right. In fact, we have two relatively normal curves that overlap each other, resulting in an overall bimodal pattern that is highly non-normal. The historical inequalities and the dual economy phenomenon reflected in the above graph have immense implications for any analysis of South African schooling data (as we shall see in this thesis), and for inter-country comparisons (which are not attempted in the thesis in any depth).

Whilst visual inspection of the distribution of values is useful, a more precise and consistent indication of normality is obtained through a statistic such as the Jarque-Bera test statistic (Gujarati, 2003: 886). This statistic requires four other statistics: the mean, variance, skewness and kurtosis. These statistics will be discussed in relation to the reading score for South Africa.

The mean reading score is 38.04. The variance would be calculated as follows:

$$\sigma^2 = E(X-\mu)^2 \tag{5}$$

where μ is the mean of the reading scores. Essentially we calculate a new value for each learner equal to the squared difference between the learner's score and the overall mean, and then calculate the mean of all the new values. The above gives us a result of 271.47 for South Africa. Skewness is calculated as follows:

$$S = \frac{E(X-\mu)^3}{\sigma^3} \tag{6}$$

This gives us 0.674. Zero would mean normal skewness, a positive value means a long 'tail' to the right (this is what we have in the case of South Africa) and a negative value a long tail to the left. Kurtosis is calculated as follows:

$$K = \frac{E(X-\mu)^4}{\sigma^4} \tag{7}$$

giving us 2.508. A value of 3 means a normal vertical shape, less than 3 means a flatter than a normal curve, and greater than 3 means more vertically elongated than a normal curve. We have a flatter than normal curve in the case of South Africa.

The Jarque-Bera statistic is calculated as follows:

$$JB = n\left(\frac{S^2}{6} + \frac{(K-3)^2}{24}\right) \tag{8}$$

The value we obtain for South Africa's reading score is 271.5. Zero would have meant a completely normal distribution. The Jarque-Bera statistics for the reading scores of all SACMEQ countries are provided in table 3.

Table 3: Normality of reading score distribution in SACMEQ countries

country	Jarque-Bera
Botswana	71
Kenya	76
Lesotho	954
Malawi	593
Mauritius	172
Mozambique	57
Namibia	1508
Seychelles	95
South Africa	272
Swaziland	52
Tanzania (mainland)	96
Uganda	203
Zambia	563
Zanzibar	29

Source: SACMEQ, 2000.

South Africa's distribution is the fifth least normal. In all the four countries with less normal curves, there is a long tail on the right, and in the case of two of the countries, Namibia and Uganda, we can clearly observe the higher peak on the left and the lower peak on the right, similar to what we saw in figure 4. The precautions one would need to take in analysing the data from a highly unequal country such as South Africa, would apply to a number of other SACMEQ countries too.

Comparing variances across different variables is sometimes included in the preliminary examination of a dataset. For this, we need a comparable variance statistic such as the coefficient of variation, which is calculated as follows (the notation from the previous formulas is used):

$$cv = \frac{\sqrt{E(X-\mu)^2}}{\mu} \tag{9}$$

As there are very few ratio variables amongst the original variables of the SACMEQ dataset, we turn, prematurely, to a few of the new variables that were created from the original variables (all of the variables below are ratio variables), in order to examine variances across the explanatory variables, and to compare this to the variances of the performance scores. Whilst a coefficient of variation can be calculated for many binary unconverted nominal and ordinal variables, ratio variables yield more meaningful statistics on variance.

Table 4: Coefficient of variation of several variables

Variable	coefficient of variation	
math_score/read_score	0.440	0.438
textbooks_math/read	**0.618**	**0.448**
daily_meals	0.287	
parent_educ	0.401	
learner_ses	**0.676**	
yrs_preserv_math/read	0.033	0.033
day_inserv_math/read	**1.984**	**1.735**
teacher_ses	**0.574**	**0.518**
hrs_year_math/read	0.399	**0.476**
par_involve_math/read	**0.857**	**0.816**
school_infra	**0.738**	
yrs_preserv_prin	0.049	
prin_teach_load	**0.849**	
dist_support	**1.067**	

Source: SACMEQ, 2000.

In general, there is more variance amongst the explanatory variables than amongst the performance scores (values in bold indicate variances that are greater than the variances of the performance scores). This is what Vinjevold *et al* (2001: 28) find in another South African schools dataset, to their surprise, as they expected the variance for the performance statistics to be higher. Evidently, greater variance in the explanatory variables is not unusual. Even before we directly link inputs to outputs within a production model, aspects of the relationship between inputs and outputs are being revealed. If inequalities on the input side are greater than inequalities on the output side, this suggests that performance is relatively resilient to resourcing inequalities. Resourcing inequalities, whilst undesirable, do not necessarily imply the same degree of inequality in terms of the education that learners receive.

4 BUILDING A MENTAL MODEL

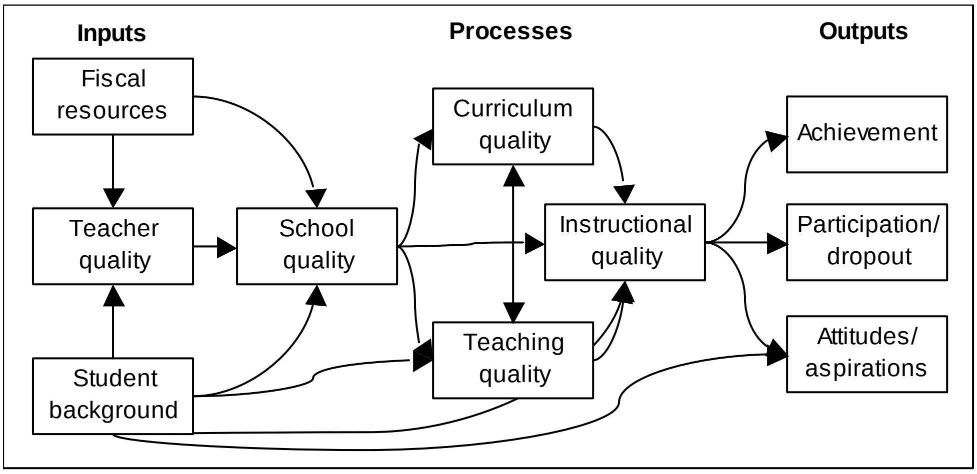
The Economist (2004: 63) argues that ‘A failure to separate statistical significance from plausible explanation is all too common in economics, often with harmful consequences’. Even reputable economists, the argument goes, are too often guilty of deceiving their audiences through over-emphasis on the outputs of an econometric model, and under-emphasis of more commonsensical and historically informed models of how particular economic phenomena work.

The term ‘mental model’, taken from Baker (2001: 83), will in this thesis refer to a non-statistical model of how school production occurs. The mental model is thus the hypothetical model that we use as a point of departure, that we attempt to prove or disprove through our data modelling, and that we adapt and modify as our understanding deepens. If it is to be adequate, the mental model must take into account three key things: (1) what the range of inputs is at the levels of the learner, the class and the school and how these inputs interact to generate education; (2) how the hierarchical nature of schooling systems might influence production, and (3) what previous school production models, and previous findings from effective schools research and other methodological approaches, indicate about the production process. These three points are dealt with in the following three sub-sections. A fourth sub-section examines mental models used in the monitoring programmes described in Appendix A and a fifth sub-section presents a policy-oriented mental that will be used extensively in the analysis contained in the thesis.

4.1 Education stocks and flows

The following diagram, taken from Kaplan and Elliot (1997: 324), represents a typical mental model of the schooling process. This model is pedagogical rather than economic in its terminology, but it can nonetheless be a useful point of departure for conceptualising a school production model.

Figure 5: A mental model of school production



Economists have in fact looked to the theory of pedagogy to obtain the fundamental elements of a school production model. Thus far, the demand for these fundamentals has not been satisfied. It is commonly argued that the theoretical basis for how education occurs is weak, making it extremely difficult to analyse the economics of education production (Hanushek, 1979: 363). Moreover, the wide range of inputs used in education production models has been criticised. Whilst the inclusion of technology as an input in the production process occurs in both educational and non-educational models, it is only in education production models that we find a great range of organisational, management and process variables being considered as inputs. In terms of figure 5, treating curriculum as an input is less problematic, insofar as the curriculum represents the technology of education, but treating the processes of the classroom, in other words instructional quality, as an input, has attracted more criticism.

Questions around what constitutes a legitimate input in a production model carry important implications for our notion of school efficiency. They even have important philosophical repercussions. For instance, if the level of willingness of educators to teach is regarded as an input, then we may reduce the degree of inefficiency in the system. Unwillingness to teach becomes a negative input, in the same way as a lack of textbooks would be, and poor learner performance becomes a predicted outcome, as opposed to an inefficiency that can be dealt with through management intervention.

If we impose some economic terminology on the model in figure 5, we could say that as inputs into the production process we have three distinct human capital stock inputs. We have teacher quality, and captured within the student background, we have the human capital of the learner's parents, and of the learner himself. What is perhaps missing from the diagram is a clearer reference to the physical stock of the school, and the physical tools that support learning, such as textbooks, library books, wallcharts, and so on. Moreover, the physical environment of the home, and in particular whether the learner has access to electrical lighting and a space to study, would be difficult to situate in the figure 5 schema.

The various inputs of the model should be considered in terms of their measurability. Clearly certain inputs, for example the condition of the school buildings, are easy to quantify. Other inputs are either present or not, for example free school lunches of a particular standard. Some other inputs can be extremely difficult to quantify, for instance teacher quality. It should be noted that in figure 5, all paths ultimately originate in one of two sources: fiscal resources or student background. According to this view, everything except for student background is generally purchased or hired. It is thus possible to quantify school inputs in terms of their prices. However, in particular in developing countries it is common for households to supplement school inputs such as stationery and books, raising the risk that a model like the one in figure 5 will under-estimate the inputs.

Turning to specific inputs, the teacher input is both quantitative (for instance in terms of contact hours and the ratio of learners to educators) and qualitative. Two key aspects of teacher quality need to be considered: ability to manage the classroom situation and subject knowledge. The latter should not be taken for granted. Harbison and Hanushek (1992: 110) demonstrate that educators often obtain alarmingly low scores in the same standardised tests used for learners in school monitoring programmes. The language skills of the educator are a key aspect of teacher quality, in particular in a multi-lingual society such as South Africa. Moreover, how switching between languages occurs in the classroom is an important determinant of instructional quality. Interestingly, the linguistic qualities of the educator are not just an issue in multi-lingual contexts. Monk (1990: 353) identifies variations in teacher verbal ability in the United States as a major factor behind learner performance. He doubts that there is much that policy can do to influence this factor, as it is linked to very individual traits of the teacher.

Harbison and Hanushek (1992: 95) emphasise the importance of the level of education of parents in the education process. Not having fairly detailed information on the skills and knowledge of parents (this is the case in most monitoring programmes) can be regarded as a serious drawback. Data on the highest level of education attained by the parents, and the level of income of the parents, are the proxies usually used in place of the more detailed data. It should be emphasised that the income of parents as such is of limited importance in any model. The issue is the parents' skills and knowledge, which income is meant to proxy.

The stock of human capital of the learner herself is obviously at the centre of the whole production process. In many developing countries, it is especially important that this stock concept should cover the physical health of the child. Typically, measures such as the height-weight-age ratio of learners and triceps skinfold thickness are used (Harbison and Hanushek, 1992: 125).

With regard to physical stock, Glewwe *et al* (2000: 15) emphasise the importance of distinguishing between the actual availability in the school of stock such as textbooks, access by learners to that stock and usage of the stock. It would be dangerous only to focus on the availability of, for instance, textbooks in a school without also focussing on how learners access and use the materials. We need to be aware that stock may be locked away in school storerooms, and hence not be a part of the production process.

The instructional quality box in figure 5 is often considered the black box of education production. How educators, learners, the curriculum and physical resources come together in the classroom is the key to understanding education production, and yet we know very little about this aspect of the process. It is possible to gain a clearer empirical picture, but the analysis process is laborious and costly, and requires skilled analysts (Monk, 1990: 312). Glewwe *et al* (2000: 21), in their time-series analysis of the impact of textbook provisioning in a Kenyan NGO project, provide a rare glimpse into the black box in a developing country context. As part of this Kenyan study, video tapes of classroom activities were analysed to create a database of 85 variables, with values for each minute of the class. Variables covered such things as the tools used by the educator (textbook, chalkboard, worksheets, and so on), the type of educator-learner interaction occurring (was the educator interacting with one learner, a group of learners, the whole class, or no-one?), level and type of engagement by learners, and the language being used by both the educator and the learners (the situation was strongly multi-lingual).

Absent from the model illustrated in figure 5 are important systemic factors, such as the administrative processes to which the school is subjected. Requirements to account to public or private funders, tax systems, policies regarding the admission, promotion and assessment of learners, and laws regarding the employment of personnel can all arguably be regarded as inputs in the production process (considering we have already created such a wide definition of what an input is). A very commonly used variable in this regard is whether a school is private or public.

This discussion of the range of inputs is highly selective and does not get close to being comprehensive. But it should provide a sense of the basic categories, and the potential complexities of individual inputs. Apart from having an adequate sense of each input, our mental model should also deal with the complexities of the inter-relationships between inputs.

Harbison and Hanushek (1992: 98) regard the relationship between the gender of the educator and the gender of individual learners as an important classroom variable. The issues relating to gender can also be said to relate to race in the South African context, at least where there is a racial mix in the classroom or the school.

We can expect the inter-relationship between inputs to be influenced by the size of the production unit, for our purposes the school. The economic orthodoxy of increasing returns to scale followed by decreasing returns to scale, as the production unit grows, was discussed in section 2.1. Our model needs to take a position in this regard. Does the economic orthodoxy apply to schools, and if so, can we speak of an optimal school size in terms of the efficiency of the production process? The common understanding is that a larger school permits a greater degree of specialisation by educators. If there are several classes and educators within one grade, for instance, there is greater scope for specialisation in terms of (1) level of performance (the school can practice streaming, or can establish parallel remedial classes more easily), (2) learning area in the curriculum (certain educators can specialise in mathematics, for instance) or (3) language of instruction. This would appear to indicate better efficiency. Monk (1990: 399), however, reminds us of the important trade-off between productivity gains relating to specialisation, and losses relating to more complex and time-consuming coordination, management and timetabling in the school (to some extent, these losses can be described as transaction costs). Burstein (1980: 166) and others often remind us of the importance of distinguishing between primary and secondary schooling. Returns to scale dynamics would clearly work very differently at these two schooling levels, especially where the curriculum in the secondary level allows for considerable diversity. We can expect increasing returns to scale to continue through to a larger size in secondary schooling than in primary schooling.

The relationship between inputs and outputs may be less unidirectional than one would expect. Monk (1990: 334) describes the simultaneity problem as one where inputs are influenced by the level of output. For example, learner achievement may well influence instructional quality, meaning there is a backward movement along the arrows in figure 5. Better test scores, in particular test scores that are better relative to those of peers in the same class or in similar classes in the same school, could lead to greater motivation in the classroom, and hence better instructional quality.

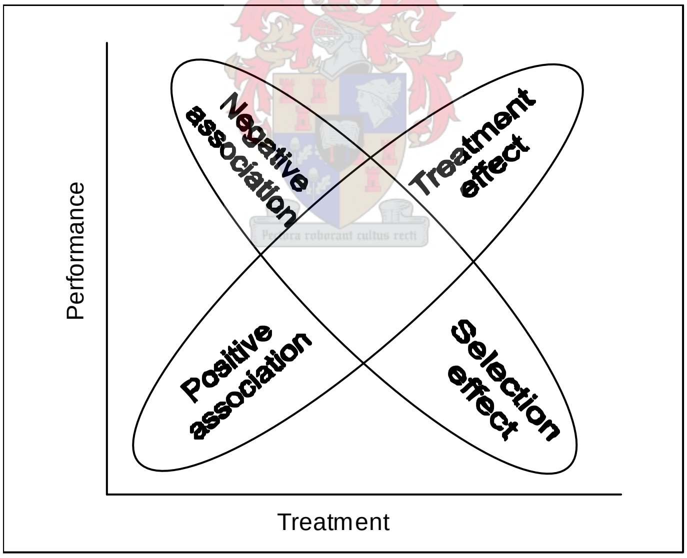
The problem Monk discusses is one of defining causality. This, in a mental model or an econometric model, is an especially important task, and one that, if incorrectly undertaken, can lead to erroneous policy conclusions. Notwithstanding some attempts at formulating methods to identify causality in econometric models, such as Pearl's directed acyclic graph, described by Berk (2004: 192), the analyst is left with virtually no econometric tools to deal with the issue of causality. As Berk (2004: 101) emphasises, '[t]here is nothing in the data by themselves that properly can be used to directly determine if x is a cause of y (or vice versa). There is nothing in simple linear regression that, by itself, will lead to causal inferences.' The implication is that information outside the econometric model, largely the logic and meaning of the mental model, must play a key role in advising on directions of causality. This is particularly so when using cross-sectional data. Causality is inextricably linked to processes over time, and time is something that cross-sectional data cannot deal with in any direct manner.

It is useful to bear in mind what would be demanded in the research for us to be truly certain that changes in the value of an X variable in equation (3) above were responsible for changes in the value of variable *Y*. We would in fact require an experiment, where we would select units, for instance learners, at random, and apply different treatments or interventions, in other words different values for one of the *X* variables, to different randomly defined groups. Depending on the resultant changes in the value of *Y*, conclusions might be drawn about the causal effects of *X*. The implication is that *X* is subject to manipulation, for instance policy manipulation. This would indeed be the case with respect to, for instance, the application of different types of textbooks to different groups of learners. Berk (2004: 82) refers to this as 'what if' causality, because we are implicitly asking 'what if' a specific intervention were applied. In contrast to this, 'but for' causality is identified where inherent characteristics, for instance the gender of learners, are said to exert a causal effect. We could randomly select learners and gauge performance differences based on gender, and say that 'but for' the learner's gender, he/she would have performed so much better or worse. In such an experiment, we would have no direct control over the assignment of gender, in other words we could not manipulate the *X* variable. Thus, even if associations between *X* and *Y* were found, we would have to be less certain about what was causing the effect. Was it some biological or psychological factor associated with gender, or were we dealing with the effects of differentiated social conditioning? These questions might remain unanswered. Causality where the *X* variable cannot be manipulated is clearly more difficult to define and to estimate.

Rarely is data available on economic experiments, so the foregoing discussion on the necessary basis for identifying causality serves more as a sobering reminder of the limitations inherent in the typical cross-sectional dataset, than as a guide on how to analyse the data. To some extent, however, cross-sectional data can be regarded as 'quasi-experimental' data in the sense that, if the sample is large enough, it may provide sufficient instances where units differ only with respect to one *X* variable, and not the others. Depending on how uniform the schooling system is, this quasi-experimental nature of the data may be relatively high. Case and Deaton (1999) argue that the schooling system for Africans in South Africa under apartheid provides data that is useful for gauging the causal effects of class size, because so many other factors, including parental choice and school conditions, were so restricted and uniform. Only class size, they argue, varied greatly across schools.

Much of the complexity with regard to understanding causality in cross-sectional data relates to the interplay between treatment effects and selection effects. In a purely experimental situation, treatments are randomly assigned to different units. What confounds the analysis of causality in a real world situation is that there is non-random assignment of treatments to units, or various selection effects. The example of in-service training of teachers as a treatment is instructive. We would expect the treatment effect to be positive – the greater the recent exposure of teachers to in-service training, the better the learner scores should be. Selection effects, on the other hand, would exist where the state or some other agent had recently concentrated in-service training in a non-random manner on those teachers most in need of training, or those in schools with the lowest learner scores. We would expect selection effects to be negatively associated with performance. The opposing directions of the two effects, based on the discussion by Berk (2004: 225) on the matter, can be illustrated as follows:

Figure 6: Treatment and selection effects



If the treatment, for instance in-service training, were randomly distributed across the schooling system, we would expect a positive association with performance, at least after the lapse of some time. The time aspect is of course crucial. When the treatment started, there should be no difference in terms of performance. It would only be after some time that the differences would emerge. Cross-sectional data generally provides some sense of time, for instance (as will be discussed in section 6) teachers are asked in the questionnaire about the in-service training they received over the previous three years. As the selection effect involves the non-random targeting of the treatment towards those most in need of the treatment, we would expect a negative association between the treatment and performance, even after some time had lapsed, because the treatment would take time before it closed the gap between the worse performers and the better performers. Importantly, for figure 6 to be meaningful, it should control for variables other than performance and treatment, in other words it should represent schools or learners that were in other respects similar, or where the performance variable had been adjusted statistically, on the basis of the other characteristics, to make the schools equivalent (this is done in section 7). Our concern in the education production function of equation (3) is clearly with the treatment effect, or the causal effect of the treatment on performance. The selection effect is what bedevils the data as far as the production function analyst is concerned. One problem is clearly that if the selection effects dominate figure 6, we cannot discern the treatment effects at all. It is possible that all schools with high performance are not receiving the treatment, and that all schools with low performance are receiving the treatment. However, even if both the selection and treatment effects co-exist, the superimposing of one effect on the other would make it difficult, though theoretically not impossible, to discern from the data how positive the treatment effect was, or what the slope of the treatment effect would be in figure 6.

The causality issues we have discussed are often referred to as endogeneity problems. The production function in equation (3) assumes that the X variables are exogenous, in other words determined outside the model, and that the Y variable is endogenous, in other words determined within the model. Clearly, there is a possibility that the X variables may not be strongly exogenous, because to some extent they are sensitive to the values of Y. They then become at least partially endogenous. What we have not discussed is the endogeneity problem where one X variable exerts a causal effect on another X variable. Clearly, there are an enormous number of such relationships in a schooling process. Class size affects who wishes to teach in a school, which affects the quality of teachers available. The tools at the disposal of the teacher affect the teaching methodology applied by the teacher. These endogeneity problems are arguably less serious than endogeneity problems involving the causal effect exerted by the output, or learner performance, as the latter type interfere more fundamentally with the idealised production function.

Turning to the outputs themselves, there is understandably much controversy around how to construct the output variable Y from equation (3). Figure 5 presents a much wider range of outputs than what is usually incorporated in econometric models of school production. These models usually cover only learner performance in the output variables, specifically results from standardised tests. Hanushek (1979) argues that the test scores approach is difficult to justify. Yet models using a narrow test score output remain the norm. Two arguments could be put forward to support this. Firstly, if test scores are sufficiently correlated to other desirable outputs, such as attitudes, aspirations and an ability to participate in the labour market, then the test scores can serve as convenient proxies of the other outputs (apart from being measures of desirable outputs themselves). One important thing that better performance scores are undoubtedly correlated to, is reduced dropping out of the schooling system – the better the scores are, the more likely it is that learners will stay within the schooling system right up to the exit point (Hanushek and Luque, 2003: 483). Secondly, if we consider the test scores, or what they represent, as sufficiently important, then models of school production focussing on just this output are valuable tools providing us with key policy guidance. Hanushek and Luque (2003: 482) seem to adopt this standpoint when they refer to 'the knowledge base and analytical skills that are the focal point of schools'. (In fact, Hanushek's clear shift on this matter between 1979, when he argued that the test scores approach was difficult to justify, and 2003 is telling.) It is likely that the arguments work differently at different levels of country or educational development. There is probably a stronger argument for focussing exclusively on standardised test scores in developing countries experiencing problems with respect to access to basic education, or in schooling systems where literacy and numeracy performance is particularly low. Moreover, as Hanushek (1979: 362) points out, it is likely that the basics of literacy and numeracy matter more in the early grades than they do in the later secondary school grades, at which point other outcomes, for instance an awareness of optimal career choices, become relatively more important.

Lockheed and Langford (1989: 24) emphasise the importance of taking 'within-pupil variance' into account when considering test scores used as outputs. Factors unrelated to the skills and knowledge of the learner, such as health and emotional state, play a role every time a learner sits for a test. An average score obtained from more than one test is an effective way of dealing with this within-pupil variance.

We now turn to a key input not referred to in figure 5: time. Time both as a medium range input (for instance an input present during a year) and as a long range input (over the educational life of the learner) is obviously a critical factor in education.

Monk (1990: 371) argues that official time specifications are notoriously bad at predicting what amount of time is actually spent on education within, say, one year. Downtime, or time intended for educational usage that for one or another reason does not get used for education, is high, perhaps as high as 40% in the USA. Even when we do have relatively accurate data on, for instance, contact time between learners and their educators, this can vary vastly due to differing qualities of the contact time. In developing countries, given large rural populations and the problem of poverty, we can assume that the quantity and quality of contact time is partly dependent on two key services: (1) scholar transport (which could be a low cost intervention such as the subsidisation of bicycles) that obviates excessive walking and (2) school feeding programmes that combat hunger and malnutrition and encourage attendance. Educational contact time between parents and learners is often difficult to pinpoint in a typical data collection exercise, and so proxies are used, in particular family structure (a higher ratio of learners to adults is assumed to reduce time investment for each learner) and the work status of parents (parents who work are assumed to have less time to spend supporting the education of learners in the home) (Harbison and Hanushek, 1992: 95). Characterising contact time becomes truly complex and interesting when we delve into the details of time flows within the classroom, as some studies, for example that by Glewwe *et al* (2000), have done.

Time as a long range education input is inextricably linked to the matter of how long a learner stays in the education system. The differentiated retention of learners in the system is often referred to as a question of selection effects. Laws regarding compulsory school attendance, the direct and opportunity costs to the household of keeping a learner in school, and pressures pushing worse performing learners out of school play a role in determining when a learner leaves the system. These are non-random factors, and it is hence common to think of these factors as constituting selection effects that muddy the waters of causality in the analysis. Typically, dropping out in secondary schools is strongly associated with lower socio-economic levels. We can imagine comparing more advantaged schools to less advantaged schools in a model, and reaching the conclusion that in the less advantaged schools, we need more public resources to compensate for the disadvantages in the home background. If we have not considered selection effects, specifically dropping out, we then ignore the fact that in the disadvantaged schools, many learners would have dropped out before the point at which the analysis occurred. Typically, those learners who did not drop out in the less advantaged schools would be less socio-economically disadvantaged on average than those who dropped out. We would then have under-estimated the production cost of education in the less advantaged schools.

4.2 The hierarchical structure of schooling systems

Schooling systems are always hierarchical. Within any country, the schooling system is typically sub-divided into administrative regions and sub-regions, and then into schools, and then into classes, before we reach the level of the individual learner. In the case of Brazil, an important hierarchical level between the school and the class is the shift.

The hierarchical structure of schooling systems has important implications for our mental model of school production. Associations between inputs and outputs occur at different levels. At each level of the system, there are important peer interactions between units within the same group, for instance between learners in the same class, between educators in the same school, or even between school principals in the same school district. The way learners and educators come to be in particular classes and schools must also be understood.

Bryk and Raudenbush (1992: 9) describe how the essentially universal positive association between SES and learner performance must be understood as existing at different levels simultaneously. In one school the relationship may be stronger than in another. In one school district, the associations may be stronger than in another. The differences between groups can be understood in terms of differences in management and in other factors apart from SES. The influence of one learner's SES on her performance at school can be understood as an amalgam of, firstly, factors relating just to that learner, secondly, the interplay between SES and performance at the learner's specific school and, thirdly, the interplay between the two variables at the level of the district. The relationship between the learner's SES and performance can thus be decomposed into individual, school and district effects. (Bryk and

Raudenbush do not bring the district level into their discussion, but doing so is completely consistent with their theory.)

Burstein (1980: 136) discusses two important kinds of peer effects. Compositional effects occur when general characteristics within the group have an effect on the performance of the individual in that group. A simple example would be the socio-economically disadvantaged learner who benefits from being in a class or school with more socio-economically advantaged learners, because private resources are shared and the advantaged home background of the peers rubs off somehow on the learner in question. Frog-pond effects, on the other hand, result from a learner's status relative to the peers, as opposed to the learner's absolute status. For example, a learner may benefit from the relatively higher status of being in a school with less advantaged learners. In such a school, the learner feels more advantaged, probably performs better relative to the other learners, and may consequently feel more motivated. It should be clear that the hypotheses underlying the compositional and frog-pond effects are contradictory in the examples provided here. The challenge in the modelling process is to determine which of the two effects is dominant in particular contexts. Peer group effects should not only be understood in terms of ratio variables. Nominal characteristics such as gender can also be linked to peer effects. Harbison and Hanushek (1992: 100) regard the gender mix of learners within the group as an important peer effect that influences learner performance.

To the selection effects dealt with in the previous section we can add the selection effects that determine which groups (which classes and schools) individual learners and educators fall into. Proximity and affordability of schools, and admissions requirements, all play an important role in determining which learners enter which schools. Within schools with several classes in each grade, placement of learners in particular classes may be random, or may be the result of a conscious streaming policy. These are the selection effects operating with respect to learners. With respect to educators, there are important selection effects that influence in which schools, grades and classes educators end up teaching. Educator distribution between public schools is often a function of both government policy and some degree of educator choice. Within schools, the management approach of the school principal is often a strong factor in the distribution of educators across grades and classes. Again, it must be emphasised that not taking cognisance of selection effects can have grave consequences for the analysis. The way educator preference influences the placement of educators in schools and classes has been highlighted as a pitfall (Hanushek, 1979: 374 and Monk, 1990: 334). For example, a positive relationship between educator age and learner performance may well be masking the fact that older educators have more choice in which classes they teach, and prefer to teach classes of better performing learners.

Whilst the selection effects relating to dropping out of learners can be controlled for in a model (this is discussed in section 5.3), selection effects relating to the placement of educators are much more difficult to deal with. In fact, solutions to the problem seem elusive. Häkkinen *et al* (2003: 330), in a study of school production in Finnish secondary schools, at least emphasise the need to deal with the problem in the design of their model, but then fail to provide an explicit solution. The solution may well have to lie at the model interpretation stage, as opposed to the model design stage. Having an adequate grasp of the policies and behaviours influencing such phenomena as teacher and learner placements seems to be an important prerequisite for the analyst.

4.3 The evidence so far

The preceding sections have examined the various inputs of the school production model in terms of their economic categories and the institutional hierarchies. Here we shall examine what the literature tells us about the relative strengths of these inputs, where strength is taken to mean proof that changes on the input side result in changes on the output side. This means that strength does not necessarily assume the meaning of economic and education policy importance (this dimension is dealt with in depth in section 8). We shall examine the issues only with respect to developing country dynamics. Production dynamics in the schooling systems of developed countries tend to be very different, as explained in section 2.2.

Hanushek (2002) provides a framework for undertaking meta-analyses of production function studies that has been deemed sufficiently useful to be reproduced in the 2005 *EFA Global Monitoring Report* (UNESCO, 2005: 65). The framework, with Hanushek’s figures, is reproduced below.

Table 5: Meta-results from developing country production function studies

	Number of estimates	Statistically significant (%)		Statistically insignificant (%)
		Positive	Negative	
Pupil/teacher ratio	30	27	27	46
Teacher education	63	56	3	41
Teacher experience	46	35	4	61
Teacher salary	13	31	15	54
Expenditure per pupil	12	50	0	50
Facilities	34	65	9	26

Copied from Hanushek (2002: 23)

Ironically, whilst Hanushek uses the above figures to support the argument that resources make little systematic difference to the quality of schooling (the title of Hanushek's article is 'The failure of input-based schooling systems'), UNESCO uses the same figures to examine what resources *do* make a difference to quality or outputs in developing country schooling systems. UNESCO's interpretation appears valid, in fact more valid than that of Hanushek. The above figures summarise a meta-analysis of 96 production function studies. Though Hanushek (2002 but also 1995) does not explicitly state that all the studies are within-country studies, the implication is that this is the case.

Hanushek (2002: 23) focuses on what he considers to be low values in the second column. He would have expected a larger percentage of the 96 production function studies to yield positive and statistically significant associations, if resources did truly matter (of course in the case of the pupil/teacher ratio, the expectation would be a high yield in the *negative* column, given the nature of this input). UNESCO (2005: 64), on the other hand, underlines the fact that in the case of teacher education, expenditure per pupil and facilities, a majority of associations come out as significant and positive, suggesting that these inputs do make a difference. Arguably, teacher experience also qualifies as an input that matters, given how few studies found a negative association between the this factor and outputs (if there was no association, and the statistically significant associations were random accidents, we would expect a more even spread across the positive and negative columns). In the above framework, teacher salary appears to matter little, and pupil/teacher ratio appears to have absolutely no systematic association with outputs.

How do we interpret the relative importance of per pupil expenditure, given that teacher salary and pupil/teacher ratio seem relatively unimportant? These three variables are strongly inter-connected, in fact expenditure per pupil is largely a function of teacher salary divided by pupil/teacher ratio. If expenditure per pupil is a significant predictor of quality, but the other two variables are not, we can probably conclude that it is non-personnel recurrent expenditure that is important. This, in turn, implies that the tools required by teachers, specifically textbooks and other teaching aids, make a difference. In summary, table 5 indicates that the quality of the human capital of the teachers and the quality of the physical capital of the school facilities do play a positive role, as do the tools required by the human capital. The quantitative aspect of the human capital, specifically the ratio of pupils to teachers, does not make a significant difference.

Limitations to Hanushek's framework include the fact that certain school factors are excluded from it, and that each association in each production function study is weighted equally, regardless of how many associations are tested in each study, the countries from which the studies originate, the methodological rigour of each study, and so on. In fact, Hanushek's framework has inspired a discourse on how best to go about this type of meta-analysis (Lee and Barro, 2000: 8).

UNESCO's 2005 *EFA Global Monitoring Report*, which focuses extensively on what school inputs make a difference to quality, emphasises a few important inputs or factors not captured in Hanushek's framework. On the basis of SACMEQ data, it is concluded that differences in the time spent by the learner on learning activities, inside and outside the school, explain much of the quality differences (UNESCO, 1995: 48). The vital issue of time is clearly not covered by Hanushek's framework. Moreover, PISA data points to the importance of the governance factor of allowing schools to take decisions for themselves. Another important factor falling somewhat outside the parameters of the typical input-output model is participation in pre-school. Some 87% of the cost of early childhood care and education (ECCE) is said to be recovered through efficiency savings in primary schooling, in the form of less repetition and dropping out (this is apart from the quality gains of ECCE) (UNESCO, 2005: 146).

A Lee and Barro (2000) study, referred to in the *Global Monitoring Report*, provides an interesting alternative to the multi-country meta-analysis represented by table 5. Lee and Barro construct a production model covering 58 countries and with 214 observations, where each observation represents a country situation with respect to a specific year (where possible, data from several points in time are used) and a specific primary school grade. Lee and Barro (2000: 2) call this a panel data approach, though their definition of what constitutes panel data is less demanding than that of Gujarati referred to in section 3 above. Lee and Barro conclude that the pupil/teacher ratio, teacher salary and learning time predominate as predictors of quality. In other words, the two inputs rejected in table 5 become pre-eminent in Lee and Barro's study. This kind of contradiction is frustratingly common in the literature. However, a closer examination of the model resolves much of the frustration. Lee and Barro (2000: 20) only include pupil/teacher ratio, teacher salary, learning time and expenditure per pupil as inputs. In contrast to Hanushek's framework, they include no variables on the quality of the human or physical capital. We are thus left with only one contradiction, namely that on the financial side Hanushek's meta-analysis emphasises expenditure per pupil, and de-emphasises teacher salary, whilst in Lee and Barro's study the reverse occurs. Importantly, and this may explain the contradiction, Lee and Barro's set of countries is a mix of developed and developing countries.

Willms and Somers (2001: 411) provide a brief review of the literature that essentially supports what has been said above. In developing countries, the learning tools are important (specifically, textbooks are important), as is the quality of the human capital of the teachers (specifically, teacher knowledge is important). Moreover, time spent learning is significant. Teacher salary and pupil/teacher ratio are said to have no impact, or a very limited one. Willms and Somers (2001: 412) also make the point that there is no conclusive evidence that national curriculum reform bears fruits in the form of better quality of learning as measured in standardised tests.

Two inputs that frequently appear in production function studies and are hardly ever rejected in the models focussing just on developing countries are, firstly, the quality of the human capital embodied in the teachers, and, secondly, the time spent learning and teaching. These two inputs are perhaps the bare bones of the education production function, the two inputs that we should at all costs attempt to include and elaborate on in our analysis. This clearly makes sense from an intuitive mental model perspective: a good teacher can achieve learning in any physical environment, and can to some extent compensate for the absence of textbooks. But a good teacher needs to spend sufficient time with her learners if quality learning is to occur.

It would have been helpful to produce a table such as table 5 concentrating only on the results of South African production function. However, this type of analysis is still in its infancy in South Africa, though it can be expected to grow judging by the commitment of the Department of Education to this methodology (Department of Education, 2005: 56). A brief overview of the key studies on South Africa follows.

Case and Deaton's (1999) input-output model uses 1993 test scores as the output, for a very small sample of 383 African learners. (This is the part of the analysis most relevant for this discussion – they also constructed other models, for instance for white learners). They conclude from their model that lowering the pupil/teacher ratio is a key lever for improving test scores. Crouch and Mabogoane (1998) constructed a production model for 303 schools, using observations at the level of the school. The main objective of their study was not to examine the production function itself, however, but to generate non-absolute and relative Senior Certificate examination results per school, using SES and school resources as conditioning variables. They demonstrate how absolute and relative results differ – according to the former no historically black schools emerge as top performers, whilst according to the latter, many historically black schools are top performers. The production model is not structured in a way that tells us what school inputs matter more, however. Crouch and Mabogoane dispute the Case and Deaton findings mentioned above due to the lacking statistical significance and data integrity of the Case and Deaton model. Crouch and Perry (2002) focus specifically on what school inputs make a difference in their regression modelling of data from 300 learners participating in a donor-funded school intervention project. Data problems result in a fairly low statistical significance. R^2 is never greater than 0.25. Regression analyses of school production using cross-sectional data frequently yield R^2 values in excess of 0.50. Nevertheless, it is noteworthy that they find teacher knowledge, measured in terms of a test that the teachers write, to be the most significant explanation for better learner performance. Once again, the importance of the quality of teachers is underlined. Van der Berg and

Burger (2003) construct a production function at the school level for 2,770 secondary schools, using Senior Certificate results as the output, and find teacher salary the strongest school input predictor. Importantly, however, teacher salary is considered primarily a proxy for teacher quality, given that salary is correlated with the teacher's level of qualifications. R^2 is around 0.56, making the Van der Berg and Burger model a particularly reliable analysis of what inputs make the greatest difference to performance. Their model again underlines the importance of the quality of teachers.

It is useful to juxtapose the evidence emerging from research oriented towards production functions (this is what has been discussed in this section so far) with evidence from the effective schools research. The latter indicates that it is less the level of inputs, or even the mix of inputs, and more the way inputs are combined, that explains higher performance in schools (Monk, 1990: 414). Put differently, effective schools research stresses the importance of school management processes, the qualities of the school principal, and the school culture or ethos. The production function approach, partly due to the structure of its fundamental approach, and partly due to the kinds of questionnaires typically used, is not good at examining the very things the effective schools approach claims are important. It is noteworthy that school management does not feature in table 5 above. The weakness of the production function approach in examining management and organisational factors, despite these factors clearly being important, is what prompts Crouch and Mabogoane (1998) to title their article *When the residuals matter more than the coefficients.* In other words, what the production function approach cannot explain is more important than what it can explain. The problem of assessing the influence of school management will receive further attention in section 7.1. What diminishes the problem somewhat in the case of developing countries, is that in relatively resource-deprived or skills-deprived schooling systems, the quantum of the inputs (as opposed to their management) becomes a relatively more important matter. Most of the evidence from the effective schools approach emanates from developed countries.

4.4 Cases of actual mental models

The official report from South Africa's Systemic Evaluation of 2001 is not intended as a formal production function study. However, the aim is partly to uncover aspects of education production:

> The purpose of the analysis was to report on: the conditions of learning and teaching at schools, the performance of learners *and the factors influencing learner performance* (Department of Education, 2003: 10). [own emphasis]

Moreover, some of the statistical analysis in the report is very much like production function analysis (this will be discussed later). The report identifies 27 input and process indicators covering all the main school and home background areas discussed earlier (these indicators are listed in Appendix A). The indicators are divided into three groups labelled access, equity and quality. This division seems insufficiently clear. For example, the learner/educator ratio is regarded as an access indicator, whilst educator qualifications is an equity indicator. The lack of clarity about this categorisation is particularly problematic given that the report presents three completely separate regression models for access, equity and quality.

The Brazilian SAEB analysis that we will limit ourselves to here is that by Ferrão, Beltrão, Fernandes *et al* (2001) using 1999 SAEB data. The analysis has a clear school production modelling focus, with an emphasis on the use of hierarchical linear modelling (they use the term multi-level model, but this is synonymous with the HLM discussed in this thesis). The analysts discuss their mental model briefly. They use the effective schools framework of Sammons, Hillman and Mortimore. In this framework, an effective school is one that maximises learner performance, given, firstly, the baseline performance of the learner at the beginning of some period and the learner's SES. According to the framework, there are eleven key indicators that should receive focus:

1. Professional leadership
2. Vision and goals shared by education stakeholders
3. Learning environment
4. Focus on the teaching and learning process
5. Clearly structured and defined teaching objectives
6. High expectations
7. Positive reinforcement of attitudes
8. Monitoring of progress
9. Rights and duties of learners
10. Collaboration between families and the school
11. Management focussed on the learning process

These indicators are rather different from the Systemic Evaluation ones, which relate more to specific factors such as the learner/educator ratio, school nutrition or the parents' level of education. The effective schools indicators focus more on cultural and management aspects of the school. But as we shall see further on in section 6, despite the rather different conceptual models, both analyses end up using very similar variables. Although Ferrão *et al* explain the statistics of their hierarchical linear model, there is no discussion of how the effective schools framework can be adjusted to take into account the schooling hierarchy, nor are the key issues relating to hierarchies and groups such as peer groups effects and selection effects discussed.

The extensive planning that preceded the data collection of the SACMEQ data included the identification of key policy concerns by participating countries. Of twenty such policy concerns, the first two deal directly with school production (this is taken from documentation accompanying the dataset):

> What were the personal characteristics (for example, age and gender) and home background characteristics (for example, parent education, regularity of meals, home language, etc.) of Grade 6 pupils that might have implications for monitoring equity, and/or that might impact upon teaching and learning?
>
> What were the school context factors experienced by Grade 6 pupils (such as location, absenteeism (regularity and reasons), grade repetition, and homework (frequency, amount, correction, and family involvement) that might impact upon teaching/learning and the general functioning of schools?

Clearly, home background is seen as an important input in the schooling process. Other policy concerns not referring specifically to input-output linkages (as the above two do) cover other inputs such as teacher quality.

4.5 A policy-oriented mental model

Given that the focus of the thesis is government monitoring and action, a policy-oriented mental model was constructed as one way of organising the school production debate, and as a framework for undertaking the analysis of the SACMEQ data. If the purpose of the school production model is to inform government policies and interventions, it seems logical to organise the inputs according to the policy levers typically available to a government. This will facilitate the translation of model findings into policy-specific recommendations, or further policy-specific cost models (as discussed in section 8). Even home background inputs, such as the educational level of parents, is at least potentially subject to government action, in this case in the form of adult education. The following mental model is thus proposed. It focuses on the complexities of the input side of the equation, using input groups and, below that, individual inputs, where each individual input is linked to an area of government policy or action. The input appears in the left-hand column, and the policy area in the right-hand column. The middle column indicates whether we would expect the input to be positively or negatively associated with school outputs.

