Measuring Improvements in Learning Outcomes

BEST PRACTICES TO ASSESS THE VALUE-ADDED OF SCHOOLS

OECD

Measuring Improvements in Learning Outcomes

BEST PRACTICES TO ASSESS
THE VALUE-ADDED OF SCHOOLS

OECD

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT

The OECD is a unique forum where the governments of 30 democracies work together to address the economic, social and environmental challenges of globalisation. The OECD is also at the forefront of efforts to understand and to help governments respond to new developments and concerns, such as corporate governance, the information economy and the challenges of an ageing population. The Organisation provides a setting where governments can compare policy experiences, seek answers to common problems, identify good practice and work to co-ordinate domestic and international policies.

The OECD member countries are: Australia, Austria, Belgium, Canada, the Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Japan, Korea, Luxembourg, Mexico, the Netherlands, New Zealand, Norway, Poland, Portugal, the Slovak Republic, Spain, Sweden, Switzerland, Turkey, the United Kingdom and the United States. The Commission of the European Communities takes part in the work of the OECD.

OECD Publishing disseminates widely the results of the Organisation's statistics gathering and research on economic, social and environmental issues, as well as the conventions, guidelines and standards agreed by its members.

This work is published on the responsibility of the Secretary-General of the OECD. The opinions expressed and arguments employed herein do not necessarily reflect the official views of the Organisation or of the governments of its member countries.

Corrigenda to OECD publications may be found on line at: *www.oecd.org/publishing/corrigenda*.

Foreword

As OECD countries seek to improve their education systems, a growing emphasis is being placed upon measures of school performance as they are central to school improvement efforts, systems of school accountability and school choice, and broader education policies. However, the value of this emphasis rests on the accuracy of the school performance measure. A number of countries have shied away from using raw test scores as a measure of school performance as these scores can reflect student background factors and poorly represent the performance of schools. This can create problems: without an accurate performance measure, equitable outcomes and efficient policy responses can be compromised as resources are not directed to where they are most needed. Policies and practices cannot be improved if it is not known what has proven to be effective. This is where value-added modelling plays such an important role as it provides a more accurate measure of school performance which overcomes many of the problems that plague other measures that can be biased against schools serving more socio-economically disadvantaged students. This increases the confidence that stakeholders can have in the system of school performance and evaluation.

Value added indicators provide an important tool for identifying good practice in the education system. However, relatively few OECD countries have established mechanisms to provide some sort of value-added information at the school level. There are many challenges in improving the design and implementation of value-added modelling in education systems. Developing the appropriate datasets, designing appropriate statistical techniques, and combining these with commensurate policy responses and school improvement initiatives are all on-going challenges. There are also a number of technical difficulties in successfully incorporating value-added modelling into education systems. These technical difficulties have substantial policy repercussions and can impose severe limitations on the applicability of value-added modelling to policy development. This report therefore devotes considerable attention to the technical aspects of the development of value-added modelling as it is essential that they are properly addressed for effective policy development in this area.

This report seeks to provide policy makers, administrators, school principals, teachers and those interested in education systems with information and direction of how a system of school-level value-added modelling can be developed to the benefit of an education system. It draws on the latest research, best practice, and the lessons learned in a number of education systems that either are implementing systems of value-added modelling or have done so in the past. The report is divided into three Parts: Part I discusses the objectives and use of a system of value-added modelling. The focus here is on the main policy and programme applications to benefit school improvement initiatives and systems of school accountability and school choice. Part II is more technical in nature and might be more suited to readers with an interest in the technical issues involved in value-added modelling. It focuses on the design of value-added models, the type of models that can be chosen and the pertinent statistical and methodological issues. Part III focuses solely on the implementation of systems of value-added modelling encompassing both the political and institutional issues that must be addressed and the technical considerations that need to be overcome. In addition, a section titled '*Implementation of a system of value-added modelling: Key steps in the implementation phase*' summarises the key points from Part III and provides a relatively quick and easy guide to the key steps facing policy makers and administrators in the implementation phase. This is provided at the end of Part I before the more technical discussion presented in Part II.

The development of this report began when the Education Committee of the OECD Directorate for Education, during their meeting on 6-7 April 2005, endorsed an initiative to examine the use of school-level value-added measures across OECD member countries that was originally initiated by the Norwegian representative. This led to a proposal for countries to consider their participation in such a project that was funded by the Norwegian Government and was developed by Ben Jensen in the OECD Secretariat. Henry Braun, a leading researcher in the field, was asked to provide an expert paper that was included in the proposal.

Value-added modelling is not used extensively in education systems across OECD member countries. They are most common in the United Kingdom and the USA. Smaller regional and pilot initiatives have also been developed in a number of countries. OECD member countries were invited to join the project in July 2006. Thirteen countries chose to participate in the project: Australia; Belgium (Flemish Community); Czech Republic; Denmark; France; Netherlands; Norway; Poland; Portugal; Slovenia; Spain; Sweden; and the United Kingdom. It was determined that the project would be steered by an expert group that would deliver a report on the development of value-added modelling. The group would be made up of experts from participating countries and steered by the OECD Secretariat.

Additional experts would also be invited to assist with the development of the report. The expert group consisted of:

Bieke De Fraine	Belgium (Fl.)	Maciej Jakubowski	Poland
Eva Van de gaer	Belgium (Fl.)	Maria Eugénia Ferrão	Portugal
Radim Ryska	Czech Republic	Gasper Cankar	Slovenia
Sine Frederiksen	Denmark	Rosario Martínez Arias	Spain
Poul Schjørring	Denmark	Anders Broberg	Sweden
Claudie Pascal	France	Andrew Ray	United Kingdom
Wim van de Grift	The Netherlands	Henry Braun	Invited expert
Torbjørn Hægeland	Norway	Ben Jensen	OECD

The group was led by Ben Jensen of the OECD Secretariat who chaired each of the four 2-day meetings that were scheduled to discuss pertinent issues and develop the report. The first meeting was held in Oslo in November-December 2006. The focus of the meeting was a discussion of value-added modelling across participating countries. Each member of the expert group prepared a background report on value-added modelling in their own country, including discussion of both the accuracy and use of value-added modelling, and the presentation of school results to further policy objectives. This highlighted the commonalities and differences across countries in the development of value-added modelling and fostered discussion about how to define value-added modelling at the school-level.

The second meeting of the expert group was held in London in March 2007. The meeting focused on the use of value-added information for school improvement purposes and the presentation of such information across countries. To assist the discussion of these issues, presentations were given by Henry Braun, Ben Jensen, and Andrew Ray. The third meeting was held in Warsaw in May 2007. The meeting focused on the statistical and methodological issues in the development and use of value-added modelling. Specific papers were presented that examined issues of the stability of school results, the sensitivity of model specifications and the use of various socio-economic contextual characteristics, and other issues such as missing data and measurement error. An additional paper was prepared that illustrated the advantages of the use of value-added information by school inspectorates. These papers were prepared by Maria Ferrão, Torbjørn Haegeland, Maciej Jakubowski, Andrew Ray and Wim van de Grift. The fourth meeting of the expert group was held in Copenhagen in September 2007. The meeting concentrated on analysis for the development of the report, discussion of methodological issues such as the relationship between value-added and growth modelling, and the potential for analysis in the OECD INES framework to examine issues related to value-added modelling. Papers were prepared and presentations given to foster discussion of these issues by Henry Braun, Maciej Jakubowski, Ben Jensen and Eva Van de gear.

Members of the expert group continued to participate in the development of the report that was led by Ben Jensen of the OECD Secretariat. Moreover, a number of edited sections of this report have either been taken from or inspired by the papers presented at expert group meetings. Particular members also volunteered to provide detailed contributions of Chapters and sections of the report. In addition, each member of the expert group acted as a reviewer in the drafting of the report. It was considered appropriate that a technical expert be appointed to review the report. A two-stage review procedure was undertaken to ensure that all issues identified in the technical review were addressed. Dr Daniel McCaffrey, a leader in value-added analysis, agreed to provide a technical review and made a substantial contribution to the report. The editing of the report was provided by Andrew Tierney. Research assistance in the development of the report was undertaken by Diana Toledo Figueroa. Administrative support was provided by Fionnuala Canning. Juliet Evans, Shayne MacLachlan and Elisabeth Villoutreix coordinated the production of the report.

Table of Contents

List of Figures

List of Tables

List of Boxes

Introduction

With education systems in all OECD countries coming under increasing pressure to enhance their effectiveness and efficiency, there is a growing recognition of the need for accurate school performance measures. Assessments of student performance are now common in many OECD countries, and the results are often widely reported and used in public debate as well as for school improvement purposes. There are diverging views on how results from evaluation and assessment can and should be used. Some see them primarily as tools to reveal best practices and identify shared problems in order to encourage teachers and schools to improve and develop more supportive and productive learning environments. Others extend their purpose to support contestability of public services or market-mechanisms in the allocation of resources, *e.g.* by making comparative results of schools publicly available to facilitate parental choice or by having funds following students. Regardless of the objectives of measuring school performance it is important that they truly reflect the contributions which individual schools make rather than merely or partly the different socio-economic conditions under which teachers teach and schools operate. If this is not the case, resources can be misallocated and perverse incentives created if, for example, schools can receive a higher performance measure through academic selection or through selecting students from privileged socio-economic backgrounds, rather than improving outcomes through investment in better instructional methods.

This report documents state of the art methods, referred to as value-added modelling, which allow users to separate the contributions of schools to student performance from contextual factors that are outside the control of classrooms and schools. The greater accuracy they provide in measuring school performance and the role they can play in the development and implementation of education policy and school development initiatives has created a growing interest in value-added modelling. A number of studies have shown that value-added modelling provides more accurate estimates of school performance than do the comparisons of raw test scores or cross-sectional contextualised attainment models (discussed in more detail below) that are often used to provide school performance estimates (Doran & Izumi, 2004). They provide a fundamentally more accurate and valuable quantitative basis than do raw test scores and cross-sectional studies for school improvement planning, policy development and for enacting effective school accountability arrangements.

Value-added models are statistical analyses that provide quantitative school performance measures (*e.g.* a school value-added score) that can be used to develop, monitor and evaluate schools and other aspects of the education system. In this sense, implementing a system of value-added modelling should be viewed as a means to an end rather than an end in itself. How value-added measures are used shall differ between education systems and these differences should inform decisions and actions undertaken in the development of a system of value-added modelling. Therefore, the development process should be shaped by the intended use and application of schools' value-added scores to achieve specified policy objectives.

Three broad policy objectives are identified in this report that can benefit from the use of value-added modelling: school improvement initiatives; school accountability; and school choice. The effectiveness of the use of performance data in decision-making concerning these policy objectives relies on the accuracy of the performance measures used. However, the growth of data-based decision-making to advance policy objectives has been stymied by the lack of accurate school performance data that is essential for educational improvements (Raudenbush, 2004; Vignoles et al., 2000). Raw test scores provide measures of student performance but there are clear problems with drawing inferences from these data about school performance. Cross-sectional contextualised-attainment models take into account contextual characteristics such as student background but are less useful in isolating the effects of individual schools upon students' education. Value-added measures are a significant advance, providing an accurate measure of school performance upon which to base decisions to advance policy objectives and lift school performance. This report illustrates how value-added information can be used for school improvement purposes, for individual programmes and policies and in decision-making at the system- and school-level.

For all *school improvement initiatives* it is important to recognise that improvement in a given activity or set of activities first requires an accurate evaluation of the current situation that, in turn, requires an accurate measure of performance (Sammons et al., 1994). It is difficult to effectively develop programs for the future if it is not possible to accurately analyse the current situation. At the system level, value-added information can be used to determine the areas of the education system and schools that are adding the most value and those areas in which further improvement is required. At the school level, the subjects, grades and groups of students can be identified where the school is adding most value and where improvement is needed. In this sense, value-added scores and information are most valuable if they not only document the current status of schools but also generate information that can support continuous school improvement. Statistical analyses of the relations between school inputs and indicators of school performance can suggest which strategies are and are not working, leading to policy adjustments and the reallocation of resources.

Value-added modelling can also be used to create projections of school performance that can assist in planning, resource allocation and decision-making. Projections can be used to identify future outcomes, for example, providing estimates if current performance trajectories were to continue, and also to set performance targets. Such targets can inform decision-making at the school level of how best to utilise resources and structure the education offered to meet specified performance targets (Hill et al., 2005; Doran and Izumi, 2004). Combined with additional information collected within schools, the projections of future student performance based on value-added estimates provide a comprehensive picture of a school's performance. School personnel then have at their disposal an information base that can serve as a foundation for planning and action.

Systems of school accountability can benefit greatly from the use of value-added modelling. Systems of accountability identify which entities are accountable to which bodies for specific practices or outputs (McKewen, 1995). Such systems might provide information to the general public: taxpayers might be informed as to whether tax money is used efficiently, and users might be able to choose educational institutions on a more informed basis. Yet the key issue remains whether the assessment of processes and of performance is accurate and fair to individual schools. This report illustrates that value-added modelling provides a more accurate, and therefore fairer, measure of school performance (as measured by increases in student performance) that can also be used to improve the evaluation of school processes. The results of value-added modelling (*i.e.* schools' value-added scores) provide measures of the extent to which schools have succeeded in lifting student performance. When used in systems of school accountability, these measures can be used effectively in school evaluations, with fairer consequences for schools and school personnel.

School choice is the third key policy objective discussed in this report that benefits from the use of value-added modelling. This data is intended to inform parents and families of the performance of different schools to aid their decision-making in choosing their school. This requires publishing the data on school results (Gorard, Fitz, and Taylor, 2001). While this does not occur in all countries, it is a growing trend among OECD member countries (OECD, 2007a). As is discussed in Part I of this report, there are numerous benefits from improved levels of school choice within an education system. Parents are able to choose schools that are better suited to their needs and resources can then flow to those schools best meeting those needs (Hoxby, 2003). However, such benefits depend upon an accurate measure of school performance, otherwise families' choices are misinformed and resources are misallocated. The greater accuracy of value-added modelling is essential to the effectiveness of a system of school choice. It allows parents a more accurate measure of school performance upon which to base their decisions and allows schools a fairer opportunity to improve their performance.

The policy considerations and political issues surrounding systems of value-added modelling can differ. Given such differences, it can be beneficial to structure the development and implementation of a system of value-added modelling to suit the prescribed policy objectives. The use of value-added modelling to advance specific policy objectives is discussed in Part I of this report and are also detailed in Part III that deals with implementation issues.

The greater accuracy inherent in value-added modelling creates greater confidence in the use of performance measures to further the three policy objectives outlined above. The greater confidence stems from the improvements made in this modelling over time and the advantages compared to other methods of estimating school performance. The modern era of ‘school effects’ research began, at least in the USA, with the so-called Coleman Report that studied the relationships of schools and families to student academic attainment (Coleman, 1966). This complemented a number of European studies that looked at issues of inequality in terms of intergenerational analyses that compared outcomes over generations (Carlsson, 1958; Glass, 1954). Subsequent school effectiveness studies also carried out quantitative comparisons of schools. In the initial phase, high-achieving schools were identified by comparing the average test scores of the students. The next step for researchers was often to select a small number of such schools for further analysis with the hope of identifying the elements of their practice that were responsible for their success. The ultimate goal was to disseminate the findings in order to effect broader school improvement. Early work in this area is reviewed in Madaus, Airasian and Kellaghan (1980).

It was recognised early on that school rankings based on students’ ‘raw’ test score were highly correlated with their students’ socio-economic status (McCall, Kingsbury and Olson, 2004). Bethell (2005), for example, discusses some of the controversies arising from the use of tables comparing raw test scores in England. Multivariate cross-sectional analyses have been used to try and overcome these problems. In the simplest version of these analyses, school average test scores were regressed on a number of (aggregate) relevant demographic characteristics of the schools’ students. The idea was to rank schools on the basis of their residuals from the regression. These residuals were often termed ‘school effects’. Schools with large positive residuals were considered to be exemplary and worthy of further study. Schools with large negative residuals were considered to be problematic and also requiring further study, although for different reasons. Alternative adjustment strategies have been proposed and the resulting differences in school rankings compared (Dyer, Linn and Patton, 1969; Burstein, 1980).

More sophisticated cross-sectional models have subsequently gained in popularity and use with methods that take into account the hierarchical structure of school systems, with students nested within classes, classes nested within schools and schools nested within districts/local areas (Aitkin and Longford, 1986; Goldstein, 1986; Willms and Raudenbush, 1989). The estimates provided by these models have grown in sophistication and have been commonly used in education analyses across OECD member countries. These cross-sectional estimations have been categorised in this report as *contextualised attainment models*. These multivariate models can be used to provide a measure of school performance but it was considered that such analyses did not contain the required analytic framework to be classified as value-added models. Contextualised attainment models estimate the magnitude of contributing factors to student performance or attainment at a particular point in time. A typical example is a regression model that regresses a vector of students' socio-economic backgrounds or contextual characteristics and a variable identifying the school each student attends against some achievement measure. The adjustment to raw scores made with the inclusion of contextual characteristics provides measures that better reflect the contribution of schools to student learning than the use of 'raw' test scores to measure school performance. The results of these cross-sectional models build upon theoretical analyses of the role of the family in shaping people's socio-economic outcomes and often find that the main contributor to the level of student attainment is parental socio-economic background (OECD, 2007b; Haveman and Wolfe, 1995; Becker, 1964). Information on the role of student socio-economic background in educational attainment, while interesting and important, often does not yield sufficient information to enable policy makers to make decisions on school accountability and school choice and to drive school improvement reforms. Nevertheless, these contextualised attainment models are a clear improvement on the use of unadjusted results and raw attainment scores to assess school performance.

A significant advance was made with the development of value-added modelling that utilised multiple measures of student performance to estimate the impact (or value-adding) of individual schools upon those student performance measures. An important assessment of value-added modelling was provided by Fitz-Gibbon (1997) who was asked to advise the British Government on the development of a system of value-added modelling. Fitz-Gibbon concluded that such a model could be the basis for a statistically valid and readily understood national value-added system. Value-added models employ data that tracks the test score trajectories of individual students in one or more subjects over one or more years (Mortimer et al., 1988; Goldstein et al., 1993; SCAA, 1994; Sanders, Saxton and Horn, 1997; Webster and Mendro, 1997; Rowan, Correnti and Miller, 2002; Ponisciak and Bryk, 2005; Choi and Seltzer, 2005; McCaffrey et al., 2004; McCaffrey et al., 2003; McCaffrey et al., 2005). Through various kinds of adjustments, student growth data is transformed into indicators of school value-added. Examples are discussed of the main types of value-added models in Chapter Five of this report.

Value-added models are a substantial improvement on many current measures of school performance. Comparisons of raw test scores provide some important information but are poor measures of school performance. They fail to take account of prior achievement levels and produce results that can largely reflect differences in contextual characteristics such as students' socio-economic background. Contextualised attainment models try to address these problems by measuring the impact of contextual characteristics upon a specific performance measure but are less useful in disentangling school effects upon student progress from other contextual characteristics and are therefore less useful in measuring school performance. Value-added models attempt to overcome these problems by incorporating student prior attainment measures and, in some cases, contextual characteristics. This enables a more refined analysis of progress in student performance that is more effective in disentangling the effects of various factors that affect student progress. These advantages allow for greater accuracy in measuring performance which then creates greater confidence in the interpretation of school performance measures.

In summary, this report argues that value-added modelling contributes to system-wide learning by accurately measuring higher and lower performing aspects of the education system; to school improvement through improved identification and analysis of 'what works'; to improved and more equitable transparent systems of school accountability and school choice that can then create well-defined incentives for schools to improve their performance; to the development of information systems that allow schools to analyse and evaluate their performance and strengthen the overall system of school evaluation; to systems of education funding that more effectively direct resources to areas of need; and, to overcoming entrenched socioeconomic inequalities that exist in societies that might be masked at the school level by indiscriminate and inaccurate performance measures.

Value-added modelling: A definition

Given the advantages of using value-added modelling, it is essential that this report distinguishes value-added modelling from other statistical approaches. Across participating countries there has been a large variation in the use of value-added modelling and statistical analyses to analyse school performance. Such variation increases the importance of defining both 'value-added' and 'value-added modelling' to clearly differentiate them from other types of statistical analyses. In this report, the value-added contribution of a school is defined as:

> the contribution of a school to students' progress towards stated or prescribed education objectives (*e.g.* cognitive achievement). The contribution is net of other factors that contribute to students' educational progress.

From this definition of value-added it was possible to define value-added modelling as:

> a class of statistical models that estimate the contributions of schools to student progress in stated or prescribed education objectives (*e.g.* cognitive achievement) measured at at least two points in time.

Particular value-added models might utilise a narrower definition of the estimation of school performance but this general definition can be applied to a variety of value-added specifications while still clearly delineating value-added modelling from other types of statistical analyses. Statistical analyses that have been undertaken in a number of countries to monitor school performance would not be considered to be value-added modelling using these definitions. Such analyses often did not include at least two measures of student performance that can be considered to be the basis of value-added modelling. These analyses have been defined in this report as contextualised attainment models. It was considered appropriate not to try to expand the definition of value-added modelling to fit the performance measures used in each participating country as it would decrease the effectiveness of the analysis.

A distinguishing feature of value-added modelling is the inclusion of prior performance measures that allow a more accurate estimation of the contribution of the school to student progress. Doran & Izumi (2004) emphasised the advantages of value-added modelling in tracking students over time compared to cross-sectional (or contextualised attainment) models that provide a 'snapshot' picture of student performance. Value-added modelling facilitates more detailed analysis of school improvement by estimating the contribution of the school to improvements in student performance over a given time period. Additionally, value-added models are able to better account for unobserved factors contributing to the initial performance measure, such as student ability that are a systemic problem in much contextualised attainment modelling (Raudenbush, 2004).

The inclusion of a prior performance measure allows a school's value-added to be estimated. The value-added should be interpreted as the contribution of the school to student performance between the two performance measures. This is an important issue as it is possible to employ different student assessments at different time intervals. Such differences need to be recognised in interpretation of the contribution of individual schools (*i.e.* a school's value-added score). A key distinction is the subject matter of the student assessments as the school's value-added is being estimated only on the subject matter included in the assessments (this is discussed further in Chapter one). A further consideration is the timing of the assessments. A number of value-added estimations estimate the contribution of the school in a given year. However, a number of education systems do not have annual assessments or a structure of assessments that would permit the estimation of a single year value-added score. This is not to say that value-added cannot be estimated over a multiple-year timeframe. On the contrary, such estimations are made in a number of education systems. But it is important to recognise that these differ from single year value-added scores so that in discussion of schools' value-added scores it is made clear the subject matter and the time-span in which value-added is measured.

The importance of multiple attainment measures raises the issue of what should be considered an appropriate prior performance measure upon which to measure progress. There is considerable debate about the comparability of test scores and the conversion of scores into meaningful and comparable scales (Braun, 2000; Dorans et al., 2007; Patz, 2007; Kolen and Brennan, 2004). Of course, many value-added models do not actually require that the test scores be vertically scaled. They simply require that scores in successive grades be approximately linearly related and, in most cases, that is a reasonable measure (Doran and Cohen, 2005). This report does not discuss the development of student assessment instruments themselves: a review of the considerable literature analysing assessment issues is outside the scope of this report. However, the definition of value-added used in this report focuses *on progress in stated or prescribed education objectives (e.g. cognitive achievement)*. This precludes some contextualised attainment models that include intelligence measures such as IQ scores that might be considered to be a measure of general ability but are less suitable as a measure of prior attainment upon which to measure progress. In discussion of schools' value-added scores it should always be clear what the prior and current attainment measures and test scores actually represent and how they should therefore affect policy actions and schools.

Even with the greater accuracy obtained with the use of value-added modelling, there remain some difficulties in measuring school performance. The interpretation of schools' value-added scores should include various caveats and cautions for correct interpretation. These issues are discussed in Part II of this report. While this discussion seeks to illustrate the various measurement issues in designing and utilising value-added modelling, it is not the intention to negate their considerable potential. To the contrary, accurate value-added estimations have great potential for use in policy development and school improvement initiatives and are a substantial improvement on alternative measures. For example, Chapter Six discusses the statistical and methodological issues that must be addressed in the development and use of value-added modelling. These issues are highlighted not to deter the use of value-added modelling in education systems but to encourage their effective development in advancing specified policy objectives. In fact, a key reason why the use of value-added modelling is encouraged is that these statistical and methodological issues often create far greater problems of misspecification with other statistical approaches and school performance measures. These alternative approaches normally provide less accurate measures of school performance and are therefore less useful for effective system and school development. The attention given in this report to statistical and methodological issues is thus done to emphasise the need to develop and provide accurate value-added measures of school performance to both inform policy development and school improvement initiatives and to gain the confidence of stakeholders.