Figure 7: Policy-oriented mental model of school inputs

INPUT		***POLICY AREA***
Educator inputs		
Quantity/quality of pre-service teacher training	+	Teacher training (pre-service)
Quantity/quality of in-service teacher training	+	Teacher training (in-service)
Educator salary and fringe benefits	+	Teacher conditions of service
Incentives for educators to perform	+	Evaluation and rewards for teachers
Learner/educator ratio	-	Teacher supply/distribution
Curriculum inputs		
Relevance/clarity of the curriculum	+	Curriculum quality
Contact time	+	School year/day
Level of learner repetition	(-)	Grade repetition
Level of stratification	(-)	School admissions and streaming policy
Learning support materials inputs		
Quality of LSMs	+	Materials development
Quantity of LSMs	+	Materials provisioning
Quantity of cutting edge LSMs	+	ICT
Infrastructure input		
Quality of school buildings and equipment	+	School construction/equipping
Management inputs		
Management capacity of school principal	+	Management training
School principal salary and fringe benefits	+	School principal conditions of service
Level of community involvement	+	Governance training
Quantity/quality of district support	+	Provincial/district support
Access promotion inputs		
Transport for remote learners	+	Scholar transport
Health of learners	+	School nutrition
Household inputs		
Educational support from parents	+	ABET
Socio-economic welfare of household	+	Poverty relief
Level of non-school education and culture facilities	+	Sports and culture

The mental model includes five educator inputs, linking them explicitly to policy areas. The curriculum group of inputs includes two variables that, as we have seen, are not unambiguously linked to outputs. The association between the repetition of learners and outputs is assumed to be negative in this model. The more repetition there is, the worse the output. Perhaps more controversially, stratification of the schooling system through mechanisms that keep the advantaged in certain schools and within-school stratification through mechanisms such as streaming are assumed to affect overall output negatively. Learning support materials inputs are split into a qualitative and a quantitative aspect, in recognition of the policy reality that very different processes influence these two aspects. LSM quantity has to do with budgets and provisioning and funding systems. LSM quality, on the other hand, is linked to materials development over which the national education authorities are likely to have fairly direct control. Information and communication technologies (ICTs) are included on a separate line in view of the special interest in these inputs by governments and educationists. School infrastructure is probably sufficiently straightforward as an influence on school outputs to be reducible to one line. School management, on the other hand, is divided into four critical areas, each with distinct policy levers. One way in which the above mental model breaks with typical mental models of school production is that it makes a distinction between the educational level of parents, and the poverty of the household, recognising that very different government interventions would deal with, for instance, literacy of parents and the provision of electricity and running water to households.

5 SELECTION OF A STATISTICAL MODEL

This section describes key statistical models available for production modelling based on cross-sectional school data. Importantly, developing country school monitoring programmes tend to collect cross-sectional data, and not time-series data, nor the combination of the two, panel data. This explains why models with a time aspect receive almost no attention here, though this is a rich area of research (Monk 1990: 330).

This section links up strongly with the following two. Whilst this section focuses on the theory behind key statistical models, section 6 deals with the selection and manipulation of the variables that the model will contain. Section 7 brings these sections together in a comprehensive and practically oriented examination of how the modelling process can proceed. Importantly, it is only in section 7 that we go beyond a purely demonstrative analysis of the SACMEQ data, and turn to a more comprehensive treatment of the data, which in turn allows us to begin extracting policy information from the data.

Matrix algebra and the related notation is used, partly because in the sections on hierarchical linear modelling this approach is almost indispensable. Moreover, an analysis approach is pursued that is fairly antagonistic towards the 'black box effect'. In other words, the trend is towards explaining explicitly how statistical outputs provided by the software packages Excel, Stata and HLM are computed.

5.1 The basic menu of econometric models

Econometric models can be distinguished by, firstly, their basic equation and, secondly, the estimation method used to arrive at values in the equation, and in the ancillary statistical outputs that accompany the model. The classical linear regression model (CLRM), the 'cornerstone of most econometric theory' (Gujarati 2003: 66), consists, in its simplest manifestation, of the following equation:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{u}_i \tag{10}$$

For optimal prediction, the CLRM requires rather stringent conditions to be met. For example, relationships between the dependent variable Y and the explanatory X variables should be homoscedastic – for instance, there should not be a stronger association between access to textbooks and learner performance amongst richer students than amongst poorer students. In practice, however, these conditions generally do not prohibit use of the CLRM as much as demand special caution in interpreting the results of the model. The CLRM receives detailed attention in section 5.2 below.

Estimation in the CLRM generally occurs through the method of ordinary least squares (OLS), though the maximum likelihood method exists as an alternative. OLS is in fact a special case of the generalised least squares (GLS) method, of which the weighted least squares (WLS) method is another special case. GLS and WLS both receive attention in section 5.6 on optimisation in the hierarchical linear model. Of relevance too for this thesis is the classical normal linear regression model (CNLRM), a sub-class of the CLRM. The CNLRM includes the additional constraint that the error terms, or the $\hat{u}$ in equation (10) should be normally distributed (Gujarati, 2003: 65, 107, 947).

Related to the CLRM are a number of other regression models, where 'regression model' can be considered a model explaining the association within a sample dataset between one or more explanatory variables and one dependent variable, as well as the inferences that can be made about the population as a whole.

Nonlinear regression models are needed to describe nonlinear relationships between the explanatory and dependent variables (Gujarati, 2003: 563). Some caution is needed when defining a relationship as nonlinear. Equation (10) represents a linear relationship, but so does:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i^2 + \hat{u}_i \tag{11}$$

Whilst the former is 'linear in the variables', the latter, whilst not linear in the variables, is 'linear in the parameters', in other words classified overall as a linear model. The following model, where one slope coefficient is squared, would be an example of a fully nonlinear model:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_i^2 X_i + \hat{u}_i \tag{12}$$

Qualitative response regression models are used when the dependent variable Y is qualitative, and not quantitative (Gujarati, 2003: 580). For example, these models would be used where Y represents a binary value such as buying property or not buying property. Education outputs are of course qualitative, and are often binary if we think of the pass versus no pass outcome. However, underlying a pass-no pass variable in the education setting is usually a fairly continuous scale that is richer in terms of information than the derived binary variable. For this reason, qualitative response regression models can be regarded as fairly marginal to the concerns of education analysis.

Similarly, time-series models seem to have limited applicability to school data as these models do not allow for distinctions between units. Panel data models, on the other hand, are potentially very powerful tools for understanding the economics of schooling. Unfortunately, however, panel data for education usage is on the whole limited to developed countries. An example of panel data modelling of a schooling system can be found in the Finnish study by Häkkinen *et al* (2003).

Fixed effects models and random effects models are often fitted to panel data. The form of the fixed effects model would be as follows:

$$Y_{it} = \hat{\alpha}_0 + \hat{\alpha}_1 D_{1i} + \hat{\alpha}_2 D_{2i} + \hat{\alpha}_3 D_{3i} + \hat{\beta}_1 X_{it} + \hat{u}_{it} \tag{13}$$

Here a three-learner system is assumed. The intercept $\hat{\beta}_0$ from equation (10) has been decomposed so that each learner has its own intercept comprising the sum of $\hat{\alpha}_0$ and, in the case of learner 1, the product of $\hat{\alpha}_1$ and a dummy variable. The name 'fixed effects model' derives from the fact that for each unit or learner, the intercept is fixed through the time periods t. The random effects model gives us:

$$Y_{it} = \hat{\beta}_0 + \hat{\beta}_1 X_{it} + \hat{\varepsilon}_i + \hat{u}_{it} \tag{14}$$

Here $\hat{\beta}_0$ from equation (10) is retained and the sum of $\hat{\beta}_0$ and a new error term, $\hat{\varepsilon}_i$, which is specific to each learner across all periods t, comprises the intercept (Gujarati 2003: 647). These panel data models are interesting for our purposes largely because they have much in common with the hierarchical linear models discussed below.

How should the analyst select a model? There is a wealth of often polemical texts answering this question. Two criteria seem to stand out, however. Firstly, reducing the amount of error in the model, as measured for instance in the sum of the squares of the error terms, is important. Secondly, the model must make ‘good economic sense’, in particular to managers and policymakers (Gujarati, 2003: 507; Fuss, McFadden and Mundlak, 1978: 220). Crucially, the modelling approach should never be mechanical or recipe-driven. The specificities of the data and the economic system at hand require the application of experience and iterative modelling and re-modelling before anything like an optimal model can be obtained.

The rest of section 5 focuses on two statistical models that have received considerable emphasis in the literature: the CLRM, which will be referred to as the ‘basic regression model’ here (sections 5.2 and 5.3), and the hierarchical linear model (sections 5.4 to 5.6).

5.2 The basic regression model

What is referred to in this section as the basic regression model is the unelaborated CLRM. The next section examines a few elaborations and adaptations. Here the meaning and calculation of the outputs of the CLRM will be examined. The explanations in this section are highly selective, and are far from a comprehensive treatment of the CLRM. The selection is strongly informed by the concepts and methodologies that must be understood when looking at the hierarchical linear model.

The relationship between one dependent variable and two explanatory variables in the SACMEQ data will be used for explanatory purposes. The dependent variable, *ratotp*, captures the reading score, and the two explanatory variables are *zpses*, an index reflecting the socio-economic status (SES) of each learner and *sres21*, indicating the no-yes (0-1) response of the principal to the question of whether the school had any computers. In the interests of clarity, no weights were used, and only observations with valid values for all three variables were considered.

The following are the averages for *sres21*. A typical mental model is confirmed – the presence of computers is associated with better learner performance.

Table 6: Mean reading score by presence of school computers

sres21	stats N	ratotp mean
no	1941	30.77743
yes	1198	49.93322
Total	3139	38.08824

The first two data columns of the next table confirm the expected relationship between SES and learner performance. For every increase in the SES index, the average reading score improves, with the exception of one contradictory trend between the index values 7 and 8.

Table 7: Mean reading score by SES of learner and presence of school computers

zpses	All		sres21=no		sres21=yes	
	stats N	ratotp mean	stats N	ratotp mean	stats N	ratotp mean
1	13	22.76923	12	23.75	1	11
2	51	28.80392	40	29.075	11	27.81818
3	134	29.00746	115	27.50435	19	38.10526
4	234	29.39744	196	**28.22959**	38	35.42105
5	284	29.96127	232	28.77586	52	35.25
6	295	31.31525	234	**29.58547**	61	37.95082
7	305	31.92787	250	30.252	55	39.54545
8	283	31.92226	214	30.07477	69	37.65217
9	316	34.78165	207	31	109	**41.9633**
10	296	38.41892	184	34.15217	112	45.42857
11	248	44.06452	114	35.39474	134	51.4403
12	218	47.80275	79	36.12658	139	54.43885
13	212	54.46698	42	39.95238	170	**58.05294**
14	212	60.16509	21	32.57143	191	63.19895
15	38	65.65789	1	45	37	66.21622
Total	3139	38.08824	1941	30.77743	1198	49.93322

Source (here and in subsequent SACMEQ tables): SACMEQ, 2000.

But what if we want to consider both explanatory variables simultaneously, or, put differently, we want to observe the associations relating to one explanatory variable, whilst controlling for the other one? We can produce a two-way tabulation, as in the last four columns of the above table. These columns confirm, for instance, that even if only schools with no computers are considered, we still find a clear association between better SES and better reading scores.

The basic regression model is essentially a more compact way of performing this analysis. Clearly, the greater the number of explanatory variables, the greater the advantages of the regression model and the impossibility of pursuing the previous approach. The basic regression model for two explanatory variables looks as follows:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{u}_i \tag{15}$$

OLS would involve the minimisation of overall error, specifically the sum of the squares of $\hat{u}_i$, or $\sum \hat{u}_i^2$.

The next table represents the output of the regression obtained from Stata. For this analysis, the values in *sres21* and *zpses*, which were either 1 or 2, were reduced by 1 in order to obtain values of 0 and 1. The resultant variables were *_sres21* and *_zpses.* The reasons for this will be explained.

Table 8: Three-variable regression (Stata output)

Source	SS	df	MS		
				Number of obs	**3139**
				F(2, 3136)	1167.82
Model	**365528.8**	2	182764.4	Prob > F	*0*
Residual	490785.8	3136	156.5006	R-squared	0.4269
				Adj R-squared	0.4265
Total	**856314.6**	3138	272.8855	Root MSE	12.51

ratotp	Coef.	Std. Err.	t	P>t	[95% Conf. Interval]	
_sres21	**13.24567**	**0.519232**	25.51	*0.000*	*12.2276*	*14.26374*
_zpses	**1.847791**	**0.075515**	24.47	*0.000*	*1.699727*	*1.995856*
_cons	**19.29085**	**0.54863**	35.16	*0.000*	*18.21514*	*20.36657*

Statistics in bold must be obtained from the SACMEQ data itself. Statistics in italics involve some looking up in standard statistical tables. All the other statistics can be calculated on the basis of other statistics in table 8.

We begin with the three **coefficients** under *Coef.,* the one **intercept** – *cons* in table 8 and $\hat{\beta}_0$ in equation (15) – and the two **slope coefficients** – *_sres21* and *_zpses* in table 8 and $\hat{\beta}_1$ and $\hat{\beta}_2$ in equation (15). The intercept represents the mean reading score of learners with zero for both variables, i.e. 'no' for *_sres21* and bottom ranking for *_zpses*. This convenient interpretation of the intercept explains the earlier manipulation of the values. The regression model gives us reading score means which are fairly close to those in table 7 above. For example, for an SES of 10 with computers, instead of 45.4, we obtain:

$$19.29 + 13.24 \times 1 + 1.84 \times 10 = 51.01 \tag{16}$$

In matrix notation, the regression model from equation (15) is (Gujarati 2003, 927):

$$\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\mathbf{u}} \tag{17}$$

The estimation of the three coefficients, if we use matrix notation, is:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{18}$$

We can demonstrate this matrix approach to estimating the coefficients by using a greatly simplified data table of just four learners, where each learner has the values 0-4, 0-6, 1-9, and 1-13 for *_sres21* and *_zpses*, and a reading score equal to the corresponding mean, rounded, from table 7 (the reading scores in bold). Moreover, we need to create a variable linked to the first coefficient, the intercept. This variable will have a value of 1 for each of the four learners. This will cause the intercept to be multiplied by 1 in the case of each learner. The following four steps illustrate the solving of equation (18) for the four learner system.

$$\hat{\boldsymbol{\beta}} = \left[\begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 4 & 6 & 9 & 13 \end{bmatrix} \begin{bmatrix} 1 & 0 & 4 \\ 1 & 0 & 6 \\ 1 & 1 & 9 \\ 1 & 1 & 13 \end{bmatrix} \right]^{-1} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 4 & 6 & 9 & 13 \end{bmatrix} \begin{bmatrix} 28 \\ 30 \\ 42 \\ 58 \end{bmatrix} \tag{19}$$

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 4 & 2 & 32 \\ 2 & 2 & 22 \\ 32 & 22 & 302 \end{bmatrix}^{-1} \begin{bmatrix} 158 \\ 100 \\ 1424 \end{bmatrix} \tag{20}$$

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 3 & 2.5 & -0.5 \\ 2.5 & 4.6 & -0.6 \\ -0.5 & -0.6 & 0.1 \end{bmatrix} \begin{bmatrix} 158 \\ 100 \\ 1424 \end{bmatrix} \tag{21}$$

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 12 \\ 0.6 \\ 3.4 \end{bmatrix} \tag{22}$$

The values 12, 0.6 and 3.4 are different estimates of the three coefficients for which we obtained the values 19.29, 13.24 and 1.84 in the regression analysis in table 8. The large differences are to be expected, considering we are considering only four learners. (Four learners is incidentally the minimum number we must have if we want the regression model to work as it should). The solving of the above matrix calculations can easily be performed in Excel using the functions TRANSPOSE, MMULT and MINVERSE. For datasets with many observations, it is important in Excel to nest the TRANSPOSE function within the MMULT function in order not to run into problems of spreadsheet space and functions not coping with the matrix size. In Stata the commands matrix accum and vecaccum should be used. It should be noted that Stata's matrix vecaccum uses $(\mathbf{yX'})'$ instead of $\mathbf{X'y}$, but the rules of matrix transposition tell us that these two forms are equal (Gujarati 2003: 919).

The coefficients are key to understanding the strength of the associations between the explanatory variables on the one hand, and the dependent variable on the other. In education terms, it is useful to think of the intercept representing effectiveness and the slopes representing equity. The more technically and allocatively efficient the schooling system, the higher the mean performance should be, and thus the higher the intercept. And the more equitable the system, the smaller the slopes should be. In a utopian system of equal SES and equal school conditions, there would clearly be no slopes at all. But even if we did have inequities in this regard, the less difference the inequities on the input side made to the output, for instance the less poverty impacted on learner performance, the lower the slopes would be (Bryk and Raudenbush 1992: 12).

We now turn to the output statistics on the top left side of table 8, or what is often referred to as the ANOVA table, or analysis of variance statistics. The **residual sum of squares** (*Residual SS* in table 8), or the **RSS**, of 490785.6 is simply the $\sum \hat{u}_i^2$ mentioned earlier – this is what gets minimised in the optimisation process. We can also think of this as the unexplained sum of squares, and we can furthermore think of this unexplained variance as being composed of two distinct things: measurement error, on the one hand, and inefficiency in the production process, or the realities that are truly 'unexplained' by the model, on the other (Bifulco and Bretschneider, 2001: 421).

In matrix algebra, **RSS** would be found as follows (Gujarati, 2003: 932):

$$RSS = \hat{\mathbf{u}}'\hat{\mathbf{u}} = \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)'\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} \tag{23}$$

Total sum of squares (*Total SS* in table 8), often abbreviated **TSS**, is the sum of squares in the simplest possible model with no explanatory variables, in other words the model where the error is the difference between actual *Y* of each learner and the grand mean of *Y*, $\bar{Y}$. In matrix notation we have:

$$TSS = \mathbf{y}'\mathbf{y} - n\bar{Y}^2 \tag{24}$$

The **explained sum of squares**, or **ESS** (*Model SS* in table 8), is simply the difference between the **TSS** and **RSS** amounts, and indicates the sum of squares explained by our model in equation (15). The *MS* or **mean sum of squares** statistics are obtained by dividing the relevant *df* or **degrees of freedom** into the *SS* values. If *n* is total number of observations (3,139 in our example) and *k* is number of variables in the model (3 in our example), then the degrees of freedom are $n - 1$, $n - k$ and $k - 1$ for TSS, RSS and ESS respectively.

The values in the ANOVA table are seldom referred to themselves in analyses, but they serve as important inputs into other statistics to the right in table 8 that are of great importance in describing the overall 'goodness of fit', or significance, of the model. The overall goodness of fit of a regression model can be understood as the degree to which all slope coefficients are not zero, or, returning to our earlier conceptualisation of equity, the degree to which we have inequity that can be explained by our model (Gujarati, 2003: 253).

The most commonly referred to indicator of overall goodness of fit is the **coefficient of determination** – *R squared* in table 8. This value is the proportion of the sum of squares that is explained by the model (Gujarati, 2003: 81):

$$R^2 = \frac{ESS}{TSS} = \frac{365529}{856315} = 0.4269 \tag{25}$$

In our SACMEQ example, the model explains around 43% of the association between the explanatory variables and the dependent variable. Had all the slopes been zero, R^2 would have been zero. Notwithstanding important provisos relating to the coefficient of determination, and warnings against 'R^2 fetishism' (Vinjevold and Crouch, 2001:

29), R^2 is justifiably used by analysts as one key indicator of how well a model describes the economic realities.

One R^2 proviso is that this statistic is not sensitive to sample size, so it cannot be used to test hypotheses about the population, in other words to evaluate the probability of all slope coefficients being zero in the *population*, as opposed to just the sample. There are four other statistics in table 8 that fill this gap, however. *Adjusted R squared* is an alternative indicator of goodness of fit that is sensitive to both the number of observations and the number of variables in the model. It is computed as follows (Gujarati, 2003: 217):

$$\bar{R}^2 = 1 - \frac{RSS/(n-k)}{TSS/(n-1)} = \frac{490786/(3139-3)}{856315/(3139-1)} = 0.4267 \qquad (26)$$

It can be seen from this equation that it is only where n has a low value that we can expect the adjusted R squared to differ substantially from the unadjusted R squared. Whether one refers to the one or the other statistic is therefore often of little consequence. In only one of the data analysis texts examined was the adjusted R^2 referred to (Van der Berg and Burger, 2003). Most analysts prefer to use the unadjusted R^2.

A key point that must be understood about econometric models is that adding new explanatory variables to a model will tend to improve the goodness of fit marginally, even if these variables on their own have little relevance. This is the reason for including the number of variables, k, in the calculation of adjusted R^2. It should be clear from equation (26) that the greater the number of observations, the greater the value of adjusted R^2, and the greater the number of explanatory variables, the lower the value of adjusted R^2. Adjusted R^2 is always less in value than the unadjusted R^2.

The mean sum of squares statistics are used to calculate the ***F* statistic**, which is the *MS* value for the model (equal to 182764 in table 8), divided by the *MS* value for the residual (equal to 156 in table 8). This statistic is similar to the R^2 statistics in the sense that a higher value means a better goodness of fit. Like the adjusted R^2, the F statistic is sensitive to both n and k. The added benefit of the F statistic is that it can be cross-referenced to a benchmark. For a given *df* in the numerator, a given *df* in the denominator, and a given level of probability, there is a threshold above which the *F* statistic should lie if we want to say that it is improbable that all the slopes in the model are equal to zero in the population In our example, and selecting a stringent level of probability of 99%, we see from the relevant statistical tables that the threshold is 4.61. The *F* value of 1168 is clearly well above this threshold, so we can be highly certain that there are significant associations between the explanatory variables and the dependent variable. Each *F* value comes with a *p* statistic (*Prob > F* value in table 8), obtainable from a statistical table, which provides the exact probability of our slopes being zero.

The value *Root MSE* in table 8 is the square root of the *Residual MS* value and is referred to as the **standard error of the regression** (Gujarati, 2003: 78). It is yet another indicator of the overall goodness of fit of the model.

Turning to the bottom section of table 8, each of the three coefficients comes with its own set of statistics indicating the goodness of fit associated with each coefficient. Each **standard error of the estimate** (*Std. Err.* in table 8) can be computed using the variance-covariance matrix of the estimated coefficients. This matrix is arrived at as follows (Gujarati, 2003: 944):

$$\text{var}-\text{cov}\left(\hat{\boldsymbol{\beta}}\right)=\hat{\sigma}^2\left(\mathbf{X}'\mathbf{X}\right)^{-1} \tag{27}$$

In other words, we take the $(\mathbf{X}'\mathbf{X})^{-1}$ matrix and multiply each term in the matrix by $\hat{\sigma}^2$, the variance of the error term $\hat{u}_i$. The $(\mathbf{X}'\mathbf{X})^{-1}$ matrix for the SACMEQ data, obtainable using Stata's `matrix vecaccum` command (to perform the matrix multiplication) and `syminv` function (to invert the resultant matrix), is:

$$\left(\mathbf{X}'\mathbf{X}\right)^{-1}=\begin{pmatrix} 0.00192 & 0.00021 & -0.00023 \\ 0.00021 & 0.00172 & -0.00012 \\ -0.00023 & -0.00012 & 0.00004 \end{pmatrix} \tag{28}$$

With respect to $\hat{\sigma}^2$, we need to be a bit careful as there is more than one way of computing the variance statistic. In this instance we are looking at the following method:

$$\hat{\sigma}^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n-k} \tag{29}$$

This is known as the unbiased estimator of $\hat{\sigma}^2$ and it is specifically designed for the error term of a regression model. The computation is slightly different from that of the general variance described by Gujarati (2003: 880), or of the variance obtained from Stata's `var` function. The variance corresponding to the above equation using the SACMEQ data is 156.5005, and this, multiplied by $(\mathbf{X}'\mathbf{X})^{-1}$, renders the following variance-covariance matrix:

$$\text{var}-\text{cov}\left(\hat{\boldsymbol{\beta}}\right) = \begin{bmatrix} 0.30099 & 0.03275 & -0.03544 \\ 0.03275 & 0.26960 & -0.01824 \\ -0.03544 & -0.01824 & 0.00570 \end{bmatrix} \tag{30}$$

The square roots of the values on the diagonal give us the standard error of the estimate values in table 8. This in turn leads us to the confidence interval statistics, for instance the **confidence interval** of 12.23 to 14.26 for the slope coefficient for *_sres21*. This confidence interval is only applicable to a 5% level of significance, which is associated with a 95% confidence coefficient. What all this means is that we can be 95% sure that the best true slope coefficient for *_ratotp* lies in the range of 12.23 to 14.26. Roughly, we can see that the standard error is doubled and then subtracted and added to the estimated coefficient to obtain the confidence interval. This rough approach would give us a confidence interval for *_zpses* that is almost identical to the actual values in table 8. But to be completely accurate, we would need to calculate δ, the margin on either side of the estimated coefficient, as follows (Gujarati 2003: 123):

$$\delta = t_{\alpha,df} \times se\left(\hat{\beta}\right) \tag{31}$$

In other words we multiply the standard error of the coefficient with a t value that corresponds, firstly, to a level of significance α (in our case it is 5%) and, secondly, to a number of degrees of freedom $n - k$ (this would be 3,137 in our example). The relevant t value can be looked up in a statistical table, or by using Excel's `TINV` function.

The ***t* statistic** is simply the coefficient divided by its standard error. Thus the smaller the confidence interval relative to the magnitude of the coefficient, the greater the t

statistic. High t values indicate more significant associations, or a low probability that the slope is zero. In fact, a rule of thumb that is often used is the '2-t rule of thumb', which says that if t is greater than 2 or less than –2, then we can be sure to a 5% level of significance that the slope in the population is not zero.

The $P>t$ values in table 8, all of which are equal to zero in our model, are **probability values**. They represent the probability that the t value obtained from the sample data should be what we see in table 8, or greater, if in fact the true value of the slope coefficient in the population were zero. The p value is found in a statistical table, with degrees of freedom and the t statistic as the input (Gujarati, 2003, 134-8).

In the data analysis texts, the way goodness of fit or significance is described varies considerably, and some of this variation is clearly no more than a matter of style. With respect to the significance of individual explanatory variables, a distinction can be made between analysts who focus more strongly on the t statistic (Crouch and Perry 2002; Fertig 2003; Van der Berg and Burger 2003), and those who focus more strongly on the confidence interval statistics (Barbosa and Fernandes, 2001; Harbison and Hanushek, 1992; Häkkinen, Kirjavainen and Uusitalo, 2003).

This section ends with a brief discussion of how weights attached to observations are incorporated into the regression model. Using weighted observations in the regression model influences the statistical outputs of the model insofar as associations of highly weighted observations come through more strongly in the coefficients. But even the goodness of fit or significance statistics are influenced by the presence of weights. Only the computation of the coefficients using weights will be explained.

If we have weight v for each observation, and the resultant column matrix $\mathbf{v}$ for the whole dataset, we create a new column vector $\mathbf{w}$ of normalised weights (i.e. weights which add up to the total number of observations):

$$\mathbf{w} = \left(\frac{\mathbf{v}}{\mathbf{1'v}}\right)(\mathbf{1'1}) \qquad (32)$$

We then create a diagonal matrix $\mathbf{D}$ with the elements of $\mathbf{w}$ along the diagonal. Equation (18) then becomes the following (Stata, 2001: 86):

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X'DX}\right)^{-1}\mathbf{X'Dy} \tag{33}$$

5.3 Elaborations on the basic regression model

This section discusses a few commonly used or methodologically interesting elaborations on the modelling described in the previous section.

Already at the mental model stage it may be clear that we ought to be dealing with not just one model, but many models. This would be particularly true where the schooling system is in fact a number of systems. South Africa presents a striking example of a schooling system that is barely one system, given the recent legacy of apartheid segmentation and inequalities. There are essentially two ways in which we can deal with the segmented system or model.

We can create qualitative dummy variables to mark each sub-model, for instance each apartheid department to which the school in question belonged. To avoid the dummy variable trap, however, we always leave out one default sub-model (Gujarati 2003: 302). This is what Van der Berg and Burger (2004), who work with school observations, do. Harbison and Hanushek (1992: 115) identify different Brazilian regional administrations in their analysis in this way.

Secondly, we can divide the data and run completely separate regression models. This Van der Berg and Burger (2004) also do. Hanushek and Luque (2003: 485) create separate models for each country in their international study. Fertig (2003: 6) divides observations up into groups according to level of output. Importantly, the dummy variable and separate models approaches are not equivalent. The first provides only differentiated intercepts. The second gives us differentiated intercepts and slopes.

Because variables such as race and former apartheid administration do not feature in the SACMEQ data, creating sub-models as in Van der Berg and Burger's analysis is somewhat complicated. Section 7.1 below will explore ways of separating historically advantaged from disadvantaged learners and schools, on the basis of non-normal distributions of values, such as those for the reading score discussed in section 3 above.

If we follow Fertig's approach and perform a segmentation according to learner performance, using quintiles, we obtain the following:

Table 9: Coefficients within performance quintiles (Stata output)

		quintile 1 (lowest)	quintile 2	quintile 3	quintile 4	quintile 5
	n	628	628	628	628	627
_sres21	Coef.	-0.0474037	0.1114665	-0.0362115	1.420069	3.500905
	t	-0.09	0.63	-0.16	3.84	4.51
	sd	0.3207704	0.3959511	0.4430166	0.4981095	0.3266342
_zpses	Coef.	-0.0465179	0.0246055	0.0812177	0.1946127	0.7788886
	t	-0.74	0.96	2.45	3.19	7.61
	sd	2.714631	2.746311	2.9623	3.021836	2.474791

Table 9 is particularly interesting as far as SES is concerned. Association between performance and SES increases the higher the performance (and more or less the higher the SES). Importantly, the fact of the slope coefficient not being significantly different from zero, as is clearly the case in quintiles 1 and 2, could be caused by two very different factors. It could be that everyone has more or less the same SES, or it could mean that SES varies, but there is little correlation with performance. For this reason, it is important to consider the variance of the explanatory variables. Hence the standard deviation (*sd*) of these variables is indicated. We see that the standard deviation for SES does not differ greatly from quintile to quintile. This suggests that although SES varies in the lower quintiles, this variance does not become translated into better performance in any substantial way within those quintiles.

What we have been doing here, and particularly in table 9, is to check nonlinear associations (where we mean nonlinear in the variables). There are more 'elegant' ways of describing such nonlinear relationships. One example is the polynomial model:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}^2 + \hat{u}_i \qquad (34)$$

Others are the reciprocal model:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 \left(\frac{1}{X_{2i}} \right) + \hat{u}_i \qquad (35)$$

the log-linear model:

$$\ln Y_i = \hat{\beta}_0 + \hat{\beta}_1 \ln X_{1i} + \hat{\beta}_2 \ln X_{2i} + \hat{u}_i \qquad (36)$$

and the piecewise linear model:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 (X_{2i} - X_*) D_i + \hat{u}_i \qquad (37)$$

In this latter model, X_* is a threshold value for the variable X_2 or the point at which the slope changes significantly, and D is a dummy 0-1 variable that acquires the value 1 if X_{2i} exceeds the threshold X_* (Gujarati, 2003: 175, 317).

Of the four forms just mentioned, the one that best captures the nonlinearity we saw with respect to the SES-performance relationship in table 9, is the piecewise linear model of equation 37 with X_* equal to 8. This model yielded an adjusted R^2 value of 0.455. The polynomial model in equation 34 was the only other model to yield an adjusted R^2 value greater than that obtained in our original table 8 model. Single-equation nonlinear forms such as those in equations (34) to (37) are not common in education production models. In none of the texts that were analysed did they appear.

Another approach that involves manipulating the input variables in some way is the creation of one or more interaction terms. Interaction terms are typically the product of two explanatory variables. This is particularly useful if one or both variables are 0-1 dummy variables. This is best demonstrated with our SACMEQ demonstration data if we introduce another 0-1 dummy variable, apart from *_sres21*. We added *_zsloc,* which assumes a value of 0 if the school is rural, and 1 if the school is urban. The model with no interaction term is as follows:

Table 10: Four-variable regression (Stata output)

Source	SS	df	MS		
				Number of obs	3139
				F(2, 3136)	779.7
Model	365905.57	3	121968.52	Prob > F	0
Residual	490408.99	3135	156.4303	R-squared	0.4273
				Adj R-squared	0.4268
Total	856314.56	3138	272.88546	Root MSE	12.507

ratotp	Coef.	Std. Err.	t	P>t	[95% Conf. Interval]	
_sres21	12.66697	0.639154	19.82	0	11.41377	13.92018
_zpses	1.817713	0.0779462	23.32	0	1.664882	1.970543
zsloc	0.9769957	0.629512	1.55	0.121	-0.2573016	2.211293
_cons	19.19042	0.552311	34.75	0	18.1075	20.27335

With the product of *_sres21* and *zsloc* added as a new explanatory variable, called *_interaction,* we obtain the following model:

Table 11: Regression model with interaction term (Stata output)

Source	SS	df	MS		
				Number of obs	3139
				F(2, 3136)	589.49
Model	367658.2	4	91914.551	Prob > F	0
Residual	488656.35	3134	155.92098	R-squared	0.4293
				Adj R-squared	0.4286
Total	856314.56	3138	272.88546	Root MSE	12.487

ratotp	Coef.	Std. Err.	t	P>t	[95% Conf. Interval]	
_sres21	2.952848	2.966848	1	0.32	-2.864315	8.77001
_zpses	1.790393	0.0782447	22.88	0	1.636977	1.943809
zsloc	0.6014461	0.6383904	0.94	0.346	-0.6502595	1.853152
_interaction	10.21441	3.046631	3.35	0.001	4.240815	16.188
_cons	19.47073	0.5577136	34.91	0	18.37721	20.56425

Including an interaction term adds marginally to the goodness of fit as measured by adjusted R^2. The first of these two models allows us to gauge just the additive effect of the presence of computers and being in an urban centre. Because we are dealing with 0-1 dummy variables, we would simply add the two slope coefficients, 12.67 and 0.98, in order to obtain an overall advantage of 13.64 associated with the concurrent presence of the two inputs. The second model allows us to also gauge the multiplicative effect of these two inputs through the slope coefficient of the interaction term. If we add 2.95, 0.60 and 10.21, we obtain 13.77. If we take into consideration the multiplicative effect of the two inputs, we detect a greater advantage, in other words. We also see that the advantage of having just computers, but being in a rural area, is greatly reduced, from 12.67 to 2.95, as is the advantage of just being in an urban area but having no computers. What the second model is therefore telling us is that the performance advantage occurs when *both* inputs are present together. They interact with each other (Blalock, 1979: 492).

In our mental model, we could imagine the advantage of computers only being fully realised if the school has access to a reliable power supply and computer repair companies, and learners have access to computers in the home and in Internet cafes, all features commonly associated with urban centres. Of course an opposing mental model is possible, whereby computers, which we assume permit access to the Internet, make the largest difference in rural areas because they facilitate access to information otherwise only available in urban centres. A negative slope coefficient on the interaction term would have supported this opposing mental model.

If possible, a model should take into account the 'within-pupil variance' referred to in section 4.1, in other words the fact that any learner will display varying performance across a range of tests depending on a range of relatively random factors such as the mood of the learner. The SACMEQ data contains both reading and mathematics scores. If we calculate the mean score for each learner using the two scores, we should reduce within-learner variance. In fact, the table 8 model with output changed to the mean of the two scores does yield a marginally greater level of explanation – R^2 increases from 0.4269 to 0.4288.

Many regression analyses of schooling systems are based on data containing school observations, where learner attributes are collapsed into mean values per school, as opposed to data containing learner observations, where school attributes are repeated across the learner observations pertaining to that school (Van der Berg and Burger, 2003; Crouch and Perry, 2002). The school level equivalent of equation 15 would be as follows (*j* refers to school, and X_2 is assumed to be a learner level variable such as SES):

$$\overline{Y}_{\cdot j} = \hat{\beta}_0 + \hat{\beta}_1 X_{1j} + \hat{\beta}_2 \overline{X}_{2\cdot j} + \hat{u}_j \qquad (38)$$

It seems this approach is more a matter of necessity than choice, and applicable to situations where details per learner are not available. It is important to understand how collapsing learner observations into school observations affects the outputs of the model. Not taking this into account can have dire consequences for the interpretation of the model.

If we collapse the data used for the model in table 8 to the school level, we obtain the following outputs:

Table 12: Collapsed school-level regression (Stata output)

Source	SS	df	MS
Model	21091.516	2	10545.758
Residual	10368.774	164	63.224233
Total	31460.291	166	189.51982

Number of obs	167
F(2, 3136)	166.8
Prob > F	0
R-squared	0.6704
Adj R-squared	0.6664
Root MSE	7.9514

ratotp	Coef.	Std. Err.	t	P>t	[95% Conf. Interval]	
_sres21	9.372338	1.615654	5.8	0	6.182173	12.5625
_zpses	3.116637	0.3110611	10.02	0	2.502436	3.730837
_cons	11.33434	2.076338	5.46	0	7.234539	15.43414

Both unadjusted and adjusted R^2 and the standard error of the regression (*Root* MSE) point to a better goodness of fit than was obtained in the table 8 model. Other goodness of fit statistics, the F statistic and, at the level of individual explanatory variables, the t statistics, point to a worse goodness of fit. All of this is a typical result of collapsing the data to a more aggregated level. On the one hand, having individual learner details in the data, as opposed to means of these details at the school level, automatically incorporates more extremes, and hence more variance. This explains the 'better' coefficient of determination R^2 in the school level model of table 12. On the other hand, having much fewer observations, 167 as opposed to 3139, reduces the goodness of fit in the sense of the power of the model to describe the population. This explains, for instance, the 'worse' F statistic. We should bear in mind that the models in tables 8 and 12 are based on exactly the same dataset. Clearly, the unit described by each observation makes a great difference to the statistical outputs. Often, a comparison of the goodness of fit between one model using school observations and another using learner observations, would be meaningless.

The slope coefficients are very different in the two models. It is in fact typical for the slope coefficient of the learner level variable (in our case *_zpses*) to become greater and for the school level variable (in our case *_sres21*) to become smaller when we create a model on a dataset where values have been collapsed to the school level (Burstein, 1980: 132). This difference is entirely attributable to the error values $\hat{u}_i$ from equation 15. If we artificially eliminate them, by making the value of *ratotp* equal to only the explained part of equation (15), and then collapse the data to the school level, we find that we obtain exactly the same intercept and slope coefficient values in the two models, though these values would not be equal to either those in the table 8 model or those in the table 12 model. However, the deviation from the artificial no-variance slope coefficients would be smaller in the case of the learner level model (table 8) than in the case of the school level model (table 12). We can take this as an indication that the learner level model provides a more accurate picture of the association between inputs and the output than the school level model. In particular, the problem with the school level model is that it inflates the slope coefficients linked to the learner details (such as the SES of learners) strongly. (These conclusions are based on repeated modelling of both original and manipulated SACMEQ and other data. The other data used was the PISA 2000 data pertaining to Brazil.)

The value-added variant of the basic regression model for education production deserves mention. This involves inserting an earlier performance score, Y_{0i}, as an explanatory variable, where that score is comparable to the later performance score, Y_{1i}, which becomes the output in the model. Equation (15) thus becomes:

$$Y_{1i} = \hat{\beta}_0 + \hat{\beta}_1 Y_{0i} + \hat{\beta}_2 X_{1i} + \hat{\beta}_3 X_{2i} + \hat{u}_i \qquad (39)$$

Whilst the data for this model barely qualifies as time-series data (only one variable needs a time series, and only for two points in time), it provides great analytical opportunities. In this model, selection effects in the form of learners dropping out are dealt with – only learners with both a prior and a later score would be included. Arguably, the model eliminates the need for learner background data, as Y_{0i} would capture the advantage or disadvantage linked to for instance, SES, and presumably these factors would not have changed substantially over time. This allows the model to focus on school inputs, the inputs that are presumably more likely to change from one year to the next for each learner (in particular the teacher input is likely to change) (Hanushek, 1979: 363; Harbison and Hanushek, 1992: 86). A counter-argument would be that even if Y_{0i} captures learner background in an overall sense, it would be important to reflect within the model the degree to which different aspects of learner background impact differently on the improvement between Y_{0i} and Y_{1i}.

To conclude, the point is often made that the uses of the basic regression model described in this section are far from optimal, in the sense that they are not based on a more sophisticated formal model of how education works. Essentially, various inputs that seem important to the analysis are gathered as variables on the right hand side of the equation, without a clear sense of the hierarchy or inter-connectedness of these inputs. This has been referred to as the 'kitchen sink' phenomenon in education production modelling. The problem is not just a paucity of empirical research in the economics of education field, but also in the fields of psychology, sociology and pedagogy (Monk, 1990: 324).

5.4 Form of the hierarchical linear model

In the previous section, constructing different models for different sub-systems was discussed. The sub-models gave us differentiated intercepts and slope coefficients for each sub-system. But can we consider each school to be a sub-system? The school as a sub-system within our mental model is certainly possible, for instance we can suspect that associations between learner characteristics and learner performance may work differently for each school. However, when it comes to econometric modelling, we would not want create a model per school – 168 in the case of the SACMEQ sample – for a variety of reasons. We should rather make use of the hierarchical linear model. HLMs allow us to differentiate intercepts and slopes by school within just one econometric model (Bryk and Raudenbush, 1992: 9).

The HLM has many basic forms. The more common ones will be considered here. We begin by converting the model from equation (15) to the following variant of the HLM. We assume that X_1 is a variable describing the school, whilst X_2 is a variable describing the learner.

$$Y_{ij} = \left(\hat{\alpha}_{00} + \hat{\alpha}_{01} X_{1j}\right) + \hat{\beta}_2 X_{2ij} + \hat{u}_{ij} \tag{40}$$

Here we have $\left(\hat{\alpha}_{00} + \hat{\alpha}_{01} X_{1j}\right)$ taking the place of the intercept $\hat{\beta}_0$ from equation (15). The intercept has potentially a different value for each school, depending on the value of the school variable X_1. Of course different schools may have the same intercept. This would depend on the range of possibilities for variable X_1.

The notation used in equation (40) must be explained. The alphas refer to coefficients which, though constant in value for the whole schooling system, are used to determine differentiated school-specific coefficients. The alphas are commonly referred to as level 2 coefficients. The two digits subscripted to the right of each alpha indicate which variables they relate to. The first of the two digits refers to the learner level, whilst the second one refers to the school level. Hence in $\hat{\alpha}_{01}$ the subscripted 0 indicates that this coefficient is linked to the intercept at the learner level (originally this was $\hat{\beta}_0$) whilst 1 indicates that the coefficient is simultaneously linked to the school level variable X_1.