Format of this report

This report is divided into three parts that might be suitable to slightly different audiences. Part I discusses the objectives and use of value-added modelling. This includes a discussion of the policy objectives (discussed in Chapter One) that can be advanced with value-added modelling. Linked to this issue is a discussion of how value-added information and school scores can be presented to different stakeholders, distinguishing between the presentation of value-added information for internal purposes, for public consumption, and presentation in the media. A number of examples are provided of effective presentation methods in countries in Chapter Two. The discussion of the presentation of value-added information for internal purposes focuses upon the application of value-added for modelling for school improvement purposes in Chapter Three. Central to this discussion is how the information can play a key role in fostering data-based decision-making in schools that utilise accurate performance measures to develop and monitor school improvement initiatives. This discussion views schools as learning organisations that undertake and benefit from analysis of different aspects of school and student performance. Focus is given to the targeted use of value-added modelling for: specific sub-groups of the student population and specific aspects of schools; setting performance targets and performance projections; identifying students in need of special assistance and early interventions; and, improving the overall system of school evaluations.

Part II discusses the design of value-added models and focuses upon the technical aspects of value-added modelling. Chapter Four discusses key design considerations in developing a system of value-added modelling and identifies the key issues that need to be addressed. Examples of the main types of value-added models are presented in Chapter Five to provide some tangible examples and to illustrate their various requirements, and how they might be adapted to particular settings. Chapter Six discusses the key statistical and methodological considerations in the development of value-added modelling. These are emphasised in order to assist in the identification of the key criteria with which to choose a preferred value-added model(s) in an education system. A number of issues are presented with supporting analysis from participating countries discussed to highlight the steps that can be taken in choosing the appropriate value-added model. The point is made that a key aspect of this issue for administrators is to decide upon what is the most appropriate model to meet the objectives and planned use of value-added modelling.

Part III discusses the implementation of systems of value-added modelling in education systems. This discussion provides policy makers and administrators with guidance on how to implement a system that best meets their needs. Again, the experiences from participating countries are drawn upon to illustrate the key issues and potential strategies that can be employed. Chapter Seven focuses upon the initial steps that need to be taken in the development of the system leading up to, and including, the pilot phase of implementation. Chapter Eight discusses the ongoing development, with considerable attention given to the development of a communication and stakeholder engagement policy. This engagement policy should accompany the introduction of a system of value-added modelling and include training for pertinent users. The actions and consequences for school principals, teachers and other stakeholders will need to be clearly articulated to not only build confidence in a new system but also to assuage fears of the introduction of a system that can be perceived as potentially lacking in fairness and transparency. Specific strategies will need to be developed that explain the system and educate stakeholders in how value-added scores are calculated and how they will be used. As is illustrated in Part III, successful strategies have been developed that highlight the benefits of value-added modelling compared with other performance measures. In a number of countries, stakeholders have welcomed the development and use of value-added modelling: its greater accuracy provides a fairer measure of school performance that creates more equitable systems of school accountability and school choice and fosters more accurate and therefore effective school improvement initiatives.

Also included is a discussion of the main steps that need to be undertaken in the implementation of a system of value-added modelling. The discussion of these steps is not meant to provide an exhaustive list of all activities that need to be undertaken but should assist policy makers and administrators who hope to gain a quick understanding of the process required in the implementation of a system of value-added modelling. This is presented as a small separate section at the end of Part I to emphasise the importance of implementation issues and their connection to specific policy objectives and uses of value-added modelling.

Part I

Objectives and Use of Value-Added Modelling

Chapter One

Policy Objectives of the Development of a System of Value-Added Modelling

The current focus upon school performance in many countries is driven by questions about the effectiveness of investments in schooling, coupled with widespread concerns about national economic competitiveness. Given the central role of human capital in the modern economy (Friedman, 2005; OECD, 1994, 1996, 2001), a nation's schools are seen as a potential source of competitive advantage. A related worry is that the existence of substantial levels of heterogeneity in school performance together with meaningful differences in education outcomes for recognisable subgroups in the population can lead to societal strains and create economic inefficiencies, (OECD, 2008; Lucas, 1988; Romer, 1994). To properly address these issues methods for accurately measuring school performance are required to effectively evaluate investments in schools, identify best practices and to highlight areas where improvements need to be made. Such a system should adequately convey this information to illustrate how such improvements can be made to enhance the performance of all schools.

A value-added analysis is designed to evaluate schools on the basis of what their students have learned whilst they have been enrolled in the school. School value-added scores are aggregates of individual student performance trajectories that might be influenced by a number of factors in addition to the influence of the school itself. Value-added approaches therefore try to isolate the school's contribution to student learning from other factors that are associated (in a statistical sense) with student learning, such as students' socio-economic background. Whatever the ambition driving the development of a system of value-added modelling, there is a need to construct accurate measures of school performance; measures that reflect the real performance of a school and not factors that are more or less beyond the school's control, such as differences in student composition or 'random noise'. Value-added models can provide measures of school performance that for most education systems will greatly improve the data and information currently used to inform decision-making. Importantly, value-added measures provide accurate measures of the contribution of the school to student performance that overcome many of the problems with current school performance measures. More accurate information on school and student performance facilitates more targeted and well-defined initiatives that can yield sustained improvements.

Value-added models can be used to focus attention on particular education programmes or groups of students that are found to be low- or high-performing. This information can be utilised by policy makers, administrators, and school principals and teachers to better identify performance issues, and guide the development and evaluation of school programmes. Policies and programmes aimed at increasing performance require a form of evaluation that identifies both high-performing areas and those areas that are in need of improvement. Value-added models can provide accurate quantitative indicators of performance that facilitate the identification of areas for improvement within schools and school systems, permit the creation of performance benchmarks, and facilitate learning within and between schools. Value-added modelling can also be used to increase the effectiveness of existing institutions such as school inspectorates and enable more informed judgements to be made about schools.

This report focuses on three broad policy objectives that provide the impetus for the development of value-added modelling in an education system: school improvement initiatives; school accountability; and school choice. These policy issues are outlined in this chapter. They are further discussed in subsequent chapters through illustration of various applications, at both the system- and school-level, of value-added models to advance policy objectives. While these three broad policy objectives differ in their focus and in the development of programmes to achieve them, all have the overall objective of improving standards in school education systems. The discussion of these policy objectives is relatively brief given their breadth and complexity. Greater attention is given to the presentation and application of value-added models to further these policy objectives in Chapter Two and Chapter Three as this is considered to be of greater relevance and interest to policy makers.

While each of the three main policy objectives is presented below, it is clear that for most education systems the implementation of a system of value-added modelling has multiple objectives. For example, in England value-added modelling is now used:

- in school Performance Tables[1] to provide information to parents and hold schools to account;

[1] Performance tables in England are now referred to as Achievement and Attainment Tables (AAT).

- in systems for school improvement, where data is used for self-evaluation and target-setting;
- to inform school inspections that are incorporated into the broader school improvement process;
- to help select schools for particular initiatives; and
- to provide information on the effectiveness of particular types of school or policy initiatives.

These multiple objectives illustrate the importance of obtaining accurate performance measures in a number of areas of education systems. It should also be recognised that specific programmes might serve multiple policy objectives. Many of the programmes and initiatives bundled together under a policy of school improvement would also be applicable for school accountability purposes. In some instances, initiatives to promote informed school choice would impose a form of accountability upon schools, teachers and school administrators.

Further complexities arise in a report such as this that is aimed at policy makers and educators operating in different political and cultural contexts. Such differences might lead to divergent interpretations of the objectives of particular programmes. The history of an education system, the interaction of education institutions and the current climate of system development all affect how a particular policy or initiative can be viewed. For example, consider the development of a system that applies value-added modelling to the issue of improving school choice. In this situation, schools’ value-added results can be made available to parents and published on a centralised website that allows parents and students to learn more about school performance. The extent to which this is also considered a form of school accountability can differ according to the context and historical development of the school education system. If this policy were enacted in a system where little performance information had been analysed previously then there is a greater likelihood that it would be perceived as the implementation of a form of school accountability than it would in a system where information about schools’ and students’ performance has routinely been made available to the public. What might be considered commonplace in one education system might represent a significant change in another education system. The impact of the policies upon schools, school administrators and teachers would vary accordingly. The discussion of the use of value-added modelling to further policy objectives presented in this report does not include estimates of the impact upon various policies. The discussion focuses on using value-added information and schools’ value-added scores as a basis for action: the development and monitoring of initiatives and practices that can be implemented under multiple policy objectives.

Use of value-added modelling for school improvement purposes

Value-added models provide accurate performance indicators and information that can be used as a basis for action that advance school improvement objectives. These actions will differ between education systems and can include a number of initiatives that vary in size and specific intent. They are most valuable if they are not only able to document the current status of the system but also generate information that can support continuous improvement, particularly if subsequent, more detailed analyses are carried out. For instance, at the policy level, value-added models can be used to identify schools that are high- or low-performing and direct attention and funding to where they are most needed. Moreover, statistical analyses of the relations between school inputs and school performance can suggest which strategies are more effective, leading to ongoing policy adjustments and reallocation of resources.

By creating accurate measures of school performance, value-added models empower schools and administrators to make more informed decisions to improve school performance (Saunders, 2000). Such information allows for more detailed, and often more targeted, development of school improvement initiatives. Moreover, value-added information empowers schools and policy makers to monitor and evaluate such initiatives. Such data-based management and decision-making overcomes many of the difficulties in assessing whether resources are being effectively utilised and should therefore enable the continuous development of efficiency and quality improvements.

Data-based decision-making

Using value-added measures to advance school improvement requires a greater focus on data-based decision-making within schools and school education systems. In the past several years, education policy makers in a number of countries have seen a groundswell of interest in data-based decision-making and there have been a number of initiatives to create scalable models for the use of data to support school improvement (Saunders, 2000). Data and measurement play critical roles in guiding strategy and monitoring progress toward policy objectives (Atkinson Review, 2005). In this decision-making context, there is an explicit focus on the use of comparative data to identify, for example, areas for potential improvement and to set meaningful goals.

The changing structure of education systems has also increased focus on the allocation and use of resources and the performance of schools. There has been a shift towards greater school autonomy with less centralised regulation of inputs and processes (OECD, 2004). With less centralised control, however, a system is required to measure the performance of schools and for these measurements to be made available in a systematic manner. For example, Ryska (2006) discusses how in the Czech Republic before 1990 the characteristic features of the school system included centralised governing, uniform education and a rigid supervision of teachers. The Czech Ministry of Education played a decisive role, using direct governing and controlling instruments, while inputs and processes were prescribed in detail. School inspection was the main instrument of supervision at the school level. School directors and teachers had little freedom to act with regard to compliance with curricula, in terms of both the content and methods of instruction. At the student-level, teachers' assessments constituted the main assessment method and focused primarily on assessing the knowledge acquired under the prescribed curriculum. The standards and performance of the evaluation system as a whole were neither monitored nor evaluated. More recent changes have led to greater decentralisation in the school system with a greater emphasis on performance measures and the efficient use of resources. A more decentralised school system has presented all parties involved with a new situation. Schools realise the need for structured and systematic evaluation providing feedback of 'what works' at all levels of the education system (Ryska, 2006).

Data-based decision-making should not only be the domain of policy makers: practitioners at all levels of the school education system can use the data. School principals and teachers are able to utilise data on inputs, processes and outputs to analyse resource allocations and the effectiveness of various policies, programmes and managerial decisions (Odden and Busch, 1998). It is important to view schools as learning organisations in the same manner as other public or private sector organisations (Caldwell and Spinks, 1998). However, data alone cannot lead to the success of more comprehensive data-based decision-making approaches. There must be a systemic approach to improving performance that utilises accurate performance measures and aligns efforts to stated objectives.

Many OECD member countries have, in recent years, attempted to shift the focus of the public sector from inputs to outputs in order to improve public sector performance (Eurostat, 2001). In terms of school systems, the data has traditionally focused on resources (Atkinson Review, 2005) and the information available in official statistics and administrative systems has mainly been related to school inputs rather than outputs. Policy makers often have access to detailed information on inputs into the education system. Financial information on capital and current expenditures has often been divided into expenditure in different schools and school inputs such as buildings and maintenance and on salaries to teachers and other school personnel (OECD, 2007a). This information can sometimes be further broken down to analyse expenditures at different levels including centralised, regional, programme, school, and student expenditure. The benefits of more comprehensive data-based decision-making are maximised with combined analysis of input, process and output data. Relying on just one type of data can lead to misleading conclusions and actions. Combining data on inputs with data on school processes permits more extensive analysis of both the allocation and use of resources. Information on school processes has been collected in numerous countries for many years. The most common method for collecting this information is through a specified evaluative framework that, in most OECD member countries, includes the use of school inspectorates (OECD, 2007a). The focus of these evaluations and the information collected varies across countries but normally the focus is on school-level processes and ensuring an adherence to school regulations and procedures. Information is often collected on the form and structure of teaching and specific problems within schools. In a number of countries, school inspections evaluate school performance against pre-determined criteria in these areas (OECD, 2007a).

While these analyses are more extensive than a monitoring system focused solely on inputs, decision-making in this context is restricted by a lack of performance data. Decisions on inputs and processes cannot therefore be analysed in terms of their effect on performance and the optimal allocation of resources and 'mix' of policies and programmes affecting school processes cannot be effectively analysed. Once value-added information is obtained, decision-makers can better analyse how to adjust resources and enact the appropriate school processes to improve student performance. The inclusion of value-added information allows organisational learning of what best contributes to performance improvements.

Accuracy of performance measures

Given an increasing need for analysis of school performance, an accurate performance measure is required for measuring progress in student performance and the effect of the allocation and use of resources in the education system. Clearly, the accuracy of such a measure is paramount if it is to be used in the evaluation and development of the school education sector. In a number of countries, measures of school performance have concentrated on unadjusted test scores or student attainment measures; for example, average scores on standardised tests or the percentage of students in each school progressing to higher levels of education. However, there is growing recognition that there are problems in using these as measures of school performance. These measures often do not take other factors that influence educational achievement into account, such as: the innate ability of students; their socio-economic background; the influence of peers and individuals in and outside school; various events and situations that occur outside the school that might affect student learning; and general randomness in student assessments.

In countries that have decentralised the structure of the school system and introduced a new focus on school accountability, there is a recognition that an emphasis on performance data can give rise to concerns of fairness in the absence of value-added measures (Jakubowski, 2008; Hægeland, 2006). Such concerns have caused consternation among education stakeholders in a number of countries (Linn 2004, 2005). School principals and teachers can perceive that their performance is being unfairly judged with the imposition of accountability measures based on factors outside the control of the school. These concerns can spread to communities, families, parental associations and education unions (Bethell, 2005). A common concern is that a student's standing or attainment at a given point is a function of their cognitive development prior to school entry, as well as their growth during all their years in school. Indeed, a student's development will have been influenced not only by his or her previous schooling but also by out-of-school experiences and support from family and community over time. Holding the current school solely responsible for the results is neither defensible nor fair. For example, many studies have shown that student attainment is strongly correlated with family and community characteristics, further undercutting the credibility of using only data on current student attainment as a basis for school accountability. McCall, Kingsbury and Olson (2004) reported on correlations between school test score means and the percentage of students eligible for free or reduced price lunch (a crude measure of the poverty level of a student population in the USA). Data were obtained from hundreds of schools located in a number of different states. Student performance was based on scores on the Measures of Academic Progress test administered by the Northwest Evaluation Association in 2002 and 2003. Summarising the results for grades 3 through 8, correlations in reading scores ranged from (-0.54 to -0.66), and in mathematics scores ranged from (-0.51 to -0.59). When school means were replaced by a fairly simple measure based on changes in student test scores, the correlations in reading ranged from (-0.07 to -0.27), and in mathematics ranged from (-0.02 to -0.24). The correlation between changes or progress in students' scores and free school meals was therefore much lower than the correlations between this socio-economic status measure and raw scores. This indicates that school performance can be more easily isolated from other factors in analysis of student progress rather than relying on student performance at a single point in time. Further evidence for preferring indicators based on student growth to those based on student attainment is offered by Zvoch and Stevens (2006) who analysed data for three successive cohorts in a large school district in the USA. Such findings provide indirect support for approaches to school value-added that use student score trajectories as input variables. Jakubowski (2008) illustrates that in Poland there exists a strong conviction that unprocessed external examination results are of little value in assessing the quality of teaching or school performance. Also in Norway, there were misgivings about making judgements on school performance based on measures that did not account for the variety of factors outside the control of schools that can affect student performance (Haegeland, 2006). Given the objective of presenting indicators that reflect school performance, it is clear that unadjusted school-level averages of individual student achievement are insufficient measures, since they are influenced by many important factors that are either beyond the control of schools or are unevenly distributed between schools.

In England in the early 1990s, the new emphasis on performance data to hold schools accountable gave rise to concerns that schools could not be judged fairly in the absence of value-added measures. At the same time, the development of the Key Stage tests offered the possibility of calculating value-added scores for each school based on progress between each Key Stage once national data were available for the relevant cohorts of students. This meant that schools could be evaluated on their students' performance in national tests in English, mathematics and science at ages 11 and 14, in national exams in all subjects at ages 16 and 18, and the progress students made between these tests[2] (Ray, 2006). Value-added modelling further developed over time. In the early years of using of school performance measures, school results were reported in terms of the proportion of students that exceeded the relevant threshold in each subject. In this sense, schools were compared with a standard by ordering them by the proportion of students meeting the standard. These so-called 'league tables', which are of much interest to the public, did not, when first published, involve considerations of individual student growth. As discussed above, comparisons based on raw scores can be counter-productive if no account is taken of either this aspect or of the school context. As Jane Davidson, former Education Minister for Wales said in 2002:

> 'I don't need a league table (that were then based on raw scores) to tell me that performance will be better in one of our richer communities than one of our poorer ones.' (Bethell, 2005: p. 8.)

A value-added analysis, on the other hand, provides a comparative measure of school performance. That is, each school is compared with the

[2] Value-added is modelled in England on student assessments at the end of each of the 4 Key Stages of school education. The assessments and the national curriculum are maintained by the Qualification and Curriculum Authority. Key Stage 1 covers Year 1 and Year 2 in primary schools, with pupils assessed at the end of Year 2 when most are 7 years old. Key Stage 2 covers Year 3 to Year 6 which is usually seen as the end of 'primary education. Key Stage 3 covers Year 7 to Year 9; the first three years of secondary schooling. Key Stage 4 covers the final two years of secondary school with most of the assessment falling at the end of the final Year (Year 11). The main qualification is the GCSE (General Certificate of Secondary Education).

average of all schools included in the analysis with the comparison based on students' test score changes over time. A value-added analysis is designed to evaluate schools on the basis of what their students have learned while enrolled in the school rather than inadvertently measuring what students already knew before entering the school. This is considered a fairer basis for comparing schools that serve different student populations with different skill and knowledge levels. Feedback from a teacher training programme that accompanied the introduction of the use of value-added modelling in Poland in 2006 illustrates how greater accuracy translates into increased fairness in the system (Jakubowski, 2007; see also Chapter Eight for a more detailed discussion). A number of key areas were highlighted in the feedback received from training participants that supported the introduction of value-added modelling. Teachers highlighted:

- the benefits of the objectivity of value-added results that highlight good schools working with disadvantaged students and fighting invalid comparisons based on raw test scores;
- the accuracy of quantitative assessments and statistical methods;
- the greater transparency and comparability with value-added methods of school assessment;
- the potential for improved internal school evaluation of students' progress, especially through additional school-level analysis (*e.g.* analysing value-added scores for specific groups of students); and
- the benefits of extensive training and public consultations before the live implementation of the system of value-added modelling.

Increased transparency and the greater accuracy of value-added estimates were very important in the minds of teachers and other stakeholders. Levels of confidence in the system increased once teachers were trained to compute value-added estimates. Some teachers who initially were afraid of a new measure that could be used for school accountability became enthusiasts of value-added modelling when they realised that it was a much fairer evaluation than the system that had already been in use for several years in Poland (Jakubowski, 2007).

Use of socio-economic characteristics in value-added modelling

A key argument for constructing value-added models, rather than simply using raw test scores as measures of school performance, is that raw test scores are the cumulative outcome of students' learning experiences and are influenced by many factors beyond the schools' control. Perhaps the key 'external factor' is the distribution of social and economic characteristics within and between schools that are related to student performance. Such socio-economic factors have been shown in numerous studies to influence student performance and outcomes (OECD, 2007c). Performance tables that rank schools by raw test scores or entrance to a higher level of education do not take into account the numerous factors that can affect disadvantaged students and thereby unfairly compare schools that educate these students.

In value-added modelling, a school with a student population of lower socio-economic status than average could receive a value-added estimate near zero (*i.e.* average) or above, even though the mean absolute performance of its students might be well below the mean for all students within the school system. This point highlights the importance of undertaking a value-added analysis. When making inferences about schools' performance, it is important to take into account the reality that different schools face very different challenges in educating students. This analytical power can be increased with the inclusion of contextual socio-economic characteristics in value-added modelling. These models have been termed contextualised value-added models in this report. The use of relevant background characteristics can result in school-level value-added indicators that are both more accurate and more credible.

Given the need for more credible and accurate results, countries are increasingly collecting and utilising socio-economic data for value-added modelling and other measures of school performance. However, it should be noted that schools' contextualised value-added scores might not be suitable for all policy decisions. There can be concern that contextual variables can mask low levels of student performance and thereby skew incentives and decision-making that might actually reinforce existing disadvantage in schools with a high proportion of students with lower socio-economic status. This can have an impact on the schools themselves and also on policy development. Important objectives of the use of value-added modelling for school accountability and school improvement purposes include the incentives created to lift student and school performance and the use of data at the school-level. Countries that have implemented systems of value-added modelling have done so, at least in part, to provide a more meaningful incentive to lift student performance (Bourque, 2005; Ray, 2006). By publishing schools' value-added results, incentives can be created for school administrators, teachers and other stakeholders to lift the performance of schools in that measure. The incentive to lift performance might be lowered in schools that have substantially higher contextualised value-added scores that take account of differences in socio-economic status. This might lower expectations and reduce incentives even in schools where the proportion of students with low absolute performance is worryingly high. Therefore, the introduction of contextual variables into the value-added model might have undesired consequences for the incentive effects upon schools. Also from the perspective of students and their families, school value-added measures might be of less interest compared to measures of students' absolute performance or individual student progress.

The use of socio-economic characteristics in contextualised value-added models can also have a negative impact on equity and the efficiency of decision-making, however much of this depends on how value-added information is used. There are benefits to using a number of value-added measures (and models) to make more informed decisions that serve distinct policy purposes. Consider the situation of schools with a combination of a high proportion of students with lower socio-economic status characteristics and low academic performance (as measured in test scores). Value-added models (without socio-economic contextual characteristics) might show these schools as achieving a relatively low value-added score.[3] The inclusion of socio-economic characteristics in a contextualised value-added model might show that some of these schools have high contextualised value-added scores. For this group of schools, the raw test scores of students are low and the school value-added score is also low. However, the contextualised value-added score is higher and might be much closer to the average. In determining whether the use of socio-economic contextual characteristics in the value-added model facilitates the advancement of stated policy objectives, each objective should be considered.

Analysis of value-added information at the system-level can assist decision-making in the allocation of resources in the school education system. Many education systems provide equity funding so that lower-performing schools receive additional funds. In this example, the allocation of funds would be very different if contextualised value-added modelling as opposed to value-added modelling was used. For these schools, their low value-added scores might be a signal to policy makers that additional resources would be required to assist students in these low-performing schools. However, analysis of contextualised value-added scores (that were higher for these schools) would indicate that these schools do not require additional resources despite the very low overall performance of students in these schools. The inclusion of socio-economic contextual information leaves students in these schools worse off in this scenario. It can be beneficial therefore necessary to analyse results from modelling including and excluding socio-economic characteristics.

Value-added analyses can also assist decision-makers at both the system and school level in identifying effective schools, policies and programmes. The use of value-added results that do not include socio-economic contextual characteristics would be misleading. The use of a contextualised

[3] As illustrated in Chapter Five, some of these schools might achieve high value-added scores but for this example onsider those schools that achieved a low value-added score.

value-added model better identifies those schools with a greater proportion of students from disadvantaged backgrounds that were able to lift student performance. Such an analysis would not be possible with the value-added model that did not include socio-economic contextual characteristics and would be even less likely with the use of just raw test scores. For a system of school accountability, it would therefore seem to be more equitable to utilise contextualised-value-added scores as the main indicator of school performance. Given the advantages and disadvantages of these applications of value-added and contextualised value-added models, it might be optimal to use a variety of measures designed to serve distinct policy purposes as long as users were trained to correctly interpret differences in school results across different models. Information concerning the socio-economic characteristics of students, raw test scores, and both value-added and contextualised value-added school results would enable a more detailed analysis on which to base a range of decisions.

The scenario presented above assumes that there is a substantial difference between schools' valued-added scores and their contextualised value-added scores. As discussed in Chapter Six, this is not always the case. It has been argued that in systems with more frequent student assessments that are included in value-added modelling, the quantitative importance and statistical significance of socio-economic contextual characteristics decreases to the point where they have a negligible effect upon school value-added results. This issue is discussed further in Chapter Six but it should be noted here that it will be important in the implementation of a system of value-added modelling to analyse the degree to which such differences exist across schools and the school education system.

The use of contextualised value-added models can help gain the confidence of key stakeholders who are concerned with the treatment of schools and teachers trying to educate socially and economically disadvantaged students. The inclusion of these variables can not only produce more accurate models but also send a signal to these stakeholders. The importance of communicating the message that a contextualised value-added model adequately compensates for the additional difficulties in educating students from lower socio-economic backgrounds can be vital. As discussed in Part III, a number of important steps have been undertaken by Governments to build the confidence of teachers, school administrators, parents and other key stakeholders in systems of value-added modelling. These steps focus on aspects of the design and use of value-added modelling including how the results are presented and how stakeholders are assisted in the proper interpretation of school results. In building the knowledge base of the use of value-added modelling, stakeholders have greater confidence in value-added results and the system that utilises such results. This could ease many of the initial concerns (Jakubowski, 2008).

Enhancing school accountability through the use of value-added modelling

Over the last decade, the adoption of accountability systems for schools has become more common across countries (OECD, 2007a; Kane and Staiger, 2002; Goldstein and Spiegelhalter, 1996; Hanushek and Raymond, 2004; Braun, 2006a; Taylor and Nguyen, 2006). This development might be seen as part of a broader international trend towards establishing systems that measure public sector performance in terms of effectiveness and efficiency. The aim of such systems is to facilitate comparisons of resource use, results and productivity across institutions in sectors such as health and education (OECD, 2008).

Efforts to implement accountability systems are often driven by concerns that there exists substantial heterogeneity in school performance together with meaningful differences in education outcomes for recognisable subgroups in the population (OECD, 2007b). The upsurge of interest in value-added modelling is a consequence of this renewed emphasis on holding schools and teachers accountable for their performance. Value-added modelling is one way of implementing what is often referred to as test-based accountability. Although school accountability is – or should be – broader than just test-based accountability, the latter can frequently play a dominant role. This is due, in part, to the relative cost-effectiveness of testing and, in part, the apparent objectivity of test data. However, it can be difficult, if not impossible to incorporate all aspects of school performance into a single indicator and some aspects might be inherently immeasurable (Dixit, 2002). Value-added models utilise student assessment scores so performance in these assessments would be the focus of a system of school accountability based solely on value-added scores.

School accountability is a component of the system-level monitoring and regulatory functions that are carried out by an agency in a national or state education Ministry. Regulations govern, among other things, how each component of the system operates, the credentials demanded of the different professionals in the system, and the requirements for curriculum and assessment at each level. Monitoring refers to the various mechanisms by which the authorities monitor system functioning, as well as how the findings are reported internally, to various stakeholders, and to the general public (Caldwell, 2002). Initially, accountability focused on whether schools were complying with regulations governing different aspects of school functioning, such as the number of days of instruction, class size, teacher credentials, textbooks used, as well as various matters related to financial management. In short, the focus was on input and process. It is now becoming more common to consider school accountability in relation to output measures. To assert that 'schools should be held accountable for their performance' should entail more than requiring a simple description of what transpired in schools over a designated period of time. Accountability now might require that schools provide a justifying analysis or explanation of their results. One aspect of accountability with respect to education quality is whether students are making satisfactory progress with each year of instruction. Another aspect is whether they are meeting the standards set by the authorities. With respect to the goal of equity, are all groups of students achieving the quality goals in roughly equal proportions? It might be possible to evaluate the goal of efficiency by asking whether schools operating in more challenging environments are functioning as effectively as schools with relatively fewer challenges. This last question can be addressed by relating outputs to inputs.

The focus of systems of school accountability differs across countries. Some countries put much greater focus on the performance of individual schools while in other education systems there is a system-level focus and there are relatively few references to school accountability and sometimes relatively few methods for evaluating school performance (OECD, 2007a). A number of OECD member countries have, in recent years, begun to develop systems of school accountability. As an example, in Norway the concept and measurement of school accountability has changed somewhat over the last few years. Based on the recommendations from a government commission, a national school accountability system was established in 2004. The central element of this system is an open-access website containing detailed information for all Norwegian schools. In addition to basic administrative information, the website contains a large number of indicators on resource use, the learning environment and results. The school performance indicators are basically raw school-level averages or distributions from national tests or centrally administered exams. The objective of the system is to improve the type and level of information on school performance for a number of different stakeholders, including the government itself. There are no direct sanctions or rewards attached to any of the indicators. It is intended to be a central tool in the process of school development, through identifying good practice in schools that do well and identifying schools that have potential for improvement. Central authorities might use the system to monitor the general level of development in the school sector, and local authorities and school owners might monitor the level of development in their own schools compared with other schools. Finally, parents, students and the general public now have comprehensive and standardised information on aspects of schools that are of particular interest to them, instead of having to rely on more anecdotal and unsystematic evidence (Haegeland, 2006).

The construction and publication of performance measures alone might provide implicit and indirect (including monetary and non-monetary) incentives to school principals and teachers (Glenn and de Groof, 2005). In addition, public sector accountability systems might also include explicit sanctions and rewards, both to institutions and their employees. The government, by disclosing information about school performance and perhaps connecting rewards to performance, might induce teachers and administrators to respond by increasing their efforts to increase performance (Bourque, 2005). To analyse the type and effectiveness of incentives created through the development and use of value-added modelling in education systems, the incentives created for teachers and school principals must be analysed. In value-added modelling, the school is the unit of accountability and therefore improvements in learning are aimed at improvements in the school as an organisation. However, it needs to be recognised that the impact on the organisational learning of schools is primarily achieved through the impact upon teachers in the classroom (OECD, 2005). While schools are the unit at which output is measured, it is for individual teachers and school principals that incentives are created either collectively or individually. Such incentives have two main effects: the incentive effects created for teachers and school principals; and, potentially equally important, the sorting or selection that occurs in the labour market in these professions in response to these effects (Lazear, 2000).

The structure of incentives influences the actions of organisations and workers throughout public and private organisations (Ballou, 2001; Doeringer and Piore, 1985). There is no reason to believe this would not also be the case in the education sector. Teachers and school principals should be expected to respond to both positive and negative incentives that might influence the education that students receive. Lavy (2002) illustrates the positive effects of an experiment run in the Israeli education system whereby teachers were rewarded for increases in student test scores. In a carefully designed incentive structure, teachers were rewarded with a variety of monetary incentives for increasing student performance. Substantial positive effects accrued with increases in student performance reflecting the creation of incentives. Moreover, the incentives were structured in such a manner as to reward teachers of lower-performing or more disadvantaged students. Greater rewards were offered to teachers who achieved increases in the performance of students who were either previously relatively low-performers or considered more disadvantaged. Positive outcomes were evident with low-performing students achieving significant gains. This illustrates the possibilities for policy makers in designing incentive structures to achieve gains in student performance in areas where they are considered to be most valuable.

The most direct incentives that can be created with the use of value-added modelling are those that identify the value-added of individual teachers and provide commensurate rewards. School-level value-added models differ in their focus but still provide a variety of incentives to teachers and school principals. These incentives can employ both monetary and non-monetary outcomes and can have a variety of effects. School-level value-added scores can be used to provide incentives for all teachers within a school or the scores can be disaggregated to identify particular groups of teachers (*e.g.* teachers of a particular subject). It should be noted that an additional layer of complexity is created in identifying incentives from the school-level value-added modelling that are the focus of this report as the unit of measurement (the school) can differ from the target of the incentives (teachers). This places greater responsibility upon the leadership of schools to ensure that all teachers and staff members are working together to achieve school objectives.

As mentioned above, incentives can take a variety of forms for teachers and school principals and differ with the level of outcomes, intended or otherwise, of a system that utilises value-added modelling. The outcomes from the development of systems utilising value-added modelling can, in general, be placed into four broad categories:

1. *Direct monetary outcomes:* These would take the form of rewards and sanctions that alter the financial compensation received by teachers and school principals. Examples would include financial bonuses or increases in salary received by teachers from a positive or high school value-added score (OECD, 2007a; Figlio and Kenny, 2006).

2. *Non-monetary outcomes*: These mainly consist of rewards such as additional professional development and changes in work responsibilities and the personal rewards from working in a successful school. This might lead to greater job satisfaction and the prestige that comes from an increased standing in the professional community. It should be noted that these outcomes are often evident in conjunction with outcomes in the other three categories (OECD, 2005).

3. *Workplace and school outcomes*: A variety of rewards and sanctions can be placed upon schools that can have a large effect upon teachers and school principals. Common examples these can have an impact upon school autonomy with high-performing schools being granted additional autonomy and low-performing schools placed on probation and/or receiving interventions from school inspectorates (or national equivalents). This can create a significant incentive to increase outcomes given both the stigma of being on probation and the desire for greater autonomy for teachers and school principals. At their most extreme, these sanctions can also result in school closures and the involuntary dismissal of staff (Ray, 2006; van de Grift, 2007).

4. *Career outcomes*: These can include both monetary and non-monetary outcomes. They accrue through the course of teachers' and schools principals' careers from the benefits of working in a high-performing school. This is dependent upon the interpretation and use of school-level

value-added information in the labour market that should have a beneficial impact on future pay and promotion prospects for those staff associated with high-performing schools (Ladd and Walsh, 2002).

These four effects can exist together or operate somewhat independently. As value-added modelling can focus on a variety of aspects of school performance, the models can be structured to focus on particular outcomes depending upon the objective of the system, with the strength of the incentive dependent on the size of the outcomes or rewards and sanctions.

While much focus is on the effects of direct incentives, research has shown that an equal or greater effect upon organisational productivity can occur through the sorting and selection effects within the labour market for teachers and school principals (Lazear, 2000). The effects of sorting and selection operate slightly differently from direct incentive effects. Direct incentive effects of systems utilising value-added performance measures focus on the change in the work of, and instruction provided by, existing teachers while sorting and selection effects focus on the impact in the labour market of people who choose to become teachers and those who leave the profession. An analysis of the effects of the introduction of a system of value-added modelling should include both the direct incentive effects and those of sorting and selection in the labour market.

The analysis of incentive effects focuses upon the incentives for teachers and school principals to increase the value-added scores of students and schools. An incentive is created that seeks to cause teachers to alter work behaviour to increase student performance. Sorting and selection effects occur as these incentives will attract those individuals into the profession who believe they can increase a school’s value-added score. Intuitively, this would affect the composition of new entrants into the teacher labour market. Correspondingly, it would affect the composition of retention of existing teachers and the teachers who are least able to contribute to the value-added of schools would be relatively more likely to leave the profession (Lazear, 2000). These teachers would then be replaced by new entrants who would believe that they would be able to contribute to schools’ value-added scores. Theoretically, the size of these effects will depend heavily on the size of the incentives. For example, if career progression relied heavily upon schools’ value-added scores *and* there were substantial monetary and non-monetary benefits to such career progression then both the incentive and sorting and selection effects would be magnified. However, the effectiveness of these changes relies on accurate and transparent indicators and evaluation of performance and how they are incorporated in to the broader system of school and teacher evaluation.

The use of raw test scores can provide unintended incentives given the inaccurate relationship between raw test scores and the performance of schools. Value-added scores provide a more accurate measure of school performance that would improve information flows in the labour market. It is therefore possible to shape incentives so that their impact advances desired policy outcomes. A prime example of this is to structure incentives so that the greater part of their effect is directed to disadvantaged or low-performing students. For example, incentives could be created for higher-performing teachers and school principals to move to low socio-economic status schools where improvements in value-added receive greater rewards. In this way, the system might be able to counter a trend in many education systems where more experienced teachers are more likely to work in schools with higher socio-economic status students (OECD, 2005). Few education systems currently link teacher and school principal remuneration directly to outcomes in value-added modelling. However, it should be noted that the affect of sorting and selection in the labour market can be as important as direct incentives. This effect can also have a longer timeframe than direct incentives. As an example, consider a school principal who is at a relatively early stage in his or her career and is principal of a school in a relatively low-socio-economic status community. Now, consider a system of value-added modelling that utilises student tests in the language of instruction, science and mathematics in Years 3, 5 and 7 in the school. Even if this system has no direct link to the principal's remuneration, there is a clear career incentive to improve performance on these tests as a considerable portion of his or her career as a school principal remains. If the principal is successful in lifting the school's value-added score, this achievement can be used in the job market. After five years at that school, it is possible to enter the job market citing the value-added scores that show the ability to lift student performance in a low-socio-economic status school. The principal have a relative advantage over other people competing for the position and could thereby expect to be commensurately rewarded to the extent that the labour market for school principals can offer such rewards. This incentive would be increased if additional resources were allocated to reward school principals and teachers of these students. However, Ladd and Walsh (2002) illustrate that if school performance measures are misspecified and incentives not properly structured then the reversal of this pattern can occur with teachers moving to schools that serve more socially advantaged students.

The size of these incentive effects is affected by the structure of the labour market for teachers and school principals which vary considerably across countries. For instance, an education system with a more flexible labour market and relatively higher degree of school autonomy might be able to create greater career incentives. Another key factor is the extent to which value-added information is made available and is able to be utilised by both employers and employees to inform hiring, firing and general mobility between schools within the labour market. However, the four categories of incentives listed above illustrate the point that incentives to lift student and school performance can be created in education systems that do not offer direct monetary incentives or merit pay to teachers. Non-monetary, workplace, and career incentives can facilitate increasing school performance. This is particularly important given that relatively few OECD member countries offer performance-based pay to teachers (OECD, 2007a)

Incentives and sub-optimal outcomes

Whenever a performance measure is created there is the potential for negative or sub-optimal outcomes if processes or even outcomes can be manipulated to erroneously create a positive performance measure. The manipulation can be a direct result of perverse incentives created through the setting of the performance target. Such perverse incentives can arise when the performance measure has both a large impact upon actors and focuses on an aspect of schooling that does not reflect the true or overall purpose and objectives of schools. Unfortunately, this can be common in school performance measures if the performance measure is too narrowly defined such as focus on a specific subject or on a specific performance level or the measure does not accurately measure school performance.

Clearly the choice of assessments used for value-added modelling creates an incentive to increase performance in those assessments. A perverse incentive can potentially lead to a sub-optimal outcome if resources are devoted to increase performance in those specific assessments at the expense of other areas of schooling (Nichols & Berliner, 2005). However, it should be noted that this is only a sub-optimal outcome if this is not the intended purpose. A greater emphasis on the assessments that create the school performance measure may be an intended consequence and design feature of the performance management system. The same incentives can exist if a specific performance level is the focus of the performance measure. For example, if value-added is calculated for students reaching a specific benchmark literacy level, then the incentive is created to focus on a particular sub-group of students at the expense of other students. Great care therefore needs to be taken in using value-added scores to identify schools as low or high performing. Unless specific objectives such as minimum literacy levels are explicitly identified with the consequences detailed, then a school performance measure should focus on the performance of students of all abilities.

As discussed above, the size of the incentive created depends upon the actions stemming from the measure of performance. The larger the impact upon schools and teachers (*e.g.* financial rewards and sanctions), the larger the incentives created. In addition, the degree to which teaching practices and the curriculum can be altered with the implementation of a system will depend upon the degree of autonomy that schools and teachers possess. While most education systems give significant degrees of autonomy to schools and teachers in regard to the teaching practices they employ, many have a prescribed curriculum (OECD, 2007a). However, within a prescribed curriculum there is usually scope to allow schools and teachers to emphasise certain aspects and to fashion practices such as student assessment to focus on particular measures. An often-cited example of the impact of school performance measures is the 'teaching to the test' in systems with high-stakes testing (Haney and Raczek, 1993; Kohn, 2000).

An additional issue is the potential for the narrowing of the curriculum. Many systems do not incorporate student assessment in all subjects. The viability of such an undertaking and various resource constraints might preclude such a structure of student assessments. Instead, assessments in only a few core subjects are normally used (see Table 4.1). Narrowing the number of subjects assessed might create an incentive to adapt the school curriculum and teaching practices to achieve higher performance measures in the subjects that are the focus of the value-added performance measure, potentially reducing an emphasis upon the full range of subjects available to students. This effect of the narrowing of focus applies to the use of all types of performance measures, not just value-added scores.

Most countries only include two or three subject areas in their student assessments that are suitable for value-added modelling. These are most commonly the language of instruction, mathematics and science (see Table 4.1). School principals and teachers therefore have an incentive to focus more heavily on the subjects included in the performance measurement. However, it is important to note that there is no systematic evidence of a narrowing of the subjects taught in schools that are subject to such performance measures (Jacob, 2002). However, in a study of schools in the USA, O'Day (2002) found that the test specification used in high-stakes testing became the curriculum specifications for a number of schools.

Incentives that focus on more narrowly defined performance measures should not be viewed solely in a negative context. A greater focus upon particular student assessment outcomes might have a positive effect, particularly if it is considered that schools or systems suffer from misaligned objectives. This might be particularly true if a greater focus on the areas of assessment have a positive follow-on effect upon other instruction and learning areas that are not included in the output measure. For example, a system that provides an incentive to increase the focus upon student performance in particular measures of mathematics might have a positive effect upon student learning in other areas. This might occur for two reasons. First, improvement in the measured aspects of mathematics might facilitate student learning in other areas of mathematics and in other subjects. Second, a greater focus upon improving student performance in mathematics might encourage other areas of the school to learn from these experiences and increase effectiveness across the school. This might have a flow-on effect to student performance in non-measured areas.

The impact on the curriculum can be a direct policy choice but both the intended and unintended impacts should be assessed to avoid the unintended consequences of such choices. Given that these consequences can be both positive and negative, it seems appropriate for policy makers to monitor these outcomes through the development of value-added modelling in their education system. This would further add to the information in the education system that might aid school and system development. The emphasis upon the choice of output measure is made here to illustrate the point that if the use of value-added models will have an effect upon schools then the choice of the subject areas that are to be assessed matters. Similar issues exist with the process of how schools' value-added scores are calculated across multiple assessments. While a value-added score can be calculated for assessments in each subject, if a single school value-added score is going to be applied in a system of school accountability, then a choice needs to be made between the value-added scores of different subjects. As discussed above, specific subjects can be chosen if emphasis needs to be given to that area of learning. Alternatively, an average of a number of subjects can be calculated and used. However, in such circumstances the average might conceal differences between subjects (Wilson, 2004). It should also be noted that the choice of the assessment measures used in value-added modelling should not obscure the need to utilise other measures in making decisions about school improvement and other policy objectives. These measures could include data on school inputs and various measures of school processes.

Improving school choice with value-added information

The effectiveness of school accountability decisions rest to a large degree on the accuracy and appropriateness of the performance measure to which schools are held accountable. Value-added information therefore needs to be both accurate and transparent, both of which are increased with the publication of schools' value-added results. This information could also advance school choice. However, it should be noted that in a number of countries school choice does not exist. Families do not have the right to choose the school to which their child will be sent. In most of these systems, the child will simply attend the local school regardless of the wishes of the family (OECD, 2006). In other countries, school choice exists with limits placed upon the schools that students can attend and entrance requirements to particular schools acting as a further impediment to free school choice. In addition, countries might place no legal or administrative requirement upon school choice but the geographic proximity of schools and the capacity of schools to meet high-demand can limit the extent to which freedom of school choice truly exists.

Much has been written about school choice and how it improves education systems by allowing students and families to choose the school that best suit their needs (Hoxby, 2003). Through this mechanism, education is enhanced through students' learning needs being better met (Levacic, 2001). Families choose the school for their children based on a number of reasons: geographical proximity; the programmes offered by the school; the peer group into which their child would integrate; and religious orientation are just some of the reasons upon which families might base their decisions concerning school choice. Schools' value-added scores would also become an important factor for families and students choosing the school they wish to attend (OECD, 2006).

The signals sent by students and families choosing the schools that best suit their needs are key elements of the proposed benefits of increased school choice within education systems. As students and families move to those schools that better meet their educational needs this will provide schools, administrators and policy makers with clear information about which schools parents and families consider to be most effective (Hoxby, 2003). This can inform decisions of resource allocations, the processes and programmes that are offered and enacted within schools, and should also feed into system-level learning. A key aspect of the provision of information to inform school choice is that stakeholders are informed of school performance (OECD, 2006). While this has clear implications for the accountability of schools to such stakeholders, it can also facilitate the involvement of stakeholders in improving school performance. Once stakeholders have access to reliable information and accurate measures of school performance they are empowered to engage with schools in efforts to lift performance. To do this, stakeholders must be able to properly interpret value-added information. This is discussed in Chapter Two and Part III of this report.

The use of school evaluative and performance information differs across OECD member countries. In a few countries there is relatively little information on student performance in national examinations or national assessments. Approximately two-thirds of OECD member countries make school inspection and evaluation information available to the general public. Just under half of these countries have reported doing so in order to improve decision-making in a system of school choice (OECD, 2007a). Since 2001, information about individual school results and other facts have been published on a national basis by the Swedish National Agency for Education (Antelius, 2006). The purpose is to facilitate the identification of the factors that influence school results and to contribute as background for discussions and analysis of opportunities, processes and results in schools. The Swedish National Agency for Education also publishes expected school results for each individual school.[4] The expected school results are estimated

[4] This applies only to schools at the compulsory education level.

with/using linear regression analysis.[5] The residual, calculated as the difference between schools results (in terms of average grade points) and the expected result of the school, is then used as an indicator of school performance given the composition of students across schools. However, these are not value-added measures and therefore do not estimate the individual school's contribution to students' progress over time.

In France, the French Education Ministry publishes school performance results that measure the performance of students in schools in achieving the *baccalauréat*. These are not value-added measures but the French Education Ministry's purpose in publishing the *lycée* performance indicators each year is to make information available of the performance of national public education services and to give the heads of educational institutions the proper tools to help them improve the effectiveness of their policies and programmes (MNEHER, 2006). The publication of results is sensitive as there is no single definition of what constitutes 'good results' for an individual *lycée*. For example, the question remains unanswered as to what criteria should be adopted to evaluate a *lycée*'s results. In this case, student and parent objectives might differ. Some put emphasis on obtaining the *baccalauréat* in a given series and are therefore willing to repeat a year or to change institutions to do so, while others prefer to complete their entire education in one *lycée*. Others simply want to get their *baccalauréat* as quickly as possible. Overall, it is considered that it serves little purpose to establish a list or ranking of top-performing *lycées*, and any number of indicators could be determined to correspond to the various expectations of different people. As a result, two guidelines have been drafted for drawing up the *lycée* performance indicators:

- giving complementary points of view on *lycée* results;
- offering a relative assessment of the institutions' contribution, taking into account the characteristics of their students.

It is assumed that parents, national education personnel, journalists, and a host of public and private actors are all interested in evaluating each individual *lycée*'s performance and the contribution it makes to the initial level of the students taught there. By publishing the *lycée* performance indicators every year, the Ministry is attempting to provide information to help answer this rather sensitive question (MNEHER, 2006).

In England, raw test scores were used to facilitate school choice prior to the development of their extensive system of value-added analyses. In 1992,

[5] An Ordinary least squares regression model used the schools average grade as the dependent variable and gender, foreign background and parents' education as independent variables.

the Performance Tables[6] for schools were introduced with the aim of informing parents in their choice of school and providing schools with an incentive to raise their standards. The first tables showed results in the GCSE exams taken by 16-year-olds (along with one indicator for A-levels taken by 18-year-olds). In 1996, the first tables for primary schools were produced with results for the new Key Stage 2 tests taken by 11-year-olds. Over time, the tables have come to include more indicators, partly as a result of the greater quantity of information available at national level. The first value-added scores for all secondary schools were included in 2002, with value-added scores for primary schools following a year later. The objectives of the tables remain to provide consistent and accessible national data on the performance of schools, to provide information to parents and the public more generally, and to ensure that schools are accountable for their results (Ray, 2006). The tables are resource-intensive to produce accurately every year and are deliberately restricted to a limited range of key indicators. They therefore do not, for example, provide results or value-added for every subject taken at Key Stage 4. Users are directed to the National School Inspectorate's inspection reports for a fuller picture of a given school. Users are also told that value-added measures represent a better estimate of school performance than the raw results that take no account of prior attainment. As noted above, the new School Profiles also include the Performance Tables value-added measures. The presentation of these performance tables is discussed in Chapter Two.