However, equation (40) must be expanded to the following to illustrate more comprehensively the form of the typical HLM.

$$Y_{ij} = \left(\hat{\alpha}_{00} + \hat{\alpha}_{01} X_{1j} + \hat{\varepsilon}_{0j}\right) + \hat{\beta}_2 X_{2ij} + \hat{u}_{ij} \qquad (41)$$

Equation (41) can be rewritten as:

$$Y_{ij} = \left(\hat{\alpha}_{00} + \hat{\alpha}_{01} X_{1j}\right) + \hat{\beta}_2 X_{2ij} + \hat{\varepsilon}_{0j} + \hat{u}_{ij} \qquad (42)$$

Here the single error term $\hat{u}_{ij}$ from equation (40) has been split into $\hat{\varepsilon}_{0j} + \hat{u}_{ij}$ in equation (42). Importantly, $\hat{u}_{ij}$ from equation (40) is equal to $\hat{\varepsilon}_{0j} + \hat{u}_{ij}$ in equation (42), meaning the value of $\hat{u}_{ij}$ is not the same in the two equations. In equation (40), $\hat{u}_{ij}$ captures all the difference between performance explained by the model, and the actual performance of each learner. In equation (42), on the other hand, we can roughly describe $\hat{u}_{ij}$ as the difference between the mean actual performance at school *j* and the actual performance of learner *i*. The term $\hat{u}_{ij}$ is thus said to deal with level 1 error. The term $\hat{\varepsilon}_{0j}$, on the other hand, captures the difference between the mean performance of the school explained by the fixed part of the level 1 intercept, in other words $\left(\hat{\alpha}_{00} + \hat{\alpha}_{01} X_{1j}\right)$, and the actual mean performance at the school. The term $\hat{\varepsilon}_{0j}$ thus captures level 2 error. As we shall see further on, things are not quite as simple as this, mainly because what we refer to as the mean performance of the school here poses a number of problems. But roughly, this is how error in the HLM is broken up. How this is expressed in the statistical outputs of the model will receive attention in section 5.5 below.

The random effects model in equation (14) is in fact closely related to the HLM. Using the commonly employed terms ‘random effects’ and ‘fixed effects’, we can say that in equation (41), $\left(\hat{\alpha}_{00} + \hat{\alpha}_{01} X_{1j} + \hat{\varepsilon}_{0j}\right)$ represents a random effect, whilst $\hat{\beta}_2$ represents a fixed effect.

In order to distinguish between the many structural variants of the HLM, it seemed convenient to capture key differences between variants in the following small schema:

L1	Intercept			Slope for X_{ij}			u_{ij}
L2	α_{00}	$\alpha_{01} X_j$	ε_{0j}	α_{10}	$\alpha_{11} X_j$	ε_{1j}	

Here the schema captures the structure of equation (42). We have a random effect for the level 1 intercept based on a level 2 variable, hence the α_{00}, $\alpha_{01} X_j$ and ε_{0j} areas are highlighted, but we have a fixed effect for the level 1 slope, hence only α_{10}, representing $\hat{\beta}_2$ from equation (42), is highlighted in the slope section of the schema. Finally, u_{ij} is highlighted as we have some residual learner-specific error that remains.

Barbosa and Fernandes (2001), in analysing SAEB 1997 data, use a model with random effects but no level 2 variable in the level 1 intercept. Their model could be expressed as follows:

$$Y_{ij} = \left(\hat{\alpha}_0 + \hat{\varepsilon}_{0j}\right) + \hat{\beta}_2 X_{2ij} + \hat{u}_{ij} \tag{43}$$

and in the schema as:

L1	Intercept	Slope for X_{ij}	u_{ij}

L2	$\boxed{\alpha_{00}}$	$\alpha_{01}X_j$	$\boxed{\varepsilon_{0j}}$	$\boxed{\alpha_{10}}$	$\alpha_{11}X_j$	ε_{1j}

However, Barbosa and Fernandes in fact use an HLM with three levels, corresponding to learner, shift and school. Instead of equation (43) we should therefore have:

$$Y_{ikj} = \left(\hat{\alpha}_0 + \hat{\varepsilon}_{0j} + \hat{\upsilon}_{0kj}\right) + \hat{\beta}_2 X_{2ikj} + \hat{u}_{ikj} \qquad (44)$$

where k refers to shift, and $\hat{\upsilon}_{0kj}$ is the error specific to shift.

The fullest possible HLM using just two explanatory variables, one at the school level and one at the learner level, would be the model with all seven blocks in the schema highlighted as follows:

L1	Intercept			Slope for X_{ij}			$\boxed{u_{ij}}$
L2	$\boxed{\alpha_{00}}$	$\boxed{\alpha_{01}X_j}$	$\boxed{\varepsilon_{0j}}$	$\boxed{\alpha_{10}}$	$\boxed{\alpha_{11}X_j}$	$\boxed{\varepsilon_{1j}}$	

The equation is:

$$Y_{ij} = \left(\hat{\alpha}_{00} + \hat{\alpha}_{01}X_{1j} + \hat{\varepsilon}_{0j}\right) + \left(\hat{\alpha}_{20} + \hat{\alpha}_{21}X_{1j} + \hat{\varepsilon}_{2j}\right)X_{2ij} + u_{ij} \qquad (45)$$

or, with the error terms gathered towards the right:

$$Y_{ij} = \left(\hat{\alpha}_{00} + \hat{\alpha}_{01}X_{1j}\right) + \left(\hat{\alpha}_{20} + \hat{\alpha}_{21}X_{1j}\right)X_{2ij} + \hat{\varepsilon}_{0j} + \hat{\varepsilon}_{2j}X_{2ij} + \hat{u}_{ij} \qquad (46)$$

There is a potentially confusing variety of other forms, some of which have their own names. Bryk and Raudenbush (1992: 17) deal with the **one-way ANOVA with random effects**, which our schema would represent as:

L1	Intercept			Slope for X_{ij}			$\boxed{u_{ij}}$
L2	$\boxed{\alpha_{00}}$	$\alpha_{01}X_j$	$\boxed{\varepsilon_{0j}}$	α_{10}	$\alpha_{11}X_j$	ε_{1j}	

the **means-as-outcomes regression**:

L1	Intercept			Slope for X_{ij}			$\boxed{u_{ij}}$
L2	$\boxed{\alpha_{00}}$	$\boxed{\alpha_{01}X_j}$	$\boxed{\varepsilon_{0j}}$	α_{10}	$\alpha_{11}X_j$	ε_{1j}	

and the **random coefficients regression model**:

L1	Intercept			Slope for X_{ij}			$\boxed{u_{ij}}$
L2	$\boxed{\alpha_{00}}$	$\alpha_{01}X_j$	$\boxed{\varepsilon_{0j}}$	$\boxed{\alpha_{10}}$	$\alpha_{11}X_j$	$\boxed{\varepsilon_{1j}}$	

Goldstein (1995, 18) discusses the **variance components model**:

L1	Intercept			Slope for X_{ij}			$\boxed{u_{ij}}$
L2	$\boxed{\alpha_{00}}$	$\boxed{\alpha_{01}X_j}$	$\boxed{\varepsilon_{0j}}$	$\boxed{\alpha_{10}}$	$\alpha_{11}X_j$	ε_{1j}	

Apart from selecting one of the above forms, or some other configuration possible within the schema, the analyst using more than one school variable and one learner variable must decide which school variables to link to which learner variables, and for which level 1 (learner level) coefficients, error terms should be attached. The possibilities are almost endless.

Having described the form of the HLM, we can now turn to the analytical benefits of this model. This assists us in determining which of the above forms are most pertinent to the analysis at hand. Bryk and Raudenbush (1992, 5) explain how the HLM essentially adds analytical power in three respects. Firstly, the HLM allows us to obtain relatively accurate models for individual groups, in our case schools. If we run 168 separate regression models for each of the schools in the SACMEQ data, we run into problems of insufficiently large numbers of observations, resulting in a poor goodness of fit. However, we could take equation (45) from above, and construct a model specific to school *j* that estimated the performance of new learner *i*:

$$\hat{Y}_{ij} = \left(\hat{\alpha}_{00} + \hat{\alpha}_{01}X_{1j} + \hat{\varepsilon}_{0j}\right) + \left(\hat{\alpha}_{20} + \hat{\alpha}_{21}X_{1j} + \hat{\varepsilon}_{2j}\right)X_{2ij} \tag{47}$$

All the terms on the right-hand side would have known values – the alphas and the error terms would have been estimated through the HLM estimation methodology. The advantage with the equation (47) model over a model using only data from school *j*, is that associations in other schools, some of them similar to school *j*, will have been taken into account in the estimation process. This strengthens the reliability of our prediction relating to school *j*. In terms of the kind of government monitoring dealt with in this thesis, this benefit of the HLM is of limited value.

Secondly, and this is important for our policy concerns, the HLM allows us to examine cross-level effects. For example, equation (45) allows us to examine how the school variable X_2 influences the association between the learner variable X_1 and performance Y. The model might tell us how school size influences the link between learner SES and learner performance. It may be that this link is systematically stronger, or weaker, in larger schools. The random coefficients model referred to above would also yield this analytical advantage, whilst a model such as the one in equation (41) would not.

Thirdly, and again this is important to us, the HLM allows us to split variance into a between-school component and a within-school component, using the relationship between $\hat{\varepsilon}_{0j}$ and $\hat{u}_{ij}$ in, for instance, equation (41) above. This is extremely important in terms of our understanding of how inequities in the system work. More will be said on this benefit in section 5.5.

5.5 The statistical outputs of the hierarchical linear model

The previous section explained the form of the HLM, a form that is clearly based on the form of the basic one-level regression, though the use of multiple levels brings in considerable complexity. Below, we shall examine the statistical outputs obtained from a statistical package that models data within an HLM. The package used is HLM for Windows Version 6.0. (A student version of this is available as freeware from the Internet. The URL is given at the beginning of the thesis). The outputs produced by this package are inserted into a new text file each time a model is run. We shall only discuss what HLM 6.0 refers to as the HLM2 model, meaning that other modelling options provided in the software, such as the more sophisticated three-level HLM3, will not be used.

Most of the discussion in this section is based on an analysis of the HLM outputs, and outputs from the one-level model described in section 5.2 above and a number of dummy models that were constructed to illustrate certain points. This type of discussion is more or less absent in Bryk and Raudenbush (1992) and the user's guide for the HLM software. This seems unfortunate.

The HLM statistical outputs that we shall discuss are the following:

- **Fixed effect coefficients.** These are the estimated values of the intercepts and slope coefficients, being the various values of $\hat{\alpha}$ and $\hat{\beta}$ obtained for a model such as the one in equation (42). These are referred to as the **fixed effect** coefficients in the HLM software output.

- **Error term values.** These are the values of the estimated error terms $\hat{\varepsilon}$ and $\hat{u}$ in a model such as that of equation (42). Using equation (42) as our example, this would involve as many $\hat{\varepsilon}$ values as there are groups, or schools, and as many $\hat{u}$ values as there are learners. We should bear in mind that the HLM may have several level 2 error terms $\hat{\varepsilon}$. For instance, equation (45) has two such error terms. In such a case, the number of values obtained would be the number of schools multiplied by the number of error terms in the equation. Normally, we would not request the values of all the error terms in our statistical output. But we may, and these values, in the case of the HLM software, are placed in separate data tables (we requested Stata .dta files), one file for level 1 values, and one for level 2 values. We could request this from Stata in the estimation of a one-level regression model, but we did not do that in our discussion in section 5.2 because it was not necessary in order to illustrate the model. However, in dealing with the HLM, it does become necessary to pay more attention to the error terms if we want to explain the statistical outputs on a more technical level.

- **Random effect variance statistics.** This is the variance associated with each error term $\hat{\varepsilon}$ or $\hat{u}$. More or less following Bryk and Raudenbush (1992, 29), the variance of $\hat{u}$ will be denoted by $\hat{\sigma}^2$ whilst the variance of $\hat{\varepsilon}_{0j}$ is denoted by $\hat{\tau}_0$, the variance of $\hat{\varepsilon}_{2j}$ by $\hat{\tau}_2$, and so on (using the notation in equation (45) as our point of departure). It should be pointed out that there is just one value for $\hat{\sigma}^2$, just one value for $\hat{\tau}_0$, and so on, given a particular model. The variance statistics are referred to in the HLM software output as the **random effect** in the model.

It is convenient to begin the discussion with the so-called null model or null form, an HLM form that is so basic it did not receive any attention in the previous section. It is a reduced version of equation (41) that looks as follows:

$$Y_{ij} = \left(\hat{\alpha}_0 + \hat{\varepsilon}_{0j}\right) + \hat{u}_{ij} \qquad (48)$$

It has no explanatory variables, and hence can have no slope coefficients. As we shall see in section 7.2 below, this null model is commonly used as a point of departure in multilevel analyses of schooling systems. The model has two error terms and, by implication, two variance statistics.

To begin the illustration of the HLM, we shall use the dataset that was used in section 5.2 with *_zpses* indicating the learner's socio-economic status (SES), *_sres21* indicating the school's possession or non-possession of computers, and *ratotp* indicating the reading score of each learner. For the HLM dataset, we would also need a variable with the identifier of the school. The outputs we obtain if we use just the reading score *ratotp*, and the school identifier variable, from this dataset, to run the null model in equation (48) is as follows:

Table 13: Null model (HLM output)

dependent var: ratotp	Level 1 units	3139
	Level 2 units	167
Fixed effect		
	coefficient (t stat)	
For intercept β_0		
intercept α_0	37.93 (35.7)	
Random effect		
	variance (p value)	
For intercept β_0	184.7 (0.000)	
Level-1	88.3	
Details on error terms		
	mean	*mean of squares*
ε_{0j}	0.000	180.1
$(\varepsilon_{0j})_i$	0.155	180.0
u_{ij}	0.000	83.8
$u_{ij}+(\varepsilon_{0j})_i$	0.155	272.9

It should first be clarified why β_0 appears in the statistical outputs. The term β_0 is actually the level 1 intercept within which we find the level 2 intercept α_0 – see discussions relating to equation (40) above. The value of the intercept we see in the output above, 37.93, is similar to the actual mean of the reading score, of 38.09. The fact that it is not exactly the same is the first of many counter-intuitive phenomena in the HLM statistical outputs that are related to the weighting of schools, which depends on school size, meaning the number of learners per school. This weighting is calculated by the HLM software. The SACMEQ weights have not been used in the above model, as we are still just illustrating the model (clearly introducing the SACMEQ weights, which is done further on, makes the HLM estimations even more complex). If we manipulate the SACMEQ data so that all schools have the same number of learners, we obtain an intercept in the output which equals the mean of the reading score exactly.

We now turn to the random effect statistics provided by the HLM software. The sum of the two variance statistics in table 13 is 273.1, which is close to the actual variance of the reading score, 272.9, but not exactly equal to it (it would not be exactly equal if we used the equal school size dummy data either).

The calculated mean and the mean of squares of three variables appearing in the error term tables produced by HLM are given above. One of these variables, ε_{0j}, is from the level 2 table (meaning there is one value per school). Two, $(\varepsilon_{0j})_i$ and u_{ij}, are from the level 1 table. The variable $(\varepsilon_{0j})_i$ is simply ε_{0j} repeated across each learner. The mean and mean of squares of the composite level 1 error term $u_{ij}+(\varepsilon_{0j})_i$ is also given.

Of note is the fact that the mean of all the squares of the composite error term $u_{ij}+(\varepsilon_{0j})_i$, which equals 272.9, is equal to the overall variance of the reading score calculated in the normal way (the mean of squares was calculated using 3139 minus one as the denominator, to be in line with the variance calculation). This is what we might expect. However, there are two phenomena we may not expect. Firstly, the mean of the level 2 error term repeated for all learners, the statistic $(\varepsilon_{0j})_i$, is not zero. Instead the mean is 0.155. This results in the mean for the composite error term also being 0.155 (the mean of the level 1 error term is zero). In a one-level regression model using unweighted data, we would always obtain a mean of zero for the error term values. However, the fact that the mean of ε_{0j} (the level 2 error term without repetition across learners) is zero, should indicate to us that the phenomenon is a function of having a different number of learners in each schools. If we use the manipulated dataset where each school has the same number of learners, the mean of the composite error term, and the mean of $(\varepsilon_{0j})_i$, become zero.

The second phenomenon is that the means of squares of each separate error term $(\varepsilon_{0j})_i$ and u_{ij}, that is the values 180.0 and 83.8, are substantially lower than the variances we were given in the HLM software output of 184.7 and 88.3. This is quite easily explained, however. At the level of each learner, one of the basic quadratic identities of algebra (Sydsæter and Hammond, 2002: 11) can be expressed as follows:

$$\left(\left(\varepsilon_{0j}\right)_i + u_{ij}\right)^2 = \left(\left(\varepsilon_{0j}\right)_i\right)^2 + 2\left(\varepsilon_{0j}\right)_i u_{ij} + \left(u_{ij}\right)^2 \tag{49}$$

As long as the mean of $2(\varepsilon_{0j})u_{ij}$ across all learners is not equal to zero, the mean of the left hand side of the identify (this corresponds to the composite error term referred to earlier) will not equal the mean of the sum of squares of the two error terms (the right hand side without the middle term). The average of the middle term for the actual SACMEQ dataset is in fact 9.1, which is the difference between the sum of 180.0 and 83.8, on the one hand, and the 272.9 mean of squares of the composite error term mentioned earlier, on the other hand. Relatively low mean of squares values for $(\varepsilon_{0j})_i$ and u_{ij} are a constant feature of the HLM residual outputs. However, as we shall see below, in a 'crude model', the sum of the middle term across all learners is always zero.

The standard ratio used to compare the variance statistics of the two levels is the **intra-class correlation coefficient**, which is calculated as follows for our null model:

$$\rho = \frac{\tau}{\tau + \sigma^2} = \frac{184.7}{184.7 + 88.3} = 0.677 \tag{50}$$

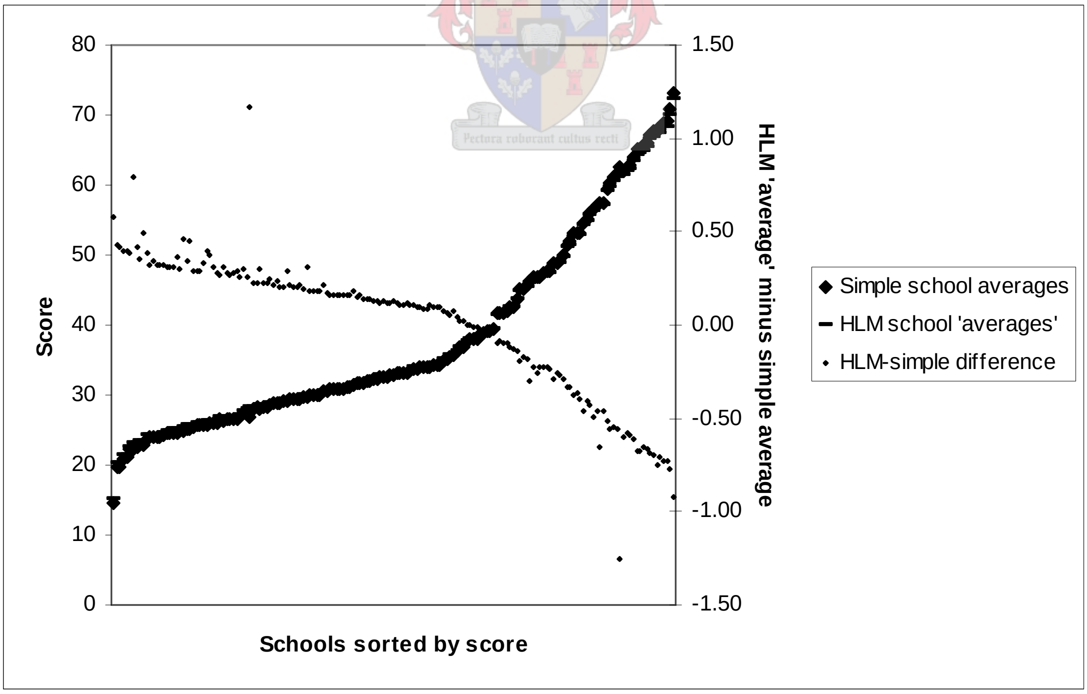
We would obtain a very similar coefficient of 0.682 had we used the variance values 180.0 and 83.8 derived directly from the error terms.

We obtain almost exactly the same two variance statistics of 184.7 and 88.3 by using loneway in Stata, which estimates what is referred to in Stata as the variance components model. We speak of the overall variance of a variable, in this case the reading score, being split into various components, hence the name of the model. We can refer to the variance of 184.7 as being the **between-school** variance, and the variance of 88.3 being the **within-school** variance, or the variance attributable to differences between learners within their schools. Thus the intra-class correlation coefficient is simply between-school variance as a proportion of total variance (Bryk and Raudenbush 1992, 30). The intra-class correlation coefficient is often referred to as rho (ρ is the Greek letter rho).

It is instructive to compare the HLM outputs of the null model against what we would obtain if we used a very crude, but rather intuitive approach to splitting the total variance across two levels. In this crude approach, we would obtain the simple average score of each school, measure the difference between each individual learner's score and the school average and call it u and measure the difference between each school's average score and the overall simple average for the system, and call it ε. This approach, when applied to the 167 schools of table 13, would result in a level 1 variance of 189.2, a level 2 variance of 83.7. Here the sum of the two variance statistics equals 172.9, or the total normal variance of all the reading scores. This implies that in the crude approach, the sum of the middle term $2(\varepsilon_{0j})u_{ij}$ is equal to zero.

The intraclass correlation coefficient obtained using the crude approach variance statistics of 189.2 and 83.7 is 0.693. This is substantially higher than the coefficient obtained in equation (50). Why this should be so is partly explained by the following graph.

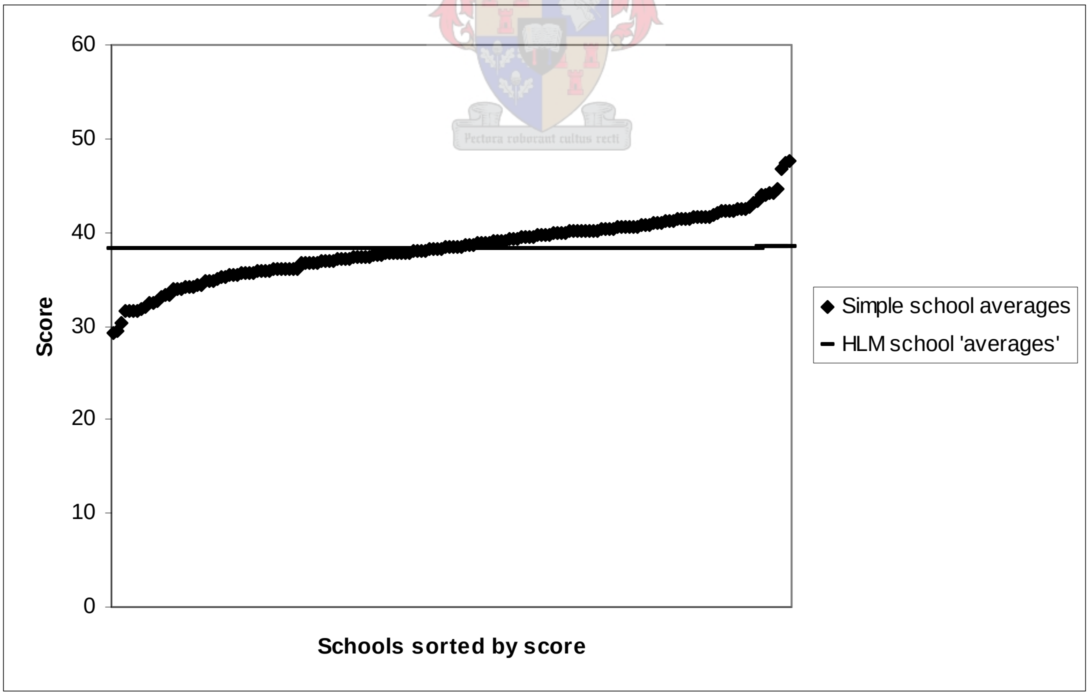
Figure 8: Simple school averages and HLM school 'averages'



The HLM school 'averages' can be understood as $\alpha_0 + \varepsilon_{0j}$. This HLM school 'average' is not the same as the simple school average of *ratotp.* In the above graph, the difference between the two is too small to be discerned, and hence the two sets of points are superimposed. However, the points indicating the difference between the two (these points should be read against the right-hand Y axis) show that where the simple school average is low, HLM provides a slightly higher value, and where the simple school average is high, HLM provides a slightly lower value. The end result must be that we will have less between-school variance if we use the HLM figures than if we use the simple school averages. This explains why the intraclass correlation coefficient using HLM (see equation (50)) is lower than what one would obtain using the crude approach. Essentially, HLM reduces differences between schools, and hence makes between-school variance seem a smaller part of overall variance.

The reduction in the between-school variance brought about by the HLM estimation methodology is extreme where all learner scores are random. The following graph indicates the differences between the simple school averages and the HLM school 'averages' in such a random situation (the actual SACMEQ mean and standard deviation for *ratotp* were used to generate completely random scores with a normal distribution).

Figure 9: Simple school averages and HLM school 'averages' in a random context



Although the scores are random in the artificial data, not all simple school averages are the same. Hence the simple school averages produce an inverted S curve in the above graph when sorted from lowest to highest. The HLM school 'averages' are all essentially the same, however, and equal more or less the average for the system as a whole. The HLM algorithm is in other words adjusting for the fact that we are dealing with a completely random dataset, and that differences between schools are not statistically significant. If we go back to figure 8, we can think of the slight reduction in between-school variation brought about by HLM as being a recognition of the fact that some of the between-school variance would be random, and not systematic.

Moving beyond the null model to a relationship between two inputs and one output, we can in fact use the HLM software to run a one-level model, if we specify the model with only one error term, as in equation (40) above. The three coefficients we would obtain are those we would obtain from Stata using a one-level model, in fact those of table 8 above. This would be a rather nonsensical use of the HLM software, but the point that the estimation of equation (40) would render the same results, regardless of the software used, is important.

The benefits of the HLM software are realised when we estimate the following model that includes a second error term, based on the earlier equation (42):

$$ratotp = \left(\hat{\alpha}_{00} + \hat{\alpha}_{01}\_sres21\right) + \hat{\beta}_2\_zpses + \hat{\varepsilon}_{0j} + \hat{u}_{ij} \qquad (51)$$

The output for this model is as follows:

Table 14: Two-input model (HLM output)

dependent var: ratotp	Level 1 units	3139
	Level 2 units	167
Fixed effect		
	coefficient (t stat)	
For intercept β_0		
intercept α_0	25.25 (33.5)	
slope *_sres21*	16.57 (9.9)	
For slope *_zpses*		
intercept	0.87 (8.6)	
Random effect		
	variance (p value)	
For intercept β_0	78.5 (0.000)	
Level-1	85.4	
Details on error terms		
	mean	*mean of squares*
ε_{0j}	0.000	73.7
$(\varepsilon_{0j})_i$	0.051	75.2
u_{ij}	0.000	81.1
$u_{ij}+(\varepsilon_{0j})_i$	0.051	156.3

To understand the above outputs, we should compare them to the HLM results in table 13 and to the results we obtained from the one-level model in table 8. A comparison with table 8 indicates that the fact of having a two-level model with two error terms changes both the intercept and the slope coefficients fairly substantially. The intercept is now 25 and not 19, the school computers slope coefficient is now 10 and not 13, and the learner SES slope coefficient is now 0.9 instead of the original 1.8. The overall picture has not changed radically, but the individual coefficients have shifted fairly substantially. We should bear in mind that had we inserted just one level error term, the level 1 error term u, the HLM software would have yielded exactly the same coefficients as those in table 8. The structure of the error terms in the HLM is clearly an important determinant of how the model will describe the relationships.

If we compare the table 14 results to those in table 13, we see that with respect to the random effects, both the variance values have dropped, but in particular the level 2 variance. A drop should be expected, as the variance reported in the statistical outputs is the residual variance, or the unexplained variance remaining after the model has explained what it can. Clearly, the introduction of the two explanatory variables reduces the amount of residual variance. The amount of variance left unexplained is not that different in the one-level and two-level models. The sum of the two variance statistics under the random effect heading above, equalling 163.9, is not that different from the residual mean sum of squares of 156.5 that we saw in table 8. And the total residual variance we would obtain if we used the mean of squares values for $(\varepsilon_{0j})_i$ and u_{ij} from the above table would be 156.3, which is very close to the 156.5 residual from the one-level model.

What are these statistics saying about the schooling system? They are saying that the explanatory variables of the SES of learners and the presence of computers in schools explain more of the between-school variance than the within-school variance. In fact, the model with the two explanatory variables has cut the between-school variance by more than half. We should not make the mistake of believing that because *_sres21* is the school-level variable, this is the variable solely responsible for explaining the between-school variance. The SES of learners explains both within-school and between-school variance. In fact, if we had run the model with only *_zpses*, and not *_sres21*, then the residual between-school variance would have been 141.2, which is a substantial reduction from the 184.7 variance we obtained in the null model. What this means is that differences in the average learner SES per school is associated with differences in the reading score, quite distinct from associations that may exist between the SES of an individual learner and his or her reading score. Put differently, individual effects are distinguished from compositional effects in the model.

What is striking is the absence of an overall goodness of fit coefficient, such as the R^2 statistic, in the typical discussion about and outputs from the HLM. To obtain an R^2 statistic that is comparable to that of the typical one-level model, we would need to use the variances obtained from the error terms. In the case of the above table, the total residual variance obtained from the error terms is 164.8, which when compared to the total variance of the reading score, 272.9, yields a proportion of explained variance, in other words an R^2 statistic, of 0.396, which is somewhat lower than the 0.427 obtained from the one-level model (see table 8).

As we saw in the previous section, there is considerable choice in the arrangement of level 2 of the HLM, both in terms of the positioning of error terms and the positioning of the level 2 variables. We need to be particularly careful when interpreting a model that has level 2 error terms attached to the level 1 slope coefficients. The following model illustrates the point:

$$ratotp = \left(\hat{\alpha}_{00} + \hat{\alpha}_{01}\_sres21 + \hat{\varepsilon}_{0j}\right) + \left(\hat{\alpha}_{10} + \hat{\alpha}_{10}\_sres21 + \hat{\varepsilon}_{1j}\right)\_zpses + \hat{u}_{ij} \qquad (52)$$

The HLM software produces the following output for this model:

Table 15: Expanded two-input model (HLM output)

dependent var: ratotp	Level 1 units	3139
	Level 2 units	167
Fixed effect		
	coefficient (t stat)	
For intercept β_0		
intercept α_0	26.67 (38.6)	
slope *_sres21*	8.97 (5.1)	
For slope *_zpses*		
intercept	0.62 (5.9)	
slope *_sres21*	0.79 (4.1)	
Random effect		
	variance (p value)	
For intercept β_0	28.2 (0.000)	
For slope *_zpses*	0.5 (0.000)	
Level-1	82.8	
Details on error terms		
	mean	*mean of squares*
ε_{0j}	0.000	17.7
ε_{1j}	0.000	0.3
$(\varepsilon_{0j})_i$	0.026	18.1
$(\varepsilon_{1j})_i$	0.004	0.3
u_{ij}	0.000	77.9
$u_{ij}+(\varepsilon_{0j})_i+(\varepsilon_{1j})_i(\_zpses)_i$	0.526	157.6

It would be an easy mistake to make to just add the mean of squares from the three error terms, and get a total residual variance of 96.3 (using the variances for $(\varepsilon_{0j})_i$, $(\varepsilon_{1j})_i$ and u_{ij}), and on the basis of this assume that the above model explains much more variance overall than the model from table 14. The mistake lies in ignoring that the error term $(\varepsilon_{1j})_i$ is multiplied by the value of *_zpses* for each learner. The last row of the above table provides what the correct overall unexplained variance would be, and it is almost the same as that in table 14. The mistake could just as easily be made if we used the three variances given under the random effect heading without making the necessary adjustment to the error term linked to the slope of *_zpses*.

5.6 Optimisation in the hierarchical linear model

Section 5.2 explained what was optimised in the basic one level regression model, and what computations can be used to obtain the statistical outputs of the model independently of any statistical software. This section explains what is optimised in the HLM, but not the more complex matter of how the computations, or estimations, to achieve that optimisation work. Moreover, this section will take as a given the HLM ‘average’ for each school, already introduced in the previous section, without explaining how this statistic is computed – the HLM school ‘average’ values are given in the residual files produced by the HLM software. This section is thus very far from being a complete exposition of the inner workings of the HLM. Such an exposition is not presented for two reasons. Firstly, given the complexity of the HLM, such an exposition must necessarily be long. Secondly, attempts at replicating the computations of the statistical software, in this case HLM 6.0, in Excel were only partially successful. What could be replicated almost perfectly, with the assistance of the Excel Solver facility, was the optimisation, on condition that the HLM school averages were already known, and this is one reason for the focus of this section. Another reason is that understanding the optimisation structure of any model is particularly important.

A major reason why the HLM computations are so complex is that the model makes use of Bayesian statistics. The study of statistics and econometrics can follow one of two basic methods, the classical and the Bayesian one (Gujarati, 2003: 12). The classical method, which is used in the estimation of the one-level regression model, involves solving a number of unknown terms from a set of known values according to a formula in a rather linear fashion. The Bayesian method is used when there are not enough known values to compute the unknown terms unambiguously. What then happens is that one or more plausible values are put in the place of some of the missing values, and the formula is applied as if the plausible values were real. This results in the solving of the remaining missing values. One or more tests are then applied to check whether the plausibility of the imputed missing values could be improved. If this is the case, then the formula is applied again. The process is repeated iteratively until an optimal situation is reached. This is a crude rendition of what occurs in Bayesian statistics, and in the HLM (Bryk and Raudenbush, 1992: 230). The HLM outputs presented in the previous section were all the result of an iterative estimation procedure. For example, the two-input model of table 14 involved five iterations.

Bryk and Raudenbush (1992: 32) explain that there are three types of statistics that are estimated through the HLM's Bayesian estimation procedures:

- The **fixed effects**, for instance the α_{00} or β_2 coefficients of equation (51). The optimisation that was replicated in Excel, and is explained in this section, results in these fixed effects.

- The **random level 1 coefficients**, or what has been referred to so far as the ‘HLM school averages’. There is one such statistic for each group, or school. In this section, it will be assumed that these statistics have already been calculated.

- The **random effects**, or the between-school and within-school variance statistics reflected in table 14. These statistics are not discussed in this section.

We saw that optimisation in the one-level regression model is aimed at the minimisation of the total sum of squares of the residual error term, or $\sum \hat{u}_i^2$ in equation (15). This might lead one to expect the HLM to optimise the intercept and slope coefficients through the minimisation of what has been called the composite error term, or $u_{ij}+(\varepsilon_{0j})_i$ from table 14. However, this is not the case. The HLM incorporates a weighting system which, firstly, weights u_{ij} differently from ε_{0j} and, secondly, uses number of observations per school to weight each school.

Key to understanding the HLM weighting system is the weighted least squares (WLS) optimisation methodology, which differs from the ordinary least squares (OLS) methodology described in section 5.2. In a one-level WLS model, the following equation replaces equation (15) (Gujarati, 2003: 397):

$$w_i Y_i = w_i \hat{\beta}_0 + w_i \hat{\beta}_1 X_{1i} + w_i \hat{\beta}_2 X_{2i} + w_i \hat{u}_i \qquad (53)$$

The variable w_i is a weight, which may differ across observations, applied to all the terms in the equation.What is minimised is $\sum (w_i \hat{u}_i)^2$, in other words the sum of the squares of the *weighted* residual term. Equation (53) will yield different β coefficient values from those obtained from equation (15). Why would the analyst use WLS? Essentially, this is to deal with heteroscedacity, a situation in which there is a non-constant distribution of variance, for example more variance, or higher absolute values of u, at higher values of Y, X_1 and X_2 (Gujarati, 2003: 387). The solution lies in weighting observations more in the regression computation if they have lower error term values in absolute terms. Through this approach, outliers come to count for less in the estimation of the β coefficients.

We shall demonstrate the construction of the weight w in the HLM through reference to the structure we had in equations (42) and (51). Equation (42) is reproduced below.

$$Y_{ij} = \left(\hat{\alpha}_{00} + \hat{\alpha}_{01} X_{1j}\right) + \hat{\beta}_2 X_{2ij} + \hat{\varepsilon}_{0j} + \hat{u}_{ij} \tag{54}$$

The explanatory variables X_{1j} and X_{2ij} represent the same adapted SACMEQ variables *_sres21* and *_zpses* used in the previous section. What we know are the school-level values for X_{1j} and the learner-level values for X_{2ij} and Y_{ij}. However, we also know the school-level values for the HLM school 'average', represented by β^*_{0j} in Bryk and Raudenbush (1992: 39), because we have decided to accept these values a priori from the HLM level 2 residuals file. The unknowns are the three values corresponding to the coefficients α_{00}, α_{01} and β_2, and the two error terms ε_{0j} and u_{ij} from the school and learner levels respectively. Values for the three coefficients were obtained in Excel by setting up the formulas explained shortly, and getting the Excel Solver facility to minimise an overall variable statistic containing the two error terms. As we shall see, the formulas involved having a form of a WLS equation at school level.

The term β^*_{0j} does not appear in equation (54), but it is implicitly part of this equation. The following equation explains how β^*_{0j} fits in to the previous equation.

$$\beta^*_{0j} = \hat{\alpha}_{00} + \hat{\alpha}_{01} X_{1j} + \hat{\varepsilon}_{0j} \tag{55}$$

Logically, the following equation should also hold given the previous two equations:

$$Y_{ij} = \beta^*_{oj} + \hat{\beta}_2 X_{2ij} + \hat{u}_{ij} \tag{56}$$

What should be clear is that both of the error terms rely completely on the value of β^*_{0j}. The variance of the level 2 error term ε_{0j} is denoted by τ and is called the **parameter variance**, using the language of Bryk and Raudenbush (1992: 33). The variance of u_{ij}, which is one value for the whole dataset, is divided by the number of observations per school to create an **error variance** statistic for each school, which is denoted by V_j:

$$V_j = \frac{\operatorname{var}\left(\hat{u}_{ij}\right)}{n_j} \tag{57}$$

Clearly, the error variance would differ across schools if there is a different number of learners per school in the dataset (this is the case with the SACMEQ dataset).

An overall variance statistic that is sensitive to the variance at both levels is then constructed as follows:

$$\omega_j = \frac{1}{\tau_{00} + V_j} \tag{58}$$

This overall variance statistic ω_j is also a school-specific weight, and it is referred to as the **precision statistic** by Bryk and Raudenbush. Clearly, the greater the overall variance for a school, the lower the value of ω, and the less importance we would want to attach to the school in question in the estimation procedure. This takes us back to the WLS regression equation, equation (53) above. The inverse of the weight ω replaces w, and we construct the following equation at level 2 (using the terms from equation (55)):

$$\frac{\beta^*_{0j}}{\omega_j} = \hat{\alpha}_{00}\left(\frac{1}{\varpi_j}\right) + \hat{\alpha}_{01}\left(\frac{X_{1j}}{\omega_j}\right) + \frac{\hat{\varepsilon}_{0j}}{\varpi_j} \tag{59}$$

If we examine equations (55) to (59), we see that we have now linked the two error terms to the three unknown coefficients we want to estimate. We minimise the sum of squares of $\frac{\hat{\varepsilon}_{0j}}{\varpi_j}$, or $\sum\left(\left(\frac{\hat{\varepsilon}_{0j}}{\varpi_j}\right)^2\right)$, by testing various combinations of values in the coefficients α_{00}, α_{01} and β_2. Because we are following a 'lazy' approach, we do not pursue the iterative estimation procedures of Bryk and Raudenbush (1992). Instead, we let Excel Solver do the work. The Excel Solver results were very similar to the HLM software results:

Table 16: Optimised fixed effect coefficients from the HLM

	HLM software output	*Excel Solver results*
intercept α_0	25.24580	25.23879
slope *_sres21* (or X_{1j})	16.57203	16.58625
intercept for slope *_zpses* (or X_{2ij})	0.86958	0.86930

Given how close the results were, it was concluded that the Excel simulation was a true reflection of the optimisation of the HLM.

6 INITIAL VARIABLE SELECTION AND MANIPULATION

In section 5 above, we examined how to model associations between an output, in our case a school performance score, and two explanatory variables in such a way that we could take into account the interaction between the two explanatory variables, or gauge the *net* effect of one explanation whilst ‘controlling’ for the existence of the other. That examination was good for an illustration of the theory, but a more realistic model of school production clearly requires more than two explanatory variables. Very often, there are hundreds of variables in a school dataset that might explain learner performance. In the case of the SACMEQ dataset, there are 381 variables derived from the three questionnaires. In order to avoid what Gujarati (2003: 508) refers to as specification errors, we need to (amongst other things) ensure that we do not include unnecessary or irrelevant explanatory variables in our model. Moreover, we occasionally need to manipulate the original variable values in order to make them more usable for the modelling process. It may be beneficial to combine several of the original variables into one new variable in order to simplify and clarify the model, and to maintain the overall goodness of fit of the model as reflected in the adjusted R^2 statistic. This section is about the prioritising of original variables, and the manipulation of those variables, in order to come up with a new, reduced set of variables that we can use as our basic set of ingredients in our model construction work (discussed in section 7 below). The discussion here is organised into three parts. Firstly, data mining as a way of prioritising variables is discussed. Secondly, the generation of new variables through a few methods, including factor analysis, receives attention. Thirdly, the matter of multicollinearity between explanatory variables is examined.

Data mining involves the use of some initial regression modelling in order to determine what variables to include and what variables to exclude from a final regression model. The approach is either bottom-up, or top-down. The former involves starting with no explanatory variables, and including significant variables one by one that comply with some minimum criterion. The latter involves starting with a model that includes all explanatory variables, and excluding variables one by one on the basis of some minimum criterion. The top-down approach is regarded as more credible (Gujarati 2003: 515). The overall credibility of data mining is far from secure, however, largely due to the fact that the technique has been used irresponsibly in violation of well-informed understandings of the object of study. Baker (2001: 82), who is less critical of data mining than Gujarati, differentiates between responsible data mining and 'data dumping', in other words reckless use of statistical method. Any data miner should clearly be wary of the temptation of being carried away blindly by the data and the statistics. The mental model must guide.

In Stata, a key data mining tool is the stepwise command, `sw`, which has four variants: the backward selection, backward stepwise, forward selection and forward stepwise approaches. All of these approaches involve the use of the *p* value as a threshold for inclusion into or exclusion from the model. For example, the backward selection approach begins with a first step where all explanatory variables are included in the model, and the one variable with the highest *p* value, if greater than the minimum threshold set by the analyst, is excluded. In the second step, the variable with the highest *p* value above the minimum threshold of all remaining variables, is removed. And so on until no variable has a *p* value exceeding the threshold. Importantly, in each step the *net* association of each explanatory variable and the dependent variable is being gauged, because we are working with a multivariate model.

Baker (2001, 82) clearly opposes the use of stepwise techniques in selecting variables, and he is not alone in this regard. Berk (2004: 132) is ambivalent about the utility of these techniques. Despite the opposition and reservations, however, analysts make use of these techniques (see for instance Crouch and Perry, 2002: 7). This seems understandable, given the simple yet highly informative nature of stepwise techniques. As always, what should be avoided is some slavish adherence to the statistical technique at the expense of an informed understanding of the real world system at hand.

Data mining should occur using variables that are fit for the modelling purpose, generally ratio or interval variables. Nominal and ordinal variables can generally not be used without some manipulation. Specifically, nominal and dummy variables must be converted to dummy variables, binary variables that take on a 0 or 1 value. A SACMEQ example would be the nominal variable relating to type of floor in the learner's home, which has five possibilities. This variable would need to be converted to four dummy variables indicating the presence or absence of four of the options. The fifth option is implied by a 0 value in all the four variables. Five dummy variables should be avoided. This constitutes the 'dummy variable trap' and leads to anomalies in the model outputs (Gujarati 2003: 302).

The data mining of the SACMEQ variables began with the manipulation of a few variables so that they became more relevant for the question of school production. For example, number of permanent teachers in the school was converted to percentage of teachers in the school who were permanent. Thereafter, a program constructed in Stata was used to examine two things with respect to all variables: strength of the association with learner performance within a one-to-one bivariate regression analysis, and the strength of this association in net terms when many explanatory variables are included in the model simultaneously. Importantly, the program did not ultimately select what variables to include in the new set of SACMEQ variables. The program simply provided background information (reflected in Appendix B) which informed a policy-driven approach to variable selection.