The use of value-added modelling of school performance enables school choice to be based on more accurate measures. It should therefore enhance the effectiveness of a system of school choice to the extent that school performance determines the choice of the most appropriate school. Improvements in decision-making derive from parents being better informed of the performance of schools. Effective school choice could be further facilitated if value-added information and scores were provided for different groups of students (Wilson, 2004). This would enable parents and students from these groups to better choose the school that meets their educational needs. As discussed above, decision-makers can utilise information garnered from observations of the schools that families choose to best suit their needs. If families' choices are better informed through the use of value-added modelling then the decisions made through the school system are also better informed. This increases the efficiency of the system in two ways: with families able to send their children to schools that best suit their educational needs; and the school system able to learn from these choices and develop school practices that lead to the increased performance. School choice has a reduced positive impact in an education system that does not have

[6] Now called the School and College Achievement and Attainment Tables, but for brevity referred to in this paper as 'Performance Tables'.

meaningful indicators of school performance. Parents and families cannot make informed choices, schools and policy makers cannot make performance-enhancing responses to a changing pattern of demand based upon accurate measures of school performance, and schools cannot be adequately rewarded for their performance.

The provision of value-added information can foster a culture of data-based decision-making that fosters school improvement. Such decision-making would enable effective responses to changes in the demand for school education. It can be beneficial to provide more than a single performance measure to inform school choice. The provision of value-added data alongside 'raw' test score data provides parents and families with additional information upon which to aid their decisions concerning school choice. In deciding the school which best suits their needs, families might be as interested in the overall performance of students in schools as in differences in value-added measures of school performance. Efforts to educate the families and the general public in how to interpret value-added measures and the differences with raw attainment scores will prove beneficial to the system of school choice. Initiatives to inform and educate users of value-added date have been considered crucial in a number of countries. They are discussed in further detail in Part III of this report that looks at the implementation of value-added models.

Conclusion

Key policy areas of school improvement, school accountability and school choice are presented separately here but they can often be considered to be complementary objectives, especially given increasing levels of school autonomy in a number of systems. Intuitively, the greater levels of accuracy achieved with the use of value-added estimates as school performance measures increases the efficiency impacts of decentralisation initiatives in the education system. As decentralisation shifts decision-making responsibilities to the school-level, the use of value-added information enables such decision-making to be made on an informed basis. It can empower schools to more efficiently allocate resources and alter the education they offer to improve their value-added results. But such decision-making requires a degree of school autonomy that allows schools to alter the education they provide to better meet the demands of students and parents in a system that emphasises greater school choice.

In some education systems, decentralisation of the school education system, the system of school choice, and the funding mechanisms for schools are combined to provide an incentive for schools to compete for students and therefore greater budgetary resources. The development of a system of value-added modelling would increase the effectiveness of such a system. Decentralisation enables schools to respond to changes in the demand for school education to attract greater numbers of students (Sandstrom and Bergstrom, 2005). For the additional students they attract, schools also receive greater resources from the central administrative unit as funding is provided on a per student basis. This relies on a system of school choice that enables parents and families to choose the school that best suits their needs. Such choices require the information upon which families can base their decisions to be made available. As value-added modelling provides more accurate measures of school performance, decision-making would improve and students would choose those schools with higher value-added scores. These schools would then be properly rewarded for their better value-added performance. The increased effectiveness of using value-added information to promote school choice thereby improves the effectiveness of the allocation of resources in the school education system.

Chapter Two

Presentation and Interpretation of Value-Added Models

As defined in the Introduction, value-added models are a class of statistical models that estimate the contributions of schools to student progress in stated or prescribed education objectives (*e.g.* cognitive achievement) measured at at least two points in time. Value-added modelling can produce comparative results that do not provide absolute measures of progress in student learning but measures of the relative contributions of schools to student learning, when learning is measured by changes in test scores over time. The outputs of value-added models vary with the model employed and the presentation of the results can be structured by varying the unit or level of analysis to suit the intended purpose and audience. Value-added measures can be calculated and presented for individual students, subject areas, grade levels, and schools. It is also possible to calculate and present value-added measures for regional and more local areas. However, it should be noted that composite value-added scores that present a single measure for groups of schools can lead to problems of interpretation if the intention is to analyse aspects of school performance and if there is variation in school performance within the specified regions or local areas.

This chapter discusses how value-added models can be presented to assist in effective interpretation that advances desired policy objectives. This includes an analysis of the advantages and potential hazards in classifying the performance of schools (*e.g.* high- and low-performing). Several examples are provided of how value-added information can be presented to assist in accurate interpretation. These examples illustrate the advantages of developing a comprehensive system whereby schools' value-added scores are used, for example, to create benchmarks and standards as a basis for actions to advance policy objectives. The chapter ends with a discussion of the presentation of value-added information in the media and the need to ensure that this coverage does not distort information flows and impede actions for school improvement.

Value-added modelling can be used to classify schools as high- or low-performing (or somewhere in between). Obviously, such classifications need not be made and value-added information can be assessed and used without placing schools into particular categories. It can be politically difficult for administrators, policy makers and stakeholders to classify a school as low or under-performing and it is important they are based upon statistical or valid conceptual criteria (*e.g.* value-added score statistically significantly different from the mean) and are not chosen arbitrarily. It is an important decision that can have a major impact upon schools and, depending on the structure of the school system, upon the level of their resourcing and development. The difficulty for administrators and policy makers might come from opposing pressures. On the one hand, a classification might need to be made in a time-efficient manner both so that appropriate actions can be implemented to remedy problems and so that issues such as under-performing students and schools are addressed as soon as possible. In these instances, value-added information needs to be translated into meaningful actions so that, for example, schools with value-added scores statistically significantly below the average for 2 years running are classified as low-performing schools that quickly translate into actions of a school evaluation and additional resources being invested in the teaching of their low-performing students. On the other hand, working in a school that has been classified as low-performing can have negative effects upon school principals, teachers, students and other stakeholders. The extent and impact of these negative effects depends on the structure of the system in which the classification is made and the actions that stem from such a classification. There is substantial pressure therefore to ensure that accurate measures are obtained so that the classification of schools as low- or high-performing is fair and accurate.

Numerous statistical and methodological issues are identified in Chapter Six that need to be considered in the development of value-added modelling and the interpretation of their results. These issues include the potential for various measurement errors and potential sources of bias in estimations. These, and similar issues need to be considered in the difficulty of classifying schools as low- or high-performing. The statistical caveats that are discussed in interpreting such classification mean that it is easier to identify when *not* to classify a school as low- or under-performing than when it is appropriate to do so through the use of value-added modelling. This difficulty needs to be balanced with the imperative to utilise the results of value-added models as a basis for action that might include classifying schools and then implementing the required actions. Stability of results from one year to the next is also discussed in Chapter Seven and Part III of this report. This discussion includes the recommendation that a three-year moving average of value-added results be used as the main value-added indicator for schools in the application and presentation of results.

It is important that the focus of a discussion of statistical and methodological issues does not remain solely on the caveats of a value-added model or that policy makers are too cautious in classifying schools as requiring specific actions, then it can potentially undermine the reasons for implementing a system of value-added models. Utilising value-added models to create a system for school improvement or school accountability requires that schools be evaluated and consequent decisions made. These decisions should, as part of overall policy objectives, be aligned with the goal of improving the school system. If too many obstacles are placed in front of administrators or policy makers before a school can be classified, then it can prevent the required actions being taken. For example, consider a system in which schools that are classified as low-performing receive extra evaluations and assistance. Once a school is classified as low-performing, it receives an evaluation from a school inspectorate and undertakes a self-evaluation to analyse the reasons for the low performance. Depending on the outcomes of this additional evaluation, additional assistance or resources might be provided, professional and organisational development undertaken and a monitoring system developed to track future performance. Impediments or resistance to the classification of low performance would therefore also impede the application of initiatives to improve the performance of these schools and students.

Presenting value-added information

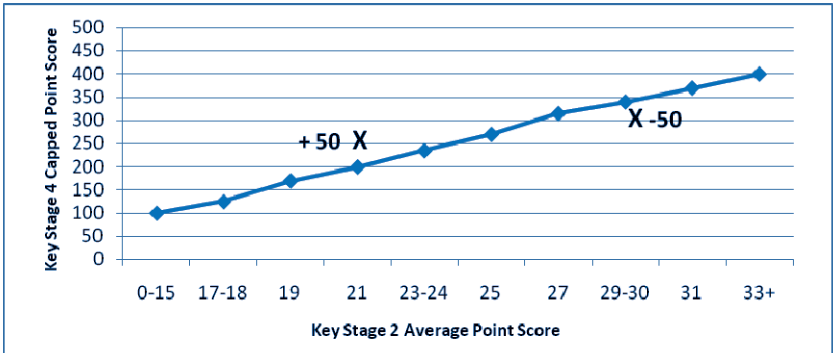
Presentation of value-added information and schools' value-added results must take into account the issue of how to best present statistical issues that can be complex to a non-statistical audience. Part III of this report emphasises the importance of stakeholder education and training in implementing a system of value-added modelling. It is also important that the presentation of value-added information is clear and transparent to stakeholders to maximise the benefits of implementing a system of value-added modelling. As illustrated in the examples provided below, numerous steps can be taken to ease the problems of interpretation and facilitate an effective understanding of what value-added scores represent and their use in advancing stated policy objectives. Clarity is required to achieve such aims and it is important to note that despite the complexity inherent in value-added modelling, simplified presentations of value-added models and related information can be effective in achieving clarity.

Value-added results are numerical and can be presented in numerous forms such as a continuous score or one that is specified as being either above or below an expected performance outcome for the school. Depending on the structure of the database and the kinds of analyses performed, school value-added estimates can be presented by subject, grade and student characteristics. The relative positions of different schools can serve as a useful starting point for discussions regarding school development. Such discussions could also take into account other school characteristics such as the profile of the teaching staff, student mobility, and particular local and community issues. As is evident in their use across education systems, a number of possibilities exist for the presentation and use of value-added information.

A number of different value-added models have been calculated in England as the system has developed over time. Distinct models have also been used to analyse particular aspects of the school education system. A contextualised value-added model is the more complex model used in England that controls for the influence of various socio-economic characteristics upon changes in student performance. It also includes a number of other characteristics that influence student progress that are outside the control of schools such as students' birth month (see Table 4.2 for more information). This is analysed alongside the median method of presenting the results that was designed and used to illustrate in a simplified manner the calculation of a school's value-added score. An example of this median method is provided below. It has been used in England to illustrate the fundamentals of value-added modelling and explain how they should be interpreted so that they can easily be utilised by schools. School results of the value-added 'median method' have been published in the form of charts showing median outcomes from each prior attainment point. This was adapted for the calculation of school scores, which are derived as the average for each school of the differences between each student's actual result and the national median result for students with their prior attainment score.

The median method was designed for its simplicity and clarity and because it could be easily integrated into the production cycle for performance tables that were, and continue to be, used in England. The method also had to allow schools to be able to calculate their own value-added scores with reference to information on the national expected results. Rather than use a regression model, the method was based on the median lines familiar to schools from previous developments in this area. In this system, a school can look at the prior attainment of each student and compare it with the median line, the difference being the contribution of that student's value-added score to the value-added score of the school. Figure 2.1 provides an example calculation using a median line. One student has achieved 50 points more than 'expected' at Key Stage 4, given their prior attainment at Key Stage 2. Another student has achieved 50 points less than expected. The sum total of the vertical distances to the median line, divided by the total number of students, is the school's value-added score.

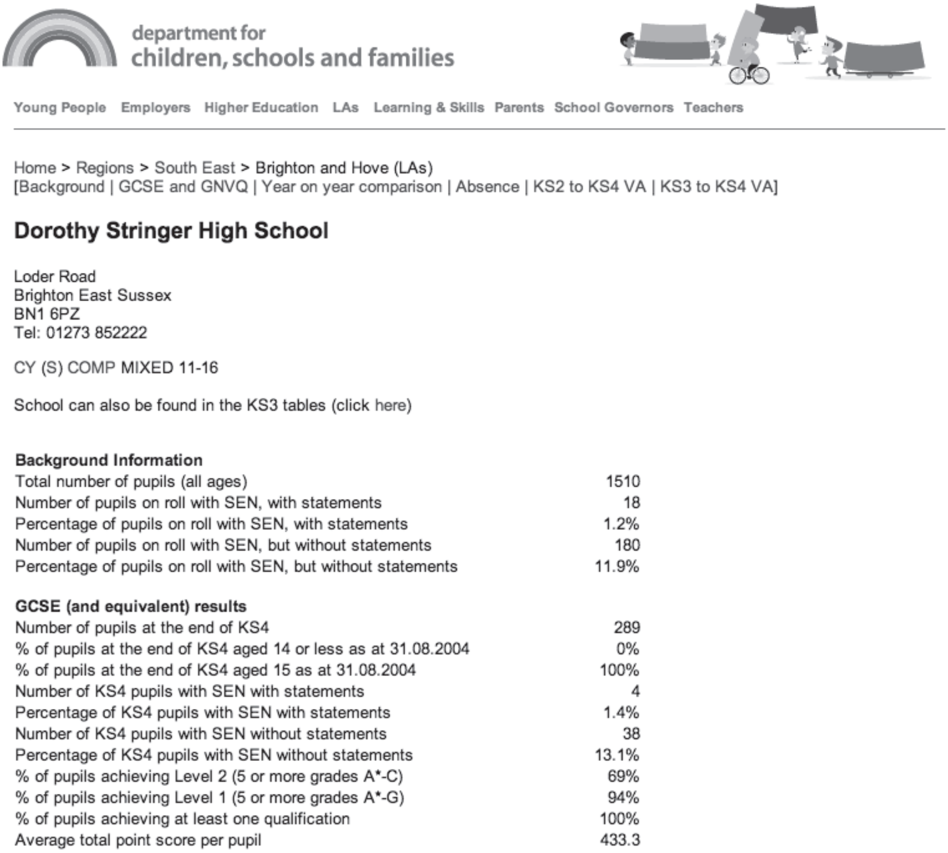
Figure 2.1. Example median line calculation of value-added in England



Schools can therefore easily calculate and check their own value-added scores with reference to the 'expected results' along a national median line. The use of a simple ordinary least squares regression model, that provides a formula for the calculation of 'expected' results, could also be used by schools to calculate and verify results. The main reason for the use of the median method in England is its simplicity of interpretation and understanding. One method that illustrates the results of a typical pupil while employing a regression formula for schools to calculate their value-added scores was considered not to be as easily utilised by a non-statistical audience. Value-added information is presented with School Performance Tables and has been developed and published both online and in booklets for each Local Authority. The Performance Tables include a limited range of statistics on schools. Value-added data is presented alongside facts about overall attainment and school context. Figure 2.2 shows how the 2005 value-added scores, based on the median method and prior attainment only, would be presented for an example secondary school.[7] The value-added result is included alongside raw results and some contextual information. Here, the Key Stage 2-4 score of 989.8 means that the students in this school achieved, on average, 10.2 value-added points less than the median students for each prior attainment level.

[7] See http://www.dfes.gov.uk/performancetables/schools_05.html

Figure 2.2. Screenshots showing value-added on the website for Performance Tables



Home > Regions > South East > Brighton and Hove (LAs)
[Background | GCSE and GNVQ | Year on year comparison | Absence | KS2 to KS4 VA | KS3 to KS4 VA]

Dorothy Stringer High School

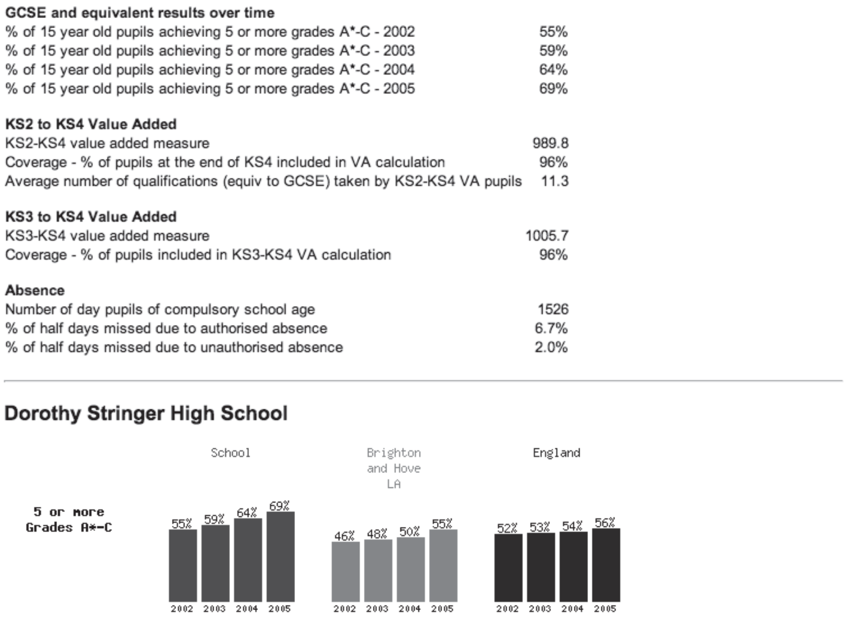
Loder Road
Brighton East Sussex
BN1 6PZ
Tel: 01273 852222

CY (S) COMP MIXED 11-16

School can also be found in the KS3 tables (click here)

Background Information	
Total number of pupils (all ages)	1510
Number of pupils on roll with SEN, with statements	18
Percentage of pupils on roll with SEN, with statements	1.2%
Number of pupils on roll with SEN, but without statements	180
Percentage of pupils on roll with SEN, but without statements	11.9%

GCSE (and equivalent) results	
Number of pupils at the end of KS4	289
% of pupils at the end of KS4 aged 14 or less as at 31.08.2004	0%
% of pupils at the end of KS4 aged 15 as at 31.08.2004	100%
Number of KS4 pupils with SEN with statements	4
Percentage of KS4 pupils with SEN with statements	1.4%
Number of KS4 pupils with SEN without statements	38
Percentage of KS4 pupils with SEN without statements	13.1%
% of pupils achieving Level 2 (5 or more grades A*-C)	69%
% of pupils achieving Level 1 (5 or more grades A*-G)	94%
% of pupils achieving at least one qualification	100%
Average total point score per pupil	433.3



GCSE and equivalent results over time	
% of 15 year old pupils achieving 5 or more grades A*-C - 2002	55%
% of 15 year old pupils achieving 5 or more grades A*-C - 2003	59%
% of 15 year old pupils achieving 5 or more grades A*-C - 2004	64%
% of 15 year old pupils achieving 5 or more grades A*-C - 2005	69%
KS2 to KS4 Value Added	
KS2-KS4 value added measure	989.8
Coverage - % of pupils at the end of KS4 included in VA calculation	96%
Average number of qualifications (equiv to GCSE) taken by KS2-KS4 VA pupils	11.3
KS3 to KS4 Value Added	
KS3-KS4 value added measure	1005.7
Coverage - % of pupils included in KS3-KS4 VA calculation	96%
Absence	
Number of day pupils of compulsory school age	1526
% of half days missed due to authorised absence	6.7%
% of half days missed due to unauthorised absence	2.0%

Dorothy Stringer High School

Source: Department for Children, Schools and Families, United Kingdom (2008).

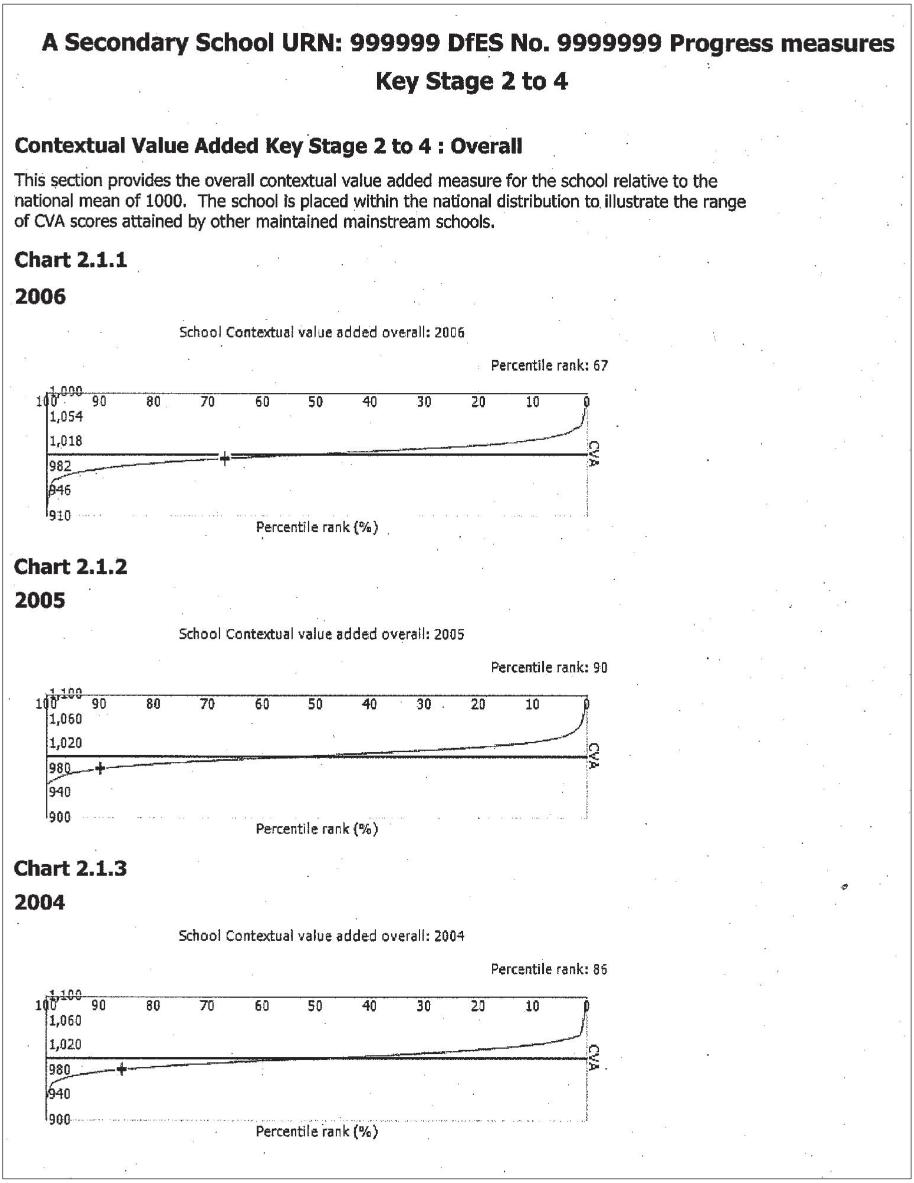
The use of school Performance Tables has been expanded with the development of RAISEonline, an interactive software program that provides analysis of school and student progress data. The program is a prime example of how value-added scores and other information can be presented to facilitate analysis at both the school- and system-level. The presentation of value-added information and the utilisation of an interactive interface in England has been a key step in facilitating the use of data at the school-level and in empowering stakeholders to use the data to advance policy objectives. The key objectives of the introduction of RAISEonline were: to enable schools to analyse performance data in greater depth as part of a school self-evaluation process; to provide a common set of analyses for schools, local authorities, school inspectors and School Improvement Partners; and to better support teaching and learning (Ray, 2006). A considerable amount of information is available to primary and secondary schools and the interactive elements of the software enable users to drill-down into the data to better analyse student and school value-added performance. Key features of RAISEonline include:

- reports and analysis covering the attainment and progress of students in Key Stages 1, 2, 3 and 4, with interactive features allowing the exploration of hypotheses about student progress;

- contextual information about the school, including comparisons with schools nationally;
- question-level analysis, allowing schools to investigate the performance of students in specific curriculum areas;
- target-setting that supports schools in the process of monitoring, challenging and supporting student progress; and
- a data management facility providing the ability to import and edit student-level data and create school-defined fields and teaching groups.

This information can be accessed through the RAISEonline website with school principals provided with a specific login and password to ensure that only they can access their school's information (Ray, 2006). An example of the presentation of value-added information that can be utilised by schools is provided in Figure 2.3. The example provides contextualised value-added information for an English secondary school. Information is presented both graphically and in tabular form to facilitate the interpretation of school results with comparisons with previous performance. As shown in the charts, this secondary school had contextualised value-added scores below the national average for each of the three years presented. For illustrative purposes, the national average is set at a score of 1 000. It was found that setting the average score at zero was not preferable as it meant that schools that performed below the average would receive a negative score and it was felt that the connotations of a negative score might have adverse effects upon stakeholders. In addition, the presence of negative scores can complicate interpretation as they can be interpreted as showing a drop in the overall performance of students (Ray, 2006). Schools' value-added scores are relative to the performance of all schools and, therefore, a value-added score that is negative (below the average) does not necessarily imply that overall student performance has decreased. To help avoid this misinterpretation, the average value-added score was set at 1 000. This secondary school had a contextualised value-added score of 994.5 in 2006 for all subjects. This was below the national average but an improvement upon their contextualised value-added score of 980.9 in 2005.