The program required the user to manually input which variables were nominal or ordinal, and hence which variables required conversion to dummy variables. It then automatically assigned an R^2 coefficient for each of the 381 variables. This corresponded to whichever of the two performance scores produced the best overall goodness of fit in a bivariate regression analysis (or multivariate analysis, where the original variable was broken down into dummy variables). With the conversion of nominal and ordinal variables to dummy variables, the total number of variables increased from 381 to 831, a number too high to run a successful stepwise analysis in Stata. The program selected around 200 of the 831 variables, based on the R^2 values obtained. Stepwise analysis, using the backward selection method, then occurred using the around 200 variables. This process occurred twice, once for the mathematics score, and once for the reading score. The end result of this stepwise analysis was that around 25 variables were identified as having particularly strong net associations with the performance scores. The p value threshold used in the analysis was 0.0001. The R^2 coefficient for each variable, whether each variable passed the stepwise process or not, and the actual question asked in the questionnaire were tabulated. The table constituted a useful basis from which to make decisions around the selection of a new, substantially reduced set of variables (see Appendix B).

The results of the bivariate analyses, and the results of the stepwise analyses, were clearly saying slightly different things. The variables prioritised by the stepwise analyses were not always the variables with the highest R^2 coefficients in the bivariate analyses.

As long as the mental model is used as a guide, it should be relatively safe to collapse several variables into one, where these several variables deal with the same basket of inputs or explanations in our mental model. For example, we would want to collapse the variables from the learner questionnaire in SACMEQ dealing with the presence of various household items such as a cassette player, a refrigerator and a telephone into one new variable that would deal with the learner's material standard of living. Two approaches to collapsing original variables into new variables are discussed here, and both were applied to the SACMEQ data.

The first approach is conceptually easy once we have understood the basic regression model. We can simply take the slope coefficients obtained from a multivariate regression that treats the learner score as the dependent variable, and use these slope coefficients to construct a single new variable. The following operation is performed to produce the new variable **z**.

$$\underset{n\times1}{\mathbf{z}} = \underset{n\times v}{\mathbf{X}}\ \underset{v\times1}{\boldsymbol{\beta}} \tag{60}$$

The matrix **X** of n observations and v original variables is multiplied by the slope coefficients **β** obtained from a prior multivariate regression analysis. This approach, not dealt with in any of the textbooks consulted, seems defensible, though we should be conscious of some problems. Above all, by using a performance score to influence the values of the new variable, it could be argued that we are artificially enhancing the significance of the new explanatory variables. However, the approach is essentially no different from the common approach where the analyst attaches an importance weight to each original variable, on the basis of the analyst's mental model of how production occurs. Such an approach is taken by Hungi (2005, 2) in the weighting of different degrees of textbook availability. What the approach of equation (60) is doing, is to obtain the importance weights statistically, using the slope coefficients **β**. The approach seems justified where it is not possible or relevant for the study at hand to rely on theory or some hypothesis. For example, a study focussing on the economic linkages in school production cannot be expected to pay too much attention to the theory behind the classroom pedagogy, and so the use of slope coefficients may be permitted in weighting each of several classroom methodologies pursued by the educator. This explains the use of the regression approach in the construction of the new variables *class_meth_math* and *class_meth_read* dealing with classroom practice.

The second approach discussed here is arguably more defensible in that it does not rely on performance scores at all. This second approach is the factor analysis approach. This approach is still promoted guardedly due to a bad reputation gained prior to the existence of computers, and associated enhancements to the technique. If used in a manner informed by a clear mental model, factor analysis is a powerful tool. In Stata, a new variable based on several original variables is easily obtained through the use of the factor and score commands. The underlying statistical method is fairly complex. A simplified explanation follows of a common factor analysis method, the principal component method. In this explanation, it is assumed that we only want to obtain one new variable, and that the original variables have all been standardised so that their mean is zero and their variance 1.0. The explanation is based on both Johnson and Wichern (2002: 477) and the relevant Stata manuals.

$$\underset{n\times 1}{\mathbf{z}} = \underset{n\times v}{\mathbf{X}}\,\underset{v\times 1}{\mathbf{s}} \quad (61)$$

The new variable, $\mathbf{z}$, referred to as the factor, is obtained by multiplying the values of the original variables with scoring coefficients, represented by the column matrix $\mathbf{s}$ above. The scoring coefficients are derived from the correlation matrix, $\mathbf{\Sigma}$, of the original variables and as many factor loadings as there are variables.

$$\underset{v\times 1}{\mathbf{s}} = \underset{v\times 1}{\mathbf{f}}{}'\,\underset{v\times v}{\mathbf{\Sigma}} \quad (62)$$

The correlation matrix can be obtained using, for instance, the pearson function in Excel. To obtain the factor loadings, it is illustrative to first obtain the covariance matrix of the original variables, which can be calculated using mmult in Excel and the following equation (the method is made easier by the fact that we are using standardised variables):

$$\underset{v\times v}{\mathbf{V}} = \underset{v\times v}{\mathbf{X'X}} \quad (63)$$

We then obtain the eigenvalues and eigenvectors associated with the covariance matrix **V**. The non-transparent but simpler way of doing this is to use the command `symeigen` in Stata. The factor loadings are then calculated as follows:

$$\underset{v\times 1}{\mathbf{f}} = \sqrt{\lambda}\, \underset{v\times 1}{\mathbf{e}} \tag{64}$$

where λ is the first or highest eigenvalue from the covariance matrix, and **e** is the eigenvector associated with that eigenvalue.

Factor analysis has been used in the analysis of Brazil's SAEB data, specifically to obtain improved variables indicating learner SES (Barbosa, Fernandes, Dos Santos *et al*, 2000a). Willms and Somers (2001: 415) do the same in their analysis of Laboratio data, and Hungi (2005: 2) follows this approach in dealing with the Kenya SACMEQ data.

Appendix C explains what methodologies were used in deriving each of the 21 variables in the new, reduced set of SACMEQ variables for South Africa. Both factor analysis and the use of slope coefficients as weights were employed. The 'true' approach is one where an actual metric was (more or less) maintained, for example in the calculation of number of meals eaten per day. The mental model used to guide the process is the policy-oriented one explained in section 4.5 above. The aim was to produce a new variable for each policy area. The data seemed to permit this for 18 of the 22 policy areas of the mental model. In some cases, the link between policy area and the new variable is fairly indirect. For example, the school principal's teaching load is used as an indicator of the principal's salary and fringe benefits, the logic being that a lower teaching load is an incentive to attract better managers into the job. The teaching load variable can be regarded as a proxy or instrumental variable used as a substitute for the more directly relevant variable, which would be the salary and fringe benefits of the school principal (Gujarati, 2003: 527).

Apart from the 18 new variables selected to describe individual policy areas, three new variables of a more generic nature were included due to their significant association with learner performance, and the emphasis they receive in existing production models. These are the variables relating to teacher latecoming, the learner's age, and the learner's gender. All the 21 variables are discussed briefly below. Clearly the variable selection and manipulation stage already begins to render interesting policy information.

Full details on how the new variables relate back to the original variables and original questionnaire items is provided in Appendix B.

The following are the new learner-level variables, meaning variables that could differ in value from one learner to the next within the same class and school:

- **Number of years repeated** (*repetition*). This new variable contains the total number of years that a learner has repeated.

- **Textbooks per learner** (*textbooks_math/read*). This variable uses data from the learner questionnaire. An increase in the ratio of textbooks per learner is most markedly associated with an increase in learner performance below the 0.5 textbooks per learner level, in other words when two or more learners share the same textbook. Above this level, the association is relatively weak. Glewwe *et al* (2000: 3) make reference to a similar 0.5 threshold in a study on Philippine schools. For this reason, this variable was changed to 0.5 at all values above this threshold to improve the sensitivity of the model to differences below the threshold. This effect of this is akin to that of the piecewise linear model of equation (37). This variable was duplicated to deal with mathematics and reading textbooks separately.

- **Average number of meals per day** (*daily_meals*). As having regular suppers was associated more strongly with performance than having regular breakfasts, suppers received a weight of 1.5 and breakfasts a weight of 0.5. Lunch remained with a weight of 1.0.

- **Years of schooling of parents** (*parent_educ*). The mother's education emerges as twice as powerful a predictor of performance as the father's education. The use of English is also a powerful predictor, and hence data on this was worked into the new variable. For the 12% of learners who had only one parent and the additional 18% of learners who only knew about the educational attainment of one parent, schooling of just one parent was considered. For the 2% of learners who said they had no parent and additional 13% of learners who knew about the educational attainment of neither parent, the school mean of parents' years of schooling was used. As Harbison and Hanushek (1992: 95) remind us, years of schooling of parents is a proxy variable for the real variable of concern, which is the level of educational support received from other members of the household, including siblings. It is important to keep the caveats in mind (with respect to this variable and other variables), both when we need to explain why our production model is not explaining more, and when we need to draw policy conclusions. (The Brazil SAEB learner questionnaire is interesting with respect to educational support in the home in that it elicits more responses than the SACMEQ learner questionnaire regarding educational activities in the home, and the structure and relationships of the household.)

- **Learner SES** (*learner_ses*). Six features in the learners home (type of lighting, type of floor, type of walls, and the existence of a cassette player, telephone and refrigerator) were used to construct an index of the learner's SES. This is one of the three variables where factor analysis as described above was employed.

- **Learner age** (*learner_age*). What seemed surprising is the number of highly over-aged Grade 6 learners. Altogether 57 learners are above age 16 – one is age 25. The distribution of the data suggests weakly that these high values are valid, and not errors. Therefore no cleaning of the high values took place.

- **Learner gender** (*learner_gender*). Although this variable does not appear to be an important explanatory variable according to the bivariate regression and stepwise tests performed, the emphasis that gender receives in education planning generally and in the research prompted a closer look at the significance of this variable in the dataset. The gender variable is in fact significant in several multivariate regression analyses, despite the low explanatory power of this variable when taken on its own.

The following are the new educator-level variables (these eight variables were duplicated to deal with the responses of the mathematics and reading teachers, so with the duplicated textbooks variable, there is in fact a grand total of 30, not 21, variables in the new reduced set of variables):

- **Years of pre-service training** (*yrs_preserv_math/read*). This is the number of years of schooling plus training achieved by educators. The relevant data from the principal questionnaire relating to educators in the school as a whole turned out to be a more powerful predictor of learner performance than the data relating to the specific educator of the learner. This suggests the presence of strong compositional effects with respect to the level of education of educators in a school. The school data received a weight of 0.75, and the data on the individual educator received a weight of 0.25.

- **Days of in-service training** (*day_inserv_math/read*). Though there are questions in the educator questionnaire relating to the quality of in-service training (the educator is asked to evaluate the training received), the data from these questions was not used due to the type of correlation with learner scores. Basically, educators with better performing learners gave a lower rating to the in-service training they received. This is noteworthy, but there are several possible interpretations. It seemed impossible to gauge the quality of the training programmes from the questionnaire data. Therefore, only the quantitative matter of number of days of in-service training received in the last three years was taken into account.

- **Teacher SES** (*teacher_ses_math/read*). The ideal would have been a variable on educator salary. There is a question in the educator questionnaire in which the educator is asked to gauge the importance of salary, but the question is couched in such a way that it is not possible to ascertain whether the educator thinks her salary is adequate or not. A variable on teacher SES was constructed instead, using factor analysis, and two teacher household variables.

- **Intensity of evaluation of educators** (*teacher_eval_math/read*). In place of a variable on educator incentives, as specified in our mental model of figure 7, a variable on the evaluation of educators was constructed. There was no system of educator incentives in place in 2000 in South Africa, and educator incentives are not dealt with in the questionnaire. Constructive criticism and evaluation can, however, be regarded as a non-material and psychological incentive. The frequency of professional advice from the school principal reported in the teacher questionnaire was used as the best indicator of the intensity of evaluation.

Interestingly, the relationship between the frequency of the principal's advice and learner performance is convex. No advice and the most frequent advice (once or more a month) are associated with the lowest performance, whilst advice on an annual basis is best. This finding emerges whether one controls for other effects or not. It should not be surprising if we consider the interaction of selection effects and treatment effects discussed in section 4.1. On the one hand, a school principal can be expected to direct more frequent advice to educators who do not perform well, and whose learners perform poorly. This would account for the selection effect, and the association between a high frequency of advice and low performance. On the other hand, we might expect more frequent advice from the school principal to improve educator performance, and hence learner performance. This would be the treatment effect, and could account for the association found between zero advice and low performance. A production model is primarily concerned with the treatment effect, and not the selection effect. We want to find out how well the intervention improves output, not how effectively actors in the system react to poor output through targeted interventions. How can we disentangle the two effects? We can use the apparent distribution of the two effects across the range of intervention intensity to inform our variable construction. The fact that there is a positive correlation between the intervention and performance at low levels of the intervention, in other words at the zero to once a year level, suggests that it is within this range that we would need to seek out the treatment effects. In this range, the selection effects appear to be weaker. Of course, it is possible that the dynamics are even more complex than a 'simple' dichotomy between treatment and selection effects. The treatment effect may itself be non-linear. It is plausible that advice on a monthly basis may have a lower impact than annual advice, because it is less thoroughly planned advice. The approach taken in constructing the variable was to follow the approach of equation (60). Specifically, an index was created that gave the strongest weighting to what appeared to be the most effective intervention frequency, namely the annual one.

- **Class size** (*class_size2_math/read*). The values in the original class size variables in the dataset were squared in order to cater for the non-linear relationship between performance and class size. Essentially, the larger the class, the more the addition or subtraction of a learner influences mean performance. Put differently, there are increasing negative marginal returns as class size grows. Squaring class size provides us with a more predictive variable, as in the polynomial model of equation (34).

- **Value of class methodology** (*class_meth_math/read*). Here the results of some regression analyses were used to weight a few classroom practices that were associated with better learner scores. In mathematics, allowing learners to work on their own, interacting on a one-to-one basis with individual learners, assigning homework, and getting parents to sign homework books came out as valuable practices. In reading, promoting listening skills and having parents sign homework books came out as important.

- **Teacher hours in a year** (*hrs_year_math/read*). The best proxy for hours of contact time in the year for learners was hours in a year educators would be expected to teach. This variable used teaching time data reported by the educator and the school principal’s report of days lost in the year due to, for instance, bad weather. Given that in Grade 6, educators are responsible for one class for the entire year, this variable would reflect contact time before educator or learner absenteeism had been taken into account.

- **Level of parent involvement** (*par_involve_math/read*). The frequency of meetings between parents and the educator displayed a non-linear relationship with learner performance similar to that found for professional advice from the principal. Best was meetings once a term, worse was more than once a month, and worst was never. We can be certain that we are dealing with intertwined selection and treatment effects. Educators or parents select more interactions when a learner performs poorly, but at the same time, more pro-active interaction between the educator and the parents results in better performance through the treatment effect. The approach taken was to construct a variable where the most effective treatment, which was once a term meetings, was weighted the strongest.

The following are the new school-level variables:

- **Level of school infrastructure** (*school_infra*). The ratio of flush toilets to learners, as well as the presence of a school library, a school hall, a staffroom, an office for the principal, a photocopier, a computer and a tuckshop were incorporated into this variable.

- **Principal's years of pre-service training** (*yrs_preserv_prin*). This variable was constructed much like the similar variable referring to educators. It appeared to be the best indicator of the principal's capacity to manage, and the policy area management training from the figure 5 mental model. There is a question in the school principal questionnaire that asks whether the principal has participated in school management training, and if so, for how many weeks. Data from this question was not deemed appropriate, given the low association with learner performance. The principal's years of pre-service training, on the hand, correlated well with learner performance (see table 36 in Appendix B).

- **Principal's teaching load** (*prin_teach_load*). In the absence of any data on the SES or the income of the principal, the management constraint experienced by the principal in terms of hours of teaching per week, was considered.

- **Intensity of district support** (*dist_support*). Frequency of departmental visits to schools is positively correlated with learner performance (both with and without controls for school and home background effects). This could mean two things: either visits are causing performance to improve, or visits are merely targeted at schools which perform better anyway, possibly because they are not remote (and hence not disadvantaged) schools. One would actually expect the correlation to be negative, given the emphasis placed on supporting worse performing schools in Department of Education plans, but such a hypothesis is not supported by the data. The existence of a resource centre near the school resulted in an addition to the intensity of district support value.

- **Proximity to urban facilities** (*ruralness*). The original variable indicating how urban or rural the surroundings of the school are was more or less retained, except that four categories were reduced to three. Where the school principal described the school as 'isolated' or 'rural', the variable *ruralness* assumed a value of 1 – these two descriptions were associated with virtually identical learner performance levels. The description 'in or near a small town' resulted in a value of 2, and 'in or near a large town or city' resulted in a value of 3. The positive association between the variable *ruralness* and learner performance seen in table 38 of Appendix C thus indicates that that a less rural school is associated with better performance.

Turning to the subject of multicollinearity, Gujarati (2003: 341) explains: 'There is no pair of words that is more misused both in econometrics texts and in the applied literature than the pair "multicollinearity problem"'. This 'problem' arises when explanatory variables are correlated to each other. The correlation coefficient with respect to two variables is the covariance of the two variables divided by the square root of the product of the two separate variances (Blalock 398).

$$r = \frac{\sum\left(X_1 - \overline{X}_1\right)\left(X_2 - \overline{X}_2\right)}{\sqrt{\left(\sum\left(X_1 - \overline{X_1}\right)^2\right)\left(\sum\left(X_2 - \overline{X_2}\right)^2\right)}} \tag{65}$$

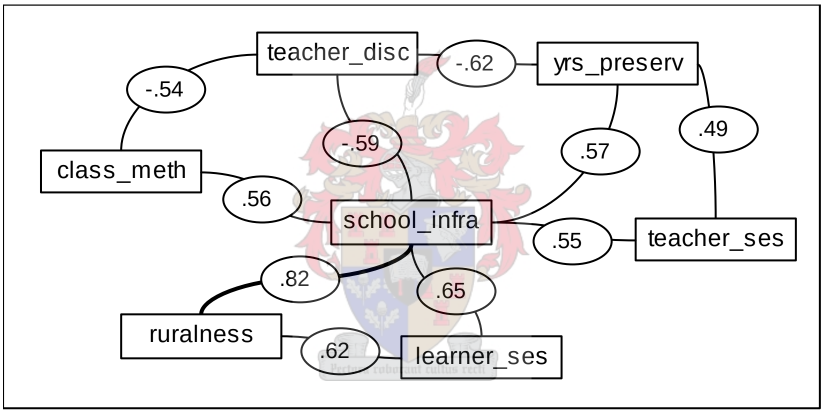
Multicollinearity can be said to be a serious matter wherever a pair of explanatory variables has a correlation coefficient exceeding a value of approximately 0.8 (Gujarati, 2003: 341). The argument against placing two explanatory variables with such a high correlation into the same model is that the model becomes confusing and unclear. Imagine two variables, *A* and *B*, which separately have a significant association with the output or dependent variable. If *A* and *B* are strongly correlated to each other, then in the multivariate model with both *A* and *B,* the significance of each individual variable, as measured by for instance the *t* statistic, might be very low, despite the fact that the overall goodness of fit as measured by R^2 is high. This would render the slope coefficients, reflecting the net effect of each variable, insignificant and meaningless.

It is important to realise that the low *t* statistics obtained in the multivariate model are essentially a result of the sample being too small. The larger the sample, the better the *t* statistics, even if the correlation between *A* and *B* is consistently high (this is assuming that that *A* and *B* are in fact distinct explanations, and not effectively the same thing).

Gujarati and others (Hanushek, 1979: 374) emphasise that multicollinearity should not stop the analyst from inserting correlated explanatory variables into the same model. The problem ought to be reduced by maximising sample size and trying to combine similar effects into single variables, but ultimately multicollinear explanatory variables may still be inevitable. The analyst would then need to admit that separating out the effects of multicollinear variables is not possible, that we are not able to distinguish between the effects of *A* and *B*. This would simply be a part of the conclusion.

Table 38 of Appendix B contains a correlation matrix indicating the correlation between all the possible pairs of variables. We can ignore the high correlations between the corresponding variables from the reading and mathematics educator questionnaires, as the one is not used with the other in any of the models presented further on. This leaves us with one correlation that is worryingly high, namely that between *school_infra* and *ruralness*. The correlation here is 0.82.

Figure 10: Multicollinearity in the new SACMEQ variables



Source: SACMEQ, 2000.

Moreover, there is a cluster of the new variables with high correlations (greater than or equal to 0.55) with *school_infra* and with varying correlations, sometimes high, between each other. The above diagram displays the cluster of highly correlated variables. The correlation between *school_infra* and *ruralness* is indicated with a thicker connector line, as this is by far the greatest correlation. Clearly, we can anticipate some difficulty in distinguishing between the effects of these explanatory variables. This, as we shall see, limits our ability to draw conclusions about the relative strength of the explanatory variables in the production model.

7 ITERATIVE MODELLING

7.1 A one-level model of the SACMEQ data

In this section, the theory and guidelines for the one-level regression model dealt with in previous sections are brought to bear on a policy-oriented analysis of the South African SACMEQ data. The aim is partly to validate and perhaps qualify the theory and guidelines, and partly to arrive at conclusions that can answer key policy questions.

The mental model of school production used in the analysis is the policy-oriented one presented in section 4.5 above. The new set of variables constructed from the original SACMEQ variables is used, and the reading score, the mathematics score, and an average of the two are used as the dependent variable. In view of the duality of South Africa's schooling system (discussed in section 3 above), separate models for the historically advantaged and disadvantaged are constructed.

Tables 17, 18 and 19 provide key statistical outputs from three regression analyses that involved the use of all the South African observations, which were weighted. Variables were selected through an 'intelligent stepwise' approach, meaning all the new variables were inserted, and then variables were rejected on the basis of the unadjusted slope coefficients, the standardised slope coefficients (or the beta coefficients), the *t* statistic and some mental model considerations (in particular the need to model the effect of inputs over which the government has some leverage). Standardised slope coefficients were regarded as important in assessing the feasibility of the input adjustments required in order to improve the scores. This use of standardised slope coefficients is followed, for instance, by Hungi (2004: 6). The *t* statistic was chosen as the preferred indicator of variable significance due to its compactness and simplicity. Any explanatory variable with a *t* value of less than 2 was automatically rejected. The level of significance associated with the *t* statistic was 0.05. Even variables with a *t* value higher than 2 were rejected for various reasons. For example, equivalent variables from the mathematics and reading teacher questionnaires were never both retained in the model. The possibility was however contemplated that a mathematics teacher variable could be a stronger predictor of the reading score than the reading teacher variable, and vice versa, but the significances never clearly indicated this was a reality so, for instance, the mathematics score appears in a model with mathematics explanatory variables only. School averages of the explanatory variables from the learner questionnaire were constructed to investigate the possibility that the compositional effect, or the peer effect, was stronger than the individual effect, and in some cases this was indeed the case. It should be kept in mind that the warnings about better significance statistics emerging from school-level regression models, discussed in section 4.2 above, are not applicable here, as the dependent variable was the score of the individual learner, not the average school score. What was important when comparing the strength of the learner-level variable to its school-level equivalent, however, was to focus on the *t* statistics and not the standardised slope coefficients, as the latter would not be comparable in this situation. The averaging effect would necessarily reduce the variance of the school-level variable, and hence make the standardised slope coefficient of this variable larger.

In the next three tables, variables are grouped according to the level of the variable, and under each level heading (for instance *learner vars* for the learner level), variables are sorted in descending order according to the standardised slope coefficient (see heading *beta coeff.*).

Table 17: The 'best' reading model (Stata output)[1]

R^2=0.631	n=3045	F=323	level=0.95
dependent var: read_score			
	coefficient	*beta coeff.*	*t stat*
learner vars			
learner_ses	0.66	0.131	8.0
learner_age	-1.12	-0.104	-8.8
daily_meals	1.11	0.047	4.1
textbooks_read	3.63	0.038	3.2
educator vars			
yrs_preserv_read	3.07	0.089	5.9
hrs_year_read	0.00	-0.071	-5.4
teacher_eval_read	0.35	0.065	5.1
class_meth_read	0.32	0.041	3.3
par_involve_read	0.21	0.034	2.7
day_inserv_read	0.01	0.023	2.1
school vars			
school_infra	1.23	0.212	9.1
teacher_disc	-5.04	-0.106	-7.0
ruralness	1.07	0.053	2.6
learner vars (school mean used)			
repetition	-6.21	-0.151	-10.3
parent_educ	0.58	0.112	6.5
learner_gender	14.20	0.091	7.9
_cons	-15.34		-1.9

Excluded: teacher_ses_read, class_size2_read, yrs_preserv_prin, prin_teach_load, par_involve_read, dist_support.

[1] The SACMEQ raw scores, and not the derived scores with mean 500 for the whole SACMEQ programme, were used in this analysis.

Table 18: The 'best' mathematics model (Stata output)

R^2=0.550	n=3005	F=244	level=0.95
dependent var: math_score			
	coefficient	*beta coeff.*	*t stat*
learner vars			
parent_educ	0.18	0.091	6.1
learner_ses	0.23	0.079	4.4
learner_age	-0.27	-0.042	-3.2
educator vars			
yrs_preserv_math	2.69	0.132	7.5
class_meth_math	0.50	0.104	6.5
teacher_eval_math	0.24	0.069	4.9
day_inserv_math	-0.01	-0.029	-2.3
school vars			
teacher_disc	-5.12	-0.184	-10.2
school_infra	0.51	0.150	7.5
dist_support	-0.05	-0.076	-5.1
prin_teach_load	0.09	0.060	4.4
learner vars (school mean used)			
repetition	-4.24	-0.175	-11.0
daily_meals	1.96	0.076	5.2
learner_gender	4.20	0.045	3.6
textbooks_math	2.51	0.036	2.7
_cons	-23.76		-4.2

Excluded: teacher_ses_math, class_size2_math, hrs_year_math, yrs_preserv_prin, prin_teach_load, par_involve_math, ruralness.

Table 19: The 'best' mean score model (Stata output)

R^2=0.660	n=2968	F=383	level=0.95
dependent var: mean_score			
	coefficient	*beta coeff.*	*t stat*
learner vars			
learner_ses	0.48	0.126	8.0
learner_age	-0.84	-0.103	-8.4
daily_meals	0.92	0.052	4.6
educator vars			
yrs_preserv_math	3.71	0.141	9.0
class_meth_math	0.71	0.116	8.1
teacher_eval_read	0.22	0.053	4.3
day_inserv_read	0.02	0.045	4.1
school vars			
school_infra	0.78	0.177	10.0
teacher_disc	-4.37	-0.122	-7.9
dist_support	-0.05	-0.048	-3.8
learner vars (school mean used)			
repetition	-5.28	-0.170	-12.0
parent_educ	0.58	0.149	8.0
learner_gender	9.53	0.080	7.0
learner_age	1.31	0.077	4.6
textbooks_math	4.00	0.045	3.7
_cons	-49.50		-7.1

Excluded: teacher_ses, class_size2, hrs_year, yrs_preserv_prin, prin_teach_load, par_involve, ruralness.

The next two tables show the results of a segmentation process according to historical advantage, with the output being the reading and mathematics scores respectively. As the previous table, which reports on a model using the mean of the mathematics and reading scores as the output, was not deemed to add any substantial analytical insights, it was not repeated for the HD and HA segments. The school average of learner SES was used to divide the 20% most advantaged learners from the remaining 80% of learners. Weighted learners were used in this segmentation, and whole schools were placed in the one or the other segment. The cut-off used for the average SES value was 8, which corresponded to a dip in the frequency distribution of these values separating the smaller advantaged curve from the larger historically disadvantaged curve (see discussion in section 3 on the non-normal distribution of values for South Africa). The point should be made that demographically, the socio-economically most advantaged 20% would by 2000 comprise a mix of races. White learners would comprise only slightly more than a quarter of these learners (in 2000 only around 6% of Grade 6 learners in the country were white), whilst Africans would comprise more or less half of these learners (author's own querying of the Annual Survey of Schools database). In the segmented models, the standardised slope coefficients, whilst allowing for comparison of variables within one segment model, are not useful for comparisons across the segments, due to there being different variances in different segments. Hence this coefficient was replaced by the coefficient of variation, which is important in that it allows us to assess why the strength and significance of the explanatory variables differ across segments. Essentially, strength or significance could be low because although the level of the input varies within a segment, it does not make a difference to performance (this would result in a high coefficient of variation and a low slope coefficient and a low *t* statistic), or because the magnitude of the input does not vary across schools within the segment (all three statistics would then have low values).

Segment models were also run for each of five performance quintiles, as in table 9 above. It was decided not to include these outputs in the present analysis due the very low significance values obtained in those five segment models. We can expect lower significance values if we segment a sample that is already as small as the SACMEQ sample, and if each of the segments is relatively homogenous. The *t* statistics in table 20 are both lower than the corresponding *t* statistics of table 17 in eleven out of sixteen cases. This loss in model significance is greater if we segment the overall model into five, as opposed to just two, segments.

Table 20: Reading score model by historical disadvantage (Stata output)

	Historically disadvantaged			*Historically advantaged*		
level=0.95	*R^2=0.325*	*n=2514*	*F=75*	*R^2=0.483*	*n=531*	*F=30*
dependent var: read_score						
	coefficient	*c.v.*	*t stat*	*coefficient*	*c.v.*	*t stat*
learner vars						
learner_ses	0.54	0.54	6.3	1.52	0.03	5.4
learner_age	-1.07	0.02	-8.3	-2.93	0.00	-4.9
daily_meals	0.91	0.10	3.4	3.79	0.02	3.3
textbooks_read	4.86	0.24	4.1	-0.46	0.08	-0.1
educator vars						
yrs_preserv_read	1.96	0.00	3.5	2.80	0.00	1.0
hrs_year_read	0.00	0.27	-5.8	0.00	0.07	-0.5
teacher_eval_read	0.26	2.88	2.8	0.43	0.50	2.8
class_meth_read	0.55	0.16	5.0	-1.07	0.07	-3.6
par_involve_read	0.17	0.86	2.0	-0.33	0.18	-1.1
day_inserv_read	0.01	3.12	2.1	0.01	1.60	0.5
school vars						
school_infra	1.17	0.62	8.0	-0.01	0.03	0.0
teacher_disc	-1.22	0.02	-0.8	-6.97	1.80	-4.5
ruralness	1.54	0.22	3.3	1.36	0.02	1.0
learner vars (school mean used)						
repetition	-5.22	0.24	-8.5	-2.61	1.09	-0.5
parent_educ	0.33	0.06	3.3	2.17	0.01	5.6
learner_gender	11.33	0.04	5.7	39.12	0.04	5.7
_cons	-0.25		0.0	-24.40		-0.5

Excluded: teacher_ses_read, class_size2_read, yrs_preserv_prin, prin_teach_load, par_involve_read, dist_support.

Table 21: Mathematics score model by historical disadvantage (Stata output)

	Historically disadvantaged			Historically advantaged		
level=0.95	*R^2=0.163*	*n=2479*	*F=32*	*R^2=0.431*	*n=526*	*F=26*
dependent var: math_score						
	coefficient	*c.v.*	*t stat*	*coefficient*	*c.v.*	*t stat*
learner vars						
parent_educ	0.09	0.18	3.4	0.42	0.03	3.4
learner_ses	0.08	0.54	1.6	1.41	0.03	5.6
learner_age	-0.20	0.02	-2.8	-1.44	0.00	-2.8
educator vars						
yrs_preserv_math	1.03	0.00	3.1	4.44	0.00	1.9
class_meth_math	0.35	0.17	5.0	0.32	0.06	1.0
teacher_eval_math	0.25	4.07	4.2	0.08	0.98	0.6
day_inserv_math	-0.01	4.01	-1.5	0.00	1.32	-0.1
school vars						
teacher_disc	2.10	0.02	2.6	-5.61	1.80	-4.5
school_infra	0.44	0.62	7.3	-0.38	0.03	-0.8
dist_support	-0.02	1.08	-1.8	-0.03	0.57	-1.2
prin_teach_load	0.00	0.61	0.2	0.47	1.19	5.6
learner vars (school mean used)						
repetition	-3.17	0.24	-9.2	0.49	1.09	0.1
daily_meals	1.12	0.03	3.5	24.71	0.00	4.5
learner_gender	3.26	0.04	2.9	-7.51	0.04	-1.5
textbooks_math	3.52	0.18	3.8	3.36	0.11	1.2
_cons	-2.84		-0.5	-103.30		-3.0

Excluded: teacher_ses_math, class_size2_math, hrs_year_math, yrs_preserv_prin, prin_teach_load, par_involve_math, ruralness.

It should be emphasised that model outputs often yield strange statistics, so we cannot take everything we see at face value, even where significance levels are high (Willms and Somers 2001: 433). Therefore the approach taken in the interpretation that follows is to regard as significant for policy purposes only those patterns that are repeated across at least two of the models appearing above, and preferably across all of them.

We begin with the policy areas relating to educators. Of the educator level variables, the variables referring to years of pre-service education and training have the highest beta coefficients in the reading and mathematics general (that is, unsegmented) models. In the segment models dealing with historically advantaged (HA) schools, the high levels of the unstandardised slope coefficients are more or less maintained though the values are lower for the historically disadvantaged (HD) than the historically advantaged, especially with regard to mathematics, suggesting that years of pre-service training makes less of a difference to the scores in the historically disadvantaged schools. The unstandardised slope coefficients are telling us that an extra year of pre-service training raises the mean reading score by between 2 and 3 additional points, and the mean mathematics score by beween 1 and perhaps as much as 4 additional points. We need to take cognisance of the country's apartheid history in interpreting the pre-service training variables. If we focus on just mathematics, we find that in the HD segment, 79% of the system has 15 years of education and training, 16% has 14 years and 5% has 16 years. In the HA segment, 70% has 16 years and 30% has 15 years – 14 years of training does not exist. (The percentages here and in much of this section refer to the percentage of the *system* in the sense of weighted learners. Clearly, this is not the same as the percentage of educators, as educators with smaller classes would be weighted less than educators with larger classes. The approach taken here has the advantage of making the percentages comparable throughout the discussion, but it would be important to keep in mind that arriving at percentages of educators or percentages of schools would require an adjustment.) The HA and HD segments are substantially different in terms of years of training, but the figures mask an even greater disparity, namely that the apartheid teacher training system provided qualitatively different training to different race groups, so the 15 years in each of the two segments would not represent the same quality of training. Hence it should not surprise us that an additional year of training makes a larger difference in the HA segment than in the HD segment, judging from the slope coefficients in tables 20 and 21. We can in fact adapt the general mathematics model from table 18 so that these qualitative differences are modelled within one unsegmented model. This is done by creating separate variables for the pre-service training of educators in HA and HD schools. The slope coefficient for advantaged pre-service training becomes 1.9, whilst it is 1.6 for disadvantaged. This should be compared to the general slope coefficient of 2.7 obtained in table 18 – we would expect this high value as that slope of 2.7 was capturing much of the difference between the two segments. We can use the new slope coefficients we obtained (plus similar ones obtained for the reading model) to simulate what would happen if quantitatively and qualitatively the pre-service training levels of teachers in HD schools were raised to the level of the HA schools.

Table 22: Simulation of teacher training improvement

	Mathematics		*Reading*	
	HD	*HA*	*HD*	*HA*
Existing mean score	19.3	35.8	33.1	61.4
Existing pre-service mean	14.9	15.8	14.9	15.7
New pre-service mean	15.8	15.8	15.7	15.7
Recalculated pre-service β	1.60	1.90	2.42	2.67
New mean score	25.5	35.8	38.9	61.4

The new mean score for disadvantaged schools is the result of, firstly, a downward adjustment to remove the effect of existing pre-service training and, secondly, an upward adjustment that both equalises the number of pre-service years to that of the advantaged segment, and equalises the pre-service training qualitatively by using the recalculated β slope coefficient for the advantaged sector. Formally, the calculation is:

$$\overline{Y}_{HD2} = \overline{Y}_{HD1} - \beta_{HD}\overline{X}_{HD1} + \beta_{HA}\overline{X}_{HA1} \qquad (66)$$

where $\overline{Y}_{HD1}$ represents the current mean score for HD schools (in mathematics or reading), β represents the recalculated slope coefficients for HD and HA schools, and $\overline{X}$ refers to the average level of training for the HA and HD segments.

The resultant mean scores for the HD segment are substantially higher, at least 5 points higher. In percentage terms, we see an average increase in the mean scores of around 25% for the HD part of system (it is important to consider the percentage rise given that the baseline reading scores are higher than the mathematics scores). This percentage increase, and the hypothetical policy intervention associated with it, are entered in table 28 at the end of this section. This table also reflects the anticipated performance impact of a number of other hypothetical policy interventions discussed in this section. The aim of that table is not to capture the impact beyond a particular quantitative threshold, but rather to capture the impact of a variety of interventions that are commonly believed to be important. Hence it is possible that of two interventions with the same impact receiving attention in the discussion, only one would be entered in table 28.

The simulation emphasises the importance of teacher quality and training in bringing about changes. However, the inequalities are great even after the adjustment. Teacher training can narrow the gap substantially, but most of the gap would remain even after a very radical (and, in the short to medium term, practically impossible) upgrading of teachers. Clearly, many other factors exert a strong influence.

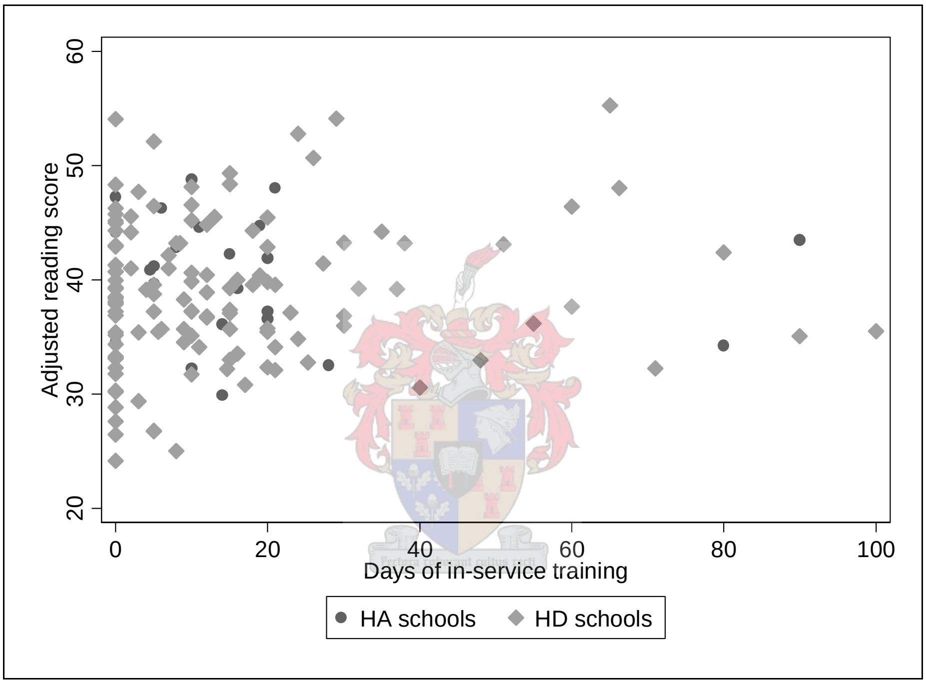
A far more modest simulation was run where educators with only 14 years of education and training behind them were upgraded to 15 years, using the slope coefficient for the HD portion of the system. This more realistic intervention, affecting around 15% of the system where educators have less than the 15 years, yielded an improvement for the HD schools of not 25%, but 1%. The range of effects on performance if we upgrade teachers is obviously very large, and this depends on how ambitious the teacher upgrading is. A third and intermediate intervention was simulated where the training levels of those educators in the half of the system with the greatest training deficit were raised by the equivalent of one year of pre-service training, of the quality received by educators from the HA schools. The results from this intervention were entered in table 28.

Given the potential impact of improvements in the human capital of educators, and the impossibility of taking all educators through pre-service training, in-service training should be a key concern. Days of in-service training received, captured in the variable *day_inserv* is significant enough to be retained in all the three general models. However, its impact is the opposite for reading and mathematics – more days of training are associated with better scores in reading and fewer days of training is associated with better scores in the mathematics models. The segmented models weakly agree with the general models in this regard. We can view this as the outcome of two opposing effects. On the one hand, more in-service training should improve scores. On the other hand, there is a vital selection effect whereby the state targets more needy teachers, in other words those achieving lower results, with more in-service training. We should thus not be surprised to see a negative relationship. A graph following the structure of figure 6 from a previous section was constructed in order to examine whether the treatment and selection effects could be separated out through a visual inspection of the distributions. To construct the graph, it was necessary to recalculate the score (the reading score was chosen), controlling for the effects of the explanatory variables other than in-service training. This was done by using the statistical outputs from the general reading model of table 17. Each learner's score was taken to be the overall mean of the reading score, plus the predicted effect of the in-service training, plus the error term for each learner emerging from the table 17 model (the latter statistic is given by the Stata command predict). This has the effect of making the learners comparable to each other. The equation is as follows (more decimals appear in the slope coefficient for in-service training than is reported in table 17):

$$Y_i = \overline{Y} + .0113295 * X_i + \hat{u}_i \qquad (67)$$

The distribution is as follows:

Figure 11: Days of in-service training and reading scores



Source: SACMEQ, 2000.

There is clearly no neat pattern in the above graph that allows us to discern the treatment and selection effects separately, in line with the idealised pattern of figure 6. In fact, figure 6 is almost exclusively heuristic. It would be very rare indeed to actually see such a pattern in a graph based on data from the real world.

We must probably resign ourselves to the fact that the data do not allow us to establish a net impact of days of training on performance, as the selection and treatment effects cannot be disentangled. To succeed in disentangling the two, we would probably need questions relating to whether the in-service training is provided by the state, or initiated by the school or teachers. This would at least allow us to measure the impact of training that is not subject to some deliberate targeting driven by the state. Moreover, given the importance of training, maybe some more questions on the type of training received would be required.

But even if we cannot establish a net production function for in-service training, the data allow us to draw some general conclusions about the effectiveness of in-service training in two respects: firstly, with respect to optimality of targeting, and secondly, with respect to the perceptions of teachers regarding the training.

Table 23: In-service training recipients and mean scores

	Mathematics		*Reading*	
	HD	*HA*	*HD*	*HA*
No training	18.6 (28)	35.2 (29)	32.3 (30)	58.8 (25)
Up to 20 days	19.9 (48)	36.2 (62)	33.6 (43)	63.7 (58)
Over 20 days	18.7 (24)	34.0 (8)	33.5 (27)	55.5 (16)

In the above table, the mean scores associated with various in-service training categories are provided – the proportion of the segment appears in brackets. The training days reported are for the preceding three years. Whilst a large proportion, over two-thirds, of the HD segment has received some in-service training, the figures suggest that the targeting is not fully effective in the sense that those excluded altogether have a particularly low mean score. If anyone is left out, it should be those with higher scores. The figures indicate that there is definitely a targeting problem, and perhaps that in-service training causes better scores. But we cannot be sure of the latter, because, as argued earlier, we are unable to disentangle the selection effects from the training (or treatment) effects. We are not sure whether those who receive training produce better results because of the training, or because in-service training programmes cover more advantaged, and better performing, parts of the HD segment better. At the high end of training days received, we see an interesting phenomenon. In all four columns, receiving more than 20 days of training is associated with slightly lower scores (relative to receiving up to 20 days of training). Again, this could be a selection effect, meaning educators with the poorest results in HA and HD schools are targeted for more intense training. However, this could also point towards a third effect, namely the negative impact of large amounts of training time on contact time between educators and learners.