Figure 2.3. Example of contextualised value-added information provided for an English secondary school



A Secondary School URN: 999999 DfES No. 9999999 Progress measures Key Stage 2 to 4

Contextual Value Added Key Stage 2 to 4 : Overall

This section provides the overall contextual value added measure for the school relative to the national mean of 1000. The school is placed within the national distribution to illustrate the range of CVA scores attained by other maintained mainstream schools.

Chart 2.1.1

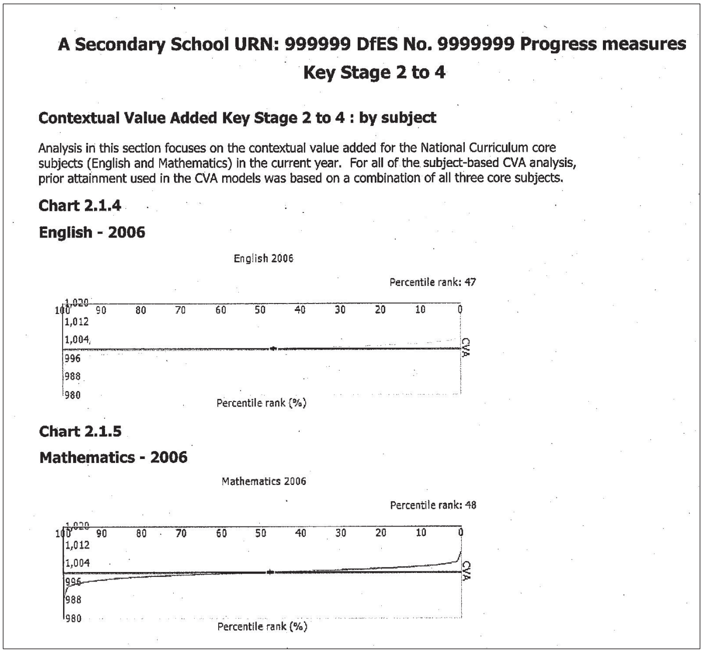
2006

Chart 2.1.2

2005

Chart 2.1.3

2004



A Secondary School URN: 999999 DfES No. 9999999 Progress measures Key Stage 2 to 4

Contextual Value Added Key Stage 2 to 4 : by subject

Analysis in this section focuses on the contextual value added for the National Curriculum core subjects (English and Mathematics) in the current year. For all of the subject-based CVA analysis, prior attainment used in the CVA models was based on a combination of all three core subjects.

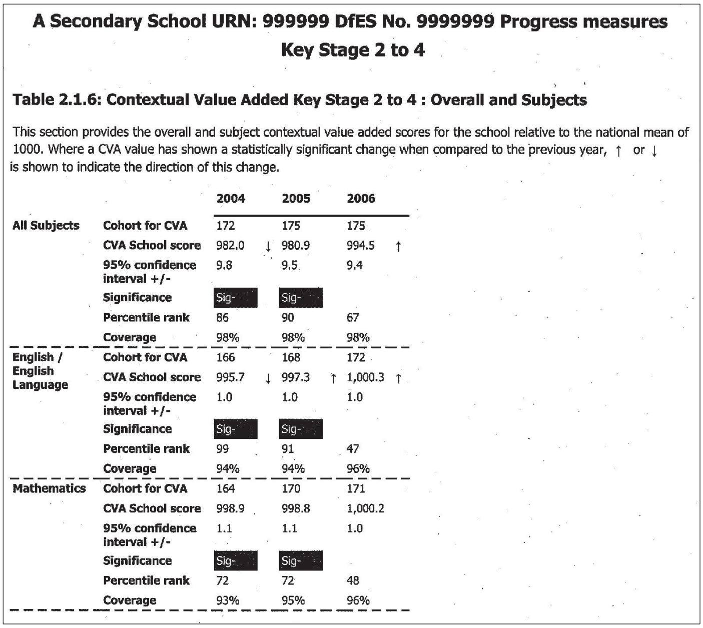
Chart 2.1.4

English - 2006

Chart 2.1.5

Mathematics - 2006

Mathematics 2006
Percentile rank: 48
100
90
80
70
60
50
40
30
20
10
0
1,020
1,012
1,004
996
988
980
CVA
Percentile rank (%)



A Secondary School URN: 999999 DfES No. 9999999 Progress measures Key Stage 2 to 4

Table 2.1.6: Contextual Value Added Key Stage 2 to 4 : Overall and Subjects

This section provides the overall and subject contextual value added scores for the school relative to the national mean of 1000. Where a CVA value has shown a statistically significant change when compared to the previous year, ↑ or ↓ is shown to indicate the direction of this change.

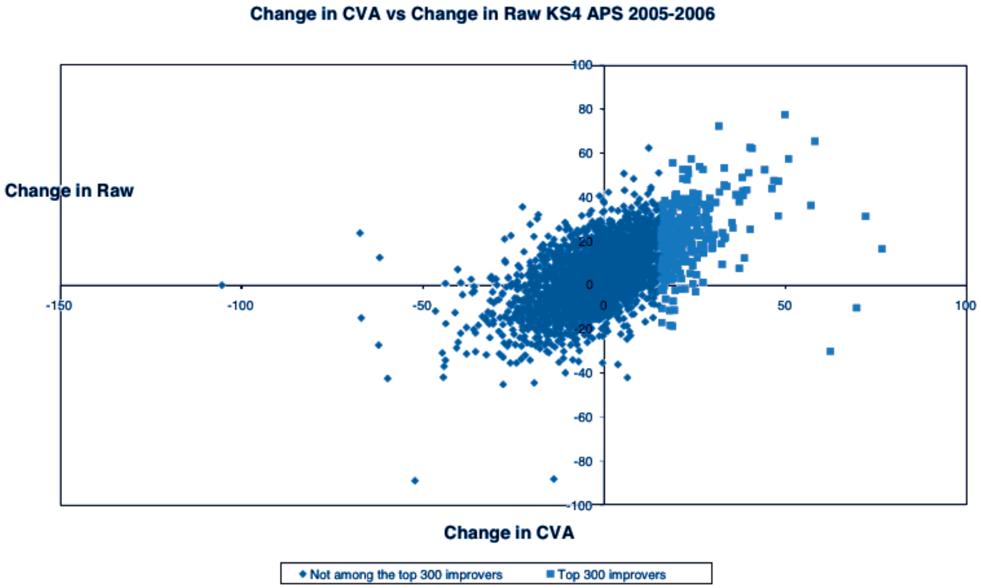
		2004	2005	2006
All Subjects	**Cohort for CVA**	172	175	175
	CVA School score	982.0 ↓	980.9	994.5 ↑
	95% confidence interval +/-	9.8	9.5	9.4
	Significance	Sig-	Sig-	
	Percentile rank	86	90	67
	Coverage	98%	98%	98%
English / English Language	**Cohort for CVA**	166	168	172
	CVA School score	995.7 ↓	997.3 ↑	1,000.3 ↑
	95% confidence interval +/-	1.0	1.0	1.0
	Significance	Sig-	Sig-	
	Percentile rank	99	91	47
	Coverage	94%	94%	96%
Mathematics	**Cohort for CVA**	164	170	171
	CVA School score	998.9	998.8	1,000.2
	95% confidence interval +/-	1.1	1.1	1.0
	Significance	Sig-	Sig-	
	Percentile rank	72	72	48
	Coverage	93%	95%	96%

Source: Department for Children, Schools and Families, United Kingdom.

The discussion of school improvement in Chapter One highlighted the opportunities for schools to analyse value-added results to identify variation in performance within schools. As illustrated in the above Table, the contextualised value-added score for all subjects in 2006 (994.5) was below the score for English (1 000.3) and Mathematics (1 000.2) that were both just above the national average. While this should not be seen as conclusive proof of low performance in other subjects, it is an indication that performance in these areas should be investigated. Further analysis of specific value-added information might illuminate the causes of these differences and internal evaluation might provide useful insights for school improvements in these areas.

As is discussed in Chapter Three, RAISEonline enables schools to conduct a variety of analyses of their performance, including analysis of the performance of individual students. Schools can compare changes in their students' contextualised value-added scores with their raw scores. This comparison can also be undertaken at the system-level to analyse the relationship between the contextualised value-added progress made by schools and the raw results of students. An example is provided using the RAISEonline software in England in Figure 2.4. These figures show that it is possible to identify those schools whose contextualised value-added scores went up without their raw results improving. These schools might be more effective in 2006, managing to maintain standards with a less able intake. The different types of improvement/decline could be categorised according to the possible changes in prior attainment (up/down/flat) and value-added (up/down/flat). Bryk et al. (1998) discuss this as different 'grade productivity profiles' and is a further illustration of how specialised comparisons among schools, as well as longitudinal comparisons for a particular school, can be powerful agents for focusing the attention of a school's staff.

Figure 2.4. Contextualised Value-Added changes compared to raw attainment changes



Source: Ray, A. (2007).

Identifying significant changes in school performance

It is important to note that users are able to identify whether statistically significant changes have occurred over time. Value-added scores that are significantly above or below the average of all schools provide a sound mechanism for classifying schools as high- or low-performing. In the example presented in Figure 2.3, significant negative changes in value-added were seen between 2004 and 2005. This is evident when looking at performance in all subjects and at performance in English and in Mathematics. In addition, the 95% confidence interval was published to illustrate the distribution of scores within this confidence interval. For 2006, the 95% confidence interval indicates a range of 9.4 points above and below the contextualised value-added score of 994.5 (985.1 – 1003.9). As the upper limit of the confidence interval exceeds the national average of 1 000, the school's contextualised value-added score is not statistically different from the average. In Poland, the development of value-added modelling has led to discussion of whether to publish confidence intervals around schools' value-added scores. It was considered that there were two key advantages to publishing confidence intervals so that value-added is reported as an interval estimate. First, it would reduce the ease with which school rankings could be produced that could be considered as a negative consequence of value-added modelling. Second, it would assist in value-added information being used not only as a method of self-assessment and school development but also as a method of evaluating educational policy and programmes at the local or regional levels (Jakubowski, 2007). For similar reasons, confidence intervals have been published in Norway (Hægeland, 2006).

Clearly, the use of confidence intervals requires greater communication and training for stakeholders. In the publication of School Performance Tables in England, RAISEonline has published guidance on how to use and interpret the value-added measures. For example, the 2005 website includes the message reproduced below, designed to assist interpretation and educate stakeholders in the increased validity of the use of value-added scores compared with raw test scores. The reference to statistical 'significance' is needed because the value-added scores are not in all cases accompanied by confidence intervals: the website sometimes gives guidance on the range of scores that can be considered 'broadly average' depending on the size of school.

> The value-added measures give the best indication in these Tables of schools' overall effectiveness. But the significance that can be attached to any particular school's value-added measure depends, among other things, on the number of pupils included in the value-added calculation. The smaller the number of pupils, the less confidence can be placed on the value-added measure as an indicator of whether the effectiveness of a school is significantly above or below average.

Such statements have the ambition of informing stakeholders how to interpret value-added scores and how they can be used to better inform decision-making (*e.g.* for school improvement purposes if being used by school principals and teachers or to aid school choice if parents are accessing the website). Statements such as these also provide clear statements of the limitations of the use of value-added results. This can assist policy makers in the use of school results and assuage some of the concerns of education stakeholders of how the data can be applied, particularly for school accountability purposes.

Creating standards and benchmarks with value-added information

The shift in public and governmental concern away from mere control over the resources and content of education toward a focus on outcomes has, in many countries, driven the establishment of standards for the quality of the work of educational institutions. The approaches to standard-setting that countries pursue range from the definition of broad educational goals and areas of competency to the formulation of concise performance expectations in well-defined subject areas. Some countries have gone beyond establishing educational standards as mere yardsticks and introduced performance benchmarks that students at particular age or grade levels should reach. It is in this context that value-added measures can play a particularly important role. The application of value-added models to trigger specific actions requires the performance of schools to be measured either against each other, or against a pre-determined standard. With respect to student growth, a standard can be defined directly in terms of average growth exceeding a pre-defined threshold. Another approach is to set growth targets for each student based on their current status, their position relative to the current attainment standard and, perhaps, historical data on the distribution of gains for similarly situated students in previous years. For example, one indicator of school performance would be based on a comparison of actual and target student gains that would provide incentives for school staff to attend to the needs of all students (variants of such schemes can be found in McCall et al. (2004) and Doran and Izumi (2004)). Alternatively, growth in different regions of the scale can be differentially valued. Hill et al. (2005) describe a methodology for building 'value tables' that capture policy makers' beliefs about student progress that can then be used to set performance standards. A number of alternative approaches to growth can be implemented but these often fall outside value-added modelling given the nature of the growth projections.

Once standards are determined for each criterion, a decision matrix can be devised to guide specific actions. Suppose, for example, that thresholds for (non-)satisfactory and exemplary performance are set with respect to each of three criteria (current status, change over time, or some combination of the two) and that the analysis is carried out only for the entire school. The combination of performance standards yields nine distinct categories with a value-added score placing a school in one of these categories. The decision matrix specifies the treatment triggered by scores in each category. For example, schools could be rewarded if they achieve the exemplary level on all three criteria for two years running. On the other hand, schools failing to achieve the satisfactory level on two or more criteria in a given year could be subject to external review.

The above examples illustrate how value-added results can provide a basis for actions. In systems of school development and school improvement, there are benefits to specifying such actions and the classifications of school performance that trigger such actions. Pre-determined cut-off scores could be used as trigger points for actions such as a school self-evaluation or an inspection by the school inspectorate as occurs in the Netherlands (van de Grift, 2007). To fashion such a scheme requires an analysis of the distribution of value-added scores in each school education system. Such an analysis in England illustrates how value-added results can be categorised. Five possible categories were considered where schools might be identified as improving if they:

- have one of the largest increases in the value-added scores (*e.g.* top 100 or top 10%);
- make a statistically significant change (at the 95% confidence level);
- move between different parts of the distribution (*e.g.* from 'low' (bottom quartile) to 'average' (middle half));
- move between different parts of the distribution defined in terms of standard deviations from the average, or from 'significantly below' to 'significantly above'; or
- improve above some pre-defined threshold.

These categories can be particularly useful in classifying schools, for activating policy and programme responses and for assessing overall school and system performance. Analysis was undertaken of the number of schools making a statistically significant change in their performance between 2005 and 2006 (as in option ii above). A comparison of school-level value-added scores in the 2005 and 2006 contextualised value-added model employed in England is presented in Table 2.1. This table utilises information from RAISEonline on whether a school's contextualised value-added score increased or decreased significantly between 2005 and 2006. On this website, information of statistical significance surrounding schools' value-added point scores is also presented graphically. Charts are provided showing contextualised value-added scores for successive years with confidence intervals around the contextualised value-added scores so that small changes are not over-interpreted. The model compares outcomes at age 16 with prior attainment at age 11 and takes into account a range of contextual data. It uses a multi-level model that shrinks the scores in smaller schools that has the advantage of reducing instability in the model. The first column is the score for the overall value-added model based upon average point score in all subjects. The other columns are models for outcomes in English and Mathematics (using the same set of input contextual variables). This illustrates the proportion of schools in which significant changes occur in a given year and informs the planning of policies and programmes enacted in response to such changes. The comparison provides for more informed decisions of resource allocations to be made and provides an overall picture of how value-added scores can be used to classify school performance. When establishing performance categories, it is beneficial during the pilot phase of implementation to analyse the number of schools that would be categorised in each performance classification.

Table 2.1. Number of schools by year on year significant change between 2005 and 2006 Key Stage 4 Contextualised Value-Added scores

	Value-added in all subjects	Value-added in English	Value-added in Mathematics
Significant increase in comparison to 2005	318	696	452
Significant decrease in comparison to 2005	430	481	422
No significant change from 2005	2337	1908	2211
Missing data	27	27	27
Total number of schools	3112	3112	3112

The table shows that for three-quarters of schools there was no significant change in the value-added for all subjects between 2005 and 2006 but in English this was only evident for 60 per cent of schools. More schools made a statistically significant improvement (22 per cent) than a statistically significant decrease in their contextualised value-added score (15 per cent). Larger year-on-year changes were also more evident in English than in maths. This is consistent with findings in England for raw attainment scores and with the value-added results in Slovenia and Poland that showed more stability in schools' mathematics and science value-added scores than in language and humanities.

It is clear that such an interpretation of contextualised value-added results provides a tangible basis upon which to trigger, for example, school improvement actions. This is an important point given that performance management systems implemented in some education systems, particularly those based on raw test scores, can provide less accurate measures and are therefore less able to distinguish between statistically significant differences in school performance (Ladd and Walsh, 2002). It should also be noted that this analysis focuses on year-on-year changes. As discussed in Part II and Part III, this report emphasises the benefits of providing three-year moving averages of schools' value-added scores to properly control for random instability in school's value-added estimates.

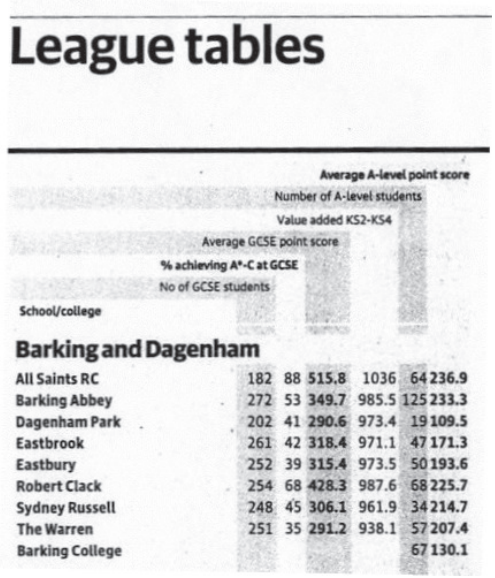
Presentation in the media

Given the impact that media coverage can have upon both the development and reception of education programmes and policies, it is important to determine both the type of media coverage afforded to value-added information and how such coverage can be managed for the effective implementation of systems of value-added modelling. In systems where families are able to choose to send their children to specific schools, the provision of the results of value-added models assists effective school choice.

Publishing results can also have an impact upon teachers and school principals and is often an integral part of a school accountability system. This might be particularly apparent if the publication of results takes the form of school rankings upon which a system of school-based rewards and sanctions is based, and also if the rankings attract substantial media attention.

In some countries, many parents will first become aware of value-added results through the media. In England, there has been considerable media attention on school performance and the publication of school results. There have also been considerable efforts to improve the interpretation of value-added results. As an example, Figure 2.5 is an extract from *The Guardian* (19/1/06) that, along with the other 'broadsheet' newspapers, published the school figures for each Local Authority in alphabetical order (although it should be noted that the title is 'League Tables'). Articles such as these in newspapers also provide a key to explain the figures, based on information published on the Performance Tables website. *The Times* (19/1/06) published one 'league table' where schools were ranked (Figure 2.6), showing the schools with the highest Key Stage 2-4 value-added (many of which were small independent schools that, as discussed above, can have larger variation in value-added cores). This could be considered a significant advancement upon the publication of league tables based on raw test scores and illustrates the progress that can be made in the presentation of school performance measures with the use of value-added modelling.

Figure 2.5. Extract from *The Guardian* newspaper (19/1/2006) showing value-added and other data

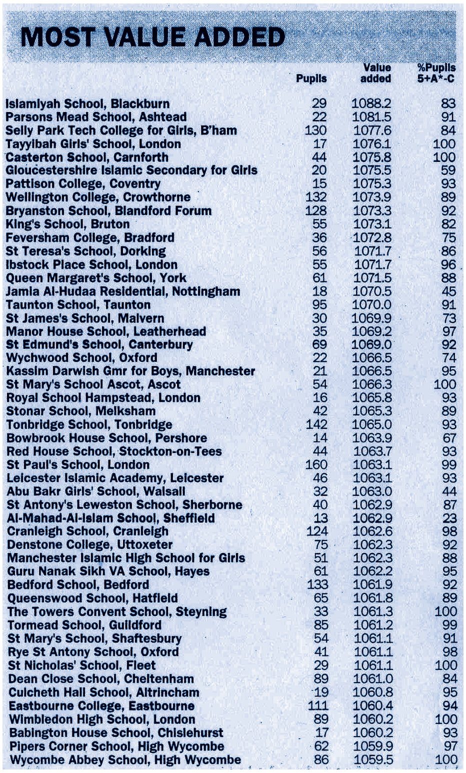


League tables

School/college	No of GCSE students	% achieving A*-C at GCSE	Average GCSE point score	Value added KS2-KS4	Number of A-level students	Average A-level point score
Barking and Dagenham						
All Saints RC	182	88	515.8	1036	64	236.9
Barking Abbey	272	53	349.7	985.5	125	233.3
Dagenham Park	202	41	290.6	973.4	19	109.5
Eastbrook	261	42	318.4	971.1	47	171.3
Eastbury	252	39	315.4	973.5	50	193.6
Robert Clack	254	68	428.3	987.6	68	225.7
Sydney Russell	248	45	306.1	961.9	34	214.7
The Warren	251	35	291.2	938.1	57	207.4
Barking College					67	130.1

Source: Copyright Guardian News & Media Ltd 2006.
Based on Ray, A. (2006).

Figure 2.6. Extract from *The Times* (19/1/2006) showing a value-added ‘league table’



MOST VALUE ADDED

	Pupils	Value added	%Pupils 5+A*-C
Islamiyah School, Blackburn	29	1088.2	83
Parsons Mead School, Ashtead	22	1081.5	91
Selly Park Tech College for Girls, B’ham	130	1077.6	84
Tayyibah Girls' School, London	17	1076.1	100
Casterton School, Carnforth	44	1075.8	100
Gloucestershire Islamic Secondary for Girls	20	1075.5	59
Pattison College, Coventry	15	1075.3	93
Wellington College, Crowthorne	132	1073.9	89
Bryanston School, Blandford Forum	128	1073.3	92
King's School, Bruton	55	1073.1	82
Feversham College, Bradford	36	1072.8	75
St Teresa's School, Dorking	56	1071.7	86
Ibstock Place School, London	55	1071.7	96
Queen Margaret's School, York	61	1071.5	88
Jamia Al-Hudaa Residential, Nottingham	18	1070.5	45
Taunton School, Taunton	95	1070.0	91
St James's School, Malvern	30	1069.9	73
Manor House School, Leatherhead	35	1069.2	97
St Edmund's School, Canterbury	69	1069.0	92
Wychwood School, Oxford	22	1066.5	74
Kassim Darwish Gmr for Boys, Manchester	21	1066.5	95
St Mary's School Ascot, Ascot	54	1066.3	100
Royal School Hampstead, London	16	1065.8	93
Stonar School, Melksham	42	1065.3	89
Tonbridge School, Tonbridge	142	1065.0	93
Bowbrook House School, Pershore	14	1063.9	67
Red House School, Stockton-on-Tees	44	1063.7	93
St Paul's School, London	160	1063.1	99
Leicester Islamic Academy, Leicester	46	1063.1	93
Abu Bakr Girls' School, Walsall	32	1063.0	44
St Antony's Leweston School, Sherborne	40	1062.9	87
Al-Mahad-Al-Islam School, Sheffield	13	1062.9	23
Cranleigh School, Cranleigh	124	1062.6	98
Denstone College, Uttoxeter	75	1062.3	92
Manchester Islamic High School for Girls	51	1062.3	88
Guru Nanak Sikh VA School, Hayes	61	1062.2	95
Bedford School, Bedford	133	1061.9	92
Queenswood School, Hatfield	65	1061.8	89
The Towers Convent School, Steyning	33	1061.3	100
Tormead School, Guildford	85	1061.2	99
St Mary's School, Shaftesbury	54	1061.1	91
Rye St Antony School, Oxford	41	1061.1	98
St Nicholas' School, Fleet	29	1061.1	100
Dean Close School, Cheltenham	89	1061.0	84
Culcheth Hall School, Altrincham	19	1060.8	95
Eastbourne College, Eastbourne	111	1060.4	94
Wimbledon High School, London	89	1060.2	100
Babington House School, Chislehurst	17	1060.2	93
Pipers Corner School, High Wycombe	62	1059.9	97
Wycombe Abbey School, High Wycombe	86	1059.5	100

Source: Ray, A. (2006).

As shown above, successful efforts can be made to reduce the exclusive focus on raw test results. In addition, statistical issues can be emphasised in the discussion of the publication and graphical presentation of value-added results. Any discussion of confidence intervals around schools’ value-added scores should include discussion of the implications for the formation of league tables by members of the media. It was thought that the publication of confidence intervals would reduce the extent to which league tables could be misinterpreted. However, it has been found that once a point estimate is produced or schools’ value-added information is presented, then there is always the possibility/tendency that league tables would be created. In publishing value-added information, greater emphasis can be placed on particular aspects. For example, the development of contextualised value-added modelling in England has been reflected in the media coverage. The BBC UK website enables users to look up the latest school league tables for

English schools. Box 2.1 below shows the presentation of results for a particular secondary school in London. As can be seen, great emphasis is placed on the 2007 contextualised value-added scores for the school and this is the first school performance measure highlighted on the website. An extensive description of how the contextualised value-added scores should be interpreted is also available. This description includes the following:

Box 2.1. Description of Contextualised Value-Added in English media

The results incorporate a complex Key Stage 2 to Key Stage 4 contextual value-added (CVA) score designed to show the progress children have made.