The next table deals with educators' perceptions of the effectiveness of the in-service training received (where such training was received).

Table 24: In-service training satisfaction and mean scores

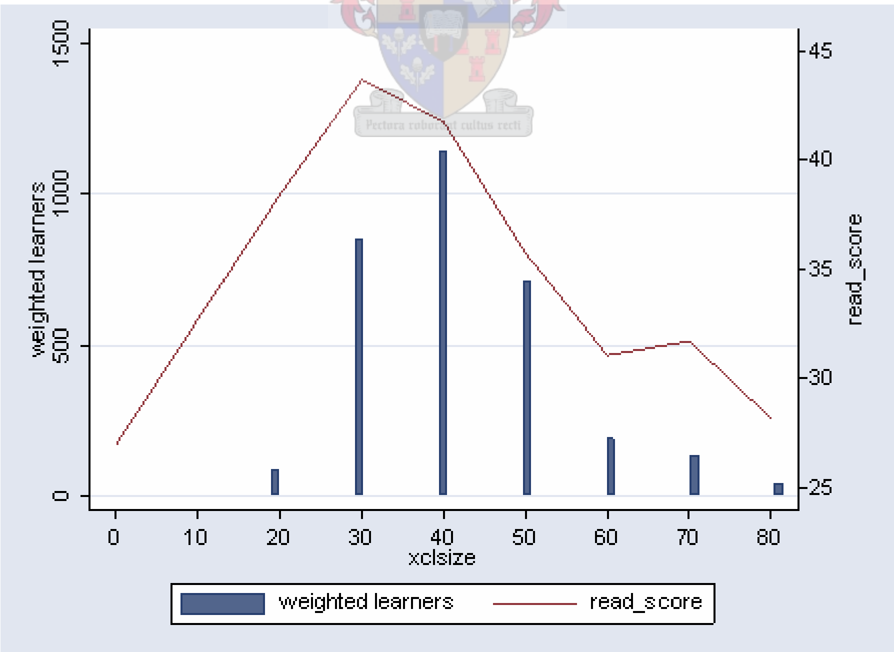
	Mathematics		*Reading*	
	HD	*HA*	*HD*	*HA*
not effective	27.6 (1)	36.3 (16)	27.7 (2)	65.1 (14)
reasonably effective	20.1 (44)	35.7 (53)	34.2 (48)	62.4 (59)
effective	19.0 (33)	35.8 (31)	35.1 (30)	56.6 (22)
very effective	18.4 (22)	(0)	30.8 (21)	62.1 (5)

Very few educators from HD schools classified the training received as 'not effective'(percentages of the segment appear in brackets). This seems important and good, even if substantial numbers of educators from advantaged schools are complaining that the training they receive is not effective. However, there is a pattern for educators, both from HA and HD schools, to view the training received more negatively, the better their learners' scores. There are two possible interpretations. Educators could be receiving training which, whilst good, is geared at too low a level. In other words, it is not effective for their level of competency. Alternatively, the training could, on average, be of a poor quality, and better performing educators, who are likely to be better judges of good and bad training, are better at detecting the true quality of the training. It does not seem as if we can gauge the relative correctness of the two interpretations, given the data we have. What is missing in the questionnaires is a question that explicitly asks educators why they do not think the training is effective. We would need a question eliciting responses such as 'the training is designed for teachers who have more serious problems than I have' and 'I do not think the training courses are of a good quality'. It would also be useful to have questions where educators gauge the value of in-service training relative to other teacher upgrading activities, for instance individual studies by correspondence, in-house professional meetings organised by the educators themselves, or the school principal, and television and radio broadcasts dealing with pedagogic matters. But even if our conclusion is that either training is being offered at a level that is too low (resulting in close to half of educators from HD schools finding the training to be below their level) or that the training being offered is of a poor quality, overall there appears to be a problem with the training that is being offered, and some policy revision is required.

The teacher evaluation variable, *teacher_eval*, is sufficiently significant to be retained in all three of the general models. It also appears as significant in the models for HA and HD schools, though the mathematics teacher evaluation variable is not significant for the HA segment. This variable takes on a value between 0 and 10, depending on the effectiveness of professional support coming from the school principal – once a year and once a term advisory meetings were weighted more than more frequent meetings, given that the former were clearly associated with better scores. Across 5% of the HD segment, the once a year advisory meeting option (the option associated with the best performance) is followed, whilst in around 25% of the HD segment the slightly less effective option of a meeting once per term is pursued. In HA schools, the corresponding figures are 21% and 9%. The finding that less frequent meetings are the best option for learner performance should clearly not be taken without some interrogation. For instance, it would be important to examine the degree to which the finding relates to the particular application of specific programmes, for instance the Developmental Appraisal System, a national teacher appraisal system that was first piloted in the mid-1990s. It could be the type of evaluation, as opposed to the frequency of evaluations, that is the key matter. In any event, the finding that somehow the frequency of evaluations by the school principal is associated with performance improvements is clearly supported by the model. If the practice in all schools were to be adjusted so that the best practice were applied across the system, we would, using a conservative slope coefficient for *teacher_eval* of 0.25 (as in the general mathematics model), achieve an increase to the scores of around 8%. The expected increase in the mean reading or mathematics score, in terms of points, is simply the increase in the mean value of *teacher_eval* multiplied by 0.25. Raising the impact of evaluations in HD schools to that in HA schools would improve the scores in HD schools by about 4%.

From a policy perspective, it can be just as important to note the non-significance of certain variables as the significance of other variables. Two educator input variables that were not retained in any of the general models due to lacking significance are the ones describing class size and the SES of educators (used as a proxy for teacher salary). Though they display significant associations with learner performance when considered on their own, their net effect in a multivariate model is negligible. This finding with respect to class size would run counter to what is generally understood about what makes a class work. The following graph indicates, firstly, that very large classes are fairly prevalent in the system (19% of learners are in reading classes that have more than 50 learners) and, secondly, that in a bivariate analysis, larger classes are certainly associated with lower scores (class size has been rounded off to the nearest 10 to make the graph more illustrative). Should we, on the basis of the models we constructed, conclude that reducing class size should not be a policy priority? It would certainly be irresponsible on the part of the analyst to reject such a measure in any hard way. There are at least two reasons for this. Firstly, the relatively strong correlation between class size and school infrastructure (see table 38) means that the effects of class size may not be correctly disentangled from the web of variables reflected in figure 10. Secondly, this is an instance where some common sense could serve the analyst and his overall package of policy advice well. School classrooms in South Africa are not designed to accommodate classes as large as 60 learners, and South African educators say they regard large classes as an impediment to effective teaching. Given this physical limitation, and this perception, it is very unlikely that over-sized classes are not negatively affecting the education production process.

Figure 12: Class size and reading scores



Source: SACMEQ, 2000.

Educator salaries in South Africa have been greatly equalised since 1994. Hence one may not expect educator SES differences to be strongly correlated to learner performance differences. The data in fact reflects a teacher corps that is not poor: 51% have cars, over 90% have a refrigerator and 92% have a television. Only 4% have none of these three items in their home. The data does not support the hypothesis that low salaries in segments of the teaching force cause poor learner performance.

Turning to curriculum concerns, variables were constructed for three of the four policy areas in the policy-oriented mental model. Variable *class_meth* captures the effectiveness of classroom methodology, *hrs_year* captures the total contact hours in the school year, and *repetition* captures the number of years the Grade 6 learners in the sample have repeated a grade. Whilst this last variable may be subject to practices very specific to individual schools or localities, it is clearly also subject to system-wide rules dealing with promotion from one grade to the next, in other words the national curriculum.

The variable reflecting classroom methodology is a variable of average (positive) strength in the general unsegmented models, in terms of the standardised beta slope coefficients and the *t* statistics. In the segment models, the positive association is clearly retained in the HD segment, but not in the HA segment. We can probably regard the negative association in the HA segment with respect to the reading score as an anomaly we need not concern ourselves with. As explained in section 6, the mathematics methodology variable places the most value on allowing learners to work on their own, interacting on a one-to-one basis with individual learners, assigning homework, and getting parents to sign homework books. The reading methodology variables places the most value on promoting listening skills and having parents sign homework books.

If we simulate a raising of the mean value of the classroom methodology variable in HD schools so that it equals that of HA schools, we obtain a much greater impact on the mathematics scores than on the reading scores (using the different classroom methodology variables dealing with the mathematics and reading teachers respectively). There are two reasons for this. Firstly, the slope coefficient for mathematics is larger, implying that improving classroom practice along the 0 to 10 scale we have constructed has a particularly positive impact on the scores. Secondly, the difference between the HA and HD schools with respect to the classroom methodology values is greater when it comes to mathematics than when it comes to reading. The simulation yields a 7% improvement for the mathematics scores of HD schools and a 2% improvement for the reading scores of HD schools. It should be kept in mind that the impact is net of other, related factors, such as the training background of educators. What the model is saying is that even if we do not improve the training levels of educators, there is considerable scope for improving performance, especially on the mathematics side. Clearly, training is important, but an improvement in scores using the existing human and physical capital is certainly possible, on the basis of what we observe in the current system.

The contact time variable, *hrs_year*, is retained in the general reading model as a negative explanatory variable. It is a particularly strong variable in terms of the standardised slope coefficient, but it is dropped in the mathematics and the mean score general models, due to lacking significance. The variable appears as a significant predictor for HD schools, and it is a negative predictor, but it loses its significance in the HA segment. This variable is constructed from information in the teacher questionnaire – periods taught was multiplied by the duration per period – and the school principal questionnaire – days lost to, for instance, bad weather, were deducted. The fact that there should be a negative association between this contact time variable and learner performance (in the case of the reading model) is surprising. In the bivariate analyses, the variable correlated positively with performance, on both the reading and mathematics sides. The negative association seems particularly surprising given the great range of values in the variable – as shown in table 37, the values for the 25th and 75th percentiles were 581 hours and 1,000 hours respectively. It could be that the assumption, put forward in section 6, that each educator would be a full-time class teacher responsible for all the teaching time during the year of the class, is incorrect, and that some schools are in fact assigning a mix of educators to each class. The variable was reconstructed at a learner level, using the *hrs_year* values as already explained, and subtracting estimated number of days absent per learner, from a question where the learner specifies how many days she was absent during the preceding month (the monthly figure was multiplied by ten to make it correspond roughly to the school year). This did not solve the puzzle, and the results remained more or less unchanged. Should we regard the result from our model as an anomaly that is best ignored? It seems so, partly because the result is so counter-intuitive, and partly because it arises only in the reading model, and not the mathematics or mean score models. What we can conclude, however, is that as far as we can tell from the available data, there is no reason to believe that some schools are performing better than others due to having more contact time.

The variable *repetition* has a net association with learner performance that is more significant than that of any other explanatory variable. This is confirmed in all the three general models, and in the models for the HD segments in tables 20 and 21. The association is always negative, meaning more repetition is associated with lower learner performance. Moreover, the variable is more significant when the school-level average is used, than when the learner-level value is used. In other words, compositional effects appear stronger than individual effects. The slope coefficient is around 5, so if on average all learners repeat an extra year, the score drops by 5 points. If half of the learners repeat an extra year, then the score drops by 2.5, and so on. The following table indicates the distribution of the repeater values relative to the reading scores.

Table 25: Distribution of repetition by reading score deciles

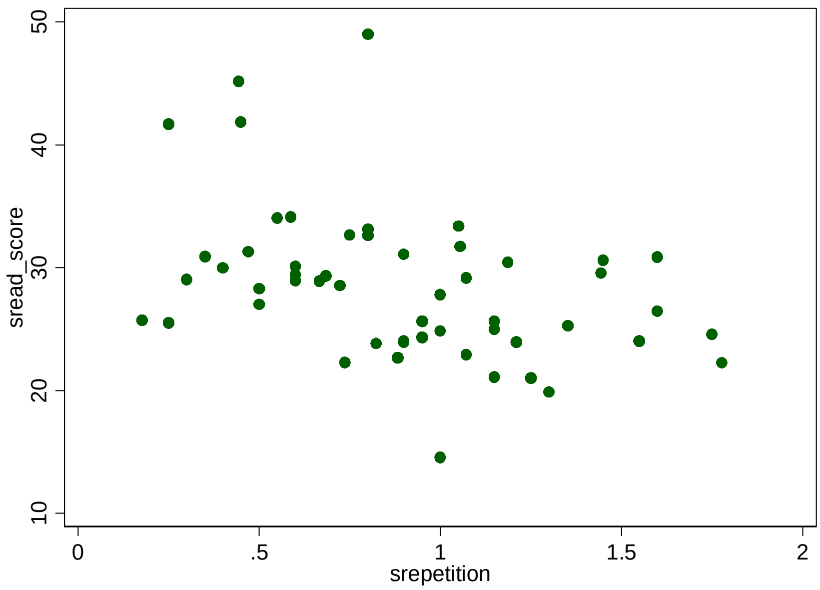
Reading score decile (school mean used)	*Mean repetition overall*	*% of learners with zero repetition*	*Mean repetition of repeating learners*	*Dropping out problem (1=never; 2=sometimes; 3=often)*
1	1.1	35%	1.7	2.1
2	0.9	44%	1.6	2.1
3	0.7	50%	1.4	2.1
4	0.8	48%	1.6	2.2
5	0.8	48%	1.5	2.0
6	0.6	52%	1.3	2.2
7	0.6	59%	1.4	2.2
8	0.4	69%	1.3	1.9
9	0.2	83%	1.2	1.6
10	0.2	89%	1.4	1.9

What stands out is that in the worst performing schools, two-thirds of learners have repeated at some stage, and the average for them is 1.7 years, whilst in the best performing schools, only 11% of learners have ever repeated, and those learners have repeated less (1.4 years on average). The pattern is fairly constant across all the performance deciles, though strongest at the very bottom end (deciles 1 to 3) and in the best performing half of the system. The question is what would have happened to the performance scores if no learners, or a very small margin of learners, had repeated. The impact of grade repetition is one of the more complex issues in the planning of education systems, and the management of schools. Whilst many parents and educators believe that repetition, even repetition of the magnitude represented in the previous table, is a necessary aspect of quality schooling, the received wisdom that has galvanised over the last decades indicates that repetition is both economically inefficient and pedagogically unsound. There is almost certainly some entanglement of opposing effects in the statistics. We cannot, on the basis of what we see in table 25, conclude that repetition is only associated with a negative impact on scores. There could well be an optimal point which allows for some repetition. Promoting all learners all the time is not necessarily the optimal point. Some studies have shown that especially in developing countries, a degree of repetition is efficient (King, Orazem and Paterno, 1999). The implication is that the treatment effect of repetition is non-linear, which obviously makes the prospect even more remote that one would be able to separate out the treatment and selection effects.

The last column of the previous table indicates the mean values for the principal's response regarding the problem of dropping out in the school. What stands out is how little the mean deviates from the overall mean of about 2.0. Clearly, across the range of schools dropping out is regarded as a problem occurring, on average, 'sometimes'. There is also not a very clear decrease in the problem, within the narrow band, as we move from worse performing schools to better performing schools.

The next graph illustrates the relationship between the mean reading score of schools and the mean years of repetition of learners in the socio-economically most disadvantaged quarter of the schooling system. The relationship is negative, though it is not strongly so. The graph indicates that it is very possible for a poor school to obtain good scores whilst keeping Grade 6 repetition low – see the schools on the top left of the graph. There are no schools with a high rate of repetition and high reading scores – such schools would have appeared on the top right.

Figure 13: Performance and repetition in poorest quartile of schools



Source: SACMEQ, 2000.

A maximum threshold for repetition of 0.5 years, beyond which repetition would begin to have a negative impact on average learner performance, was hypothesised, fairly arbitrarily, but to some degree based on the above graph. Using the slope coefficient from the general reading model, a simulation that took schools currently above the 0.5 threshold, down to this threshold, resulted in a 4% increase in the mean reading score. If we took all schools down to the average repeater level for HA schools, which is 0.17, the overall improvement in reading scores would be in the magnitude of 7%. A similar simulation with respect to mathematics yielded very similar results. These calculations do not take into account the important fact that each repeating learner represents a high opportunity cost – the system is missing the opportunity of smaller classes, or perhaps more expenditure on materials, flowing from the budgetary savings obtained through a reduction to the repeater rate, and hence a shortening of the years per learner in the schooling system. These expenditure dynamics will be explored further in section 8 below.

Turning to learning support materials (LSMs), we have only one variable, dealing with quantity of LSMs (it was not possible to construct a variable relating to the quality of the LSMs in the sense of the quality of the texts). The LSM quantity variable seems important. It is retained in all of the general models, and in the HD segment models it is clearly significant. It loses its significance in the HA segment. This is understandable, given that the new variable was constructed in such a way that all values greater than 0.5 textbooks per learner were truncated to 0.5 as beyond the 0.5 level, there was a diminishing impact on performance. Just under half of the learners across the system stated that they had access to their own textbook (in reading as well as in mathematics), so a large portion of the HD schools would not be under-resourced in terms of textbooks. A simulation was performed which ensured that each learner who had below 0.5 textbooks, was raised to the 0.5 level. The learner and not the school was used as the level of analysis, though similar results would be obtained using the school level (the general models indicate that both the individual and compositional effects are important). With regard to reading, the simulation meant raising the mean value of textbooks per learner from 0.39 to 0.50 (the maximum mean possible, given that the variable is capped at 0.50), and the resultant improvement to the reading score was around 1%. On the mathematics side, the mean value was raised from 0.34 to 0.50, resulting in a performance improvement of 2% overall (and 3% for just HD schools). This improvement strikes one as low, given the emphasis placed on textbooks in much of the literature (see discussion in section 4.3). However, much of that emphasis presupposes the introduction of textbooks into a situation in which there are no textbooks at all, as opposed to a raising of the textbook to learner ratio. Glewwe *et al* (2000) find an even less impressive impact of textbooks in their study, which focuses specifically on textbooks in a developing country context.

The physical infrastructure variable *school_infra* has a stronger association with performance in the general reading model than any other variable, in terms of the beta coefficient and the *t* statistic. The variable also comes out as important in the other general models, and in the models for the HD segment. The variable is clearly not an important one amongst HA schools. In the simulation, the more conservative slopes for the mathematics scores were used. In this conservative simulation, taking all the HD schools to the HA level in terms of physical infrastructure resulted in an increase of 13% and 15% for mathematics and reading respectively for HD schools. What should be kept in mind with regard to the *school_infra* variable, however, is the exceptionally high correlation between this variable and other variables, in particular *ruralness* (discussed in section 6 above). It seems possible that the great impact that school infrastructure appears to exert is in fact a product of other effects associated with life in rural areas, for example an inability to attract better teachers, long distances to schools, and so on. Nevertheless, the issue of school infrastructure seems to warrant much closer scrutiny than is offered here. The SACMEQ data should allow considerable further analysis, especially into the matter of what aspects of physical infrastructure seem to make a difference. (The variable *school_infra* combines availability of toilets, a school hall, a library, a staffroom, a storeroom, a photocopier, at least one computer, and a fence around the school.)

Crouch and Mabogoane (1998: 11) emphasise the importance of good management and 'more work in the sense of using more imagination, enthusiasm, and being more accountable' as important factors determining quality in post-apartheid schools. We would thus expect the management variables to come out as significant predictors of better performance in schools. This does not seem to be the case, however. That this should be so seems more a result of limitations in the SACMEQ data than limitations in Crouch and Mabogoane's conclusion. The variables relating to the principal's education and training background (*yrs_preserv_prin*), and to the teaching load of the principal (*prin_teach_load*), were not retained by the models. The variable relating to visits from the Department (*dist_support*) was retained, but it has such a weak slope (even when standardised), and the direction of the association is counter-intuitive (more Departmental visits are associated with lower performance) that it seemed unwise to read anything into this aspect of the model. The variable relating to parent involvement (*par_involve*) is retained in some of the general models, but the associations are weak. It should be pointed out that the parent involvement variable relates to contact with the teacher, and not to the involvement of parents in the governance of the school (though one may expect some link between the two types of involvement), underlining the limitations of the dataset with regard to school management information. If parent involvement (as measured in the questionnaires) were improved in HD schools so that it equalled the level in HA schools, the reading score would improve by 2% (there would be no impact on the mathematics score). As a result of the weakness of school management and parent involvement as explanatory variables in our models, these interventions are not reflected in table 28 below, which aims to capture only those interventions with a clearly substantial impact on performance.

The variable *ruralness* was retained in the reading models, but not the mathematics models. The partial non-retention of this variable, whilst the closely correlated variable *school_infra* is retained, is in fact a result of the finer calibration of *school_infra.* If *school_infra* is calibrated like *ruralness*, so it takes on the value of 1, 2 or 3, the ruralness variable is in fact slightly stronger in the general mathematics model than the school infrastructure variable, both in terms of the *t* statistic (both variables pass the 2-*t* rule of thumb) and in terms of the slope coefficient. This confirms the fact that school infrastructure and ruralness are closely intertwined. Even with the recalibration, however, we cannot be very sure of their separate effects due to the high level of multicollinearity. We can nevertheless regard it as likely that performance is affected adversely by the longer distances to school experienced in rural areas. As there are no variables dealing directly with scholar transport in the SACMEQ data, no simulations regarding this policy intervention are possible.

The variable relating to the number of daily meals eaten by learners, *daily_meals*, is retained in all the models. However, whilst it fitted in best as a learner-level variable in the general reading model, it worked best as a school-level compositional variable in the mathematics model. What is striking with regard to this variable is its strength in the HA segments in tables 20 and 21. A closer analysis of the data, however, reveals that this is a result of a few outliers in the HA segment. There may well be a problem with a minority (some 4%) of learners in the HA segment who are undernourished. The overall picture with regard to meals is as follows (here the original SACMEQ values are used, and not the slightly adjusted and condensed values of *daily_meals*).

Table 26: Distribution of meal values in HD schools (% of learners)

	Breakfast	*Lunch*	*Supper*
Not at all	11%	8%	5%
1 or 2 days per week	13%	14%	10%
3 or 4 days per week	9%	13%	7%
Every day of the week	67%	65%	78%

Table 27: Distribution of meal values in HA schools (% of learners)

	Breakfast	*Lunch*	*Supper*
Not at all	5%	2%	1%
1 or 2 days per week	7%	5%	1%
3 or 4 days per week	11%	9%	3%
Every day of the week	76%	84%	94%

Inadequate eating of lunch in HD schools should be a policy concern. This is the meal over which the schooling system has the most influence. In HD schools, this is the meal that is least eaten every day, whilst in HA schools, breakfast is the meal least eaten every day. In 2000 there was still clearly a need for further expansion of the country's School Nutrition Programme. If we simulate a situation in which the value for *daily_meal* assumes the maximum value 3 for all learners (currently 51% of learners enjoy this maximum level – 48% in HD schools and 66% in HA schools), we obtain an overall improvement in the mathematics and reading scores of around 2% and 1% respectively. Importantly, the average improvement per learner in HA schools would be around one-third of the improvement obtained in HD schools. This suggests that school meals should perhaps not be limited to HD schools only, but also that an inadequate diet amongst non-poor learners could be a problem (which in turn would imply a need not only for school meals, but ongoing education regarding the importance of nutrition).

The level of education of parents (*parent_educ*) is a prominent explanatory variable in all the general models, and in the HA and HD segments. The importance of parent education as a determinant of learner performance is emphasised in most of the economics of education literature, as discussed in section 2.2. If we simulated an improvement such that the level of the 20th percentile, from the bottom, of the variable *parent_educ* was made the minimum (in other words all parents below this level would be brought up to this level), we would obtain an overall improvement in both the mathematics and reading scores of around 1%. If we used the 40th percentile instead of the 20th percentile as our parent education standard, the improvement in scores would be around 3%. The simulation manipulated the school mean of *parent_educ*, meaning the approach of the mathematics models was followed (in the reading models, the variable fitted best at the learner level).

The variable *learner_ses*, dealing with the socio-economic level of the household apart from the parents' level of education, is retained in all the models, and is significant even within the HA segment. If we perform a simulation similar to the one we performed for *parent_educ*, but using the learner-level values, we obtain an overall improvement to the scores of less than 1% when we use the 20th percentile as our standard, and of around 1% when we use the 40th percentile as the standard. It is rather striking that the home background variables *parent_educ* and *learner_ses* should not exert a larger effect on learner performance, given the emphasis placed on home background variables in other studies. The overall effect of the home, relative to the effect of the school, will receive more detailed attention in the HLM discussions in the next section.

A variable, *teacher_disc*, that reflected the school principal's perception of the level of latecoming of educators appeared to be an indicator of some importance in the variable selection process. The variable is retained in all the models, and passes the 2-*t* rule of thumb in all the models except for the HD model dealing with the reading score. The variable is a 0-1 binary variable, where 1 indicates that the principal views teacher latecoming as a problem. The perceived problem arises in 96% of the HD system and in 36% of the HA system. If we use the slope coefficients from the general models and simulate the complete removal of the problem from all schools, the mathematics score rises by around 18% and the mathematics score by 11%. The apparently strong impact of this variable needs to interpreted with care, mainly because there are so few HD schools not experiencing the perceived problem, so being in the historically disadvantaged segment almost automatically goes together with experiencing the problem of educator latecoming. As a result, the variable *teacher_disc* could well be masking other differences between the HD and HA segments. Nevertheless, the fact that this variable should display a clearly negative association with performance when we consider the HA schools on their own suggests that the issue of teacher discipline has considerable importance independently of other school variables.

Table 28 summarises the estimated impact of the various hypothetical changes that have been discussed. The overall hypothetical effect is indicated at the bottom of the table. These overall figures are simply the sum of the various partial interventions. Because the *net* effect of each partial intervention is calculated, it is correct to simply add the parts to obtain the whole effect. As an example, the 62% improvement for the HD mathematics score would take the mean mathematics score for HD schools from the current 19.3 to 31.3, which is just over the value of the current 75th percentile *in the entire dataset.*

The overall improvements are interesting more from a theoretical point of view than from a practical point of view. Whilst the estimated impacts of the various partial interventions are fairly realistic, the overall estimates are not. This is because the coefficients from the regression models we have constructed are better for estimating the impact of improvements on the margin, in other words improvements resulting from moderate changes to the current system, than estimating the impact of a complete overhaul of the system. Specifically, our models do not take into account the common phenomenon of diminishing returns to scale discussed in section 2.1. Essentially, the more we intervene, the smaller the impact of each successive intervention effort. Importantly, it is more a matter of diminishing returns to *intervention effort* and not to interventions themselves. Implementing individual interventions in isolation from other interventions is neither possible nor desirable. For example, it is not realistic to imagine an intervention that introduces more textbooks into deprived classrooms that does not also involve the deliberate or non-deliberate improvement to classroom methodologies. To some extent, the latter is an inevitable outcome of the former. What is realistic is to imagine a simultaneous implementation of several interventions, with some more conscious weighting of certain interventions than others. It is the *total effort* that is typically subject to diminishing marginal returns.

Table 28: Summary of simulated interventions

Variable	*Hypothetical change*	*Approx. net effect on HD scores*	*Approx. net effect on overall scores*
yrs_preserv	Raise the training level of educators in the half of the system with the greatest deficit by the equivalent of one year of pre-service training.*	+5%	+3%
	Raise educator training of HD part of system in quantitative and qualitative terms to that of HA part of system.	+25%	+18%
teacher_eval	Raise the level of effectiveness of teacher evaluations by the principal in HD schools to that in HA schools.*	+4%	+3%
	Raise the teacher evaluation index to the maximum for the whole system.	+9%	+8%
class_meth_*math*	Raise the average classroom methodology indicator in HD schools to that of the HA schools *with respect to mathematics*.*	+7%	+5%
class_meth_*read*	Raise the average classroom methodology indicator in HD schools to that of the HA schools *with respect to reading*.*	+2%	+1%
repetition	Decrease the average learner years of repetition in the 61% of the system where schools exceed the 0.5 level, to 0.5.*	+6%	+4%
	Decrease the average learner years of repetition in the 89% of the system where schools exceed the average level for HA schools (0.17), to this HA level.	+12%	+8%
textbooks_*math*	Raise the average number of *mathematics* textbooks per learner so that each learner enjoys a ratio of at least 0.5 per learner.*	+3%	+2%
textbooks_*read*	Raise the average number of *reading* textbooks per learner so that each learner enjoys a ratio of at least 0.5 per learner.*	+2%	+1%
school_infra (N.B. closely correlated to ruralness)	Raise the level of physical infrastructure of all schools to the present average for HA schools.*	+14%	+10%
daily_meals	Raise the intake of daily meals so that all learners receive all their daily meals (currently some 51% of learners do).*	+3%	+2%
parent_educ	Raise the level of education of the least educated 20% of parents to the level of the 20th percentile.	+1%	+1%
	Raise the level of education of the least educated 40% of parents to the level of the 40th percentile.	+4%	+3%
learner_ses	Raise the SES of the least advantaged 40% of learners to the level of the 40th percentile.	+2%	+1%
teacher_disc	Remove the problem of perceived indiscipline of educators from all schools.*	+20%	+15%
OVERALL	Apply all of the school policy interventions marked * simultaneously *in the mathematics model* (reading variables excluded).	+62%	+42%
OVERALL	Apply all of the school policy interventions marked * simultaneously *in the reading model* (mathematics variables excluded).	+56%	+39%

Many observations and interpretations have been made in this section. The next section will explore some of the issues in greater depth, in an HLM context. A synthesis of the issues is presented in sections 8 (in particular where budget implications are involved) and in the conclusion in section 10.

7.2 A hierarchical linear model of the SACMEQ data

This section builds on the one-level analysis of the SACMEQ data presented in the previous section by using the hierarchical linear model (HLM) already introduced at a theoretical level in sections 5.4 to 5.6. The HLM will permit a better analysis of performance inequalities, and the effects of SES and school inputs on these inequalities, through the distinction made between within-school variance on the one hand, and between-school variance on the other. Moreover, interactions between effects at the learner level and effects at the school level will be studied in a way that was not possible in the one-level analysis. Part of the focus is on using the HLM to verify the magnitudes of the hypothesised interventions presented in table 28 above.

The aim is not to repeat all the steps taken in the one-level analysis of the previous section. For instance, no models focusing on just the HA segment of the schooling system are presented here. Nevertheless, there are enough points of comparison between this section and the previous section to draw key conclusions about the relative efficacy of the two models.

To a large extent, this section uses as a point of reference the HLM analyses contained in three different texts: Ferrão, Beltrão, Fernandes *et al* (2001) using 1999 SAEB data from Brazil; Willms and Somers (2001) using data from the Latin American Laboratorio programme; and Hungi (2005) using the 2000 Kenya SACMEQ data.

Ferrão *et al* (2001) construct an HLM consisting of two levels. On the basis of other texts dealing with the multi-level analysis of SAEB data, for instance Barbosa and Fernandes (2001), it seems as if the software used by Ferrão *et al* was MlwiN. Willms and Somers (2001: 415) employ a two-level model, and explain that they rejected a three-level model with country at level 3 due to the instability of the model. It is not clear what software Willms and Somers used. Hungi (2005) constructs a three-level model, the levels being the pupil, the school and the province. He uses Bryk and Raudenbush’s HLM.

The analysis that follows uses a two-level HLM, level 1 (L1) being the learner, and level 2 (L2) the school. In order to use the HLM software, an .mdm data file had to be created. Due to limitations in the HLM software, the variable names used previously had to be shortened to a maximum of eight characters. However, in the reporting that follows, the same variable names used in earlier sections are used for the sake of consistency. The only change effected to the values of the variables referred to in the one-level model of the previous section was the school averaging of educator-level variables. This change was necessary to allow educator effects to be counted as school effects in the HLM. This change had a minimal overall effect on the data as in 153 of the 169 schools there was a one-to-one correspondence between school and educator (with respect to the reading educator).

The variables containing school-level averages of learner values, for instance *slearner_ses* containing the school average for *learner_ses,* were used in the HLM analysis, as their inclusion in HLM models can have an important effect. (It may be tempting to believe that the HLM, by explicitly catering for groups, has no need for group average variables, but this is not the case.)

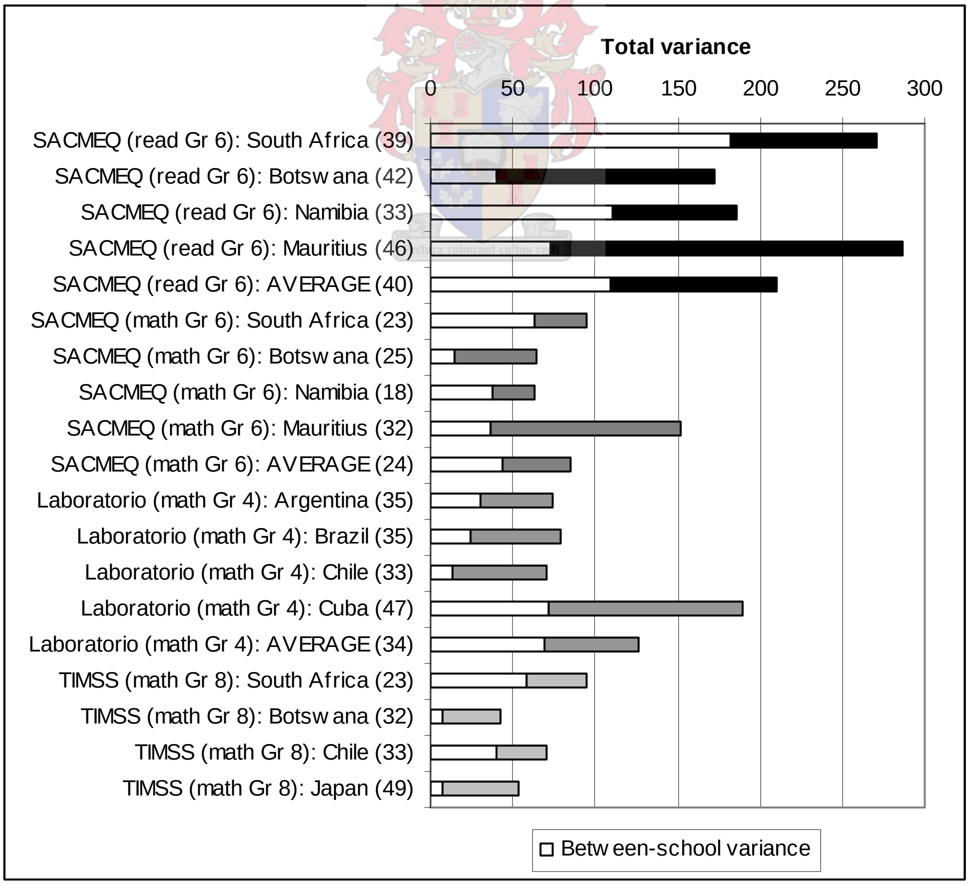
The HLM software does not handle missing data well. Whilst a more developed software package such as Stata will automatically exclude observations with missing data in the model variables from all output statistics, the HLM software always uses the total number of observations in the .mdm data file, even if they cannot be used in the model due to missing data, in determining the degrees of freedom. For this reason, it is safest to create a new .mdm data file for each new combination of variables, ensuring that each file contains no missing data. This approach was followed in the various analyses that follow.

A more serious problem with the HLM software is that when observations are weighted, no random effects, in other words variance statistics, are provided in the output. The reason for this is not explained in the user's guide. The approach was taken to use weights (the variable *pweight2* was used) whenever it was not important to know the random effects. This means that much of the analysis occurred with unweighted data.

Ferrão *et al* (2001), Willms and Somers (2001) and Hungi (2005) all begin their HLM analysis with an examination of the null model, or the model without any explanatory variables, in order to check the uncontrolled or unconditional partitioning of the overall variance across the two or (in the case of Hungi) three levels. A null model for reading has already been presented in table 13. The L1 and L2 variance statistics for the reading model are 88.1 and 182.7 respectively (figures given here differ slightly from the table 13 figures due to the fact that the latter model excludes two schools). The L1 and L2 variance statistics for the mathematics model are 31.7 and 63.2 respectively. These results are from unweighted data[2].

Of importance is both the amount of overall variance, and the partitioning of this variance across the two levels. To fully understand the statistics, it is necessary to compare the South Africa results to the SACMEQ results of other countries, or to non-SACMEQ results that have been rescaled to look like the SACMEQ results. Dolata, Ikeda and Murimba (2004) provide a useful graphical way of representing the statistics of several countries, and this approach is repeated here, with some alterations. SACMEQ 2000 and TIMSS 2003 Grade 8 data were modelled using the HLM software, and Laboratorio 1996 variance statistics as reported by Willms and Somers (2001: 419) were used. The mean score per country is indicated in brackets.

Figure 14: Variance partitioning in several countries



---

[2] The SACMEQ raw scores, and not the derived scores with mean 500 for the whole SACMEQ programme, were used in this analysis.

Sources: SACMEQ, 2000; TIMSS, 2003 (author's own calculations); Willms and Somers, 2001.

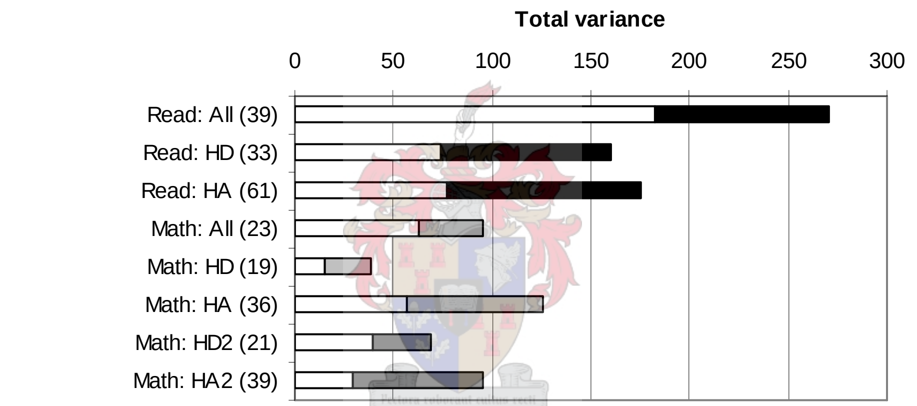
The SACMEQ reading variance works according to a different scale from the SACMEQ mathematics variance. However, the Laboratorio and TIMSS mathematics variances have been rescaled to the SACMEQ mathematics one by using South Africa to link SACMEQ to TIMSS and Chile to link TIMSS to Laboratorio.

Of importance is the total variance, and the proportional split across between-school and within-school variance. The four SACMEQ countries shown in the graph were selected because they represent four rather different variance profiles. In each case, the profile is easily attributable to a specific policy background. Mauritius has high total variance, and much within-school variance, due to an exceptionally strong streaming policy (Kulpoo and Soonarane, 2005). South Africa and Namibia are characterised by greater between-school variance than within-school variance due to their history of apartheid, which magnified inequalities between the population that could vote, and the population that could not vote. The fact that Namibia's overall variance is lower than South Africa's can probably be attributed in part to the smaller proportion of whites in that country – 8% against South Africa's 11%. Botswana's low overall variance relative to other countries within the same test, and the fact that such a small part of this is between-school variance, can be attributed to the relative equality of society in this country, which in turn can be linked to the fact that Botswana has been a stable democracy since independence in 1966.

Turning to the non-SACMEQ programmes, the fact that the South Africa TIMSS variance partitioning should be so much like the SACMEQ one verifies this pattern. The variance statistics need to be interpreted relative to the overall mean score. Mauritius and Cuba may have more overall variance, and hence inequality, than South Africa in mathematics performance, but the fact that South Africa's average score is around half of that in Mauritius or Cuba clearly refutes any argument that the situation is somehow better in South Africa. An ideal to be aspired to as a policy objective should be high average performance and low inequality or variance overall, but in particular low between-school variance. In this sense, Japan is clearly the best performing country in the above graph. (The fact that Chile's variance partitioning pattern should be reversed when we compare Laboratorio to TIMSS is striking, and could point towards a large difference between Grade 4 and Grade 8 with respect to school admissions policies. Grade 8 may be subject to selection processes for particular schools that do not exist at the Grade 4 level.)

What are the implications for the school production model? One is that the approach in the previous section to construct separate historically disadvantaged (HD) and historically advantaged (HA) models seems justified. The apartheid legacy has clearly created a very non-typical partitioning of variance profile in South Africa and Namibia. If we ignore the Chile TIMSS profile, only South Africa and Namibia in the above graph have more between-school variance than within-school variance. The next graph illustrates the variance profiles for the separate HD and HA models in South Africa.

Figure 15: Variance partitioning in SACMEQ across HD/HA segments



	Math: HA2 (39)	Math: HD2 (21)	Math: HA (36)	Math: HD (19)	Math: All (23)	Read: HA (61)	Read: HD (33)	Read: All (39)
Within school variance	66	29	69	24	32	98	86	88
Betw een-school variance	29	40	57	15	63	77	74	183

Sources: SACMEQ, 2000.

With respect to the reading test, the separate HD and HA sub-systems appear to be more typical, with within-school variance exceeding between-school variance. Moreover, the overall variance of each sub-system is well below the overall variance for the country. This supports the argument that the South Africa statistics are the result of two systems with rather different dynamics. The finding is slightly different with respect to the mathematics test, however. The HA system is more unequal overall than the country. It was hypothesised that for mathematics, the border between the two sub-systems perhaps lay at a higher point on the socio-economic status (SES) scale, and so HA2 and HD2 models were created which transferred the least advantaged half of the HA system to the HD system. Even after this adjustment, the overall HA variance equalled the overall variance for the country. One would have expected it to decline below the country level. It is clear that the high overall variance at the top end of the SES range with respect to mathematics is due to a large degree to within-school variance. This could be indicative of apartheid-like inequalities existing within individual historically advantaged schools, with white learners performing better than their black peers. In the 10% of schools that were white schools under apartheid, today one-third of learners are not white. However, the hypothesis put forward is highly speculative, and it is not possible to draw any hard conclusions from the data as to whether within-school variance with respect to mathematics in historically advantaged schools relates to race-based inequalities. It should be emphasised that a similar pattern for reading does not occur, so whatever the effect is that produces the mathematics inequalities would have to be specific to mathematics. The fact that the variance patterns for HA and HD schools with respect to the reading scores should be so similar would moreover go against an argument that variance (at any level) is systematically greater in socio-economically advantaged sectors of the schooling system.

Ferrão *et al* (2001) and Willms and Somers (2001) move from the HLM null model to a model that uses learner SES in the broad sense as the controlling factor. Hungi's (2005: 6) approach is different in that he moves from the null model to a model in which all the pupil variables are used as controlling variables – this would include some non-SES factors, for instance grade repetition and homework corrected. The purpose of producing a model in which the effect of SES is controlled is to obtain, in a sense, the variance or inequality that the school authorities are directly responsible for tackling. As figure 1, using TIMSS 2003 data showed, a positive correlation between SES and learner performance is typical, even in countries that are socio-economically relatively equal, such as Japan. In particular, less educated parents can provide less educational support to their children, and there are clear limitations to government's ability to deal with this problem in any direct way. However, if we control for SES, then the remaining inequalities are, to a large extent, an indication of how effective the education system itself is in realising educational equity.

The approach taken below was to bring in as controlling factors all variables clearly not managed by the school authorities. The variables (from the same set as the one used in the previous section) were: *learner_ses* (which captured just physical conditions in the home), *parent_educ,* and *learner_gender*. School mean variables for each of these learner-level variables were also considered. It was decided not to include learner age as an SES variable, although Willms and Somers (2001: 422) do include this, as a learner's age is to a large degree the result of the schooling system's age of admission policy in Grade 1 and repetition practices. Moreover, *daily_meals* was not considered an SES variable, as arguably it is within the power of the schooling system to ensure that learners receive at least two adequate meals a day.

Three challenges in the construction of the null model stood out (these challenges would apply to the construction of any multivariate HLM model):

- What method should be used to determine the inclusion or exclusion of individual variables?
- In which part of the L1 model should the L2 variables be inserted, and should the same L2 variable be repeated in several places in the L1 model? Equation (45) provided an example of an HLM with the same L2 explanatory variable appearing more than once.
- Where should the L2 error terms be placed? The fact that different options are possible was explained in section 5.4, as well as the fact that the inclusion or exclusion of L2 error terms does make a difference to the fixed effect and random effect output statistics.

It is perhaps surprising that Willms and Somers (2001) do not explain how any of the three problems were handled. Essentially, they provide the output statistics without any discussion of how they arrived at them. Ferrão *et al* (2001) provide some detail with respect to the second and third problems, but not the first one. Hungi (2005: 5) explains that variable selection occurred following a 'step-up' approach that would be the equivalent of the forward selection approach referred to in section 6. The placement of L2 and L3 explanatory variables within the L1 structure is illustrated diagrammatically. This placement is very simple. L2 and L3 variables are placed or nested within the L1 intercept, with one exception, which is that the L2 variable of teacher gender is nested within the slope coefficient of the L1 variable of learner age. Hungi does not repeat any L2 or L3 variable in any one model. What is not explained is where the L2 error terms are placed. However, one can be certain that one L2 error term would be nested within the L1 intercept. This leaves us with the question of whether there is an L2 error term nested within the slope coefficient for learner age.

The approach taken to produce the results in table 29 below was essentially that of Hungi. The decision was taken not to weight observations in the analysis, as we wanted to obtain variance statistics from the software (the HLM software does not produce these statistics when observations are weighted).

The steps, to some extent iterated, were as follows (these steps were also used in the construction of the full input-output model discussed further down):

1. The null model was used as a point of departure, and the insertion of all L1 variables was tested. In the case of the SES-controlled model, all three L1 variables were significant – the *t* value was at least 5. Had this not been the case, the *t* statistics would have guided the exclusion of variables to a point at which all remaining variables were significant.

2. With all the L1 variables in place, L2 variables were inserted singly within the L1 intercept, and the error term within this L1 intercept was retained at all times. The reduction in total variance, when compared to the model with no L2 variables, was noted.

3. With just the L1 variables in place, L2 variables were inserted within the various L1 slope coefficients. If the L2 slope coefficient was statistically significant insofar as its *t* statistic passed the 2-*t* rule of thumb, then this was noted. This is in line with the recommendation regarding level 2 variables provided by Bryk and Raudenbush (1992: 212), though they provide a less stringent *t* statistic cut-off of 1. (In the full model, given the great number of possible L1-L2 combinations, the mental model was used to determine which combinations were worth testing.)

4. Starting with the model with just L1 variables, L2 variables were added to the model in the L1 intercept position, beginning with the most significant variables according to step 2 above. If a newly inserted L2 variable caused a *t* statistic to drop below 2 with respect to itself or a previously inserted L2 variable, then the insufficiently significant variable was removed. If the *t* statistics of the newly inserted variable plus one previously inserted variable dropped below 2, then only the most recently inserted variable was removed. This process continued until all variables had been used up, or it was clear that no more variables of significance could be added.

5. L2 variables were now inserted into the L1 slopes, starting with the most significant variables according to step 3 above. The insertion of L2 variables already appearing in the L1 intercept position was allowed. The same selection and rejection criteria used in step 4 were used here, and rejection of L2 variables previously placed in the L1 intercept position was allowed.

6. L2 explanatory variables that were strongly correlated to each other had their positions exchanged, if they were in different L1 positions, and if this resulted in a reduction in the overall residual variance.

The following HLM output represents the final model controlling for SES with respect to the reading score.

Table 29: Reading model controlling for SES effects (HLM output)

dependent var: *read_score*	Level 1 units	3135
	Level 2 units	169
Fixed effect		
	coefficient (t stat)	
For intercept β_0		
intercept	18.44 (14.2)	
slope *slearner_ses*	16.57 (6.4)	
For slope *parent_educ*		
intercept	0.26 (5.9)	
For slope *learner_ses*		
intercept	-1.64 (-4.9)	
slope *sparent_educ*	0.19 (7.0)	
For slope *learner_gender*		
intercept	2.51 (7.1)	
Random effect		
	variance (p value)	
For intercept β_0	45.4 (0.000)	
Level-1	81.9	

The equation for the model is:

$$read\_score = (\hat{\alpha}_{00} + \hat{\alpha}_{01} slearner\_ses) + \hat{\beta}_1 parent\_educ + (\hat{\alpha}_{20} + \hat{\alpha}_{21} sparent\_educ) learner\_ses + \hat{\beta}_3 learner\_gender + \hat{\varepsilon}_{0j} + \hat{u}_{ij} \tag{68}$$

Some comments about the positioning of the L2 variables is in place. The position of *sparent_educ* (school mean of parent education) within the slope coefficient of *learner_ses* means that the effects of a learner's SES is conditional on the average parent education at the school. In other words, two learners with the same value for *learner_ses* would experience different associations between SES and performance if their school averages for parent education were different, all other variables being equal. This reflects the importance of peer effects. The HLM is particularly well suited to illustrating cross-level effects such as the one discussed here. Indeed, the explicit modelling of cross-level effects is one of the key advantages of the HLM over the one-level model, according to Bryk and Raudenbush (1992: 6).

There is a substantial reduction in between-school variance when we compare the null model to the SES-controlled model, though within-school variance remains largely unchanged. A reduction in the between-school variance that exceeds the reduction in the within-school variance when explanatory variables are introduced is in fact typical and is observed in all the three texts we have referred to. Models similar to the one in equation (68) were constructed for the HD and HA segments with respect to reading, and the process was repeated for mathematics. The model structures were not always identical to the one in equation (68). For example, cross-level effects were less common in the HA models. The results obtained, and a couple of statistics from the analysis by Ferrão *et al* (2001: 116) for Brazil, appear below:

Table 30: Partitioning of variance with control for SES

		Before SES control			With SES control			
		Between-school variance	Within-school variance	Intraclass correlation coefficient	Between-school variance	Within-school variance	Intraclass correlation coefficient	Overall variance explained by SES
South Africa: SACMEQ	Reading SA	183	88	67%	45	82	36%	53%
	HD	74	86	46%	45	80	36%	22%
	HA	77	98	44%	36	88	29%	29%
	Maths SA	63	32	67%	20	31	40%	46%
	HD	15	24	38%	13	23	36%	7%
	HA	57	69	45%	33	64	34%	23%
Brazil South (rich): SAEB		All subjects		21%			8%	
Brazil NE (poor): SAEB		All subjects		29%			17%	

Sources: SACMEQ, 2000; Ferrão *et al* (2001: 116).

The *Before SES control* statistics are taken from figure 15 and are inserted to assist comparison. The residual variance is reduced greatly when we control for SES. For both reading and mathematics, SES appears to explain around half of the total raw variance. The percentages in the last column are lower for the HD and HA segments, because the segmentation itself has explained away much of the variance. In all cases, however, within-school variance remains more or less unchanged, and between-school variance drops substantially. We nevertheless remain with much between-school inequality that is apparently not attributable to SES, and can hence be partially attributed to school factors over which the education authorities would have some degree of control. Clearly, the data would not be robust enough to account for all of the SES differentials between learners and schools, so some of the remaining between-school variance would in fact be attributable to SES differences, and not school differences. However, the matter is more complex than this, as will become clear when we discuss multicollinearity between the SES and the school variables below.

All the South African intra-class correlation coefficients are greater than the intra-class correlation coefficients for the poorest region of Brazil, the North East. In part, this is an indication of a schooling system that is systemically very unequal (even within the HD and HA segments) due to the way the authorities have organised the schooling system over the years. In part, however, the high intra-class correlation coefficient is an indication of an important opportunity for government to equalise the system. Put differently, if the between-school variance is high, then this must mean that there are examples of schools that perform substantially better than others, and hence offer instances of best practice. Arguably, a government faced with low mean scores, and a very low between-school variance, would be relatively disadvantaged with respect to knowing what improvement strategies to adopt.

What is particularly striking in table 30 is how little of the substantial raw mathematics inequality in the HD schools is explained by the SES variables. They explain only 7% of overall variance, meaning the overall variance drops by just 7% when we introduce the SES variables. This points to a systemic failing. Certain schools do worse than others not because learners are more disadvantaged, but for other reasons related to the resources the school has, or the way in which those resources are utilised. What is also noteworthy is that SES does not explain away the high within-school inequality in HA schools with respect to maths. This pattern is compatible with the loose hypothesis mentioned earlier of apartheid-like inequalities occurring within HA schools. Black learners in these schools tend to be from relatively advantaged households, so we would not expect race-based inequalities to be strongly explained by SES.

In the three texts we have referred to, treatment of the SES-controlled HLM (or the rough equivalent of this in the case of Hungi) is followed by the presentation of a full model that includes both SES and school inputs as explanatory variables. One such model is explored with respect to the SACMEQ data, namely the HD model for reading performance. This should be a model of special concern, partly because of the importance of tackling under-performance in the historically disadvantaged segment of the schooling system, and partly because from tables 20 and 21 in the previous section it was clear that a more predictive HD model for reading was possible than for mathematics.

The procedures for arriving at the full model were similar to those followed in obtaining the SES-controlled model outputs of table 29. The approach was to begin the procedures with the full set of possible explanatory variables, and not just the variables that were selected for the one-level models of the previous section. An optimal structure was first found using an .mdm file that included only observations with non-missing data across all the possible variables, and once the optimal structure had been found, this was applied to a new .mdm file that included only the variables needed for the optimum structure. The mental model was used to avoid the exploration of apparently senseless cross-level effects, such as school infrastructure influencing the impact of gender.

The final model is represented below:

$$
\begin{aligned}
&read\_score = \left(\hat{\alpha}_{00} + \hat{\alpha}_{01} school\_infra + \hat{\alpha}_{02} srepetition + \hat{\alpha}_{03} sparent\_educ\right) + \\
&\hat{\beta}_1 repetition + \hat{\beta}_2 daily\_meals + \hat{\beta}_3 parent\_educ + \\
&\left(\hat{\alpha}_{40} + \hat{\alpha}_{41} class\_size\right) learner\_ses + \hat{\beta}_5 learner\_age + \\
&\left(\hat{\alpha}_{60} + \hat{\alpha}_{61} slearner\_ses\right) learner\_gender + \hat{\beta}_7 textbooks\_read + \hat{\varepsilon}_{0j} + \hat{u}_{ij}
\end{aligned}
\tag{69}
$$

The statistical outputs follow. Fixed effects with and without the use of the SACMEQ weights are reflected (asterisks indicate outputs from the unweighted model):

Table 31: Full reading model for HD schools (HLM output)

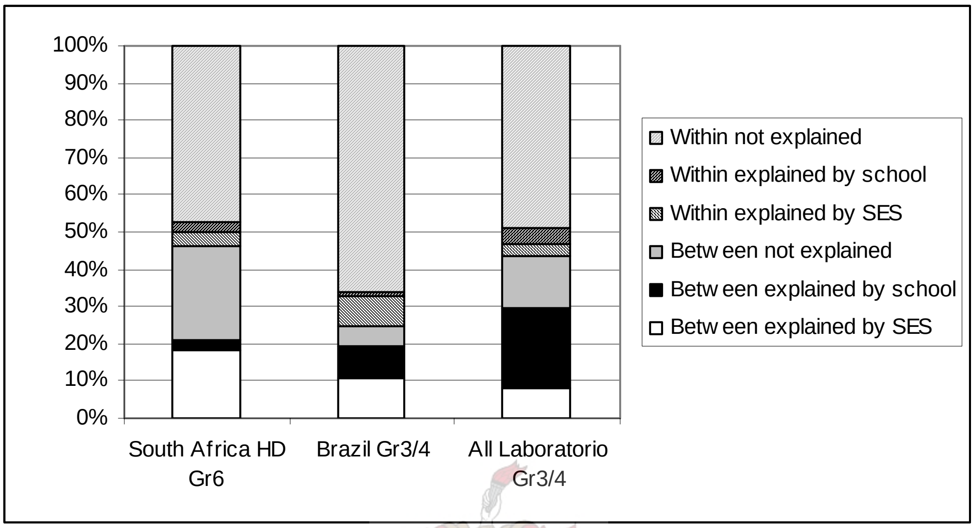
dependent var: *read_score*	Level 1 units	2588
	Level 2 units	138
Fixed effect		
	coefficient (t stat)	
	With weights	*Without weights*
For intercept β_0		
intercept	31.31 (10.3)	30.87* (9.3)
slope *school_infra*	1.17 (3.7)	1.11* (3.9)
slope *srepetition*	-4.68 (-3.4)	-3.48* (-2.4)
slope *sparent_educ*	0.42 (2.0)	0.48* (2.2)
For slope *repetition*		
intercept	-1.46 (-6.0)	-1.32* (-5.8)
For slope *daily_meals*		
intercept	1.05 (4.0)	1.06* (4.1)
For slope *parent_educ*		
intercept	0.17 (4.1)	0.15* (3.4)
For slope *learner_ses*		
intercept	0.74 (4.4)	0.81* (4.5)
slope *class_size2*	-0.00014 (-2.4)	-0.00014* (-2.3)
For slope *learner_age*		
intercept	-0.80 (-5.6)	-0.92* (-7.0)
For slope *learner_gender*		
intercept	-1.33 (-1.4)	-1.71* (-2.3)
slope *slearner_ses*	0.70 (3.3)	0.77* (4.3)
For slope *textbooks_read*		
intercept	4.03 (3.0)	4.15* (3.3)
Random effect		
	variance (p value)	
For intercept β_0	40.5* (0.000)	
Level-1	76.0*	

Outputs with * are from the model without SACMEQ weights.
Excluded: yrs_preserv_read, day_inserv_read, teacher_ses, teacher_eval_read, class_meth_read, hrs_year_read, yrs_preserv_prin, prin_teach_load, par_involve_read, dist_support, ruralness, teacher_disc.

With the introduction of school input variables, between-school variance drops from 44.5 to 40.5, and within school variance from 80.3 to 76.0, using the SES model variances in table 30 as the point of comparison. Though this is not done in any of the three texts, nothing should stop us from computing an R^2 value by calculating the difference between the original total variance without any controls (74.1 + 86.0 = 160.1), and the variance left after running the above model (40.5 + 76.0 = 116.5), and dividing this difference by the original total. The result is 0.272, indicating a somewhat less predictive model than the one-level HD model of table 20, which yielded an R^2 of 0.325.

We are now able to partition the overall reading variance into six parts, it would seem: under the between-school and within-school categories there are three sub-categories, namely variance explained by SES, variance explained by school inputs and variance not explained. The following graph illustrates this. Corresponding Laboratorio figures from Willms and Somers (2005) have been included for comparison (the Laboratorio language test results were used).

Figure 16: A six-part partitioning of overall variance



Sources: SACMEQ, 2000; Ferrão et al (2001: 116); Willms and Somers (2005).

As we shall discuss below, there are some conceptual problems with this six-part breakdown. However, first it is worth noting how little between-school variance appears to be explained by school inputs in the South African HD model. This would suggest that school inputs do not make a great difference, and that they are inefficiently utilised. In the Brazil and Laboratorio breakdowns, the part of between-school variance explained by school inputs is much larger. There are two important provisos, however. Firstly, much of the between-school variance in South Africa is not explained at all. Thus there could be important school input variables (in the broad sense of 'inputs', so this could include school management) that are making a difference, but they are not reflected.

Secondly, and more seriously, as Ferrão *et al* (2001: 118) point out, multicollinearity between SES variables and school input variables could mean that some of the between-school variance explained by SES, is in fact masking school input effects. For instance, as indicated by table 38, the variables *learner_ses* and *school_infra* are highly correlated. It is thus possible that what appears as the effects of the former, is at least partially masking the effects of the latter. The fact that between-school variance explained by SES should be *higher* in the case of South Africa than in the other two cases is telling. And the fact that we would obtain a different breakdown from the one reflected in the graph had we controlled for school inputs first, and then introduced SES into the model, should confirm the precariousness of the figures. The analysis problem discussed here is worth emphasising because the flawed logic behind figure 16 is easily succumbed to given the analysis layout and the absence of the necessary provisos in texts such as Willms and Somers (2005) and Barbosa and Fernandes (2001: 13).

There is a better answer to the important question of the relative effects of SES and school inputs on between-school variance than what is provided by the approach described above. We can use the HLM coefficients from table 31 in order to simulate a situation in which only SES effects operate, and another in which only school effects operate, and compare the residual between-school variance statistics obtained from the two simulations to the model where both types of effects operate, in order to gauge the separate strengths of SES and school inputs. In all three models, the residual variance would be based on the difference between the model reading score and the actual reading score. We should bear in mind that the coefficients from the full HLM model all reflect *net* effects, so at least some of the multicollinearity problem is dealt with. The model with only SES effects would be run as follows. All the school input variables would take on their respective grand mean values (applicable to HD schools), and the SES values (in our broad sense of SES) would be left unchanged. The model would then be run. A similar approach would be used to run the model with only school input effects. The results are as reflected below. The goal was not to estimate new coefficients, so the HLM software could not be used. Instead, the model, using the HLM coefficients from table 31, was set up in Stata. The between-school variance was computed using the variance components model and the loneway command in Stata, which, as was explained in section 5.5, provides a very close approximation of the variance partitioning of the HLM software. The results were as follows.

Table 32: Full reading model for HD schools

	Variance			
	Between-school	Between-school difference	Within-school	Total
Model with SES and school input effects	40		75	115
Model with only SES effects	53	13	78	143
Model with only school input effects	49	9	82	140

The fact that the full model statistics (first row) are virtually the same as the variance statistics in table 31 confirms the equivalence of Stata's variance components model. Allowing only SES effects to operate leaves a between-school variance of 53, implying that SES makes a difference of 13 to the original between-school variance. The corresponding difference made by the school input variables is 9. Whilst school inputs do still make a smaller difference than SES if we use this approach, the gap between the SES and school inputs effects are clearly much smaller than they were in figure 16. We do not have equivalent statistics using the same approach for another country, but it is telling that using the figure 16 approach, which is clearly biased against the effects of school inputs, school input effects still appear larger than SES effects in the overall Laboratorio results. It appears we can conclude that school inputs are indeed under-performing in the HD segment of the South African schooling system.

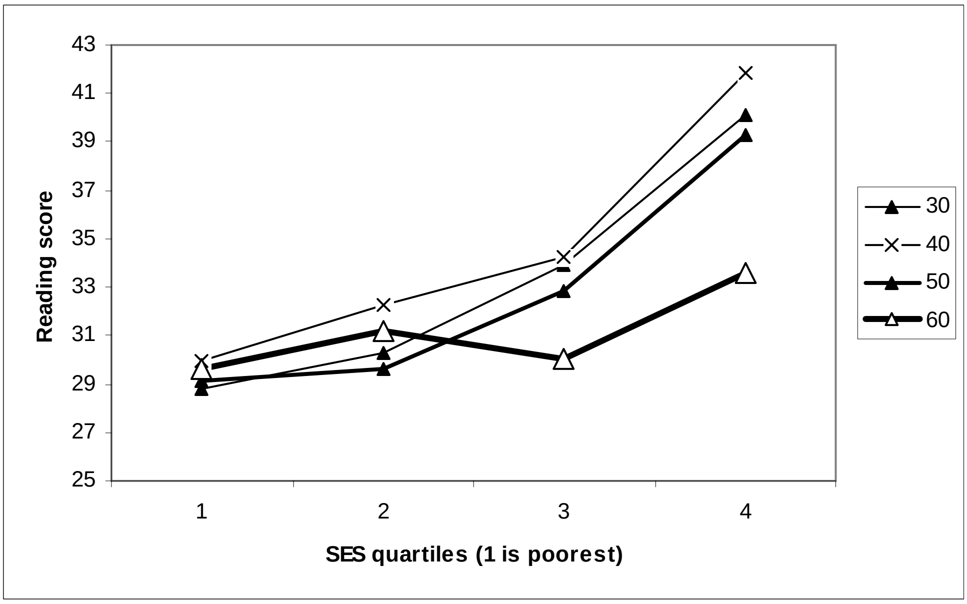
Turning to the fixed effects from table 31, it noteworthy than in the case of the SACMEQ dataset, weighting observations does not make a great difference to the slope coefficients obtained. On average, the absolute difference made to the slope coefficients is 2%. However, the difference varies. The relatively large difference with respect to the slope coefficients of the two variables referring to repetition stands out.

How does the set of retained variables in the HLM differ from that in the one-level model for HD reading reflected in table 20? Firstly, we should remember that the one-level model was not constructed separately for the HD and HA segments – the same structure as that obtained in the general reading model of table 17 was retained for the sub-models. Even taking this into account, it does seem as if learner level variables were retained more easily in HLM – all seven were included in the table 31 model, and it was more difficult to retain school level variables in the HLM model. This raises the question of whether there is something inherent in the estimation of the significance statistics in HLM that biases one against the inclusion of L2 variables relative to L1 variables. A comparison was run between the above model without the L2 variables nested within the L1 slope coefficients, in other words without *slearner_ses* and *class_size2*, and a simple one-level model that excluded these two variables but included the other variables with the exact values and exact number of observations as for the HLM model. These two models should be comparable, as the only structural difference would be the existence of two, and not one error term in the HLM model. The *t* statistics of the L2 variables were indeed markedly lower in the HLM model than in the one-level model, suggesting strongly that the HLM does carry a bias, relative to the one-level regression model, against the significance of L2 variables.

Given this bias against L2 variables, it is especially striking that the L2 variable *class_size2* is retained in the HLM model, when it was rejected by all the one-level models in the previous section. The way this variable is positioned in equation (69) is particularly useful in illustrating the effects of class size. The larger the class, the smaller the expected translation of SES into learner performance. The very low value for the slope coefficient of *class_size2* is simply a result of the fact that we squared the original class size value. This cross-level effect makes sense according to our mental model, as does the other cross-level effect in the model, namely that the disadvantage experienced by girls, is ameliorated if the mean SES of the school is higher.

Figure 12 illustrated the bivariate relationship between class size and the reading scores. The results from our HLM provide direction for some further analysis. Hungi's (2005: 12) graphical representation of the cross-level effect of teacher gender on learner age provides a useful approach. The graph in figure 17 analyses just HD schools, uses the same class size bins as figure 12 and provides separate learner SES against reading score curves for the different class size bins (the variable *learner_ses* was used for SES). The SACMEQ weighting was used in the generation of the quartiles and to calculate average SES per quartile and per class size bin. Points representing fewer than 30 weighted learners were not graphed so, for example, there is no curve for learners in classes with an average size of 20.

Figure 17: Learner SES, class size and reading scores in HD schools

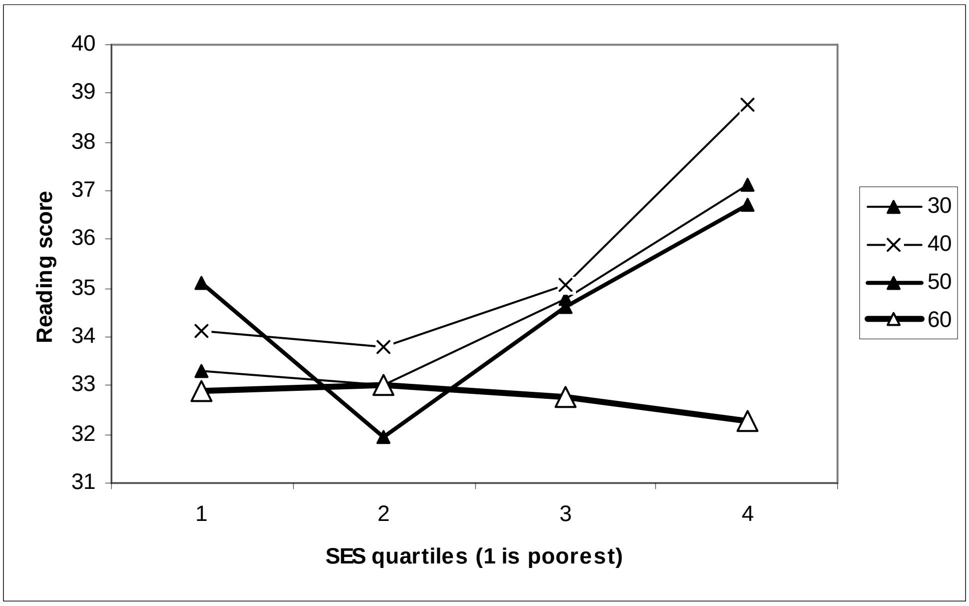


Source: SACMEQ, 2000.

The slopes for the 30, 40 and 50 class size bins are not too dissimilar, though the larger the class, the lower the score, to some degree. However, the curve for class size 60 is clearly different, and clearly indicates a lacking translation of SES advantage into performance. A strong threshold would appear to exist between class size 50 and class size 60. Below this threshold, the typical positive association between SES and performance is clearly retained, but at this threshold this association becomes, to a large extent, broken.

We can go one step further and use the coefficients from our HLM in order to examine what the above graph would look like if we controlled for all the other explanatory variables in table 31 (meaning other than *learner_ses* and *class_size2*). We do this by using the coefficients from the model (the coefficients produced from the weighted data were used) and the actual data for each learner, to create an expected reading score for each learner. We take the difference between this expected score and the actual score as an error term, and add the error term to the grand mean. The resultant score per learner is a score that captures the effects of the *learner_ses* variable and class size, whilst holding other effects constant. The following variation of the previous graph is the result:

Figure 18: Learner SES, class size and controlled reading scores in HD schools



Source: SACMEQ, 2000.

The basic pattern remains if we perform the controlling exercise, though the class size 50 curve has become unstable at the low SES end. Class size of 60 still appears as a problem in terms of SES to performance translation, in fact the relationship is slightly negative now. According to the SACMEQ dataset, 16% of HD learners were in classes of 55 or more learners (the 60 bin captures the range 55 to 65). The problem is thus fairly widespread.

It is worth taking stock of what these findings say about analysis approaches in general. Although we squared class size to make it more sensitive to increasing marginal impacts, our one-level model rejected class size. Our HLM, on the other hand, picked up class size as an important school level effect, which prompted us to examine the matter a bit further. The value of viewing the same data through different models, none of which can be regarded as particularly definitive on its own, is emphasised.

To end this section, the HLM model is used to recalculate the impact on performance of the hypothesised interventions reflected in table 28 of the previous section. The HLM coefficients produced using the SACMEQ weights were used, and only reading performance was covered. The following are the results. The HD impact figures from table 28 are repeated for the sake of comparison. Some interventions could not be recalculated, because the necessary variables did not appear in the table 31 model. Moreover, a new class size reduction intervention was introduced.

Table 33: Simulated interventions using the HLM model

Variable	*Hypothetical change*	*Net effect on HD scores from one-level model*	*Net effect on HD scores from HLM*
repetition	Decrease the average learner years of repetition in the 61% of the system where schools exceed the 0.5 level, to 0.5.	+6%	+5%
	Decrease the average learner years of repetition in the 89% of the system where schools exceed the average level for HA schools (0.17), to this HA level.	+12%	+11%
textbooks_*read*	Raise the average number of *reading* textbooks per learner so that each learner enjoys a ratio of at least 0.5 per learner.	+2%	+2%
school_infra (N.B. closely correlated to ruralness)	Raise the level of physical infrastructure of all schools to the present average for HA schools.	+14%	+17%
daily_meals	Raise the intake of daily meals so that all learners receive all their daily meals (currently some 51% of learners do).	+3%	+2%
parent_educ	Raise the level of education of the least educated 20% of parents to the level of the 20th percentile.	+1%	+1%
	Raise the level of education of the least educated 40% of parents to the level of the 40th percentile.	+4%	+4%
learner_ses	Raise the SES of the least advantaged 40% of learners to the level of the 40th percentile.	+2%	+4%
class_size2	Decrease class size to 55 in all instances where classes are larger than this.		+0.3%

Using the HLM model renders no substantial difference in expected impact when compared to the one-level model. The reliability of our earlier figures is thus strengthened. The doubling of the SES impact can probably be viewed as a result of the new structure capturing cross-effects – in both instances where learner SES appears in the HLM, it is within a cross-level effects situation. The class size intervention has a low impact overall, though it should be noted that the intervention would result in a 2% improvement in reading scores for the 16% of HD learners with classes greater than 55.

8 TRANSLATION OF THE MODEL INTO POLICY INFORMATION

This section will examine the critical link between the economic analysis of the type provided in the foregoing sections, and optimal policies, strategies and budgets. There is a special emphasis on the formulation of cost and budget implications. The widespread problems that exist in this regard within developing country education planning systems are first discussed. Thereafter, some suggestions on how to tackle the problem receive attention, with reference to cost effectiveness analysis. Finally, cost-sensitive policy recommendations flowing from the foregoing SACMEQ data analysis are put together. The suggestions are put together in such a way that they could assist the work of the South African Department of Education. This involves some cursory examination of existing costs affecting education services in South Africa. This part is cursory, however. Figures and assumptions are intended primarily to illustrate the approach, and are not the product of the detailed costing exercise one would expect from a fully-fledged report.

Translating economic research into information that ministries can use for the practical purposes of preparing plans and budgets is widely recognised as a problem (Saïdi, 2001). Penrose (1993: 3) says the following about sub-Saharan Africa, though this would apply to much of the developing world.

> [O]ne of the weakest – if not the weakest – link in the chain of policy implementation is the relation between planning and budgeting, including how budgets are made. There has been a tendency to put broad educational policy objectives on the one hand and the economic planning and management of resources on the other into two separate compartments, so that while there is no shortage of analysis of what needs to be done, the means of achieving given objectives are often unspecified.

Policy recommendations appearing at the end of economic analyses are typically short on detail, and usually not compiled by analysts with much knowledge of the policy and budget formulation systems and processes. Ferrão *et al* (2001) conclude their multi-level analysis of SAEB data with two bullet points stating, firstly, that there are significant associations between race and performance and, secondly, that having multiple shifts in the schooling system (for instance a morning and an afternoon shift) ought to be investigated further as a possible efficiency problem. In fact, what is striking about the SAEB programme is how little policy information it yields relative to the extent and cost of the programme (see Appendix A).

Hungi (2005) provides a larger volume of policy advice than do Ferrão *et al*. Although Hungi does not draw this distinction, there are clearly recommendations that require budget increases, and ones that do not. In the former group we can include the recommendations that more physical school places for learners, the electrification of schools, a larger school feeding programme, a lower pupil/teacher ratio and more in-service training for teachers will all contribute to enhancing the quality of learning outcomes in Kenyan schools. In the latter group we can include the recommendation that repetition rates need to be lowered.

The SACMEQ project documentation provides a recommended categorisation of policy recommendations. Distinctions should be drawn between those involving the collection of new data, the tightening up of existing monitoring systems, the establishment of new research programmes, investigative consultations with stakeholders, and the reform of existing practices and policies. The distinction between recommendations with and without major budget implications is also emphasised (Ross *et al*, 2004: 19).

Given the difficulties relating to the translation of research into practical budget information, it seems prudent to pay special attention to whether recommendations involve major budget advocacy and preparation work. Recommendations that do not imply any major budget preparation can be treated differently. There is less of a trade-off or opportunity cost problem because these recommendations deal with ways existing budgets are translated into educational services, not with the sizes of the budgets themselves. It is feasible to argue, for instance, that the advocacy of alternative classroom methodologies, changes to the school governance rules, a different school calendar, and different forms of pre-service teacher training should be pursued simultaneously, because none of these recommendations imply a major increase in any budget, or a major budget re-prioritisation.

The recommendations *with* major budget implications should ideally be expressed in a very specific way. It is important to realise that there is an inherent resistance to *any* budget re-prioritisation within a government planning system, and fierce competition between interest groups for any new budget amounts that become available. Actors within government are only very partially driven by concerns around the greater good. Even planners and managers with a strong commitment to the aims of the government, tend to see themselves as lobbyists for particular institutions (for instance technical schools as opposed to ordinary schools), particular inputs (for instance school buildings as opposed to textbook spending) or stakeholder groups (for instance parents as opposed to teachers). Pradhan (1996: 104) views the problem as the 'tragedy of the commons', an allusion to the problem of over-grazing on common land. There is an insufficient realisation of the overall common good, meaning that a government system, in dealing with the competing budget demands, is prone to over-spend relative to what is actually required for an efficient delivery of the services actually being delivered.

There are economic analysis tools that have the potential to place the political lobbying on a more rational and empirical footing. This can increase the likelihood that decisions and budgets will maximise social welfare and the aims of the government. We saw in section 2.2 that cost benefit analysis (or rates of return analysis) can assist in gauging the relative importance of investing in the various levels of the education system. That section also discussed the importance of production function analysis of the type pursued in much of this thesis in prioritising expenditure on different types of inputs. Production function analysis can feed into cost effectiveness analysis rather conveniently to provide important budget information.

Cost effectiveness analysis is used extensively in public health expenditure planning. This analysis involves assessing the cost of achieving the same output, often a specific number of healthy life years saved in the case of health, through different government programmes. The programme costing the least is regarded as the most cost effective (Pradhan, 1991: 60). The methodology is less commonly used in education, though, as will be indicated below, its application in education is possible.

Using the production model to calculate estimated performance gains associated with different hypothetical interventions can take us some way in assessing the relative effectiveness of various actions. Such calculations were reflected in table 28 in a previous section. The calculation can be expressed as follows:

$$\Delta\overline{Y} = \hat{\beta} \times \Delta\overline{X} \tag{70}$$

The change in the average performance, or $\Delta\overline{Y}$, is equal to the change in the average input value *X*, for instance the ratio of textbooks to learners, multiplied by the slope coefficient β. Although the production function, as expressed for instance in equation (10), deals with the input-output relationship at the level of the individual unit, usually the learner, a simple simulation in Excel can show that the production function also holds at the level of average performance and average inputs for the entire system, thus permitting us to express the relationship as in equation (70) above.

If we know the cost of one unit of *X* at the system level, for instance the cost of increasing the textbook/learner ratio by 1, we can calculate what the total cost *TC* is of achieving an increase in average performance of $\Delta\overline{Y}$.

$$TC = \frac{C_X \times \Delta\overline{Y}}{\hat{\beta}} \tag{71}$$

C_X is the cost of one unit of *X*. If we apply the above calculation to several programmes, we obtain a cost effectiveness analysis.

The following table uses the logic of equation (71) to calculate the cost of each of the interventions in table 28 where a major budget change is implied. The cost of raising the average learner performance by one percentage point is also reflected in order to allow a comparison of the cost of equal effects. The distinction between recurrent costs and capital costs is an important one. The former would need to be maintained indefinitely, whilst the latter can be completed in a limited period of time.

Table 34: Cost effectiveness of simulated interventions

Hypothetical change	*Cost (millions of rand in 2005)*	*Approx. net effect on HD scores*	*Cost per percentage point increase*
Increases in RECURRENT EXPENDITURE required (annual)			
Raise the average number of *mathematics* textbooks per learner so that each learner enjoys a ratio of at least 0.5 per learner. (1)	51	+3%	17
Raise the intake of daily meals so that all learners receive all their daily meals (currently some 51% of learners do). (2)	726	+3%	242
CAPITAL INVESTMENT required			
Raise the training level of educators in the half of the system with the greatest deficit by the equivalent of one year of pre-service training. (3)	668	+5%	134
Raise the level of education of the least educated 20% of parents to the level of the 20th percentile. (4)	8,018	+1%	8,018
Raise the level of physical infrastructure of all schools to the present average for HA schools.	6,600	+14%	471
Key cost assumptions: (1) Cost per textbook is R50. (2) Cost per meal is R1.20 (based on current Primary Schools Nutrition Programme figures). (3) Each educator requires R4,000 in part-time studies (based on post-graduate fees at UNISA). (4) Each additional year of schooling per parent implies R4,000 in adult education (based on reports of costs of ABET). (5) The cost of renovating one school is R330,000.			

The above figures should be viewed in the context of a total expenditure on public primary schools of around R32,000 million in 2005 (just over R1,000 million of this was on capital expenditure). Some key cost assumptions are noted at the foot of the table. In all cases, it was assumed that the intervention should target not just the Grade 6 level of the schooling system (SACMEQ tested Grade 6 learners), but all the grades from Grade 1 to Grade 6. This assumption was used in calculating the costs.

The figures in this table are not figures that would be used directly in budget preparation. Their function is to provide a sense of the cost effectiveness of the different interventions. The overall pattern is important, but not the exact relationships, given the fallibility of the production model, and the looseness of the cost assumptions. Thus we cannot say with any certainty that daily meals as an intervention is 14.2 times as costly as the mathematics textbooks intervention as a lever for improving the average score. However, we can be rather certain that if our aim is to raise the average score, then the mathematics textbook intervention is more cost effective than the daily meals intervention.

The cost differentials in table 34 are sufficiently great to allow for a high degree of certainty relating to how best to spend public funds in the interests of better learner performance. The teacher training intervention stands out as a highly cost effective one. This investment in the human capital of the schooling system is equal in cost to around eight years of the mathematics textbook intervention. This textbook intervention is also highly cost effective, as it implies a very low increase to the total education expenditure figure, of around 0.1%. Providing daily meals for all learners is a more costly performance enhancement intervention, which would raise total recurrent expenditure on education by 0.8%. However, given the general human capital importance of ensuring that all children are adequately nourished, this intervention is necessary and seems feasible from a budget perspective. Improving the physical infrastructure of the schooling system makes a difference to performance, but here the cost effectiveness is relatively low. Although parent education is a key determinant of learner performance, tackling the problem by targeting parents with adult education is a relatively inefficient approach. One intervention, namely the reduction of learner repetition, has been left out of table 34. This intervention involves a *negative* cost insofar as it reduces enrolment per grade over time, as flows between grades become more efficient. However, given the complex nature of this policy issue, both in terms of the performance impact of high repetition (this was discussed in section 7.1) and in terms of the effect on cost of reducing class and school sizes in a schooling system (for structural reasons cost reductions are never directly proportional to the enrolment decreases), the item was excluded from table 34. The other items in that table are all arguably of a less complex nature.

9 RECOMMENDATIONS FOR FUTURE DATA COLLECTIONS

The production modelling discussed in the foregoing sections relies strongly on the design of the school principal, educator and learner questionnaires used to gather background information in programmes such as SACMEQ, South Africa's Systemic Evaluation and Brazil's SAEB. In fact, the entire right-hand side of equation (15) rests on the way questions have been asked in these questionnaires. There are a variety of factors that influence the design of the questionnaires. The requirements of a production model is only one such factor. The questionnaires are also subject to the data needs of programme managers wanting data on their particular programme, for instance the school nutrition programme, the advice of academics not concerned with the production function as well as the political influence of the Ministry and teacher unions relating to the tone and content of questionnaire items. Nevertheless, obtaining an integral picture, or model, of how school and home factors relate to test scores is arguably the prime reason for having background questionnaires in a programme such as SACMEQ, and for this reason it seems strategic to close an analysis of the data with a set of recommendations about how questionnaires could be improved in the interests of a more explanatory production model. Some suggestions have already been made in sections 6 and 7. This section reiterates some of those suggestions, and examines practices in the SAEB 2003 questionnaires to see whether they can suggest specific improvements to the SACMEQ questionnaires.

The recommendations that follow deal with the SACMEQ school principal and teacher questionnaires. The 2003 SAEB questionnaires (available on the INEP website) were used for comparison purposes due to their focus on developing country conditions (those of Brazil), and because they appear to hold some valuable lessons for a relatively new programme like SACMEQ. It should be remembered that SAEB had by 2003 been through seven runs (the programme was initiated in 1990). It should also be noted that the three SAEB questionnaires are similar in length to the SACMEQ questionnaires in terms of pages and implied variables.

Table 35: Comparison of SACMEQ and SAEB questionnaire sizes

	SACMEQ		*SAEB*	
	Pages	*Variables*	*Pages*	*Variables*
School principal questionnaire	16	169	9	159
Teacher questionnaire	16	138*	8	138
Learner questionnaire	14	74	2	23**

* Refers only to the parts filled in by the reading teacher.
** The SAEB Grade 8 learner questionnaire.
Sources: SACMEQ, 2000; SAEB materials on www.inep.gov.br.

The school principal and teacher questionnaires from the two programmes are very similar in length (if we consider the number of variables), but the SACMEQ learner questionnaire is substantially longer than the SAEB one. There is nothing in the above figures that suggests that the SACMEQ questionnaires might be too short. The recommendations that follow should thus rather be seen as leading to revisions and replacements of existing SACMEQ questions, as opposed to additions to the existing stock.

What stands out as highly informative in the SAEB school principal questionnaire is a set of questions that essentially queries the school principal's mental model of the education production function. The opinion of the principal is elicited with respect to a number of factors typically said to influence learner performance. The structuring of the questions is arguably open to improvement, in particular a response along a multiple scale with regard to each factor, as opposed to SAEB's binary agree/disagree response, seems better. Nevertheless, directly eliciting what the principal believes makes a difference to learning seems important on a number of levels. The same questions are also posed in the SAEB teacher questionnaire. The SACMEQ school principal questionnaire elicits much less judgement and opinion from the school principal, and where it does, for instance with respect to the importance of different management activities, this is not explicitly linked to improving learner performance.

In the foregoing sections, it has been suggested that management practices in the school account for much of the unexplained variance in the models described. Such practices are typically not dealt with in any detail in the questionnaires of these kinds of programmes. Arguably, the matter could be covered better within the questionnaires. The SAEB questionnaires, whilst they still appear to fall short of an adequate treatment of school management, provide some interesting pointers. Specific information on the frequency of school governance meetings with the community is elicited. Who chooses the textbooks used in the school, and how punctually new textbook supplies arrive, are asked. The school principal is required to provide information on in-service teacher training initiated or managed by herself. In the teacher questionnaire, the teacher is asked to evaluate management systems and practices with respect to the school principal and the bureaucracy.

Selection effects were earlier identified as something that greatly complicates the modelling of input-output relationships. To some extent, the SAEB principal questionnaire deals with both learner selection effects and teacher selection effects. Questions on processes governing the admission of learners to the school, and on the system that attaches individual teachers to individual classes, are asked.

Section 4.3 referred to teacher skills and contact hours as the 'bare bones' of the education production process. Improvements to questions relating to actual hours spent per learner (or at least per class) per year in contact with the teacher, and to questions relating to the current skills and skills upgrading activities of teachers stand out as being important. In the analysis it was clear that there were some critical gaps in the questionnaires in these areas. Contact time data, which should be relatively easy to obtain through questionnaires, is not available through any direct questions on the matter, for instance through questions posed to the teacher regarding number of hours per week and weeks per year spent with the specific mathematics class. (The SAEB questionnaires also seem weak in this area.)