This is done by comparing their achievements with those of other pupils nationally who had the same or similar prior attainment in their test results at age 10 or 11 in 2002.

CVA includes nine factors known to affect pupils' attainment but outside a school's control:

Gender
Special Educational Needs
Ethnicity
Eligible for Free School Meals
First Language
Mobility
Age
In Care
IDACI (a postcode-based deprivation measure)

What CVA does is predict what a given child's attainment should be based on the actual attainment of other children with similar prior attainment and similar backgrounds.

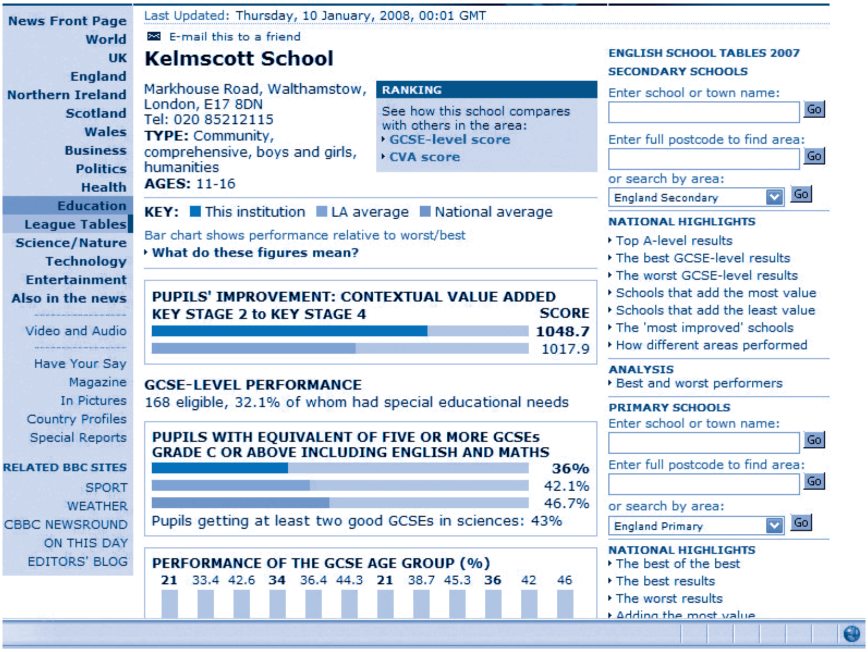
The idea is that how they actually performed - better or worse than the others - is down to the school's influence.

The pupils' individual scores are averaged to give a score for the school as a whole, to which another calculation is applied, finally producing a number based around 1000.

Source: BBC News (2007)

This information was originally provided by the UK Ministry that emphasised the importance of contextualised value-added scores in measuring school performance and highlighted the dangers of relying simply on raw test scores. The Ministry explained that the introduction of contextualised value-added scores would introduce greater equity and fairness in the publication of school performance results. This has had benefits for the school illustrated in Figure 2.7 below, that had an above-average contextualised value-added score. This is particularly important: this school did not rank as highly in the local authority in other measures such as students' performance in their Graduation Certificate of Secondary Education indicating that this school served a growing proportion of students from lower socio-economic backgrounds lowering the overall predicted results of the school. The focus upon contextualised value-added scores presented a more favourable picture of this school than would have been the case had the focus been on just raw test scores or, in this case, students' Graduation Certificate of Secondary Education grades.

Figure 2.7. An English School's value-added results available on the BBC website, 2008



News Front Page
World
UK
England
Northern Ireland
Scotland
Wales
Business
Politics
Health
Education
League Tables
Science/Nature
Technology
Entertainment
Also in the news

Video and Audio

Have Your Say
Magazine
In Pictures
Country Profiles
Special Reports

RELATED BBC SITES
SPORT
WEATHER
CBBC NEWSROUND
ON THIS DAY
EDITORS' BLOG

Last Updated: Thursday, 10 January, 2008, 00:01 GMT

E-mail this to a friend

Kelmscott School

Markhouse Road, Walthamstow,
London, E17 8DN
Tel: 020 85212115
TYPE: Community,
comprehensive, boys and girls,
humanities
AGES: 11-16

RANKING
See how this school compares with others in the area:
- GCSE-level score
- CVA score

KEY: ■ This institution ■ LA average ■ National average

Bar chart shows performance relative to worst/best
- **What do these figures mean?**

PUPILS' IMPROVEMENT: CONTEXTUAL VALUE ADDED
KEY STAGE 2 to KEY STAGE 4 **SCORE**
1048.7
1017.9

GCSE-LEVEL PERFORMANCE
168 eligible, 32.1% of whom had special educational needs

PUPILS WITH EQUIVALENT OF FIVE OR MORE GCSEs GRADE C OR ABOVE INCLUDING ENGLISH AND MATHS
36%
42.1%
46.7%
Pupils getting at least two good GCSEs in sciences: 43%

PERFORMANCE OF THE GCSE AGE GROUP (%)
21 33.4 42.6 **34** 36.4 44.3 **21** 38.7 45.3 **36** 42 46

ENGLISH SCHOOL TABLES 2007
SECONDARY SCHOOLS
Enter school or town name: Go
Enter full postcode to find area: Go
or search by area:
England Secondary Go

NATIONAL HIGHLIGHTS
- Top A-level results
- The best GCSE-level results
- The worst GCSE-level results
- Schools that add the most value
- Schools that add the least value
- The 'most improved' schools
- How different areas performed

ANALYSIS
- Best and worst performers

PRIMARY SCHOOLS
Enter school or town name: Go
Enter full postcode to find area: Go
or search by area:
England Primary Go

NATIONAL HIGHLIGHTS
- The best of the best
- The best results
- The worst results
- Adding the most value

Source: BBC News (2008).

The extent of media attention and focus on school rankings in the UK does not exist in all countries that make school performance information available. For some countries, the publication of results is a common occurrence and does not raise the interest of media to a large extent. In contrast, in the Flemish Community of Belgium the publication of results is not common. There has been considerable media interest in school results and it was the media that took the initiative to publish parts of inspectorate reports, which were accessible upon request. Partly as a reaction to this publication and media attention, a new initiative was taken in 2007 to publish the school inspectorate reports on a website (http://www.ond.vlaanderen.be/doorlichtingsverslagen/). It is difficult to determine why the publication of school results in one country does not attract the media attention that exists in other countries. One could assume that institutional and cultural factors are important, as is the history of the use of output-based performance measures. The objectives of the system and the methods with which it is introduced could also be an important factor. Value-added results that are used in a school accountability system with potentially large repercussions for school principals, teachers and families might cause a stronger reaction than a system focused upon internal school improvement. This highlights the benefit of clearly communicating how value-added school results will be used and how they are constructed. These issues are further discussed in Part III of this report.

Chapter Three

Applications of Value-Added Models for Internal School Improvement

In developing a system of value-added modelling, the objective should be to have a positive impact at the school-level in order to increase the performance of schools and the education system as a whole. The impact at the school-level will differ depending upon the intended application of the value-added information and the framework in which the value-added modelling has been developed. This chapter builds on the discussion of the presentation and interpretation of value-added models in Chapter Two. The focus here is on illustrating how value-added information can be analysed within schools or at the region- or system-level for school improvement purposes. As with much of this report, a recurring theme is the development of data-based decision-making within schools that act as learning organisations that drive system-wide improvements. These issues are first discussed before examples of the application of value-added models are provided. This discussion focuses on examples drawn from England and Tennessee in the USA that were both considered to be excellent examples of how systems of value-added modelling could be used to foster school improvements.

Schools as learning organisations

Information from value-added modelling can be utilised for a variety of school improvement purposes but only if it is utilised by actors that can influence processes and/or outcomes. In an education system, the most important actors are teachers and school principals. It is imperative therefore to ensure that they have the ability to effectively interpret and act upon value-added information. Given that the school is the unit of action, the focus of accountability and development measures are, initially at least, at the level of the school. Intuitively, school-level initiatives are likely to provide greater benefits to those schools that are best able to utilise the information to develop and implement accountability and development measures (Caldwell and Spinks, 1998). Given that the teacher in the classroom, rather than the school as an organisational unit, has the greatest effect upon student learning, it is essential that the effects of accountability and development measures are able to devolve to teachers and their actions in classrooms. This necessitates that the information is effectively conveyed to teachers and school principals and that this information continues to flow throughout schools for continual school improvement (Senge, 2000). For this to occur, it should be recognised that schools are complex organisational systems that can utilise information for school improvement. There are complexities in obtaining, disseminating and utilising information and several barriers exist that can impede the efficient use of information for school development (O'Day, 2002). If value-added information is used in a system with strong accountability measures then there exists an increased likelihood that information flows can become distorted. The presence of strong sanctions that can be placed upon schools and teachers can create the incentive to distort information as a form of self-protection from poor results (Lazear, 2000). This behaviour can then extend to distort the intervention stemming from that process. For example, placing a school on a restrictive or punitive probation that requires further information about school processes and student performance can be hampered by the distortion and restriction of information by teachers and school principals. This can severely restrict a programme of school development and hamper system-level learning.

Given these potential problems it is important to note that value-added models overcome many of the distortions associated with other school performance measures. Performance indicators that do not accurately measure student progress often suffer from undesirable consequences such as schools selecting only high-performing students to continue to later years and forcing less able students to leave the school (Meyer, 1997). This selectivity occurs because by using these measures, school performance is directly correlated to the innate abilities of students so it is of high importance to the schools which students take the test. However, with value-added modelling the focus on student progress removes many of these incentives. School performance is judged on accurate measures of progress in student performance so the incentive to only retain higher performing students is negated (Wilson, 2004).

The dissemination of information from value-added modelling should be developed in such a manner so as to take into account the complexity of information and the structure of the flow of such information within schools. In this sense, it is important to view schools as organisational units that operate within larger systems that provide resources to and set constraints upon schools. O'Day (2002: p. 294) contends that 'accountability systems will foster improvement to the extent that they generate and focus attention on information relevant to teaching and learning, motivate individuals and schools to use that information and expend efforts to improve practice, build the knowledge base necessary for interpreting and applying the new information to improve practice and allocate resources for all of the above'. The types of schools, as with other organisations, that would be best equipped to transmit school-based accountability and development measures to individual classroom teachers are those that have higher levels of peer collaboration and trust and, thus, more effective information flows. This type of school culture would be more likely to be found in those schools that already have an emphasis upon a collective responsibility for student learning that is commensurate with such collaboration and trust. Greater benefits from accountability and development measures will therefore accrue in schools with these positive organisational features. Unfortunately, poor performing schools are often those with poor levels of peer collaboration and trust and a weakened sense of a collective responsibility for student learning. A danger therefore exists that the objectives of the use of value-added modelling are less likely to be achieved in the schools that are often most in need of targeted and effective school improvement initiatives. Many schools operate with a large degree of autonomy provided to individual teachers concerning their teaching practices (OECD, 2004). This degree of autonomy can operate in organisational contexts of high peer collaboration but can also act as barriers to the flow of information and increase the complexity of implementing change within the school environment. This might help explain why some schools respond well to interventions based on outcomes of value-added models while other schools can receive poor outcomes for sustained periods, despite receiving interventions that have benefited other schools (O'Day, 2002). Overcoming such negative organisational barriers is essential to effectively disseminating and interpreting value-added models and then designing and implementing pertinent school improvement initiatives based on that information.

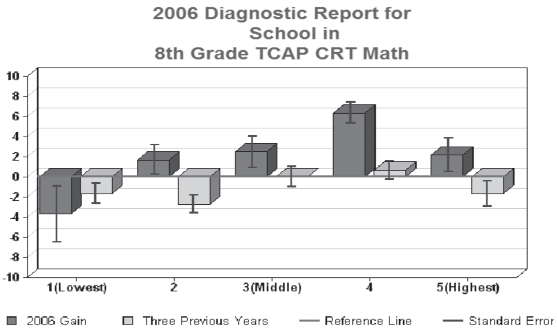
Efforts to improve the organisational aspects of schools have been an increasing focus of a number of education systems in OECD member countries (OECD, 2005, 2008). These have included efforts to foster the development of effective peer collaboration and to increase the focus of school-wide development. These efforts could facilitate the effective use of information derived from value-added modelling in addition to the fundamental benefits of improvements in peer collaboration and trust and in creating a sense of a collective responsibility of student learning. In addition to specific training to interpret value-added information, programs might be developed to facilitate peer collaboration and enhance organisational policies that facilitate effective communication between teachers, school principals and staff. These would need to recognise the complexity of both the value-added models themselves, of interpreting the results of the information obtained, and then how it can be applied in the organisational context of schools to achieve development and accountability objectives.

In some education systems, schools are placed on probation or have greater collaboration with school inspectorates or other outside agencies as a result of a low school performance measure (OECD, 2007a). These systems can be viewed as more interventionist in their efforts to improve the outcomes of schools than, for example, a system that focuses more on administrative accountability. These interventions could benefit from an increased emphasis upon organisational factors that should not only benefit school improvement efforts but also facilitate the dissemination and use of information garnered from value-added modelling. Interventions that are able to garner information from schools and investigate the causes of high or low performance might have a greater impact on school performance and, in the longer term, on the performance of the system.

Analysis of schools' value-added profiles

The analysis of value-added information for school improvement purposes will benefit from analysis of student-level data and disaggregation by student characteristics. This enables individual schools to construct or analyse their 'value-added profile'. For example, suppose all the 8th grade students in a local area or administrative unit are categorised into quintiles based on their prior performance records. A value-added model can be fitted just to the data associated with the students in a particular quintile. Applying this analysis to each quintile yields a five-component value-added profile for each school. As an example, a single school profile of this type is presented in Figure 3.1. It is an example of the system of value-added modelling that has been utilised in Tennessee in the USA that is the focus of the next section of this chapter.

Figure 3.1. Example of TVAAS school value-added profile Math



			Prior-Achievement Subgroups				
			1 (Lowest)	2	3 (Middle)	4	5 (Highest)
Math	Reference Line		0.0	0.0	0.0	0.0	0.0
	2006	Gain	-3.7	1.7	2.5	6.4	2.1
		Std Err	2.8	1.4	1.5	1.1	1.7
		Nr of Students	31	39	47	66	34
	Three Previous Years	Gain	**-1.7**	**-2.7**	**-0.0**	**0.7**	**-1.7**
		Std Err	1.0	0.9	1.0	0.9	1.3
		Nr of Students	111	110	110	115	112

Source: Reel, M. (2006).

We observe first that, based on prior achievement, this school's students are relatively more proficient than the district's student population. Moreover, for each quintile except the lowest, the school's estimated value-added is positive and exceeds the estimate based on the three previous years. For the principal of the school this is a welcome profile, although the weak result in the lowest quintile is worrisome and calls for deeper investigation. After talking with the teachers and looking through the school records, the principal learns that this group comprises students with disabilities who have formal Individualised Education Plans (IEPs) and students with a record of low achievement in mathematics but who are not clearly disabled. The disappointing (relative) gains that appear in the modelling are localised in the latter subgroup. The principal also learns that these students have been placed, so that most of the teachers have not been confronted with the challenge of teaching to a wide range of student abilities. With this information in hand, the principal can meet with the mathematics curriculum supervisor and the responsible teachers to discuss possible strategies for improving the rate of progress for low-achieving students.

Measures that are taken to improve a specific aspect of school performance can have positive flow-on effects to other areas of schools that have unrealised efficiency gains (Mante and O'Brien, 2002). For example, suppose the analysis of value-added information shows that a great portion of the poor-performance of a school derives from difficulties in teaching students with a home language other than the language of instruction. Further analysis of individual students' value-added scores shows that these students are performing, in general, well-below the level of the majority of students in different schools in the same year levels. After discussion with the school inspector or relevant administrator, value-added information is obtained from other schools to help identify schools where successful practices for students with a different home language are in use. Learning networks might then be organised with teachers and school principals from relevant schools that would allow school staff to learn from the experiences of others and share the best practices in this and other areas. Teachers and school principals might also appreciate knowing that their school is not the only one with difficulties as it reduces the sense of failing which measures of raw scores might instil and reinforces that constant learning is required and possible both within schools and throughout the school system. The education network would benefit from the use of value-added data to highlight differences in rates of progression between groups of students both within and between schools. At the system-level, an analysis of the school's results alongside the results of other schools might show a pattern that is particular to ethnic groups and that the performance of these groups has been a sustained problem. It could then be decided that greater resources need to be devoted to the education of these students and directed to the schools that require extra training and resources in teaching these students. In fact, differences in the performance of different ethnic groups from the general population are evident in a number of countries. For this reason, a number of countries include 'country-of-origin' variables in their modelling as opposed to a simple 'immigrant' variable or one identifying whether students' home language is the same as the language of instruction.

Application of value-added models to assist in school improvement initiatives

In the USA, the history of the use of value-added modelling has differed between states. It has been used for some time for both accountability and school improvement in the states of South Carolina and Florida. On the other hand, the state of Tennessee, the cities of Dallas, Texas, and Milwaukee, Wisconsin, as well as a number of school reform consortia, have made extensive use of value-added modelling for school improvement without a direct link to school accountability (Braun, 2006a). Because Tennessee has a well-established school development programme and a highly refined reporting system, the next section focuses on how Tennessee school districts employ value-added modelling in order to illustrate the potential of value-added analyses.[8]

Tennessee was the first state to formally adopt value-added analyses as part of a school development initiative. Intrigued by the work of William Sanders, who was then a professor at the University of Tennessee, the state passed legislation in 1993 requiring schools and districts to collect and transmit student data to Professor Sanders. This enabled Sanders to carry out the calculations entailed in his value-added model, termed the Tennessee Value-added Assessment System (TVAAS).[9] The legislation explicitly prohibited the use of value-added modelling results for school or teacher accountability. Rather, it was to be used exclusively for school development and, moreover, it was left to each district to decide whether to utilise the TVAAS. Input to the TVAAS is based on student performance on the Tennessee Comprehensive Assessment Program (TCAP), which consists of a battery of multiple-choice achievement tests. These tests, administered in the spring, provide both norm-referenced and criterion-referenced information. For each subject-grade combination, reports are generated at the district, school and individual student level.

From the outset, it was recognised that if TVAAS was to have the desired impact, educators throughout the state would have to go through an induction and training process. The statistical analyses are complex and are rightly viewed by many non-statisticians as a proverbial ‘black box’. Educators would first have to be convinced that the results yielded by the system were both relevant and fair. They would then have to be trained to properly interpret the results and, for this, specially designed reports were designed to facilitate

[8] It should be noted that Tennessee also uses value-added modelling to obtain estimates of teacher value-added, but an analysis of that application lies outside the scope of this report.

[9] That system is now referred to as Education Value Added Assessment System (EVAAS) and is operated by Professor Sanders and his colleagues under the auspices of a privately-held company. An abbreviated description of the EVAAS is presented in Braun (2006b).

the process and to encourage effective utilisation. Finally, there had to be support from the state's Department of Education so that school leaders would be assured both that this effort was more than a fad and that they would not be left to 'sink-or-swim' once the initial roll-out was completed.

The introduction of the TVAAS has gained much support among school leaders over its fifteen years of operation. Training for educators is ongoing since there is a continuous stream of new entrants into the state's education system. The displays and the accompanying text presented in this section are intended to give the reader a sample of the system. The introduction across the state comprised a three-phase process that involved thousands of curriculum supervisors, principals, regional directors, and state department staff. Phase I was informational, designed to provide a general introduction to the TVAAS and the structure of the reports generated by the system. Phase II was the initial implementation phase that included a review of the TVAAS, as well as a guided analysis and interpretation of local data. This phase also addressed strategies for informing parents and the wider community about the TVAAS. Phase III constituted the advanced implementation stage, wherein the TVAAS and other informational sources were integrated into a data-informed decision-making process that directly affected school personnel actions and resource allocations. The proximal goal was to facilitate the development of a culture of continuous school improvement that was based on a solid empirical foundation informed, in part, by the TVAAS results. Of course, the ultimate goal was to improve student achievement and for this a variety of indicators would be monitored and evaluated. To illustrate how the system operates, several extracts from the TVAAS report library are presented below, along with explanatory comments. The TVAAS operates a multi-subject, longitudinal, value-added model that accommodates data from four subjects (reading/language arts, mathematics, science, and social studies) from grades 3 through 8. The analyses are conducted for each school district and school reports provide results for the current year, the previous two years, as well as the three-year average.

Figure 3.2 contains part of a TVAAS report for Mathematics for a middle school containing grades 5-8. In the top panel, the estimated school effects are expressed in normal curve equivalent (NCE) units and are accompanied by estimates of their standard errors. For ease of viewing, each cell is colour-coded according to whether the estimated school effect is greater than the growth standard (zero) – blue (B), no more than one standard error below zero – light blue (LB) – or more than one standard error below zero – black (BL). In the last case, cells are labelled G* if the estimated school effect is more than two standard errors below zero. For example, in grade 6 (2006), the estimated school effect is 3.6 scale units with an estimated standard error of 0.8. Hence, it is coloured blue in the report. It is important to remember that that this estimate is an empirical Bayes estimate, so that the direct estimate of the school's mean gain has been 'shrunk' toward the district average, with the amount of shrinkage dependent on the relative precision of that estimate (see Chapter Six for further discussion of shrinkage in value-added estimates).

Figure 3.2. Example from 2006 TVAAS School report

2006 TVAAS School Report for
TCAP CRT Math

Estimated School Mean NCE Gain						
Grade:	**5**	**6**	**7**	**8**	**Mean NCE Gain over Grades Relative to**	
Growth Standard:	0.0	0.0	0.0	0.0		
State 3-Yr-Avg:	2.4	1.7	1.6	1.5	**Growth Standard**	**State**
2004 Mean NCE Gain:		3.2 B	-2.3 G*	-2.7 G*	**-0.6**	**-2.2**
Std Error:		0.8	0.8	0.8	0.5	0.5
2005 Mean NCE Gain:	1.0 B	6.2 B	-2.4 G*	2.1 B	**1.7**	**-0.1**
Std Error:	1.2	0.8	0.7	0.8	0.4	0.4
2006 Mean NCE Gain:	**-0.1** LB	3.6 B	1.0 B	2.2 B	**1.7**	**-0.1**
Std Error:	1.2	0.8	0.8	0.7	0.4	0.4
3-Yr-Avg NCE Gain:		**4.3 B**	**-1.2** G*	**0.5** B	**0.9**	**-0.6**
Std Error:		0.5	0.4	0.4	0.2	0.2
Estimated School Mean NCE Scores						
Grade:	**5**	**6**	**7**	**8**		
State Base Year (1998):	50	50	50	50		
State 3-Yr-Avg:	54.8	54.1	53.3	53.5		
2003 Mean:		48.4	52.8	51.4		
2004 Mean:		53.2	46.1	50.1		
2005 Mean:	49.3	49.3	50.6	48.2		
2006 Mean:	49.6	56.3	50.4	52.9		

🟦 B – Estimated mean NCE gain equal to or greater than growth standard.

🟦 LB – Estimated mean NCE gain below growth standard, but less than one standard error.

⬛ BL – Estimated mean NCE gain below growth standard by at least one, but less than two standard errors.

🟦 G* – Estimated mean NCE gain below growth standard by at least two standard errors.

Source: Reel, M. (2006).

The matrix structure facilitates comparisons across grades within years (horizontal) and within grades across years (vertical). For this school, there appears to be general improvement over time in each grade, with the strongest results in grade 6. At the far right of the panel, the school's results (averaged over grades) are compared both with those of the growth standard and the state.

The bottom panel of Figure 3.2 translates the school's results into mean NCE scores. This allows the viewer to consider the school's record from two different perspectives. In grade 6 (2006), the mean NCE is 56.3, corresponding to an average performance that is marginally greater than the state three-year average of 54.1. In the other grades for 2006, the school's mean NCE is marginally lower than the corresponding three-year averages for the state. (Note that comparisons to schools outside the district are always made in terms of levels of achievement, never in terms of value-added estimates.)

Application of value-added models for projections of performance

By combining observed student trajectories with a school's estimated value-added profile, it is possible to predict (project) a student's future performance. The value of such an exercise is that it enables schools and administrators to determine, given the expected growth rate of a particular group of students, what proportion of the students will meet a desired achievement standard in one or more years. This facilitates planning and resource allocation and highlights areas of low and high performing categories of students and schools. In addition, when a predicted result falls short, the school is given a clear indication of a target value-added it must aim for in order to achieve the desired level of success (Doran and Izumi, 2004; McCall, Kingsbury, and Olson, 2004; Hill et al., 2005; Wright, Sanders, and Rivers, 2006).

Target-setting in schools is an important part of the school improvement process in England. Targets are set in relation to test results rather than value-added (doing so, would mean that targets would not be straightforward as value-added scores are calculated relative to the national average and therefore it is statistically impossible for all schools to improve) but the value-added approach, taking into account student prior attainment, underpins the setting of the performance targets. Care is taken to encourage the setting of targets for students, schools and Local Authorities that are not simple extrapolations of previous performance. There are various ways to do this but the general approach has been to supply information as to the sort of results that would be expected in the future if a school, for example, improved its value-added to the level of similar schools (in terms of average prior attainment) that currently have higher value-added.

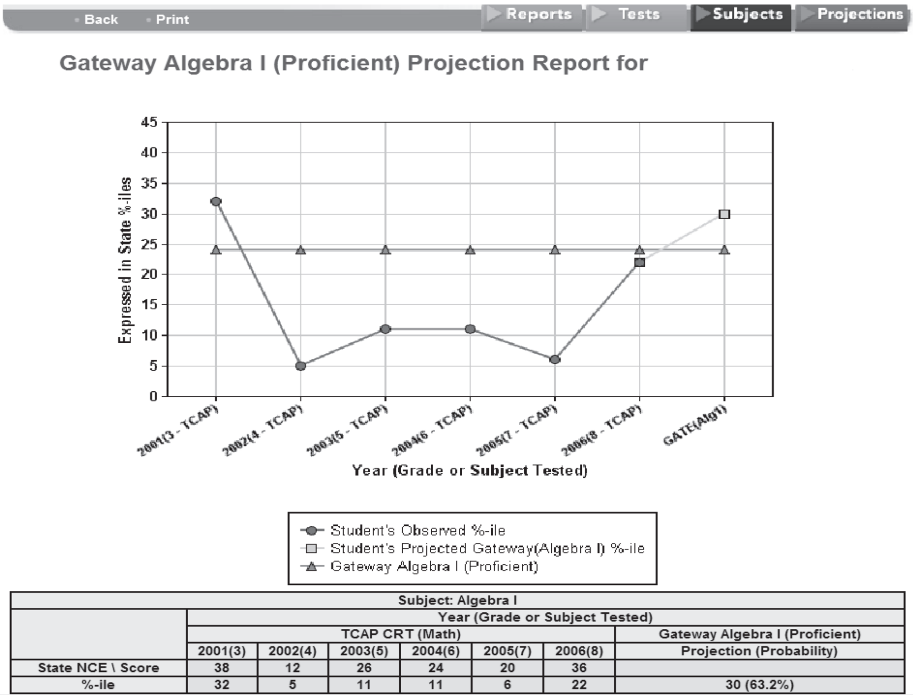
There are policy implications concerning the inclusion of more *contextual* variables in value-added modelling for setting targets. There is a risk that low expectations might be built in for students who currently make less progress on average (*e.g.* students from lower socio-economic backgrounds). On the other hand, schools with high prior attainment levels but few students from low socio-economic backgrounds could be set more stretching targets if this contextual data were included in the modelling. In Norway, contextual variables are not included in the published value-added models as it is considered it might misrepresent the intention of the programme and even act to further entrench existing inequalities.

A rather different approach involves the consideration of patterns of movement from one category to the next over the course of one or more academic years. Roughly speaking, interest centres on how successful a school has been in assisting students in moving from lower to higher categories. The relevant data is most conveniently displayed in the form of a matrix, with the rows representing the categories in the initial year and the columns representing the categories in the final year. The number of categories in the two years need not be the same. Different indices summarising the trajectories of a school's students can be proposed and the difference between the school's index value and the index value for the collection of schools is a measure of the school's value-added (for further details see Betebenner, 2007; Braun, Qu, and Trapani, 2008).

Figure 3.3 presents a Projection Report based on the TVAAS. It depicts the trajectory for a low-performing student through the 7th grade. Combining the student's record with the school's value-added for students in the lowest quintile yields projections for this student's performance in the 8th grade and in Algebra I the following year. Evidently, the student is expected to improve their relative ranking in the 8th grade to the 22nd percentile, though they will still fall below the threshold for proficiency (24th percentile). By the next year, however, they are expected to reach the 30th percentile, which would place them in the proficient category. The estimated probability that they will achieve proficiency by the 9th grade is 63.2%.

These projections, aggregated by student group, will play an important role in the reports that the state will submit to the federal government as part of the school accountability regime. However, they can also be very useful internally for school improvement. If a large number of students are projected to fall below the proficiency standard, the school has an early warning signal that it must aggressively address the factors, pedagogical or otherwise, that are retarding student progress. Even in the present instance, the projections are not guarantees. The school must work hard to sustain its positive value-added and monitor students' future performance to check that the projections are accurate, in a probabilistic sense. Clearly, such an enterprise would require training and support for the school leaders, as well as the infrastructure to support these analyses.

Figure 3.3. Example of a TVAAS Projection Report



Gateway Algebra I (Proficient) Projection Report for

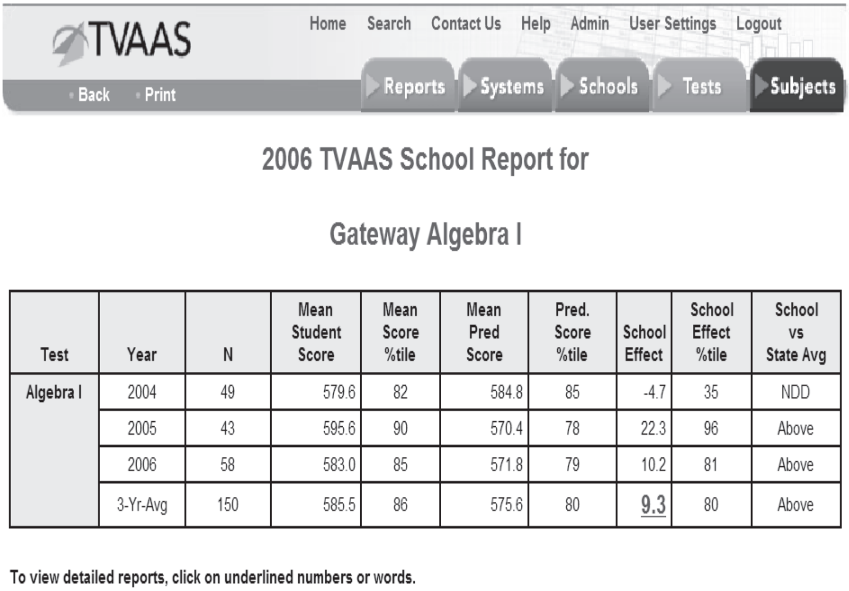
	Subject: Algebra I						
	Year (Grade or Subject Tested)						
	TCAP CRT (Math)						Gateway Algebra I (Proficient)
	2001(3)	2002(4)	2003(5)	2004(6)	2005(7)	2006(8)	Projection (Probability)
State NCE \ Score	38	12	26	24	20	36	
%-ile	32	5	11	11	6	22	30 (63.2%)

Source: Reel, M. (2006).

The ability to use value-added modelling as an early warning signal is important considering the use of alternate data. Student outcome data such as retention rates and the rate of progression to a higher level of education have inherent time-lags between falling performance and the identification of that problem through the data. Value-added modelling provides more time-responsive data as retention and progression rates are less sensitive to changes in school performance. Value-added data based upon student assessments in multiple years would provide more timely information that would allow for a more rapid identification of potential problems. It would thus facilitate actions to be put in place that address these problems.

An alternative presentation of a similar issue is provided in Figure 3.4, illustrating how performance projections can be presented and utilised by individual schools. The figure presents results for 8th grade students in a school who were taking their first course in algebra. Mean scores for the current year, the previous two years, as well as the three-year average are shown. The 'mean predicted score' (column 6) involves a calculation based on the students' score trajectories up through the seventh grade and their expected gain if they were enrolled in the typical school in the district. Thus, in 2006, the achieved mean score of 583.0 exceeded the expected mean score of 571.8 by 11.2 scale score points. The corresponding (empirical Bayes) estimate of the school's value-added is 10.2. This places the school in the 81st percentile among schools in the district with respect to value-added. It is important to note that the mean predicted score in 2006 is 13 points lower than in 2004, when the school's estimated value-added was -4.7, placing it in the 36th percentile among schools in the district for that year. This indicates that changes in the student composition of the school have lowered the predicted score and again highlights the importance of using value-added modelling as opposed to a focus on raw test scores.

Figure 3.4. Example of TVAAS School Report (Algebra)



2006 TVAAS School Report for

Gateway Algebra I

Test	Year	N	Mean Student Score	Mean Score %tile	Mean Pred Score	Pred. Score %tile	School Effect	School Effect %tile	School vs State Avg
Algebra I	2004	49	579.6	82	584.8	85	-4.7	35	NDD
	2005	43	595.6	90	570.4	78	22.3	96	Above
	2006	58	583.0	85	571.8	79	10.2	81	Above
	3-Yr-Avg	150	585.5	86	575.6	80	9.3	80	Above

To view detailed reports, click on underlined numbers or words.

Source: Reel, M. (2006).

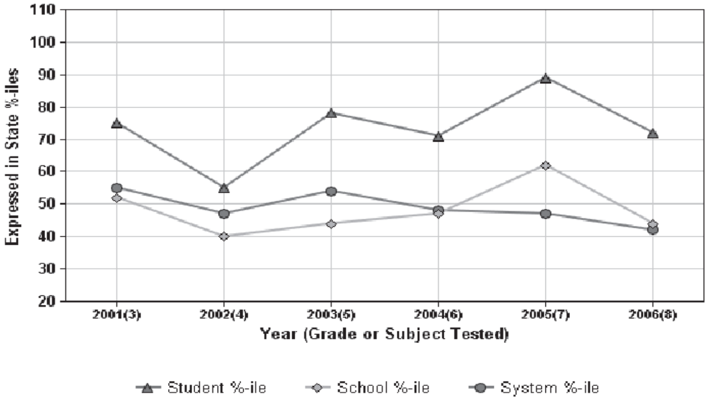
Targeted use of value-added models

Discussion of the use of value-added modelling has highlighted the advantages of focusing on particular groups of schools or students or even on particular policies and programmes. For policy makers in many OECD countries, the development of targeted policies aimed at particular groups of low- or high-performing schools and students is a priority (OECD, 2007c). Value-added scores can be used to identify specific schools in which to develop appropriate programmes, and to monitor the impact of such programmes. There are several advantages in using value-added measures rather than raw attainment scores. With raw attainment scores it is possible to identify low-performing students and the schools in which they are located. Yet this information cannot be used to analyse student progress. For example, were these students, who might be from low socio-economic backgrounds, always low-performing students? Which schools (and perhaps also programmes) have the highest value-added for these students and what can be learned from the successes? These are key equity question in education systems. They directly tackle the question of whether low-performing students are stuck at the bottom of the distribution or are able to reach high levels of proficiency.

Analysis of value-added data allows teachers, school principals and policy makers to drill down into the data for low-performing students to better understand their learning trajectories. For example, in England, policy makers analyse the data for students at specific levels of performance. Distinctions can be made between students that are actually improving over time, students that are stuck at low-performing levels, and students that are actually falling in their value-added performance measure (Ray, 2006). These are important distinctions as they not only provide considerable information about the learning and school education of these students but can also guide the development of the appropriate policy response and the programmes that might most benefit these students. After programmes have been enacted, further analysis of value-added results with subsequent assessment data facilitates the monitoring of the effectiveness of those programmes. Again, this is largely not possible with analysis of raw attainment data. The TVAAS data base that supports value-added analysis also makes possible tracking the performance of an individual student. Figure 3.5 presents a six-year trajectory for a particular student accompanied by the mean trajectories for the school and the district (system). According to the performance level indicators, this student has exceeded the threshold for "Advanced" standing since the 5th grade. At the same time, the substantial decline in relative ranking from the 7th to the 8th grade is cause for concern. Review of the trajectories of other, similarly placed students can reveal patterns across subjects and within the school that might reveal more systematic issues and provide potential answers to the problems faced by some students.

Figure 3.5. Example of TVAAS comparative performance trajectories

2006 TCAP CRT: Math Student Report for



Subject: Math						
	Year (Grade or Subject Tested)					
	TCAP CRT (Math)					
	2001(3)	2002(4)	2003(5)	2004(6)	2005(7)	2006(8)
State NCE \ Score	66	55	67	63	79	68
%-ile	75	55	78	71	89	72
Perf Level			AD	AD	AD	AD

Performance Levels
NP - Not Proficient
P - Proficient
AD - Advanced

Source: Reel, M. (2006). Note: Student's name has been withdrawn.

There is sometimes interest in the estimated school effects for a subset of the population of schools that contributed to the full value-added analysis. For example, suppose one wants to compare the apparent performance of two groups of schools, each employing a different educational program. The simplest strategy would be to extract the estimated school effects obtained from the full analysis. However, if the two groups constitute a relatively small fraction of the larger population of schools, then one might want to carry out a new value-added analysis for just the two groups of schools. The question is whether such an auxiliary analysis is necessary. Haegeland et al. (2005) carried out a study comparing the results of these two approaches on Norwegian data and reported that the differences in the comparisons were negligible. Although this is but one finding, it could be assumed to be generally the case. It should be borne in mind however, that the variances associated with estimated school effects can be quite heterogeneous and care is needed in constructing appropriate test statistics for the desired comparisons.

In England, various programmes have been targeted at groups of schools (*e.g.* the Specialist Schools programme, the Leadership Incentive Grant (Ray, 2006)). Value-added scores can be used as information to monitor policy initiatives of this kind. In addition to providing information on overall value-added, schools' value-added scores show how much between-school variation there is within the policy. Although value-added is not used directly in funding schools, it has been used as a way of selecting particular schools. For example, some schools have been designated as 'High Performing' and given additional responsibility for helping weaker local schools or engaging in other projects.[10] Schools thus identified are given additional funding to provide assistance to neighbouring schools and focus on various special activities such as vocational learning or students with special educational needs. The criteria that need to be met are based on value-added measures at different Key Stages, for the latest three years.

Targeted policies might require more complex value-added models to be developed with variables corresponding to the relevant group or sub-group of schools or students in question. In Chapter Two the discussion illustrated that simpler models are easier to present and communicate to stakeholders. A trade-off therefore exists between the desire to present a more easily communicable model and to develop a model that is more statistically robust but also more complex. As these analyses are often for an internal rather than external audience, the problem of communicating more complex models is eased somewhat. Analyses of particular programmes for policy makers and administrators do not have the same requirements for dissemination and presentation that value-added results have for school choice policies that assist parents and families in their decisions concerning school choice. The increased complexity of the statistical analyses can be more easily discussed in the form appropriate for this type of analysis (*e.g.* a report or brief) than the analyses presented in large tables presented for the general public.

More in-depth analysis can also be conducted to further analyse specific sub-groups of schools. This can be done to learn more about these schools and also to ensure that there are not misspecification problems with the simpler model. For example, if schools are classified as low-performing then the more complex model could be run to ensure the results for these schools are not a product of using the simpler model. This 'double-check' could also help in communicating the accuracy of the procedures undertaken to stakeholders. Analysis utilising more complex value-added models might be important for schools that exhibit greater instability in the results of the simpler value-added model. Again, this would have the advantage of both

[10] For more information see the section on High Performing Specialist Schools here: http://www.standards.dfes.gov.uk/specialistschools/

potentially learning more about these schools and checking to see if the lack of stability could be controlled with an alternative specification.

The analysis of more complex models that include more contextual variables can be beneficial to specific analyses of the contextual factors affecting progress in student performance. These might not be strictly value-added models but rather regression models that regress school and contextual variables on the first level of a value-added model. Clearly, there are a substantial number of opportunities to conduct more complex value-added and other multivariate estimations to analyse particular issues in the school education system.

Application of value-added models to improve the system of school evaluation

The success of initiatives to advance the policy objectives of school accountability, school choice, or school improvement rests on effective evaluations of school performance. A central message of this report is that value-added modelling provides more accurate measures of school performance than do measures based on raw scores. Yet, this report argues that such measures should complement existing methods of school evaluation. Combining value-added information with complementary information on school inputs and processes facilitates effective data-based decision-making throughout the school system. Such decision-making might then extend to discussions among teachers and principals about school-level issues such as school climate and school-level policies and programmes that might yield important information about whether particular actions are required to address the issues that might have led to a poor value-added score.

School improvement initiatives require an evaluation of the current/existing situation to identify areas in need of improvement and areas that provide examples of best practice. In a number of countries, a system of school evaluation is therefore not considered as merely another form of school accountability. Rather, it is seen as another mechanism to develop and advance school improvement initiatives. In Portugal, the use of performance measures is part of a broader system of school evaluation. The programme, Integrated Evaluation of Schools, intends to contribute to educational quality assurance, by identifying the strengths and weaknesses in the functioning of schools and in the school system in general. The main objectives of the Integrated Evaluation Programme are: to value both the learning and the quality of the school experience of students; to identify strengths and weaknesses in school performance; to induce self-evaluation processes in schools; to collect information and characterise the performance of the educational system; and, to regulate the functioning of the educational system (IGE, 2001). Efforts to reach these objectives would benefit greatly from the use of value-added modelling.

In England, value-added data for school-level Performance Tables include a limited range of statistics on schools: value-added data is presented alongside facts about overall attainment and school contextual information. For school improvement and inspection, a wider range of value-added measures, charts and other data are utilised through the use of the RAISEonline software package that was illustrated in Chapter Two. In the same year that school Performance Tables were introduced, school inspection was reformed with the creation of the National School Inspectorate. This body inspects all maintained schools and Local Authorities in England and its inspectors have access to school attainment data, in the form of the Performance And Assessment (PANDA) Reports.[11] The data in these reports has therefore played an important part in the system of school accountability as they form part of the evidence base used by inspectors to make judgements about school performance. The National School Inspectorate's inspection reports are published and schools are graded as Outstanding, Good, Satisfactory or Inadequate; schools in this last category might be put into 'special measures' or given a Notice to Improve.[12] Both school value-added scores and other types of value-added analysis have been used elsewhere: in publishing information for parents and schools; in selecting schools for particular purposes; and, as part of the approach to target-setting. RAISEonline provides a more extensive range of data than the Performance Tables, including value-added for a wider range of outcome measures and for subgroups of students within the school. The main objective of RAISEonline is to provide all schools with a free software product that allows them to analyse their own data and compare it with national patterns and the results and value-added achieved by high-performing schools. Schools use RAISEonline as part of the self-evaluation and target-setting process that they undertake with the help of School Improvement Partners. The data is also available to school inspectors for use in judging the extent to which the school is either improving or has the capacity to improve. The statistics are not made available to the public more generally.

The Dutch Inspectorate is undertaking a review of its operations to evaluate and increase school performance. While a comprehensive program of school evaluation has always been considered critical, it is considered there are benefits to focusing on specific areas to guide school evaluations and the allocation of resources to evaluate and lift school performance. This has led to a focus on school output indicators and also school organisation and process indicators. Five key output indicators have been identified:

[11] Formerly the Pre-inspection Context and School Indicator (PICSI) Report.

[12] Inspection reports can be seen at http://www.ofsted.gov.uk/reports/.

- over three years, the school's mean results at the end of the period are more than half a standard deviation below the level that should be expected of the school student population;
- more than 10 per cent of students considered to be under-performing in reading and arithmetic;
- more than 5 per cent of the students repeat a year in the school;
- more than 2 per cent of the students transfer to special elementary education or the designated expertise centres; and
- incidents of physical violence occur at least once a month in the school.

A recent study found that 24 per cent of elementary schools would have at least one of these output indicators and therefore warrant additional resources or inspections. At this stage, a lack of data prevents value-added analysis being undertaken across all school in the Netherlands but it is considered that it would greatly assist the Inspectorate in targeting schools given the greater accuracy of value-added measures and the inherent advantages in measuring if improvements are made for these low-performing students. Complementing these school output indicators is a focus on eight teaching-learning processes: the curriculum; teaching time; the nature of instruction; adaptation of teaching to accommodate differences between students; school climate; the attention given to the needs of lower-performing students; and, quality control mechanisms operating within the school. These issues are more fully detailed in Box 3.1.

By themselves, none of the issues or standards identified by the Dutch Inspectorate provides a singular indicator of school performance. Instead, as is the case in other countries, numerous indicators are combined to provide a school profile that can be used to evaluate schools and develop school improvement initiatives. In such a setting, value-added scores can serve a useful role as a 'quantitative anchor' for the development and analysis of the school profile. In this manner, the use of value-added modelling enables a more accurate evaluative framework to be constructed. Of course, the particular strategy adopted will depend on the purpose of the school evaluation, as well as the range and nature of the measures used to construct the school profile. Incorporating value-added measures into a broader school profile provides a more complete picture of school performance and, potentially, of the performance of different aspects within each school. This has flow-on effects for the quality of school improvement initiatives. Inspections can verify conclusions drawn from the analysis of value-added scores and increase the amount of information about the suitable intervention. This is particularly important given that the results of value-added models are indicators only and the information needs to be supplemented with more detailed information on school and teaching processes to determine the appropriate action or intervention.

Box 3.1. A focus on specific teaching and learning processes in school inspections in the Netherlands

In efforts to better focus the system of school evaluations that feed into school improvement initiatives, the Dutch Inspectorate has identified eight key school organisation and process standards that either measure or influence teaching and learning processes. The eight standards are:

	Standard	Indicator
1	The curriculum covers the attainment targets and is offered to all students in its entirety	The methods and materials used cover the attainment targets for the subjects of Dutch language and arithmetic/mathematics. The curriculum for the subject of Dutch language and arithmetic/mathematics is offered in its entirety to all students up to and including the level of year 8.
2	The teaching time is spent efficiently	Unnecessary loss of teaching time is prevented.
3	The teachers give clear explanations, organise the lesson efficiently and keep the students involved in their tasks	The teachers explain things clearly. The teachers organise the lessons efficiently. The teachers keep the students involved in their tasks.
4	The teachers adapt the curriculum, teaching time, instruction and time allowed for learning the subject matter to accommodate the differences between students	The teachers adapt the curriculum to accommodate differences between students. The teachers adapt the learning and teaching time to accommodate differences between students. The teachers adapt the instruction to accommodate the differences between students. The teachers adapt the time allowed for students to learn the subject matter to accommodate the differences between the students.
5	The school climate is characterised by safety and respect between people	The teachers ensure that students treat one another with respect. The school safeguards the social safety of students and staff.
6	The teachers systematically monitor the progress of their students	The school uses a cohesive system of instruments and procedures to monitor the educational performance and development of their students. The teachers systematically monitor the progress of the students.
7	The teachers provide sufficient care and assistance to students that are in danger of falling behind	For special needs students, the teachers systematically establish the pertinent issues. The school provides the care systematically. The school ascertains the effects of the care provided.

8	The school management monitors the quality of the education	Each year, the school systematically evaluates the quality of the results. Each year, the school systematically evaluates the quality of the organisation of the teaching-learning process. The school systematically works on improvement activities. The school safeguards the quality of the organisation of the teaching-learning process.

Evaluation of these standards would both complement value-added information to provide a more comprehensive school evaluation and enable analysis of the relationships between these standards and value-added scores both within and between schools. Such analysis would enable learning within schools of how these areas can be improved to lift student and school performance. It would also facilitate improvements within the Dutch Inspectorate as they could develop their performance assessments of these areas given the features more strongly associated with higher value-added scores.

Value-added information can also be used to increase the efficiency of the system of school evaluation and the institutions, such as school inspectorates, that are often at the centre of such systems. Efficiency gains can be made through both increased targeting of individual school inspections and through an improved allocation of resources that focuses upon schools in which evaluative instruments most need to be applied.

Analysis of value-added information can identify key areas upon which to focus a school evaluation to increase overall efficiency and to allow a more in-depth evaluation of key areas of school performance. Prior to inspecting a school, inspectors can have information on the value-added of the school across subject areas, year levels and for each student. Analysis allows those conducting the evaluation to focus on the key issues. An important element of the increased efficiency in England is the comprehensive nature of RAISEOnline. This interactive software enables schools and school inspectors to analyse the value-added information to, for example, identify the value-added scores of students in particular subjects and at specific year levels to gain a greater understanding of where the school is being successful and where there is a need for improvement.