Problems were noted in section 7.1 regarding the teacher's assessment (captured in the teacher questionnaire) of the quality of the in-service training received. This assessment is valuable, however the less than effective rating returned by many educators ought to have been categorised according to whether the teacher believes the training is objectively of a poor standard or whether the training is simply set at a level that is below the teacher's current skills level. However, even the data on the quantity of in-service training received is problematic due to the entangling of treatment and selection effects (this entanglement problem was also seen to occur with regard to inputs other than in-service training). The selection effect results in more treatment (or in-service training) where performance is lower, due to the deliberate targeting strategies of government programmes, whilst the treatment effect results in more treatment being associated with higher performance, because the treatment is making a difference. The solutions in terms of the questionnaires are probably not easy, though much of the answer seems to lie in controlling for the selection effect by asking respondents to provide information directly on this effect. Specifically, it seems important to split data on in-service training activities according to whether this is linked to a government programme targeting poorly performing schools or not.

The exclusion of the SACMEQ teacher test in the case of South Africa (see section 3) obviously leaves an important gap with regard to our understanding of the knowledge of the teacher. Such a gap can perhaps at least partially be compensated for through background questions eliciting the day-to-day intellectual activities of the teacher. The SAEB teacher questionnaire includes questions on the leisure-time reading habits of the teacher.

Teacher job satisfaction, and, linked to this, the professional and community identity that the teacher attaches to herself, is receiving increasing attention as a key ingredient in the schooling process (Welmond, 1999). The SAEB teacher questionnaire asks a range of questions aimed at detecting what may be missing with respect to the teacher's job satisfaction. The structure of the SACMEQ questions in this regard, as noted earlier, is problematic. For example, teachers are asked to rank how important teacher salary is in influencing the satisfaction of the teacher. This provides an idea of the teacher's 'mental model' of the link between salary and teacher satisfaction in a general sense, but does not tell the analyst whether *this* teacher feels she is under-paid or dissatisfied.

Finally, the quality of the learning support materials (LSMs) being used in the school was an area in the policy-oriented framework of figure 5 for which no data could be found in the SACMEQ questionnaires. One SAEB teacher question asks the teacher to rank the quality of the textbooks being used. This question, and perhaps a similar question asked to the school principal, though still limiting in terms of a deeper understanding of the quality of LSMs, would have allowed for the inclusion of this important input into the models.

10 CONCLUSION

In the introduction, the question was asked what the optimal practices might be for modelling school production processes in the developing country school monitoring programme context. Specifically, it was asked whether Brazil's SAEB monitoring programme might yield important lessons, given the relative maturity of this programme. And more specifically, it was asked whether the hierarchical linear modelling approach employed in analysing the SAEB data might be valuable. A set of seven steps were furthermore put forward as a framework for analysing data from a monitoring programme such as SAEB (or SACMEQ or South Africa's Systemic Evaluation). These seven steps constituted the basic structure for this thesis.

This conclusion will sum up the responses to the questions posed in the introduction. These responses have been explored from a number of different angles within the thesis. Key findings from the SACMEQ data analysis, which was partly aimed at exploring methodologies, and partly aimed at uncovering aspects of the school production function in South Africa, will be summarised. The conclusion ends by highlighting three salient findings relating to the methodology of school production modelling.

The first of the seven steps was understanding the data. It was emphasised how this step relied heavily on there being adequate documentation on the sample design, the data collection and data normalisation processes. Both the SAEB and SACMEQ documentation were found to be informative in this regard. Understanding the variance of the values in the dataset was underlined. In the case of an unequal society such as the South African one, special care should be taken to detect non-normal distributions in the output values, or test scores, as such distributions might make it difficult or impossible to conceptualise or statistically model the country as one system. The bi-modal distribution of South Africa's reading scores in the SACMEQ dataset in fact led to the development of separate historically disadvantaged (HD) and historically advantaged (HA) production models.

The second step was building the mental model. This was a step to which considerable attention was paid by the SAEB data analysts. It was seen that this step can take the researcher in many different directions, and that the abundance of texts in this field are potentially confusing due to the diversity of mental models and considerable disagreement over what the predominant production function findings in recent years have been. However, key factors that make this field a highly complex one have been clearly identified, and this facilitates the analysis somewhat. One such factor is selection effects. Neither learners nor educators are randomly distributed across schools and classes. The non-random selection effects at play clearly influence how production occurs, yet these effects can only indirectly be detected in the cross-sectional studies typically used by developing country governments. Another factor is the fact that schooling systems are multi-layered, or hierarchical, and that production dynamics are often occurring at different layers simultaneously, or between layers. And yet another factor is the common entanglement of treatment and selection effects (this was introduced in section 6). There was no obviously appropriate mental model that would serve the SACMEQ data analysis, and so, given the intention to focus strongly on the policy implications part of the analysis, a new ‘policy-oriented’ framework was constructed to guide the modelling conceptually.

The third step was the selection of a statistical model. The commonly used one-level regression model stood out as an obvious solution, given its ability to indicate the net effect of individual inputs in the school production process, in other words the effect of one input whilst controlling for the effects of other inputs. The hierarchical linear model, or HLM, has in recent years also become a preferred statistical model. Not only has it been used to model Brazil’s SAEB data, it has also been used with Laboratorio data from Latin America and recently on SACMEQ data from Kenya. Whilst the one-level regression model allows for the use of the school mean values corresponding to individual learner values in order to gauge certain school effects, the HLM allows for a much richer exploration insofar as it allows for what in some senses is a miniature regression model for each school, with for instance different slope coefficients for different schools. The discussion of the HLM revealed that these benefits of the model come at the cost of increased complexity, partly due to the inherent complexity of the statistical theory underlying the HLM, and partly due to the fact that the HLM is still at an under-developed stage. Relatively little has been written on the HLM, and the relevant statistical software packages are not yet user-friendly.

The fourth step involved initial variable selection and manipulation. Factor analysis, an approach used by, for instance, the SAEB analysts, to combine several related variables in the original dataset into one new variable, was examined and applied to the SACMEQ data. Specifically, this approach was used to consolidate information on the socio-economic status (SES) of learners. The stepwise selection approach was used to eliminate insufficiently significant SACMEQ variables as candidates for the final production model. Cognisance was taken of the considerable problems with both these approaches, especially where they are used in the absence of a guiding mental model. This step was shown to be procedurally complex. In order to deal with this, much of the data processing occurred using a computer programme, the procedures of which were described. The result of this step was the conversion of the original 381 SACMEQ variables dealing with the production inputs to a reduced set of 21 new variables.

The fifth step constituted the crux of the analysis, and was titled 'iterative modelling' given the need for repeated modelling of the data using the basic statistical models in different ways. It is clear that despite the considerable volume of texts in existence attempting the same kind of analysis in the past, there is no easy recipe or set of procedures to follow in undertaking this analysis work. Iterative running of statistical models, often very similar to each other, is inevitable, and the detection of important statistical outputs (clearly not all statistically significant outputs are important) involves a carefully combined application of the mental model, background information on the country and the schooling system and statistical analysis. A series of one-level regression models, some segmented according to historical disadvantage, resulted in likely slope coefficients, which in turn led to a series of simulated policy interventions, each with an estimated impact on average learner performance. The plausible policy interventions with the greatest impact on learner performance were found to be an improvement in teacher morale and discipline (notably, this improvement was linked to just one of the original SACMEQ variables), the reduction of learner repetition, improvements to classroom teaching practice (independently of additional in-service training), more in-service training, better advice and evaluation from the school principal to teachers, physical infrastructure improvements (particularly in rural contexts) and the elimination of textbook/learner ratios below 0.5. Various applications of the HLM confirmed many of the one-level regression findings. Moreover, the HLM added valuable new knowledge about the education production process. The HLM's ability to partition the residual variance across the learner and school levels permitted new insights into the matter of the inequality of quality in the South African schooling system. Specifically, the relative magnitudes of the between-school inequalities and the within-school inequalities could be observed and discussed. The HLM also allowed for the examination of interesting cross-level effects. Specifically, the effect of very large classes of over 55 learners was found to exert a highly destructive effect on the typical positive association between SES and learner performance. In other words, an important class size threshold not observed in the one-level model was detected by the HLM.

The sixth step involved taking the simulated policy interventions from the previous step one step further, by assessing the cost effectiveness of those interventions which clearly involved budgetary reprioritisation. For this step, there was little guidance from existing production function texts, as most such analyses stop short of an analysis of budget implications. A simple set of cost effectiveness computations was performed which revealed that procuring more textbooks and investing in more in-service training of teachers were the most cost-effective means of raising learner performance. Importantly, these two interventions are two of six interventions involving major budgetary shifts. Step five had already pointed towards interventions such as encouraging better evaluation and advice practices amongst school principals that imply little or no major budgetary reprioritisation. Very importantly, one potentially powerful intervention, the reduction of learner repetition, has *negative* cost implications.

The seventh step involved formulating recommendations for future questionnaires. The whole production model relies heavily on how effectively the questions in the questionnaires capture the various inputs. The modelling process itself can reveal important shortcomings in the questionnaires, and it seems efficient to finish the analysis with a set of recommendations in this regard.

Three salient findings on the basis and methodology of school production modelling follow:

- Despite relatively strong arguments that the typical input-output regression model on school production has serious limitations in terms of its ability to yield meaningful policy information, it continues to add valuable information to the existing stock of knowledge of what works and what does not work in a schooling system. The addition to the stock may be more piecemeal than one would wish, and a single outstanding model that captures school production integrally and comprehensively may remain elusive, but the model does nonetheless play an important role in a field where there would otherwise be even less policy direction.

- The hierarchical linear model (HLM) variant of the typical regression model is conceptually powerful, in particular in its categorisation of inequality within the schooling system. However, the HLM is still somewhat cumbersome. As this model develops further, both in terms of its theory and software application, it can be expected to form an increasingly important analysis tool within the economics of education field.

- School production analyses typically lead to findings which, whilst potentially useful, are not sufficiently processed to be easily understood by policymakers. Specifically, these findings lack an adequate cost-effectiveness framework that allows the policymaker to assess the practical implications of the findings and, above all, the relative strengths of the various policy implications being proposed. Rather than hope that the policymaker will herself undertake the cost-effectiveness analysis, the production function analyst should himself add this final aspect of the analysis. This is not a complex addition, yet it adds much policy relevance to the analysis.

BIBLIOGRAPHY

Baker, B.D. (2001). Can flexible non-linear modeling tell us anything new about educational productivity? *Economics of Education Review*, 20: 81-92.

Barbosa, M.E.F. & Fernandes, C. (2001). *A escola brasileira faz diferença? Uma investigação dos efeitos da escola na proficiência em Matemática dos alunos da 4a série*. Curitiba. Available from: <http://www.mec.es/cide/rieme/documentos/barbosa/Barbosa.pdf> [Accessed 9 July 2004].

Barbosa, M.E.F., Fernandes, C., Dos Santos, D. et al (2000). *Modelagem do SAEB-99: Relatório técnico: Análise descritiva dos dados*. Obtained through communication with INEP.

Barbosa, M.E.F., Fernandes, C., Dos Santos, D. et al (2000a). *Modelagem do SAEB-99: Relatório técnico: Redução da dimensão da base de dados do SAEB99*. Obtained through communication with INEP.

Bell, C. (2002). Development economics. In J. Eatwell (ed.).*The new Palgrave: A dictionary of economics.* New York: Houndmills.

Beltrão, K.I., Leite, I.C. & Ferrão, M.E. (2001). *O ambiente escolar no desempenho acadêmico do aluno: Criação de uma escala a partir do SAEB-99*. Available from: <http://www.dmi.ubi.pt/~meferrao/GOM.pdf> [Accessed 9 July 2004].

Berk, R.A. (2004). *Regression analysis: A constructive critique.* Thousand Oaks: Sage.

Bifulco, R. & Bretschneider, S. (2001). Estimating school efficiency: A comparison of methods using simulated data. *Economics of Education Review*, 20: 417-429.

Blalock, H.M. (1979). *Social statistics.* Singapore: McGraw-Hill.

Boston College. (2001). *TIMSS 1999: User guide for the international database.* Boston. Available from <http://www.timss.bc.edu/timss1999i/data/bm2_supplement1.pdf> [Accessed 4 July 2004].

Brown, M. (2002). Cobb-Douglas functions. In Eatwell, J. (ed.). *The new Palgrave: A dictionary of economics.* New York: Houndmills.

Bryk, A.S. & Raudenbush, S.W. (1992). *Hierarchical linear models.* Newbury Park: Sage.

Burstein, L. (1980). The role of levels of analysis in the specification of education effects. In Drebeen, A. & Thomas, J.A. (eds.). *The analysis of educational productivity: Volume 1.* Cambridge (Massachusetts): Ballinger.

Case, A. & Deaton, A. (1999). School inputs and educational outcomes in South Africa. Quarterly Journal of Economics, 114(3): 1047-1084.

Case, A. & Yogo, M. (1999). *Does school quality matter? Returns to education and the characteristics of schools in South Africa.* Washington: National Bureau of Economic Research. Available from: <http://papers.nber.org/papers/w7399.pdf> [Accessed November 2005].

Chisholm, L. (2003). *The politics of curriculum review and revision in South Africa.* Pretoria: HSRC. Available from: <http://www.hsrc.ac.za/research/output/outputDocuments/2285_ChisholmPoliticsofCurriculumReview.pdf> [Accessed July 2005].

Crouch, L. & Perry, H. (2002) *Determinants of learning and learning change in DDSP schools' Grade 3 numeracy and literacy assessment results.* District Development Support Programme, Pretoria. Unpublished mimeograph.

Crouch, L. & Mabogoane, T. (1998). When the residuals matter more than the coefficients: An educational perspective. *Journal of Studies in Economics and Econometrics*, 22(2).

Department of Education. (2003) *National report on Systemic Evaluation.* Pretoria.

Department of Education. (2005) *Strategic Plan: 2005-2010.* Pretoria. Available from <http://www.education.gov.za> [Accessed 14 July 2005].

Dolata, S., Ikeda, M. & Murimba, S. (2004). Different pathways to EFA for different school systems. *IIEP Newsletter*, XXII(1). Available from: <http://www.unesco.org/iiep>.

The Economist (2004). Economics focus: Signifying nothing? January 31st-February 6th 2004: 370(8360). London.
Ferrão, M.E., Beltrão, K.I., Fernandes, C. et al (2001). *O SAEB – Sistema Nacional de Avaliação da Educação Básica: objetivos, características e contribuições na investigação da escola eficaz.* Revista Brasileira de Estudos de População. Available from <http://www.abep.org.br> [Accessed 1 January 2005].
Fertig, M. (2003). *Who's to Blame? The determinants of German students' achievement in the PISA 2000 study*. *IZA DP*, 739. Bonn. Available from <ftp://repec.iza.org/RePEc/Discussionpaper/dp739.pdf> [Accessed 9 July 2004].
Fuss, M.A. (2002). Production and cost functions. In Eatwell, J. (ed.). *The new Palgrave: A dictionary of economics.* New York: Houndmills.
Fuss, M., McFadden, D. & Mundlak, Y. (1978). A survey of functional forms in the economic analysis of production. In Fuss, M. & McFadden, D. (eds.). *Production economics: A dual approach to theory and application: Volume 1.* Amsterdam: North-Holland.
Gillis, M., Perkins, D.H., Roemer, M. & Snodgrass, D.R. (1983). *Economics of development.* New York: W.W. Norton.
Glahe, F.R. & Lee, D.R. (1989). *Microeconomics: Theory and applications.* San Diego: Harcourt Brace Jovanovich.
Glewwe, P., Kremer, M. & Moulin, S. (2000). *Textbooks and test scores: Evidence from a prospective evaluation in Kenya.* Washington, World Bank. Available from: <http://post.economics.harvard.edu/faculty/kremer/webpapers/Textbooks_Test_Scores.pdf> [Accessed 5 June 2004].
Goldstein, H. (1995). *Multilevel statistical models.* London: Edward Arnold.
Gujarati, D.N. (2003). *Basic econometrics.* New York: McGraw-Hill.
Gupta, J. (2001). *Our simmering planet: What to do about global warming.* London: Zed Books.
Häkkinen, I., Kirjavainen, T. & Uusitalo, R. (2003). School resources and student achievement revisited: New evidence from panel data. *Economics of Education Review*, 22: 329-335.
Hanushek, E.A. (1979). Conceptual and empirical issues in the estimation of educational production functions. *Journal of Human Resources*, XIV(3): 351-388.
Hanushek, E.A. (1995). Interpreting recent research on schooling in developing countries. The World Bank Research Observer, 10(2): 227-246.
Hanushek, E.A. (2002). *The failure of input-based schooling policies.* Washington: National Bureau of Economic Research. Available from: <http://papers.nber.org/papers/w9040.pdf> [Accessed November 2005].
Hanushek, E.A. & Luque, J.A. (2003). Efficiency and equity in schools around the world. *Economics of Education Review*, 22: 481-502.
Harbison, R.W. & Hanushek, E.A. (1992). *Educational performance of the poor: Lessons from rural Northeast Brazil.* Oxford: Oxford University Press.
Hox, J.J. (1995). *Applied multilevel analysis.* TT-Publikaties, Amsterdam. Available from <http://multilevel.ioe.ac.uk/publref/papers/amaboek.pdf> [Accessed 20 December 2004].
Hungi, N. (2005). *Factors influencing Standard 6 pupils achievement in Kenya: A multilevel analysis.* Harare, SACMEQ. Available from: <http://www.sacmeq.org> [Accessed 15 April 2005].
INEP (2002). *Saeb 2001: Novas Perspectives.* Brasilia. Available from <http://www.inep.gov.br/download/saeb/2001/Miolo_Novas_Perspectivas2001.pdf> [Accessed 12 September 2004]
INEP (2004). *Resultados do SAEB 2003: Brasil.* Brasilia. Available from <http://www.inep.gov.br> [Accessed 1 January 2005].
Jimenez, E. & Patrinos, H.A. (2003). Curious George: The enduring Psacharopoulos legacy on the economics of education in developing countries. *Economics of Education Review*, 22: 451-454.

Johnson, R.A. & Wichern, D.W. (2002). *Applied multivariate statistical analysis*. Upper Saddle River: Pearson Education.

Kaplan, D. & Elliot, P.R. (1997). A model-based approach to validating education indicators using multilevel structural equation modelling. *Journal of Educational and Behavioral Statistics,* 22(3), 323-348.

King, E.M., Orazem, P.F. & Paterno, E.M. (1999). *Promotion with and without learning: Effects on student dropout.* Washington: World Bank. Available from: <http://www.worldbank.org/research/pdffiles/IEERno18.pdf> [Accessed May 2005].

Koopmans, T.C. (1965). Analysis of production as an efficient combination of activities. In Koopmans, T.C. (ed.) *Activity analysis of production and allocation.* New York: John Wiley: 33-97.

Kulpoo, D. & Soonarane, M. (2005). *Variance in pupil reading and mathematics performance: Why is it so large in Mauritius?* Harare: SACMEQ. Available from: <http://www.sacmeq.org> [Accessed October 2005].

Lee, J.-W. & Barro, R.J. (2000). *Schooling quality in a cross-section of countries.* Seoul: Korea University. Available from: <http://econ.korea.ac.kr/prof/jwlee/files/school%20quality(1).pdf> [Accessed July 2005].

Lockheed, M.E. & Longford, N.T. (1989). *A multilevel model of school effectiveness in a developing country*. World Bank discussion papers, 69. Washington: World Bank.

Mandel, E. (2002). Marx, Karl Heinrich. In Eatwell, J. (ed.). *The new Palgrave: A dictionary of economics.* New York: Houndmills.

Martin, M.O. (2005). *TIMSS 2003 user guide for the international database.* Chestnut Hill: Boston College. Available from: <http://timss/bc/edu> [Accessed July 2005].

Marx, K. (1991). *Capital: A critique of political economy: Volume three.* London: Penguin.

Mingat, A. & Jee-Peng Tan (1998). *The mechanics of progress in education: Evidence from cross-country data.* Washington, World Bank. Available from: <http://www.worldbank.org> [Accessed 15 November 2003].

Monk, D.H. (1990). *Educational finance: An economic approach.* New York: McGraw-Hill.

National Treasury (2005). *Provincial budgets and expenditure review: 2001/02-2007/08.* Pretoria. Available from: <http://www.treasury.gov.za> [Accessed October 2005].

Penrose, P. (1993). *Planning and financing sustainable education systems in Sub-Saharan Africa.* London: Department for International Development. Available from: <http://www.dfid.gov.uk/pubs/files/planfinafricaedpaper07.pdf> [Accessed November 2005].

Pradhan, S. (1996). *Evaluating public spending: A framework for public expenditure reviews.* Washington, World Bank. Available from: <http://www.worldbank.org/participation/s&pa/spathematic.htm> [Accessed 5 October 2004].

Psacharopoulos, G. (1993). *Returns to investment in education: A global update.* Washington, World Bank. Available from: <http://www.worldbank.org> [Accessed 18 October 2004].

Psacharopoulos, G. (1995). *The profitability of investment in education: Concepts and methods.* Washington, World Bank. Available from: <http://www.worldbank.org> [Accessed 18 October 2004].

Rosen, S. (2002). Human capital. In Eatwell, J. (ed.). *The new Palgrave: A dictionary of economics.* New York: Houndmills.

Ross, K.N., Saito, M., Dolata, S. et al (2004). *Chapter 2: The conduct of the SACMEQ II project.* IIEP. Unpublished technical report accompanying the SACMEQ II data set.

Saïdi, M. (2001). How ministry organization impacts educational policy design and management. In Yasmin Haq (ed.), Organization of Ministries of Education. Paris: IIEP: 108-127. Available from: <http://unesdoc.unesco.org/images/0013/001307/130735e.pdf> [Accessed November 2005].

Samuelson, P.A. (1983). *Foundations of economic analysis.* Harvard University Press.

Sen, A. (2001). *Development as freedom.* New York: Alfred A. Knopf.
Skinner, E. (2002). Smith, Adam. In Eatwell, J. (ed.). *The new Palgrave: A dictionary of economics.* New York: Houndmills.
Smith, A. (1991). *The wealth of nations.* London: Everyman's Library.
Stata (2001). *Stata 7: Reference Q-St.* College Station: Stata Corporation.
Stiglitz, J. (2002). *Globalization and its discontents.* London: Penguin.
Sydsæter, K. & Hammond, P. (2002). *Essential mathematics for economic analysis*. Harlow: Pearson Education.
UNESCO (2001). *Primer estudio internacional comparativo: Sobre lenguaje, matemática y factores asociados, para alumnos del tercer y cuarto grado de la educación básica*. Santiago. Available from: <http://www.unesco.cl/medios/biblioteca/documentos/informe_laboratorio_informe_tecnico.pdf> [Accessed July 2005].
UNESCO (2003). *Education for All global monitoring report 2003/4: Gender and Education for All: The leap to equality.* Paris.
UNESCO (2005). *Education for All global monitoring report 2005: The quality imperative.* Paris. Available from: <http://www.unesco.org> [Accessed July 2005].
United Nations (2005). *The Millenium Development Goals report 2005*. New York. Available from: <http://unstats.un.org/unsd/mi/pdf/MDG%20Book.pdf> [Accessed October 2005].
Upton, G. & Cook, I. (2002). *Dictionary of statistics.* Oxford: Oxford University Press.
Van der Berg, S. & Burger, R. (2003). *Education and socio-economic differentials: A study of school performance in the Western Cape.* University of Stellenbosch.
Vinjevold, P. & Crouch, L. (2001). *Literacy, numeracy and policy: Towards assessing learning in Grade 3.* District Development Support Programme, Pretoria. Unpublished mimeograph.
Welmond, M. *Teacher identity in Africa: The case of the Republic of Benin.* Unpublished Doctoral thesis. Stanford University, 1999.
Willms, J.D. & Somers, M. (2001). Family, classroom and school effects on children's educational outcomes in Latin America. *School effectiveness and school improvement*, 12(4): 401-445.

Organisational websites consulted

Brazil: Ministry of Education. <http://www.mec.gov.br>
Human Sciences Research Council (HSRC). <http://www.hsrc.ac.za>
International Institute for Education Planning (IIEP). <http://www.unesco.org/iiep>
Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación. <http://llece.unesco.cl>
National Institute of Educational Studies and Research (INEP). <http://www.inep.gov.br>
Organisation for Economic Co-operation and Development (OECD). <http://www.oecd.org>
South Africa: Department of Education. <http://www.education.gov.za>
Southern and Eastern African Consortium for Measuring Educational Quality <http://www.sacmeq.org>
TIMSS. <http://timss.bc.edu>
World Bank <http://www.worldbank.org>

Appendix A MONITORING PROGRAMME DETAILS

1 South Africa's Systemic Evaluation

History of the programme

In the years following the advent of democracy in 1994, a large part of the education challenge was to amalgamate the disparate sub-systems that had characterised apartheid education. The crucial Grade 12 exit point examinations were standardised across the country, and work began on nationally standardised learner assessment systems at grades below Grade 12. The first such pre-Grade 12 system to be implemented by the Department of Education was the Systemic Evaluation. It was a sample-based assessment programme that had its first run in 2001, when a representative sample of Grade 3 learners was assessed. The analysis report was released in 2003 (Department of Education 2003). In 2004, a new run took place, this time focussing on Grade 6. The Systemic Evaluation programme is still at a formative stage, in terms of its methodology, and in terms of its impact on policymakers.

Legal underpinnings

The National Education Policy Act (Act 27 of 1996), also known as NEPA, describes the important monitoring functions of the Department of Education with respect to the education system of South Africa, the pre-tertiary part of which is managed by nine provincial governments. The 1998 'Assessment Policy in the General Education and Training Band, Grades R to 9 and ABET' (Regulation 1718 of 1998) was issued as a NEPA regulation, and deals, amongst other things, with the establishment of the national Systemic Evaluation programme. The programme measures learner performance at the Grades 3, 6 and 9 levels, through the use of a 'nationally representative sample of learners and learning sites'. The critical importance of the Systemic Evaluation is underlined by the point, made in the regulation, that there will be no universal learner performance certificate below the Grade 9 level.

Institutional arrangements

The approach in the 2001 Systemic Evaluation run was to outsource the data analysis work to RIEP, a research group within the University of Free State, and the Human Sciences Research Council (HSRC), which maintains a permanent research function called Assessment Technology and Education Evaluation. The HSRC is in fact a parastatal that reports to government's Department of Arts and Culture.

Programme scope

The 2001 run in the Systemic Evaluation involved the collection of data from 1,300 primary schools. These schools made up a stratified random sample of schools. The 50,000 Grade 3 learners from these schools studied constituted about 5% of the population of all Grade 3 learners. Only public schools, and not private schools, were covered (private schools account for only 2% of enrolment, however). Questionnaires were targeted at around 2,500 educators, 50,000 learners and parents, 1,300 school principals, and 150 officials working within the education authorities. The 50,000 learners were tested in listening comprehension, literacy, numeracy and life skills.

Parallel programmes

Over the past ten years, South Africa has participated in three major international programmes gathering performance and input data from a sample of learners in the system.

South Africa participated in the 1995, 1999 and 2003 runs of the TIMSS programme. In 1995, the focus in South Africa was on Grade 12, and not on the earlier grades covered by TIMSS in some other countries. In 1999 the focus of TIMSS worldwide was exclusively on Grade 8. In 2003, South Africa participated at the Grade 8 level (and not at the other TIMSS level, which was Grade 4). Some of the background data for South Africa was missing from the international dataset due to concerns around the data quality (Boston College 2001, 8-6; Martin 2005).

South Africa took part in the 1999 international Monitoring Learning Achievement (MLA) programme of UNESCO, where the focus was on Grade 4, though not in the later Grade 8 run.

In 2000, South Africa participated in the SACMEQ programme for the first time. The focus was on Grade 6, and just under 200 schools were included in the sample.

Sampling methodology

The approach in the 2001 Systemic Evaluation was to stratify the population according to region or district, to then select schools randomly within each stratum (but only from schools with at least 30 learners), and finally to select randomly a maximum of 40 Grade 3 learners per selected school (Department of Education 2003, 9).

Data collection methodology

The methodology was tested in a pilot run that preceded the main run. In the main run of the 2001 Systemic Evaluation, field workers administered the testing of learners and the completion of the learner questionnaires, but all the other questionnaires were completed by respondents on their own. The field workers were officials of the provincial departments of education. The provincial departments of education did all the scoring of learner tests.

Emerging policy information

The 2001 Systemic Evaluation report provides profiles of the learner scores, and analyses associations between schooling inputs and home background factors, on the one hand, and learner performance, on the other, through a series of non-integrated regression models that each analyse one aspect of the system at once.

The report divides 27 input indicators up into three categories, as follows:

ACCESS INDICATORS
Parents' level of education
Availability of resources at home
Nutrition of the learners
Early childhood development
Learner:educator ratios
Utilisation of resource centres by learners
Repetition rate
Number of years to complete phase
Pass rates
Language of learning and teaching at school
Accessibility of school
EQUITY INDICATORS
Private contributions and utilisation of funds
Assistance from the Department
Educator qualifications
Functioning of SGBs

Discipline, safety and learning atmosphere
QUALITY INDICATORS
School facilities
Satisfaction rates of stakeholders
Attendance rates, contact time, time on task
Learning and teaching materials
Teaching practices
INSET and SGB training
Record keeping
Educator morale and attitude
School management and leadership
Assessment of learners and feedback procedures
Homework

Data availability

The data from the 2001 Systemic Evaluation is not freely available.

Programme advocacy

The Systemic Evaluation is still a relatively low profile government programme. It has a very limited Web presence.

Future trajectory

The latest Department of Education strategic plan indicates that the next nationwide run of the programme (following the 2004 run focussing on Grade 6), will be a run in 2008 focussing on Grade 8 (Department of Education 2005, 75).

2 Brazil's SAEB

History of the programme

In 1985 democracy was restored in Brazil after decades of mostly repressive military rule. Education and social services in general enjoyed a new prominence. In 1990, SAEB (full name *Sistema Nacional de Avaliação da Educação Básica*, official translation National Basic Education Assessment System) was launched by the federal government as part of a drive for greater central leadership in the quest for better educational equality and quality. The programme involves the collection of detailed input and output data relating to a sample of learners every two years. The 2003 run of the programme was the seventh. Extensive information on SAEB is available from the website of INEP, the National Institute of Educational Studies and Research.

Legal underpinnings

Much of the legislation currently underpinning SAEB was promulgated after the commencement of the programme. The National Education Guidelines and Framework Law (Law 9394 of 1996) formally establishes the responsibility of the federal government to monitor the schooling system.

Law 9448 of 1997 establishes the institutional and financial framework for INEP, and makes INEP the organisation responsible for managing SAEB. INEP was established in the 1930s, but the 1997 law provided a much needed demarcation of the semi-autonomous nature of the organisation. Essentially, INEP is a publicly funded body providing a range of technical and research support functions for the Ministry of Education. INEP's area of work is wide and includes the management of, amongst other things, the annual national school census, the ENEM exit point examinations for Grade 11 learners and Brazil's participation in international learner assessment programmes, in particular PISA and Laboratorio.

Brazil's national education plan, approved by the federal congress in 2000, makes explicit reference to SAEB as a monitoring programme supporting school improvement.

Institutional arrangements

SAEB is managed by a directorate within INEP dedicated to this programme. Moreover, there is extensive collaboration with a number of research institutions, including the PUC Rio University and the National School of Statistical Sciences, in the production of the SAEB data analyses.

Programme scope

SAEB focuses on inputs and outputs at the Grades 4, 8 and 11 levels of the schooling system. The 2001 run of SAEB involved the collection of input data and output data (through standardised tests) from around 360,000 learners. Moreover, around 18,000 educators and 7,000 school principals provided data on school inputs. The figure of 360,000 learners represents a sample of around 2% of all learners in the three grades. Standardised tests cover language proficiency (the only language considered is Portuguese) and mathematics.

Parallel programmes

Although Brazil is not a member of the OECD, the country participated in the 2000 and 2003 runs of PISA as one of ten non-OECD participants. PISA focuses on fifteen year old learners, making it a close equivalent of the Grade 11 component of SAEB. Brazil also participates in the Latin American regional programme Laboratorio, which resulted in, amongst other things, the collection of a major dataset of school inputs and outputs in thirteen Latin American countries in 1997.

Sampling methodology

INEP (2002, 53) describes in some detail how the sample for the 2001 run of SAEB was constructed. The population is described as follows:

> All learners enrolled during 2001 in one of the three focus grades (Grades 4, 8 and 11), in the permanent schools reflected in the 1999 School Census, with the exception of learners in the federal district, learners in rural areas, and learners in multi-grade classes. Included in the population, however, are the Grade 4 learners in rural schools in the states of the North-East Region, and the states of Minas Gerais and Mato Grosso do Sul.

Clearly, reliable data from the School Census, a massive operation covering 160,000 institutions and 55 million learners, is a prerequisite. The exclusion of many rural school learners from the 2001 SAEB population is a cost reduction measure, and it is argued that this is permissible given the results of some statistical analysis work (the details of this are not explained). The reason for the exclusion of the federal district learners is less clear.

Stratification of the population occurred along five dimensions:

- Grade – there are three possibilities: Grades 4, 8 and 11.

- State – there are 26 states.

- Owner of school – there are three possibilities: state, municipal or private (16% of schools are private).

- Location of school – there are two possibilities: state capital, or elsewhere (in the case of Grade 4, ‘elsewhere’ was divided into urban and rural, creating three possibilities).

- Number of shifts in the school – there are two possibilities: 1 to 2 shifts or more than 2 shifts. (The average number of shifts per school is 1.6. This is indicative of the particularly high prevalence of multi-shift schools in Brazil, and, indeed, across Latin America.)

The dimensions resulted in a total of 438 possible strata. Extremely small strata were excluded, resulting in a final number of 389 strata, all of which would have some learners tested. After examination of the variance of test results obtained in the 1999 run of SAEB, it was decided that 300 learners per stratum would be selected for each of the two learning areas (Portuguese language and mathematics) and each of the three focus grades, i.e. 600 learners would be selected per stratum per grade (it was assumed that non-attendance amongst learners would result in a reduction from 600 to around 500 learners).

The number of classes to be selected in each stratum and each grade in order to obtain the 600 learner target was determined. Selection of classes occurred in such a way that in larger schools two classes were tested, whilst in smaller schools one class was tested. Sequential Poisson Sampling was used to reduce the number of schools that had to be visited. Through this technique, the grand total of schools to be visited was reduced from 8,880 to 7,073 (8,880 schools would have been visited if only one grade in each school had been tested). It is important to note that in the majority of schools, only one grade was tested. This was necessary in order to maintain the randomness of the selection.

Data collection methodology

INEP employs a service provider, which is currently Fundação Cesgranrio, to implement SAEB. Testing and questionnaire completion at the school level is managed by specially trained fieldworkers from outside the school community. Starting with the 2003 run of SAEB, more intense assistance is to be offered to Grades 4 and 8 learners completing the questionnaires, given problems experienced in the past relating to inadequate levels of literacy amongst learners.

In technical documentation relating to the 1999 run of SAEB (see Barbosa *et al*, 2000), fairly comprehensive explanations of data anomalies and data normalisation processes are provided. What stands out is the high percentage of learner records with no data at all from the learner questionnaire – the figure was around 20% for the various school grades involved. Around 50% of learners, educators and school principals provided a complete set of questionnaire responses.

Emerging policy information

INEP itself publishes reports after each SAEB run that focus strongly on examining the test results (see for example INEP 2003). The INEP reports place some emphasis on inputs, but not through any rigorous modelling. Econometric modelling is carried out by external research organisations.

Data availability

The SAEB data is not available on the Internet. INEP is currently in the process of tightening up the criteria whereby analysts outside INEP gain access to the data. The SAEB data is potentially very sensitive from a political point of view, especially considering that learner performance has on average deteriorated substantially and significantly since 1995 (INEP 2003, 25). The deterioration has to be seen in the context of enormous progress in access to education, implying an increasing proportion of disadvantaged learners in the system.

Programme advocacy

There is a major public relations component to SAEB. Television programmes, the Internet and radio and television interviews are used to publicise the programme. The SAEB section of the INEP website includes a month by month list of SAEB activities, which continue even during years when SAEB is not run. SAEB staff participate in seminars on an ongoing basis to raise awareness about SAEB and deepen the understanding of the education community in this kind of evaluation. The test compilation process has led to the creation of a network of educators involved in building the item bank of questions used in SAEB. This process has also prompted enquiry into details of the school curriculum.

Future trajectory

In 2005, INEP announced that major changes would occur with respect to SAEB (see INEP website). The core programme would continue, though under a different name: Avaliação Nacional da Educação Básica (Aneb) – National Evaluation of Basic Education. However, linked to the core programme would be a much larger programme that would cover all learners in public schools in Grades 4 and 8. These learners amount to around five million learners in 43,000 schools. The larger programme would be known as Avaliação Nacional do Rendimento Escolar (Anresc) – National Evaluation of School Performance. During 2005, Anresc would focus on performance in reading. During 2006, the focus would shift to mathematics. The programme would be run annually. The key output from Anresc would be an aggregate score for every public school offering Grade 4 or 8. This score would assist in the education planning process.

3 SACMEQ

History of the programme

SACMEQ (the Southern and Eastern African Consortium for Measuring Educational Quality) was launched in 1995 after four years of planning involving a number of Southern and Eastern African Ministries of Education and the UNESCO institute IIEP. Currently, SACMEQ includes 15 countries (mainland Tanzania and Zanzibar are included as two countries, the other countries being Botswana, Kenya, Lesotho, Malawi, Mauritius, Mozambique, Namibia, Seychelles, South Africa, Swaziland, Uganda, Zambia, and Zimbabwe). SACMEQ's mission is to conduct research into quality in basic education, to develop technical skills amongst education planners, and to provide Ministries with meaningful policy guidance flowing from the research. SACMEQ I involved gathering of data during the 1995 to 1998 period, and SACMEQ II involved data collection mostly during 2000. The 2000 run of SACMEQ did not cover Zimbabwe.

Legal underpinnings

Given that this is an international programme, there is no national legal framework for the programme. SACMEQ overlaps to some degree with the Southern African Development Community (SADC) – Uganda and Kenya are not part of SADC, however, and the war-affected SADC members Angola and Congo have not been included in SACMEQ yet.

Institutional arrangements

The funding and governance of the programme is shared by the Ministries involved and the IIEP. Donor funding has also been used. SACMEQ has a programme office in Harare.

Programme scope

SACMEQ II involved the collection of schools data by means of learner, educator and school principal questionnaires, and reading and mathematics tests. Altogether some 2,500 schools, 6,400 educators and 45,000 Grade 6 learners were involved. In the case of South Africa, the sample was around 1% of the population. But this proportion varied greatly from country to country – in the case of Namibia it was a whole 30%.

Parallel programmes

SACMEQ countries that have also participated in the TIMSS programme are South Africa and Botswana.

Sampling methodology

Ross *et al* 2004 describe in detail how the sample for the SACMEQ II evaluation was constructed. The 'desired target population' is described as follows:

> All pupils at Grade 6 level in 2000 (at the first week of the eighth month of the school year) who were attending registered mainstream primary schools.

The desired target population consisted of 3.4 million learners. The sampling frame was obtained using the most recent school censuses of the respective countries. From the desired target population, very small schools were excluded in order to create the 'defined target population'. This was a cost saving measure. The cut-off for considering a school as being small was a maximum of 10, 15 or 20 learners in the whole school, with the option depending on the country concerned. The exclusion meant that around 3.6% of learners overall were excluded (this figure became 4.4% in the case of South Africa).

Stratification occurred along only one dimension, namely region (within which country is implicit). In total, there were 116 regions (each of the nine South African provinces was considered a region). Despite the use of region for stratification, specifications around adequate sample size concentrated on number of schools and learners *per country* to be included, not per region. It can thus be assumed that country samples are adequate, but not region samples.

For the country sample size to be adequate, it had to be at least as good as the following: A simple random sample of just learners (i.e. regardless of school) consisting of 400 learners, where the standard error for the learner performance score would not be greater than 0.1 standard deviations using a 95% confidence interval. This is a benchmark set by the International Association for the Evaluation of Educational Achievement.

Given that selection effects result in learners not being randomly distributed across schools, it was inevitably necessary to sample more than 400 learners per country. How many more depended on what the SACMEQ documentation refers to the 'rho' statistic. This statistic indicates the degree of variability between schools or, put differently, the degree to which the distribution of learners across schools is not random. Various previous studies as well as the SACMEQ I dataset were used to determine rho for each country.

The decision was taken to study an equal *number* of Grade 6 learners in each school as opposed to studying an equal *proportion* of all Grade 6 learners in each school. The number of learners per school was set at 20. This learner number, plus the rho for each country, then allowed for the determination of a minimum number of schools to be selected in each country. Specially designed software and sample design tables peform the calculation needed to obtain this minimum number of schools which, in the case of a country with South Africa's rho value, turned out to be 172 schools. And with 20 learners in each school, this resulted in a minimum acceptable sample size of 3,440 learners. To complete the sampling description, it should be explained that schools were selected in a way that increased the probability of selection for larger schools, relative to their larger size.

Data collection methodology

In each country, 15 to 50 data collectors were given a three-day training course. All data collection at schools was administered by the data collectors, including learner tests, teacher tests, and the completion of questionnaires. Data normalisation included the determination of weights at the level of learner records and the integration of the output scales of the various countries into one scale.

The SACMEQ technical documentation describes the extensive data verification and normalisation that took place following the collection of the questionnaires and test results. Considerable manual checking of questionnaires occurred. In some countries, data capturing occurred twice in order to facilitate the detection of capturing errors. There were several data verification cycles involving inspection and cleaning of the data within the country, submission of data to the IIEP, inspection and commenting at the IIEP, return to the country, further cleaning in the country, and so on.

A considerable amount of recoding of the data occurred. This recoding generally involved the generation of new variables after several codes or values which seemed similar had been collapsed into one code. For example, the precise number of pencils owned by a learner was collapsed into (1) having at least one pencil and (2) having no pencil. In the case of learners, 75 original variables were recoded into 58 new variables. Similar recoding occurred for educator and school principal data.

Rules were established to impute values programmatically where there were missing values in the dataset. For example, if learner questionnaire values were missing in fewer than 15% of learner records in a class, then the class mean or class mode (depending on the data type) was to be inserted. If the percentage was equal to or greater than 15%, then the school mean or mode was to be used. A limit was established: If the number of records for which there was missing values in a variable was greater than or equal to 15% for a country, no imputation would occur for the variable and the country. Unfortunately, the SACMEQ data provided to analysts has already been subjected to this imputation process, and the technical documentation does not explain to what degree missing values were replaced by imputed values.

Emerging policy information

The emphasis within the SACMEQ programme is strongly on producing policy relevant findings. SACMEQ II (the 2000 run) involved the production of national reports, all of which linked findings to policy recommendations according to a similar structure. In the South Africa report, 39 recommendations are grouped according to five categories:

- Policy suggestions concerning the monitoring of the implementation of existing education policies

- Policy suggestions that identified established practices that might need to be evaluated and reviewed in the area of policy and planning.

- Suggestions on data to be collected for planning purposes.
- Policy suggestions that called for the Department to have major consultations with communities and experts.
- Suggestions that identified educational policy research programmes for the Department.

Data availability

The SACMEQ II data is available to analysts on request.

Programme advocacy

SACMEQ has recently acquired its own web address, and the Web presence of the programme is strong. A series of academic conferences flowing from the SACMEQ II run is aimed at popularising the programme in academic and government circles.

Future trajectory

The SACMEQ website indicates that SACMEQ III may be run during 2005. However, it is not made clear whether this would be similar to SACMEQ II, would shift the focus to Grade 9, or pursue some other focus.

Appendix B THE SACMEQ QUESTIONNAIRE VARIABLES

Table 36 below contains the following:

- The column on the far left lists the variables as they appear in the original SACMEQ II database. Only the variables with data from the questionnaires are included in the list. The educator questionnaire variables are potentially from two questionnaires: that of the reading educator and that of the mathematics educator. In the case of South Africa, 23% of learners had the same educator for both subjects, so the educator data would be derived from one questionnaire. Variables beginning with ‘X’ were repeated in both the reading and mathematics educator questionnaires. Variables beginning with ‘TR’ applied to the reading educator only, and those beginning with ‘TM’ applied to the mathematics educator only. Questionnaire questions that did not elicit any response, in particular those relating to multiple sessions of classes (this is rare in South Africa), have not been included in the list. Variable names appearing in italics are those where more than 50 missing observation values were imputed within the SACMEQ data normalisation process using existing values (this is explained in section 3).

- The column headed *T* contains a code from the set R, I, O and N to indicate whether the variable contains ratio, interval, ordinal or nominal values.

- The column *Question* contains enough of the text from the questionnaire itself to provide an idea of the meaning of the variable.

- Column *A* indicates with an asterisk whether the variable was somehow transformed when the relationship between the variable and one of the two learner performance scores was examined. In the case of certain variables, we would not expect it to explain performance without some transformation. For example, number of permanent educators and number of temporary educators in the school is less likely to explain performance than the percentage of educators in the school who are permanently employed.

- Column R^2 gives the goodness of fit of a regression model involving only the variable in question (possibly transformed) and learner performance. The highest R^2 of the two obtained using the reading and mathematics score was retained. In the case of the ‘X’ variables, the R^2 yielded by the educator (reading or mathematics) giving the highest value was retained. Where variables contain ratio or interval values, the regression model was bivariate, meaning the single variable was regressed against the reading score. Where variables are ordinal or nominal, dummy variables were created to represent each of the values (minus one, to avoid the dummy variable trap), and the dummy variables were regressed against the scores. A Stata program was created to perform all these regression analyses.

- Column *S* indicates the importance of the variable as an explainer of learner performance on the basis of two stepwise regression analyses. Both stepwise analyses involved the backward selection approach of Stata, and both used 0.0001 as the maximum limit for the *p* value. One of the regression analyses used the reading score as the output, the other the mathematics score. The pool of possible explanatory variables was in both cases the ratio or interval variables, as well as the dummy variables (associated with ordinal or nominal variables), that yielded an R^2 value of 0.10 or higher in the procedure explained in the previous bullet. The dots in column *S* indicate that the variable in question (or at least one of the dummy variables derived from it) appeared in at least one of the two models obtained using the stepwise method.

- Column *New* indicates where an original variable was utilised in the formation of one of the new variables. If the name of the new variable is in brackets, the new variable is derived from more than one of the original variables. Details on the new variables are provided in the next appendix.

Table 36: Original SACMEQ variables and their questions

Original	*T*	*Question*	*A*	*R^2*	*S*	*New*
LEARNER QUESTIONNAIRE						
PBDAY	I	What is your date of birth? > Day	*	12	•	
PBMONTH	I	> Month	*	12	•	(learner_age)
PBYEAR	I	> Year	*	12	•	(learner_age)
PSEX	N	Are you a boy or a girl?		1		learner_gender
PENGLISH	O	Do you speak English outside school?		9		
PSTAY	N	Where do you stay during the school week? [parents, relatives, boarding school, by myself]		5		
PBOOKSHM	O	How many books are in the place (home) where you stay during the school week?		12		
PPOS01	O	Which of the following… in the place (home) where you stay during the school week? > Daily newspaper		4		
PPOS02	O	> Weekly or monthly magazine		7		
PPOS03	O	> Radio		1		
PPOS04	O	> TV set		9		
PPOS05	O	> Video cassette recorder (VCR)		14		
PPOS06	O	> Cassette player		14	•	(learner_ses)
PPOS07	O	> Telephone		19		(learner_ses)
PPOS08	O	> Refrigerator/freezer		21	•	(learner_ses)
PPOS09	O	> Car		9		
PPOS10	O	> Motorcycle		0		
PPOS11	O	> Bicycle		3		
PPOS12	O	> Piped water		12		
PPOS13	O	> Electricity (mains, generator, solar)		14		
PPOS14	O	> Table to write on		8		
PLIGHT	N	What is the main source of lighting by which you can read in… (home)? [Fire, … electric lighting…]		16		(learner_ses)
PLIVS1	R	Approximately how many of the following livestock are owned by the household…? > Cattle		1		
PLIVS2	R	> Sheep		1		
PLIVS3	R	> Goats		0		
PLIVS4	R	> Horses		0		
PLIVS5	R	> Donkeys		0		
PLIVS6	R	> Pigs		1		
PLIVS7	R	> Chickens		0		
PLIVS8	R	> Other livestock that can be sold for food		0		
PMEAL1	O	How often do you normally eat each of the following meals? > Morning meal/breakfast		1		(daily_meals)
PMEAL2	O	> Lunch		4		(daily_meals)
PMEAL3	O	> Evening meal/supper		5		(daily_meals)
PMOTHER	O	What is the highest… education that your mother… has completed? [… some primary,… university…]		16		(parent_educ)

Original	*T*	*Question*	*A*	*R^2*	*S*	*New*
PFATHER	O	What is the highest level of education that your father… has completed? [… some primary,… university…]		11		(parent_educ)
PFLOOR	N	What is the surface… of the floor… where you stay [… earth… carpet…]		28	•	(learner_ses)
PWALL	N	What are the outside walls… (home)… mostly made of? [cardboard… bricks…]		14	•	(learner_ses)
PROOF	N	What is the roof of… (home) where you stay during the school week mostly made of? [cardboard…tiles…]		10		
PABSENT	R	On how many school days were you absent (not present at school) during the month of…?		1		
PABWHY1	N	What was the reason for your absence? > I was not absent		2		
PABWHY2	N	> I was ill		0		
PABWHY3	N	> Family reasons (for example, funerals, weddings, illness etc.)		0		
PABWHY4	N	> I had to work		2		
PABWHY5	N	> Bad weather or floods		1		
PABWHY6	N	> I was not allowed to go to school because school fees were not paid.		2		
PABWHY7	N	> Other reasons		0		
PREPEAT	O	How many times have you repeated a grade since you started school? [… never…,… three or four times]		16	•	repetition
PREPEAT6	N	Are you repeating Grade 6 this year?		6		
PBORROW	N	Are you allowed to take… books home… school library… book corner…? [… no library… yes]		5		
PMAT01	R	How many of the following items do you have to work…? > Exercise books… marked by teacher		1		
PMAT02	R	> Notebooks (which are not marked by the teacher)		1		
PMAT03	R	> Pencils		7		
PMAT04	R	> Pencil sharpeners		0		
PMAT05	R	> Pencil erasers		1		
PMAT06	R	> Rulers		0		
PMAT07	R	> Pens or ball point pens		5		
PMAT08	R	> Files/folders (with loose sheets in them)		5		
PSIT	N	What do you sit on in your classroom? […on the floor,… at a desk]		2		
PWRITE	O	What writing place do you have in your classroom? […nowhere,… a desk or table]		2		
PHMWKDON	O	How often does a person other than your teacher make sure that you have done your homework?		5		
PHMWKHLP	O	How often does a person other than your teacher usually help you with your homework?		3		
PREAD	O	How often does a person other than your teacher ask you to read to him/her? [never, … most of the time]		2		
PCALC	O	How often does a person other than your teacher ask you to do mathematical calculations?		2		
PQUESTR	O	How often does a person other than your teacher ask you questions about what you have been reading?		0		
PQUESTM	O	How often does a person other than your teacher ask you questions about… Mathematics?		1		
PLOOKWK	O	How often does a person other than your teacher look at the work that you have completed at school?		2		
PEXTENG	O	Do you take extra tuition outside school hours in the following school subjects? > English		4		
PEXTMAT	O	> Mathematics		3		
PEXTOTH	O	> Other subjects		1		
PEXTPAY	N	Is there any payment made to the teacher …extra tuition outside school hours…?		8		
PHMWKR	O	How often are you usually given homework in Reading?		0		

Original	*T*	*Question*	*A*	*R^2*	*S*	*New*
PHMWKRC	O	How often does your teacher correct your Reading homework?		0		
PTEXTR	N	How are the Reading textbooks used in your classroom during the lessons? [… I share… by myself]		9		textbooks_read
PHMWKM	O	How often are you usually given homework in Mathematics? [once… each month,… most days]		5		(class_meth_math)
PHMWKMC	O	How often does your teacher correct your Mathematics homework? […never… always…]		3		
PTEXTM	N	How are the Mathematics textbooks used in your classroom during the lessons? [… I share… by myself]		8		textbooks_math
EDUCATOR QUESTIONNAIRE						
XSEX	N	What is your sex?		4		
XAGE	R	What is your age?		1		
XQACAD	O	What is the highest level of academic education you have attained? [… primary… secondary… tertiary…]		14	•	(yrs_preserv_read/math)
XQPROF	O	How many years of teacher training have you received altogether? [… not receive… three years…]		25	•	(yrs_preserv_read/math)
XEXPER	R	How many years altogether have you been in teaching?		0		
XINSERVC	R	After having completed your… training, how many short in-service courses have you attended…?		1		
XINSERVD	R	After having completed your… training, what is the total number of days… attending… courses…?		1		day_inserv_read/math
XINSERVE	O	Generally, do you think that these in-service courses were effective…? […not effective… very effective]		15	•	
XCLBOOKS	R	How many books do you have in your classroom library or book corner?		0		
XSIT	R	How many of the following do you have in your classroom or teaching area? > Sitting places for pupils…		0		
XWRITE	R	> Writing places for pupils		4		
XCRES1	O	Which of the following are available in your classroom or teaching area? > … writing board…		0		
XCRES2	O	> Chalk (or other markers)		1		
XCRES3	O	> A wall chart of any kind		3		
XCRES4	O	> A cupboard or locker		10		
XCRES5	O	> One or more bookshelves		27		
XCRES6	O	> A classroom library, book corner or book box		8		
XCRES7	O	> A teacher table		4		
XCRES8	O	> A teacher chair		7		
XACCESS1	O	Which of the following do you have access to in your school? > A map [Yes, No]		4		
XACCESS2	O	> An English dictionary		6		
XACCESS3	O	> Geometrical instruments (compass, protractor, etc.) for use on writing board		4		
XACCESS4	O	> Teacher's guide (English)		4		
XACCESS5	O	> Teacher's guide (Mathematics)		4		
XPERIODS	R	How many periods/lessons of actual teaching do you have in a typical school week at this school?		6		(hrs_year_read/math)
XMINUTES	R	How long are these periods?		1		(hrs_year_read/math)
XOUTWORK	R	How many hours, on average, do you spend in a typical school week working on lesson preparation…?		1		
XMEETPAR	O	How often do you usually meet with the parents…? [never… once a term…]		11	•	(par_involve_read/math)
XMEEUSUA	R	On average, what percentage of pupils have parents or guardians usually meeting with you in a year?		6		(par_involve_read/math)
XINS98	R	On how many occasions did an Inspector or Advisor… visit you…? > Inspector, 1998… occasions		0		
XADV98	R	> Advisor, 1998… occasions		0		

Original	*T*	*Question*	*A*	*R^2*	*S*	*New*
XINS99	R	> Inspector, 1999… occasions		0		
XADV99	R	> Advisor, 1999… occasions		0		
XINS00	R	> Inspector, 2000… occasions		0		
XADV00	R	> Advisor, 2000… occasions		0		
XINSP01	O	What does the Inspector, EO or DEO actually do when visiting? … > advises me		7		
XINSP02	O	> criticises me		1		
XINSP03	O	> suggests new ideas		5		
XINSP04	O	> clarifies educational objectives		4		
XINSP05	O	> explains curriculum content		2		
XINSP06	O	> recommends new teaching materials		5		
XINSP07	O	> provides information for self-development		4		
XINSP08	O	> contributes very little to my classroom teaching		1		
XINSP09	O	> makes suggestions on improving teaching methods		4		
XINSP10	O	> encourages professional contacts with teachers in other schools		4		
XINSP11	O	> provides in-service training to teachers		3		
XINSP12	O	> finds faults and reports them to my employer		1		
XADV01	O	What does the Advisor actually do when visiting? >advises me		6		
XADV02	O	> criticises me		2		
XADV03	O	> suggests new ideas		5		
XADV04	O	> clarifies educational objectives		4		
XADV05	O	> explains curriculum content		5		
XADV06	O	> recommends new teaching materials		4		
XADV07	O	> provides information for self-development		4		
XADV08	O	> contributes very little to my classroom teaching		1		
XADV09	O	> makes suggestions on improving teaching methods		5		
XADV10	O	> encourages professional contacts with teachers in other schools		7		
XADV11	O	> provides in-service training to teachers		6		
XADV12	O	> finds faults and reports them to my employer		1		
XSHADV	O	How often does your School Head advise you on your teaching? [Never… Once or more a month…]		17	•	teacher_eval_read/math
XRESCENT	O	Is there an education resource centre which serves your school? [No, Yes]		9		(dist_support)
XRESUSED	O	What exactly have you used the… resource centre for during this… year? [… no… centre…]		10		
XRESUSE1	O	> Borrow teaching/learning materials [Yes, No]		1		
XRESUSE2	O	> Make teaching/learning materials		0		
XRESUSE3	O	> Attend training courses		3		
XRESUSE4	O	> Exchange ideas with teachers from other schools		5		
XRESUSE5	O	> Seek advice from the staff of the resource center		2		
XRESUSE6	O	> Other		0		

Original	*T*	*Question*	*A*	*R^2*	*S*	*New*
XSATIS01	O	… teachers' satisfaction… How important do you think each of the following is? > Your travel distance […]		3		
XSATIS02	O	> Location of school [Not important, Of some importance, Very important]		2		
XSATIS03	O	> Quality of the school buildings		2		
XSATIS04	O	> Availability of teacher housing		2		
XSATIS05	O	> Quality of teacher housing		2		
XSATIS06	O	> Availability of classroom furniture		1		
XSATIS07	O	> Quality of classroom furniture		1		
XSATIS08	O	> Level of teacher salary		2		
XSATIS09	O	> Timely payment of salaries		3		
XSATIS10	O	> Seeing my pupils learn		1		
XSATIS11	O	> Availability of classroom supplies (e.g., books, paper, pens, etc.)		1		
XSATIS12	O	> Quality of school management and administration		1		
XSATIS13	O	> Amicable working relationships with other staff members		1		
XSATIS14	O	> Good relationships with the local community		0		
XSATIS15	O	> Expanded opportunities for promotion		5		
XSATIS16	O	> Opportunities for professional development through further study and/or training		2		
XSATMOST	N	Of the fifteen reasons listed in the above question, rank the three… most important... > Most important		12	•	
XSATSECO	N	> Second most important reason		10	•	
XSATTHIR	N	> Third most important reason		14	•	
XPOS01	O	Which of the following items do you have at home? > Daily newspaper		0		
XPOS02	O	> Weekly or monthly magazine		2		
XPOS03	O	> Radio		0		
XPOS04	O	> TV set		0		
XPOS05	O	> Video cassette recorder (VCR)		5		
XPOS06	O	> Cassette player		3		
XPOS07	O	> Telephone		7		
XPOS08	O	> Refrigerator/freezer		1		
XPOS09	O	> Car		13	•	(teacher_ses_math/read)
XPOS10	O	> Motorcycle		1		
XPOS11	O	> Bicycle		5		
XPOS12	O	> Piped water		5		
XPOS13	O	> Electricity (mains, generator, solar)		1		
XPOS14	O	> Table to write on		1		
XLIVS1	R	Approximately how many of the following livestock do you own? > Cattle		2		
XLIVS2	R	> Sheep		1		
XLIVS3	R	> Goats		1		
XLIVS4	R	> Horses		1		

Original	*T*	*Question*	*A*	*R^2*	*S*	*New*
XLIVS5	R	> Donkeys		0		
XLIVS6	R	> Pigs		1		
XLIVS7	R	> Chickens		4		
XLIVS8	R	> Other stock		0		
XLIGHT	N	What is the main source of lighting by which you can read in… (home)? [Fire, … electric lighting…]		4		
XLIVING	O	Which… reflects… the condition of your living accommodation? […poor state… good condition…]		17	•	(teacher_ses_math/read)
TREPENGL	O	Does the school report for each pupil include a specific section for comment on English?		2		
TRACT01	O	How important do you consider the following… activities…? > Listening [Not important… Very important]		4		(class_meth_read)
TRACT02	O	> Silent reading		3		
TRACT03	O	> Learning new vocabulary from a text		1		
TRACT04	O	> Pronouncing or sounding words		0		
TRACT05	O	> Reading for comprehension		2		
TRACT06	O	> Taking books home to read		2		
TRACT07	O	> Reading materials in the home		2		
TRACT08	O	> Reading aloud in class		1		
TRACTMOS	N	Of the eight activities… select the one that you consider to be the most important.		2		
TSIGNENG	O	Do you ask parents or guardians to sign that pupils have completed their home Reading assignments?		5		(class_meth_read)
TRGOAL01	O	How important do you view each of the following goals of Reading to be? > Making reading enjoyable		2		
TRGOAL02	O	> Extending students' vocabulary [Not important, Of some importance, Very important]		2		
TRGOAL03	O	> Improving word attack skills		0		
TRGOAL04	O	> Improving students' reading comprehension		1		
TRGOAL05	O	> Developing a lasting interest in reading		1		
TRGOAL06	O	> Opening up career opportunities		1		
TRGOAL07	O	> Development of life skills		1		
TRGOALMO	N	Of the seven goals listed in the above question, select the one that you consider to be the most important.		5		
TRAPPR01	O	How often do you use the following approaches…? > Introducing the background of a passage…		6		(class_meth_read)
TRAPPR02	O	> Asking questions to assess text comprehension [Never or rarely, Sometimes, Often]		0		
TRAPPR03	O	> Asking questions to deepen understanding		1		
TRAPPR04	O	> Using materials you have created yourself		1		
TRAPPR05	O	> Reading aloud to the class		1		
TRAPPR06	O	> Giving positive feedback		3		
TTESTREA	O	How often do you give your pupils a written test…? [… Once per term… three times…]		2		
TREPMATH	O	Does the school report for each pupil include a specific section for comment on Mathematics?		0		
TMACT01	O	How important do you consider the following… activities…? > Working in pairs or groups		0		
TMACT02	O	> Working alone on problems [Not important, Of some importance, Very important]		11		(class_meth_math)
TMACT03	O	> Preparing projects or posters to be shown to the class.		5		
TMACT04	O	> Using practical equipment, e.g., scales, calculators, rulers, tape measures, etc.		0		

Original	*T*	*Question*	*A*	*R^2*	*S*	*New*
TMACT05	O	> Homework assignments		3		
TMACT06	O	> Studying and interpreting graphs from magazines, newspapers, etc.		0		
TMACT07	O	> Reciting tables, formulae, etc.		4		
TMACT08	O	> Quizzes, tests, examinations, etc.		1		
TMACTMOS	N	Of the eight activities listed in the above question, select the one… most important.		5		
TSIGNMAT	O	Do you ask parents… to sign that pupils have completed their Mathematics home assignments?		7		(class_meth_math)
TMGOAL01	O	How important do you view each of the following goals…? > Basic numeracy skills		1		
TMGOAL02	O	> Problem solving (transfer of skills to everyday life and applying knowledge)		1		
TMGOAL03	O	> Thinking skills including different ways of thinking in solving mathematical problems		0		
TMGOAL04	O	> Confidence in solving Mathematics problems		2		
TMGOAL05	O	> Satisfaction from doing Mathematics		1		
TMGOAL06	O	> Opening up career opportunities		1		
TMGOAL07	O	> Development of life skills		1		
TMGOALMO	N	Of the seven goals listed in the above question, select the one that you consider to be the most important.		3		
TMAPPR01	O	How often do you use the following approaches when teaching Mathematics? > Using everyday problems		0		
TMAPPR02	O	> Teaching the whole class as a group		1		
TMAPPR03	O	> Teaching in a small group		2		
TMAPPR04	O	> Teaching individually		6		(class_meth_math)
TMAPPR05	O	> Teaching through question and answer techniques		5		
TMAPPR06	O	> Giving positive feedback		2		
TMAPPR07	O	> Relating to everyday life situations as much as possible		1		
TMAPPR08	O	> Basic skills training		4		
TMAPPR09	O	> Explaining mathematical processes		3		
TMAPPR10	O	> Using available local materials (for example, for measuring area or volume)		0		
TTESTMAT	O	How often do you give… a written test in Mathematics? [… Once per year… three times per term…]		0		
XCLSIZE	R	…the number of pupils		5		class_size2_read/math
SCHOOL PRINCIPAL QUESTIONNAIRE						
SSEX	N	What is your sex?		1		
SAGE	R	What is your age?		0		
SQACADEM	O	What is the highest level of academic education you have attained? [Primary education… tertiary…]		16	•	(yrs_preserv_prin)
SQTT	O	How many years of teacher training have you received altogether? [… did not receive… three years…]		15		(yrs_preserv_prin)
SQSPEC	O	Have you received specialised training in school management? [No… Yes, a training programme of…>]		1		
SQSPECWK	R	> weeks		1		
SEXPTCH	R	How many years altogether have you been in teaching (including… School Head)?		1		
SPERIODS	R	How many periods/lessons do you actually teach in a typical school week at this school?		7		(prin_teach_load)
SMINUTES	R	How long are these periods on average?	*	7		(prin_teach_load)
SEXPTHIS	R	How many years have you been heading this school as School Head and/or Acting School Head?		2		

Original	*T*	*Question*	*A*	*R^2*	*S*	*New*
SEXPALL	R	How many years altogether have you been a School Head or Acting School Head?		3		
STYPE	N	Is your school a government school or a private school?		0		
SESTABL	I	In what year was this school established? (When was it opened?)		0		
SFAR1	R	How many kilometres is it by road from your school to: > The nearest health centre/clinic		3		
SFAR2	R	> The nearest tarred or tarmac road		0		
SFAR3	R	> The nearest public library		16	•	
SFAR4	R	> The nearest book shop		14		
SFAR5	R	> The nearest school offering secondary grades to which most of your graduating pupils go.		1		
SFAR6	R	> The nearest shopping centre or market place with at least two shops.		3		
SLOCAT	O	Which… best describes the location of your school? [Isolated, rural… small town… large town]		37	•	ruralness
STCHPM	R	How many teachers (permanent… temporary…) are there in your school this week? > Permanent male	*	2		
STCHPF	R	> Permanent female	*	7		
STCHTM	R	> Temporary male teachers		13		
STCHTF	R	> Temporary female teachers		18		
STCHSM	R	> Student male teachers		0		
STCHSF	R	> Student female teachers		0		
STCHTOT1	R	> Total number of teachers		5		
STCHPRIM	R	How many… teachers… have completed… academic education. > Only primary school	*	0		(yrs_preserv_read/math)
STCHSECO	R	> Up to secondary school	*	6		(yrs_preserv_read/math)
STCHTERT	R	> Tertiary academic education		9		(yrs_preserv_read/math)
STCHTOT2	R	Total number of teachers		5		
STCHNOTT	R	How many… teachers… have completed… teacher training. > No teacher training		2		(yrs_preserv_read/math)
STCHSHOR	R	> Short course(s) of less than one-year of duration in total		0		(yrs_preserv_read/math)
STCH1YR	R	> A total equivalent of one year of teacher training		0		(yrs_preserv_read/math)
STCH2YR	R	> A total equivalent of two years of teacher training		8		(yrs_preserv_read/math)
STCH3YR	R	> A total equivalent of three years of teacher training		5		(yrs_preserv_read/math)
STCHMORE	R	> A total equivalent of more than three years of teacher training		31		(yrs_preserv_read/math)
STCHTOT3	R	Total number of teachers		5		
SPUPBOY	R	What is the total enrolment of your school? > Boys		1		(school_infra)
SPUPGIRL	R	> Girls	*	3		(school_infra)
SPUPBOY6	R	What is the total enrolment in Grade 6 in your school? > Grade 6 boys		1		
SPUPGIR6	R	> Grade 6 girls	*	3		
SCLASS	R	What is the total number of class groups (or classes) in your school?		12		
SCLASS6	R	What is the total number of Grade 6 class groups (or classes) in your school?		4		
SSESS1P	R	How many sessions operate in your school (excluding sessions for adults)? > Pupils… 1st session		2		
SSESS1C	R	> No. of classes 1st session		12		
SYRINSP	O	What was the last year your school had a full inspection? […never… before 1995, 1996… 1999]		0		

Original	*T*	*Question*	*A*	*R^2*	*S*	*New*
SINS1998	R	How many times has your school been visited by an inspector since January 1998?		11	•	(dist_support)
SINSP01	R	How many times has an inspector…since January 1998 for the following purposes? > Full inspection		1		
SINSP02	R	> Routine inspection		2		
SINSP03	R	> Inspection of one or more teacher – not for Promotion		0		
SINSP04	R	> Inspection for promotion of a teacher		0		
SINSP05	R	> To assist teachers to improve their teaching skills or to introduce new teachers to their work		3		
SINSP06	R	> To advise the school head and/or other senior staff on management and administration		10		
SINSP07	R	> To address a crisis or problem in the school		0		
SINSP08	R	> Only to deliver something or to make a courtesy call (that is, not for any of the purposes listed above)		15	•	
SCNTR	O	How many times have…staff of the education resource centre… visited… during this… year? [… >]		9		
SCNTRVIS	R	> times this school year		1		
SCNTRSRV	O	How many schools does the education resource centre serve? [… no resource centre… 1-5, 6-10… 16…]		10		
SACTHD01	O	In your work as a School Head, how important…? > Contacts… community [not… of some… very…]		0		
SACTHD02	O	> Monitoring pupils' progress		3		
SACTHD03	O	> Administrative tasks concerning the functioning of the school		1		
SACTHD04	O	> Discussing educational objectives with the teaching staff		1		
SACTHD05	O	> Activities aimed at the professional development of teachers		0		
SACTHD06	O	> Activities aimed at the professional development of School Heads		0		
SACTMOST	N	Of the six activities listed above, rank the three that you consider to be the most important. > Most…		6		
SACTSECO	N	> Second most important activity		12	•	
SACTTHIR	N	> Third most important activity		2		
SSCHACT1	O	Which of the following activities occur in your school? > School magazine		15	•	
SSCHACT2	O	> A public speaking day when pupils read speeches to parents that they themselves have written		4		
SSCHACT3	O	> An 'open-door policy' for parents to visit… by appointment or not by appointment		0		
SSCHACT4	O	> An 'open-day policy' where a special day is nominated for parents to visit the school head or teachers		4		
SSCHACT5	O	> Formal debates or debating contests		1		
SPUPPR01	O	About how often does the school have to deal with the following…? > Pupils arriving late […]		3		
SPUPPR02	O	> Pupil absenteeism (i.e., unjustified absence) [never, sometimes, often]		8		
SPUPPR03	O	> Pupils skipping classes		0		
SPUPPR04	O	> Pupils dropping out of school		16	•	
SPUPPR05	O	> Classroom disturbance by pupils		4		
SPUPPR06	O	> Cheating by pupils		1		
SPUPPR07	O	> Use of abusive language by pupils		2		
SPUPPR08	O	> Vandalism by pupils		1		
SPUPPR09	O	> Theft by pupils		5		
SPUPPR10	O	> Intimidation or bullying of pupils by pupils		1		
SPUPPR11	O	> Intimidation/verbal abuse of teachers/staff by pupils		3		

Original	*T*	*Question*	*A*	*R^2*	*S*	*New*
SPUPPR12	O	> Physical injury to staff by pupils		5		
SPUPPR13	O	> Sexual harassment of pupils by other pupils		2		
SPUPPR14	O	> Sexual harassment of teachers by pupils		5		
SPUPPR15	O	> Drug abuse by pupils		3		
SPUPPR16	O	> Alcohol abuse or possession by pupils		2		
SPUPPR17	O	> Fights among pupils		1		
SPUPPR18	O	> Pupil health problems		3		
STCHPR01	O	About how often does the school have to deal with…? > Teachers arriving late		35	•	teacher_disc
STCHPR02	O	> Teacher absenteeism (i.e., unjustified absence) [never, sometimes, often]		12		
STCHPR03	O	> Teachers skipping classes		9		
STCHPR04	O	> Intimidation or bullying of pupils by teachers		1		
STCHPR05	O	> Sexual harassment of teachers by other teachers		1		
STCHPR06	O	> Sexual harassment of pupils by teachers		1		
STCHPR07	O	> Use of abusive language by teachers		2		
STCHPR08	O	> Drug abuse by teachers		2		
STCHPR09	O	> Alcohol abuse or possession by teachers		3		
STCHPR10	O	> Teacher health problems		1		
SLOST	R	How many official school days did you lose… in the last school year…organization of examinations…?		4		(hrs_year_read/math)
SCLRMPER	R	How many… does your school have? > Permanent classrooms	*	2		
SCLRMTEM	R	> Temporary classrooms	*	1		
SCLRMOPE	R	> Open-air teaching areas		0		
SAREAPER	R	What is the total inside area (in square metres) of all… classrooms in your school? > Permanent		0		
SAREATEM	R	> Temporary		2		
SCONDIT	O	What is the general condition of your school buildings? […complete rebuilding… In good condition…]		22		(school_infra)
STOIBOYA	R	How many toilets or latrines does your school have? State the number of places for individual pupils. > …		28		(school_infra)
STOIBOYB	R	> Squat holes or pit toilets Boys		6		
STOIBOYC	R	> Other types of toilet or latrine Boys		0		
STOIBOYD	N	> No toilets Boys		2		
STOIGIRA	R	> Flush toilet Girls		27		(school_infra)
STOIGIRB	R	> Squat holes or pit toilets Girls		10		
STOIGIRC	R	> Other types of toilet or latrine Girls		1		
STOIGIRD	N	> No toilets Girls		3		
STOISTAA	R	> Flush toilet Staff		23	•	
STOISTAB	R	> Squat holes or pit toilets Staff		12		
STOISTAC	R	> Other types of toilet or latrine Staff		1		
STOISTAD	N	> No toilets Staff		1		
SRES01	O	Which of the following does your school have? > School library		35	•	(school_infra)

Original	T	Question	A	R^2	S	New
SRES02	O	> School or community hall		10		(school_infra)
SRES03	O	> Teacher/staff room		15		(school_infra)
SRES04	O	> Separate office for school head		9		(school_infra)
SRES05	O	> Store room (separate from head's office)		19	•	(school_infra)
SRES06	O	> First aid kit		15	•	
SRES07	O	> Sports area/Playground		7		
SRES08	O	> Piped water/Water tank/Borehole/Spring		3		
SRES09	O	> Electricity (mains or generator)		12		
SRES10	O	> Telephone		15		
SRES11	O	> Fax machine		33		
SRES12	O	> School garden		5		
SRES13	O	> Typewriter		8		
SRES14	O	> Duplicator		14		
SRES15	O	> Radio		4		
SRES16	O	> Tape recorder		26	•	
SRES17	O	> Overhead projector		30	•	
SRES18	O	> TV set		35		
SRES19	O	> Video cassette recorder (VCR)		34		
SRES20	O	> Photocopier		32		(school_infra)
SRES21	O	> Computer		39	•	(school_infra)
SRES22	O	> Fence or hedge around school borders		7		
SRES23	O	> Cafeteria/shop/kiosk		44		(school_infra)
SBORROW	O	Can pupils borrow books from the school library to take them to their homes? […no library… No… Yes]		35		
SCOMM01	O	What do parents and/or the community contribute to the school? > Building of school facilities [No, Yes]		7		
SCOMM02	O	> Maintenance of school facilities (such as classrooms, teacher houses, etc.)		0		
SCOMM03	O	> Construction or maintenance and repair of furniture, equipment, etc.		2		
SCOMM04	O	> The purchase of textbooks		13	•	
SCOMM05	O	> The purchase of stationery		15	•	
SCOMM06	O	> The purchase of other school supplies, materials and/or equipment		1		
SCOMM07	O	> Payment of examination fees		1		
SCOMM08	O	> Payment of the salaries of additional teachers		27	•	
SCOMM09	O	> Payment of an additional amount on top of the normal salary of teachers		5		
SCOMM10	O	> Payment of the salaries of non-teaching staff		11	•	
SCOMM11	O	> Payment of an additional amount on top of the normal salary of non-teaching staff		2		
SCOMM12	O	> Extra-curricular activities including school trips		1		
SCOMM13	O	> Assisting teachers in teaching and/or teach or supervise pupils themselves without pay		8		
SCOMM14	O	> Provision of school meals		1		

Original	*T*	*Question*	*A*	*R^2*	*S*	*New*
SPROBCOM	O	To what extent is lack of co-operation from the community a problem…? [Not a problem… major problem]		2		

Appendix C THE NEW DERIVED QUESTIONNAIRE VARIABLES

The table below can be explained as follows:

- The column *Input* is the link between this table and the policy-oriented mental model presented in section 4.5.

- The column *Variable* provides both a long descriptor of each variable, plus an abbreviated one. The abbreviated descriptor is what appears in the statistical outputs presented in the thesis. Wherever *math/read* appears, two new variables exist, corresponding to the mathematics and reading educator questionnaire responses.

- The column *L* indicates the level of the new variable. The codes are *L* for learner, *E* for educator and *S* for school. The level is the level at which one might expect different values for different units. Thus an *E* variable could have different values for different educators, but would not have different values for different learners with the same educator.

- Column *Meth* provides information on the general methodology used to obtain the new variable.

  T means a 'true' approach was followed, resulting in a value existing more or less along a meaningful scale, such as years of training, or number of school visits.

  T (R) means a true approach with some use of regression modelling to determine weights to assign to different characteristics.

  R means a few of the original variables were regressed against the learner scores, and the slope coefficients were used to weight characteristics. Here there is no 'true' scale for the values.

  F means a few of the original variables were condensed into one new variable using the factor analysis approach.

- Column *Range* provides the minimum value, the values at the 25th, 50th and 75th percentiles, and the maximum value, in order to provide an idea of the distribution of values by quintile.

- Column *cv* provides the coefficient of variation of the variable.

- Column *Units* provides a brief description of what the scale of the variable means.

The correlation matrix for all the variables, plus the two performance scores, appears after the table of new variables. Each correlation coefficient is multiplied by 10 in order to provide a more compact presentation. Values in bold are those with an absolute value greater than 0.5 (or greater than 5 in the format of the table).

The correlation matrix is followed by a table representing the extent of missing values in the new variables (which in turn is based on the extent of missing values in the original variables). A 1 indicates the presence of a value, a 0 the absence of a value. The first row of the table, where all values are 1, represents observations where all values are present. There are 2,968 such observations, which represent 2,907.1 weighted observations. The second row of the table indicates that there were 54 unweighted observations where *math_score* had a missing value, as well as the 16 variables listed at the top of the fourth and fifth columns.

Table 37: New derived variables

Input	Variable	L	Meth	Range					cv	Units
Quantity/quality of pre-service teacher training	**yrs_preserv_read** (years pre-service training)	E	T (R)	14	15	15	15	16	0.033	Average years of schooling/training of (mainly) educators at the school but also of learner's educator.
Quantity/quality of in-service teacher training	**day_inserv_read** (days of in-service training)	E	T	0	0	10	20	265	1.707	Days of in-service training in last three years (learner's educator).
Educator salary and fringe benefits	**teacher_ses_read** (teacher SES)	E	F	0	4	6	10	10	0.523	SES of learner's educator on 0-10 scale.
Incentives for educators to perform	**teacher_eval_read** (evaluation intensity)	E	T (R)	0	0	0	4	10	1.400	Intensity of evaluations by the school principal, with once a year evaluations being weighted more. A 0-10 scale used.
Learner/educator ratio	**class_size2_read**	E	T	16	1089	1600	2304	6724	0.598	Learners in the class squared.
Relevance/clarity of the curriculum	**class_meth_read** (class methodology value)	E	R	0	4	6	7	10	0.369	Value of classroom approaches on 0-10 scale.
Contact time	**hrs_year_read** (teacher hours in a year)	E	T	120	561	840	1000	1970	0.470	Number of hours educators teach in a year, without counting absenteeism.
Level of learner repetition	**repetition** (number of years repeated)	L	T	0	0	0	1	3	1.352	Years that a learner has repeated in the past.
Quantity of LSMs	**textbooks_read** (textbooks per learner)	L	T	0.0	0.3	0.5	0.5	0.5	0.464	Ratio of textbooks per learner for one subject, with an upper cut-off of 0.5.
Quality of school buildings and equipment	**school_infra** (level of school infrastructure)	S	F	0	2	4	7	10	0.724	A score of infrastructure (building and equipment) presence on a 0-10 scale.
Management capacity of school principal	**yrs_preserv_prin** (principal's years of pre-service training)	S	T	13	15	16	16	16	0.048	Years of schooling/training of the principal.
School principal salary and fringe benefits	**prin_teach_load** (principal's teaching load)	S	T	0	4	7	12	35	0.819	Hours per week that the principal teaches.
Level of community involvement	**par_involve_read** (level of parent involvement)	E	T (R)	0	1	3	6	11	0.865	An indicator of the value of parent-educator interactions.
Quantity/quality of district support	**dist_support** (intensity of district support)	S	T (R)	0	3	10	18	63	1.002	Number of visits by inspector the school in the last 3 years. Access to a nearby resource centre was translated into additional 10 visits.
Transport for remote learners	**ruralness** (proximity to urban facilities)	S	T	1	1	2	3	3	0.455	Scale of 1-3, with 3 meaning most urban.
Health of learners	**daily_meals** (average number of meals per day)	L	T (R)	0	2	3	3	3	0.303	Average number of meals eaten per day.
Educational support from parents	**parent_educ** (years of schooling of parents)	L	T (R)	0	9	13	16	23	0.403	Years of schooling of mother and father, with mother weighted twice father. Maximum of 7 years added depending on usage of English.

Input	*Variable*	*L*	*Meth*	*Range*					*cv*	*Units*
Socio-economic welfare of household	**learner_ses** (learner SES)	L	F	0	2	5	8	10	0.670	An indicator on a 0-10 scale of learner's SES.
Other	**teacher_disc** (degree of teacher latecoming)	S	T	0	1	1	1	1	0.409	A dummy variable indicating whether the principal believes educator latecoming is a problem.
Other	**learner_age** (learner's age in years and months)	L	T	10	12	12	14	25	0.126	The age of the learner in years and in months converted to decimal fractions.
Other	**learner_gender** (learner's gender)	L	T	0	0	1	1	1	0.977	A dummy variable with value 1 for girl.

Table 38: Correlation matrix for the new variables

	yrs_preserv_math	yrs_preserv_read	day_inserv_math	day_inserv_read	teacher_ses_math	teacher_ses_read	teacher_eval_math	teacher_eval_read	class_size2_math	class_size2_read	class_meth_math	class_meth_read	hrs_year_math	hrs_year_read	repetition	textbooks_math	textbooks_read	school_infra	yrs_preserv_prin	prin_teach_load	par_involve_math	par_involve_read	dist_support	ruralness	daily_meals	parent_educ	learner_ses	teacher_disc	learner_age	math_score
yrs_preserv_read	**9**																													
day_inserv_math	-2	-2																												
day_inserv_read	-2	-2	**6**																											
teacher_ses_math	**5**	4	-2	-2																										
teacher_ses_read	**5**	**5**	0	0	**5**																									
teacher_eval_math	3	3	0	0	2	2																								
teacher_eval_read	3	2	0	0	4	3	**5**																							
class_size2_math	-2	-1	0	0	-2	-1	-1	-2																						
class_size2_read	-2	-1	0	0	-2	-1	-1	-2																						
class_meth_math	**5**	**5**	-1	-1	3	4	2	2	-2	-2																				
class_meth_read	3	2	-1	0	2	2	1	3	0	0	3																			
hrs_year_math	2	2	-1	-2	2	2	2	1	-2	-2	2	1																		
hrs_year_read	2	2	-1	-1	2	2	2	3	-2	-2	2	2	**5**																	
repetition	-2	-2	0	0	-2	-2	-1	-1	0	0	-2	-2	-1	-1																
textbooks_math	2	1	0	0	1	1	2	2	-2	-2	2	1	1	1	-1															
textbooks_read	1	1	0	0	0	1	1	1	-1	-1	2	2	1	1	-1	3														
school_infra	**6**	**5**	-1	-1	**5**	**6**	3	4	-4	-4	**6**	3	4	**5**	-3	2	3													
yrs_preserv_prin	4	4	0	0	3	3	2	2	-1	-1	4	2	1	2	-1	1	1	4												
prin_teach_load	-2	-2	1	2	-3	-2	-2	-2	-2	-2	-2	0	0	-1	1	0	-1	-3	-3											
par_involve_math	3	3	0	-1	2	1	1	1	-2	-2	2	1	1	1	-1	1	1	3	2	-1										
par_involve_read	3	3	0	1	3	3	1	1	-2	-2	4	2	1	1	-2	1	1	4	1	-1	**5**									
dist_support	4	4	-1	-1	3	3	4	3	-2	-2	3	3	3	2	-2	2	1	**5**	2	-2	2	2								
ruralness	**5**	**5**	0	-1	**5**	**5**	2	4	-3	-3	**5**	3	**5**	**5**	-2	2	2	**8**	4	-2	3	3	4							
daily_meals	2	2	0	0	2	1	1	1	-1	-1	2	1	1	1	-1	0	1	3	1	-2	1	1	1	2						
parent_educ	3	3	0	-1	3	3	2	3	-2	-2	3	2	2	2	-2	2	2	4	2	-1	2	2	3	4	2					
learner_ses	**5**	**5**	-1	-1	4	4	3	4	-2	-2	4	3	3	3	-3	2	3	**6**	3	-3	3	3	4	**6**	3	**5**				
teacher_disc	**-6**	**-6**	1	1	-4	-4	-4	-3	2	2	**-5**	-3	-2	-2	2	-2	-2	**-6**	-3	2	-3	-3	-4	-4	-2	-3	**-5**			
learner_age	-2	-2	0	0	-2	-2	-1	-1	1	1	-2	-1	-1	-1	4	-1	-1	-3	-1	1	-1	-2	-2	-2	-1	-3	-3	2		
math_score	**6**	**5**	-1	-1	4	4	3	3	-2	-2	**5**	3	2	2	-3	2	2	**6**	3	-2	3	3	3	**5**	2	4	**5**	**-6**	-3	
read_score	**6**	**5**	-1	-1	4	**5**	3	4	-3	-3	**6**	3	2	3	-4	2	3	**7**	4	-2	2	4	4	**6**	3	**5**	**6**	**-6**	-3	**8**

Table 39: Extent of missing values in the dataset

Profile of missing values in the new variables (1=present; 0=absent)						
yrs_preserv_math day_inserv_math teacher_ses_math teacher_eval_math class_meth_math hrs_year_math par_involve_math	yrs_preserv_read day_inserv_read teacher_ses_read teacher_eval_read class_meth_read hrs_year_read par_involve_read	math_score	school_infra yrs_preserv_prin prin_teach_load dist_support ruralness teacher_disc	class_size2_math class_size2_read repetition textbooks_math textbooks_read daily_meals parent_educ learner_ses learner_age read_score	Obs	Weighted obs
1	1	1	1	1	2968.0	2907.1
0	0	1	1	1	54.0	85.3
0	1	1	1	1	54.0	74.6
1	1	1	0	1	22.0	28.7
1	1	0	1	1	23.0	23.1
1	0	1	1	1	37.0	23.0
1	0	0	1	1	2.0	18.2
0	0	0	1	1	1.0	1.7
1	1	0	0	1	2.0	1.2
				Total	3163.0	3163.0