Value-added modelling does not include financial input measures and cannot therefore provide a form of cost-benefit analysis. Analysis of the differential impact of various inputs into school education can therefore not be obtained through the use of value-added modelling. However, in providing a more accurate output measure, it is possible to undertake more extensive analysis of the impact of various resource allocations. In addition, information from year-level and subject-level evaluations could be particularly relevant if the value-added results focus on particular subjects or show that it is in particular subjects that student performance is low or high compared to other subjects. For example, if value-added results show that in the language of instruction and science students are performing at a higher level than in mathematics then this might indicate that more information at the subject-level is required. In some countries it is quite common to have subject-level evaluations instead of school-wide evaluations and these might be particularly useful in a situation such as the one described (OECD, 2007a).

Greater resources can be allocated to those schools or areas within schools that have poor value-added results. For a school inspectorate, a system of random inspections could be complemented with inspections determined by a school's value-added scores. The random component ensures that any school can still be subject to an evaluation at any time while the component determined by schools' value-added scores targets inspections at schools that are not progressing at the desired level. Efficiency gains could be increased if a particular value-added score such as one that would lead to a school being categorised as low-performing would automatically trigger a school inspection.

The evaluation of school processes is subjective by nature and complements value-added information. The quality of the subjective evaluations of school and teaching practices can be assessed and then improved with the use of value-added modelling. An accurate measure of school performance enables the further development of the subjective evaluations of 'what works' that are the basis of school evaluations. As illustrated above, numerous organisational and teaching practices are evaluated in school evaluations and by school inspectorates. These practices are often evaluated against what is considered as 'good practice'. One would assume that the definition of good practice evolves over time as understanding about effective teaching and schooling develops. It is important therefore to continually evaluate and develop what is actually considered to be good practice. It seems pertinent to incorporate some form of output measure into these decisions. As value-added estimates are more accurate measures of school performance, the results can feed into the organisational development of both school inspectorates and the conduct of school evaluations. What is currently considered as 'good practice' in schools can be analysed next to their value-added scores to assess the validity of such judgements.

Implementation of a System of Value-Added Modelling: Key Steps in the Implementation Phase

This report identifies a number of objectives for the development of a system of value-added modelling and illustrates the potential use in various applications and programmes. The following section highlights the main steps that need to be undertaken in the implementation phase. These issues are more fully discussed in Part III of this report but are presented here to highlight the importance of linking the objectives and use of value-added information with the need for a successful implementation. It also highlights the manner in which many of the more technical issues are addressed in the implementation phase. The steps discussed below are not meant to comprise an exhaustive list and the details of each activity are more fully discussed in the body of this report. It is provided here to assist policy makers and administrators to gain a quick and easy understanding of the process required in the implementation of a system of value-added modelling. It can also be used to assist in the development and review of the implementation of a system of value-added modelling.

Phase 1: Setting policy objectives and school performance measures

- Explicitly identify the policy objectives for the implementation of a system of value-added modelling. This includes a specification of the intended users of the value-added information and how schools' value-added scores can be interpreted to achieve policy objectives. This should encompass:
  - Whether schools' value-added scores will be classified into performance categories. If value-added scores will be used to classify schools as either high- or low-performing, it is necessary to determine how this classification will be determined, that is, how they relate to specific pre-determined statistical and/or valid conceptual criteria. It is necessary to identify the objectives of making these classifications including the actions that will be undertaken once a school is classified in a particular

  category. It is then necessary to identify how that classification will be communicated to the school and if it will be communicated to the public.

  - If value-added information is to be extensively used internally as a tool for developing school improvement initiatives, this will influence other decisions (such as data and model choice) and there are benefits to early planning, resourcing and designing the pilot programme to evaluate such objectives.

  - If value-added information will be published, the form in which it will be published can be further developed in the pilot stage but the parameters for publication should be established so they can be reviewed during the pilot process.

  - How value-added information will be used in the existing evaluative structures and mechanisms through which schools are already evaluated (*e.g.* school inspectorates or equivalent institutions).

- In determining the value-added measure upon which school performance is based, consideration should be given to the categorisation of the performance measure and whether a continuous, categorical, or dichotomous variable will be used in the value-added modelling. This should be linked to the actions stemming from schools' value-added scores and the incentives created within schools.

- A review of the existing structure of student assessments should be undertaken to determine whether further assessments need to be developed or the existing structure needs to be altered to fit the objectives of the value-added modelling.

- A framework should be established to clearly identify the particular student assessments upon which school performance is to be measured. The framework should enable:

  - The identification of appropriate student assessments for value-added modelling within the existing structure of student assessments.

  - The identification of the subjects and the Grade/Year levels in which the assessments should take place.

  - The identification of the focus of the student assessments (*e.g.* minimum literacy standards or continuous performance measure of all standards).

  - Consideration to be given to how such decisions about the choice of assessments could affect school performance and the incentives within schools. For example, is a focus upon numeracy too narrow to measure the performance of entire schools and would broader assessments more evenly distribute incentives to increase performance within school education?

  - Reviews and potentially further development of the assessment instruments to ensure that they can be used for value-added modelling. It is of particular importance that the scaling of the assessments allows for meaningful interpretation of performance and temporal shifts in performance measures with the longitudinal data.

Phase 2: Presentation and use of value-added information

- Given the policy objectives and the structure of student assessments supporting the system of value-added modelling, it is necessary to decide on the most appropriate method for presenting value-added information. This should be informed by a stakeholder engagement policy and through feedback from pilot schools on the most effective presentation and use of results.

- If schools' value-added results are to be published, it must be determined which particular value-added measure(s) will be used and how it will be presented (*e.g.* stand-alone or alongside other information).

- Guidelines for interpreting value-added information should be developed and these should include the categorisation of schools' scores with links established between such classifications and related policies and programmes. This might include, for example, identifying which school scores would be classified as low- or high-performing and the actions stemming from such classifications. If specific actions are to be enacted as a result of value-added results then such 'trigger points' should be identified and articulated to stakeholders.

- For specific school accountability and school choice purposes, the specific measure(s) to be used needs to be determined. For school accountability purposes, there are advantages to using a single performance measure and analysis should be conducted of the implications of these choices. For example, a measure that focuses solely on minimum literacy levels will focus the attention of schools, in both a positive and negative manner, on specific subjects and students of specific performance abilities. A focus upon specific subjects provides similar incentives. On the other hand, a value-added measure that averages value-added scores for all subjects can hide performance discrepancies across subjects.

- Given the benefits of using a three-year moving average of schools' value-added results, strategies need to be developed for the use of the interim data. This would focus on the actions stemming from the value-added results, how these actions are supported with interim data and how the interim results are published (if that is the intention). The use of interim data should ensure that issues of low school or student performance are addressed so as to lessen the impact of the delays inherent in using a three-year moving average.

Phase 3: Data quality

- There should be a review of data systems in schools and of the wider infrastructure for collecting and disseminating data to assess existing capabilities for the requirements of a fully implemented system of value-added modelling. Such a review could include an assessment of the capabilities for the use of value-added information at the school level and by other institutions (*e.g.* school inspectorates or equivalent institutions).

- Following a review of existing information systems and the structure of student assessments, a more comprehensive database might need to be established to meet the requirements for value-added modelling. The data requirements for a system of value-added modelling need to be determined and the commensurate data collection and information system designed (if necessary). This system can then be further assessed during the pilot programme.

- The sample of students to be included in the value-added modelling needs to be determined. This largely focuses on identifying the schools and students that need to be identified and, if necessary, excluded from the main sample. For example, in a number of systems, schools and students with special needs are excluded from the main sample (although much can still be learned by calculating their value-added). A further issue is ensuring that there is sufficient tracking of students to be able to identify mobility between schools between the prior and current assessment periods. In education systems with explicit tracking of students (*e.g.* between academic and vocational educational tracks) it also needs to be recorded if students move between these educational tracks as this might affect the calculation of value-added and is often related to problems of missing. These issues should be reviewed and further developed during the pilot stage of the implementation process. Such a review would include value-added analysis of the performance of specific sub-groups of the population in order to assess if they should be included with the main sample in estimating schools' value-added scores.

- Analysis should be conducted of the use of specific socio-economic contextual characteristics in the value-added modelling. This will depend upon the overall objectives of the system and the model employed, which will also be influenced by the number and frequency of student assessments and the overall performance distribution of schools.

- It should be established whether the data requirements and information system will support only the value-added modelling or also the institutions (mainly schools) that will use the information to enact the specified policies and programmes. A more comprehensive database and information system might be needed to support additional users and programme development.

Phase 4: Choosing an appropriate value-added model

- The pilot programme can be used to assess the validity of distinct value-added models. A number of value-added models will need to be estimated from the data obtained in the pilot phase (using data from pre-existing student assessments where possible). The pilot phase can then be used to assess the advantages and disadvantages of distinct value-added models and therefore inform the choice of the most appropriate model.

- In choosing a model, it is important to identify how the policy objectives and proposed use of schools' value-added scores will guide model choice. Certain policy objectives can benefit from specific modelling and these objectives need to be articulated before an analysis of different models is undertaken. This includes identifying the form of the dependent variable, how value-added information will be used, and whether school performance categories will be generated.

- It is necessary to identify the statistical and methodological criteria under which distinct value-added models will be analysed. The analysis undertaken with the pilot data during the implementation phase would concentrate on:

  - The variance in each value-added model. This should be analysed to evaluate the suitability of particular models. Specific models might be preferred if they can identify a greater number of schools statistically different from the average or some pre-determined criteria.
  - The use of socio-economic contextual data and the roles that different data components play in a value-added analysis. Analyses should be conducted to assess the impact of the inclusion and exclusion of specific characteristics upon schools' value-added scores and the value-added estimation.
  - The potential bias in the model that needs to be analysed (as well as the potential for how it can be reduced) during the pilot phase of implementation. The importance of missing data can be analysed and comparisons with existing data and analyses might prove beneficial.
  - Assumptions concerning missing data. These can be assessed against the results found in the pilot data collection. Procedures should then be developed to reduce the frequency of missing data.
  - Estimates of value-added for small schools. These can be tested and recommendations made for both the analysis and presentation of school results.
  - Stability of schools' value-added scores and how this is affected by the classification of school performance and the choice of the specific value-added model. In such analyses, it is important to consider not only the overall level of stability but changes in individual school scores. Analysis can then be conducted of the causes of such instability and to identify whether particular schools are more susceptible to instability in their school results.

- It is important to analyse the impact of different models under the prescribed policy objectives and intended use of the data. That is, it is important to analyse the impact of model choice upon different schools given the intended use of those scores. Such analysis should not just focus on the overall model (*e.g.* goodness of fit) but also the impact for individual schools. This would form the basis of the recommendation of a preferred value-added model in a pilot report.

Phase 5: Communication and stakeholder engagement strategies

- A stakeholder communication and engagement strategy should be developed that includes stakeholders in the development of the system of value-added modelling. A communication strategy can be developed that clearly articulates the objectives and rationale of the system, the value-added modelling being undertaken, and the use and interpretation of schools' value-added results.

- The focus of the communication strategy should be aligned with policy objectives. The measures on which school performance will be judged should be clearly described and the consequences for various levels of school performance articulated.

- For analysis at the school level, the appropriate infrastructure needs to be developed and guidelines and information packages should be developed for school principals and teachers concerning how to interpret value-added information and how it can be used for school improvement purposes. Similar information could be prepared for parents and the media.

Phase 6: Training

- Training programmes should be developed that target specific users. Training for school principals and teachers could focus on how value-added results are derived and how they can be used within schools for school improvement purposes. This might include training in statistical analysis and in using the required information system. Feedback from stakeholders during the pilot programme should facilitate further refinement of training programmes and highlight areas of importance to teachers and school principals.

- Training for parents and families should target the interpretation of the value-added scores presented to the general public to facilitate school choice. The publication of school results might induce a form of accountability from parents. Training and information packages can be made available to describe how the results are calculated and what they mean in terms of school performance and the education received by students. Such training can also be made available to the media and education experts.

Phase 7: Pilot programme

- The pilot programme should be structured so as to allow policy makers to assess and further develop all aspects of the system of value-added modelling and the commensurate policies and programmes surrounding the use of value-added information. This includes:

  - Operation and implementation issues ranging from the implementation of student assessments to the collection, analysis and dissemination of data and other value-added information. While a pilot programme is often conducted with a sample of schools, some education systems will have access to comprehensive student assessment data. If possible, it is beneficial to run the system on the comprehensive dataset to assess the required infrastructure, particularly if it is designed and built during the pilot phase. Estimating value-added on the comprehensive dataset would also facilitate analysis of model choice.

  - While not the focus of this report, the pilot programme should be used to further analyse the appropriateness of the student assessments used.

  - Estimations run on the pilot data that can provide the required analysis for choosing the most appropriate value-added model specification by assessing different models against pre-determined criteria.

  - If it is decided that value-added scores will be converted into specific performance categories, then the applicability of the classification scheme can be assessed. If specific categories are to be chosen (*e.g.* low-performing schools) based upon specific criteria then the number of schools falling under each performance category can be estimated with the value-added modelling under consideration.

  - The development of stakeholder communication and engagement strategies that can be informed through a review of existing strategies within schools. Input from school principals, teachers and other stakeholders should be included in such a review to elaborate on the effectiveness of various initiatives and to further develop communication and engagement strategies. Input from these groups would also assist policy makers in determining the key issues for stakeholders to be included in a system of quality control monitoring in the live implementation.

- In conducting the pilot programme, the decisions about the size and characteristics of the sample of schools should be aligned with the policy objectives of the implementation of the overall system of value-added modelling. This requires priority areas, such as schools in disadvantaged communities, to be identified and the appropriate sampling frame to be constructed.

- The pilot programme should be used to assess the actions linked to the results of the value-added modelling. Actions include the classification of schools into performance categories, the provision of rewards and sanctions, the development of specific initiatives and additional evaluations to be conducted. It should be identified how such actions will be implemented, with ‘trigger points’ (*i.e.* specific value-added scores) identified (if applicable) and commensurate actions outlined.

- The pilot programme should include a report or a series of recommendations based on the findings and experience of conducting the pilot programme. This would highlight the issues that need to be addressed prior to the live implementation. Such a report could also include the results of the analysis of the most appropriate value-added model and an assessment of the impact upon key stakeholders. It should also inform the key areas that should be the focus of a system of quality control monitoring utilised during the live implementation of the system of value-added modelling.

Phase 8: Ongoing development

- A properly resourced, quality control monitoring system needs to be established that focuses on the data collected, the capabilities of the information system utilised, the value-added modelling undertaken, the policies and programmes that it is supposed to foster, and the impact upon stakeholders.

- Such a quality control monitoring system would analyse not just the overall results of the value-added modelling but also the results of individual schools to ensure the model is still advancing the desired policy objectives. Such a system would highlight specific school’s scores (*e.g.* those with less stability across years) and analyse various sampling and data issues. It might also highlight assessment issues that need to be addressed.

- Analyses should be undertaken to continually develop the value-added model(s) that is being used. This would aim to improve the ‘fit’ of the specification and adjust to any changes in data or policy objectives. If changes are made to the underlying value-added model, then the impact upon schools should then be analysed.

Part II

The Design of Value-Added Models

Introduction

In this report, the term value-added modelling is used to denote a class of statistical models that estimate the relative contributions of schools to student progress with respect to stated or prescribed education objectives (*e.g.* cognitive achievement) measured at at least two points in time. To the extent that such progress is a desirable outcome of schooling, value-added modelling can therefore provide a valuable source of information. Indeed, as Part I makes clear, the output of value-added modelling might be used in many ways by both education authorities and school officials. There are many different value-added models in use today, each with its own advantages and disadvantages. Part II of this report identifies the key issues in the design of value-added models and then presents descriptions of some of the more common value-added models. Various statistical and methodological issues are then discussed to assist policy makers and administrators in the design of value-added modelling and in choosing the most appropriate model for school development and to monitor progress toward specified objectives in their education system.

As discussed earlier, this report maintains a distinction between value-added modelling and contextualised attainment models. The former always employ at least one measure of relevant prior academic achievement as a basis for taking account of differences in enrolled students among schools. On the other hand, contextualised attainment models do not incorporate prior achievement measures. Part II presents some empirical results concerning the advantages of incorporating prior test data into estimates of school effectiveness. Unfortunately, there is not yet universal agreement on the collection of statistical models that can appropriately be labelled "value-added". For example, suppose that there are two test scores available for each student (say scores on mathematics in successive grades). If the scores are expressed on a common scale, then one can calculate the difference (*i.e.* the individual gain score). The average gain score over enrolled students can be viewed as a measure of the school's value-added. Moreover, the difference in average gain scores between schools, or the difference between a school's average gain score and the mean over all schools of average gain scores, can be treated as a measure of the relative effectiveness of the school. Such models have problematic statistical properties because the adjustments made for the variation among schools in student intake is weak.

Accordingly, we do not consider them further. However, the reader should be aware that gain score models are discussed in the literature.

What are the basics of value-added analysis? To begin with, test score data from a large number of schools are assembled and organised according to the requirements of the model employed. At a minimum, the data base should contain for each student: the school attended; standardised test scores on at least two successive occasions; demographic and other background information.[13] Once the model is applied to the data the output is a set of numbers, one for each school. These numbers play a role that is similar to that of the residuals in an ordinary regression. That is, they represent that part of the school's outcome (*i.e.* the average student score) that cannot be accounted for by the various explanatory variables included in the model. Like residuals, these numbers average to zero. The number attached to a particular school is provisionally interpreted as a measure of the school's relative performance; that is, it is taken to be an estimate of the difference between the school's contribution to its students' learning and the average contribution to student learning of all the schools from which the data were obtained. Hence, these numbers are estimates of school value-added. Suppose, for example, that the analysis focuses on student performance for a particular examination. By construction, the residual or value-added estimate for the average school is zero. Consequently, a positive value-added estimate means that the corresponding school appears to have made a greater-than-average contribution, while a negative estimate means that the corresponding school appears to have made a smaller-than-average contribution. In the latter case, it is still possible, and even likely, that students in such a school have realised positive score gains during the period under study.

In the above example it is important to recognise that a school's value-added estimate depends on the schools that are included in the study as value-added estimates are relatively defined. That is, the model attempts to account for differences in outcomes across schools in terms of the differences in student characteristics among schools. The fitted model, and its success in explaining the variance in outcomes, will be determined by the school data that is employed. The use of another set of schools will lead to a different fitted model. The difference between a school's result and what would be predicted from the fitted model (*i.e.* the average outcome) is denoted the school value-added, since it is that part of the outcome that is not explained by the measured student characteristics. As indicated in the previous paragraph, value-added estimates defined in this way are simply residuals from a regression model and thus, are said to be relatively defined. The notion of a school performance indicator defined with respect to a

[13] Note that although most value-added models do employ non-test data, there are some that do not. The most prominent example being the EVAAS model.

particular collection of schools stands in contrast to indicators based on score gains, which are typically absolutely defined. This is not a disadvantage but must be kept in mind when interpreting value-added results. In many applications, interest focuses on those schools whose estimated contributions are substantially different from the average (*i.e.* strongly positive or strongly negative). To this end, most value-added models also generate an estimated standard error of the school's value-added estimate. The ratio of the value-added estimate to its standard error can be used to determine whether the school estimate is statistically significantly different from the average. Of course, for policy purposes statistical significance should be considered in conjunction with practical importance.

School value-added estimates can be calculated separately for each grade or year level and, if so, are especially useful for diagnostic purposes. For summary purposes, however, a composite school value-added indicator is calculated by averaging the value-added estimates for the different grades in the school. Although this is a convenient measure, it is recommended that schools with different grade spans are not compared with one another on the basis of such summary statistics as the statistical properties of the value-added estimates might vary from one grade to another. Although value-added estimates are usually called '(estimated) school effects', it must be borne in mind that even under the best of circumstances these estimated school effects can only approximate schools' 'true' contributions to student test score gains (this is discussed more fully below). The term 'effect' is taken from the statistical literature and generally does not imply a causal contribution. Equally important, statistical analysis alone cannot uncover the reasons for the (apparent) differences in school performance. Such explanations require site visits and the accumulation of much richer qualitative information on the teaching and learning activities in the school. Finally, schools have many other goals in addition to raising test scores. Accordingly, school evaluations should take into account a broad range of indicators that include, but are not limited to, test-based measures of value-added.

As indicated at the outset, value-added modelling is intended to estimate schools' contributions to student learning. The word 'contribution' denotes the part that schools play in bringing about the result of interest (*i.e.* the increase in test scores as a measure of student progress in learning), properly taking into account the roles of other factors related to this result. Thus, the intention is to endow the value-added model estimates with a causal interpretation. That is, the difference in the estimated contributions of two schools is usually interpreted as reflecting differences in their effectiveness in promoting student learning. It is understandable that policy makers would want to make such causal inferences on the basis of a statistical analysis. If one could truly isolate a school's contribution, then one would have a sound basis for actions of various kinds. Given the kind of data usually available and the realities of the constrained allocation of students across schools, however, causal inferences can be problematic. Ordinarily, causal inferences are made from large randomised experiments, such as those typically conducted in agriculture or medicine. In the simplest version, there are two groups: a control group and an experimental group. Individual units are randomly assigned to one of the two groups. Units in the first group receive a standard treatment (or a placebo), while units in the second group receive the focal treatment. The difference in the average outcomes for the two groups is a measure of the relative effectiveness of the focal treatment in comparison to the standard. The use of both randomisation and large samples reduces the likelihood that a substantial difference in outcomes is due to some combination of chance fluctuations and the action of unobserved factors.

Value-added models are an attempt to capture the virtues of a randomised experiment when one has not been conducted. In educational settings, students are rarely randomly assigned to schools, with geography and cost being the two biggest determinants. Thus, school data are considered to be the product of an observational study rather than of a statistical experiment. For that reason, simple comparisons of schools in terms of average scores or even average test score gains can be misleading. As will be seen below, most value-added modelling takes a more sophisticated approach by reporting score gains that have been adjusted for differences in a range of student characteristics. These adjustments are meant to take account of differences in the student populations across schools that might be related to those gains. The intent is to try to isolate the relative contribution of the school itself (its personnel, policies and resources) to student learning.

The proper use of value-added modelling rests on an understanding of the distinction between statistical description and causal inference (Rubin, Stuart, and Zanutto, 2004). Suppose, for example, the average gain of students over the course of a year in School Alpha is 8 points while the average gain of students in School Beta is 12 points. That is description. However, as a result of the application of a particular value-added model, we obtain estimated ‘school effects’, which we are invited to treat as indicators of relative school performance. For example, suppose the effect associated with School Alpha is 2 while the effect associated with School Beta is 5 (note that the estimated school effect will typically be different numerically from the simple average gain in the school). The desired interpretation of these effects is that if the students in School Alpha had been enrolled instead in School Beta, their average gain would have been 5 - 2 = 3 points greater. That is, the results of the value-added analysis are endowed with a causal interpretation.

However, the transition from description to statistical inference is fraught with difficulty because the students in School Alpha were not enrolled in School Beta. Moreover, the students enrolled in schools Alpha and Beta were not randomly allocated to these schools but, rather, were enrolled through a myriad of individual choices. Thus, the conditions of a randomised experiment are not fulfilled here. Interpreting differences in estimated school effects as differences in school effectiveness requires the assumption that application of the model has taken account of all relevant differences between the students in the two schools. Unfortunately, we can seldom observe or control for the factors that determine school choice. If there are unobserved factors that are determinants both of school choice and of achievement, then the straightforward causal interpretation can be problematic because the problem of the counterfactual condition has not been properly addressed. Indeed, it is the integral role of the counterfactual that distinguishes causal inference from simple description – and makes it so much more complex.

In fact, one can distinguish at least two types of causal inference in this setting (Raudenbush and Willms, 1995; Raudenbush, 2004). The first, the so-called Type A effect, is closely related to the one described above and is relevant to the situation in which parents are interested in choosing the school in which their child would do best. They can obtain a plausible answer by finding children in each school that are similar to their child and then determining which group performed better. The difference in performance would be the Type A effect in this setting. Although the observed superiority in performance might be due in part to unobserved differences between the two groups, there is no reason not to prefer the apparently more effective school. The Type A effect, however, is not a suitable instrument for evaluating school development or school accountability. The reason is that the average difference in performance between schools might be due to a combination of differences in the contexts in which the schools operate and differences in school practices. Raudenbush and Willms (1995) define 'school context' as those factors over which educators have little control, such as the demographic composition of the school and the community environment in which the school functions. They define 'school practice' as the aggregate of the instructional strategies, the organisational structures and leadership activities of the school, which, in principle, are under the control of the school staff. Although parents might be relatively indifferent to the relative contributions of the two components, Raudenbush and Willms (1995) argue that administrators and policy makers should be most interested in the contributions of school practice, as those are generally under the control of school staff. Thus, administrators and policy makers would like to disentangle the contributions of school context and school practice to the gains of the students and isolate the difference in performance due to differences in school practices. This would constitute the Type B effect.

Aside from some ambiguity with respect to what should be classified as school practice, Raudenbush and Willms (1995) find that unbiased estimates of Type B effects are essentially impossible to obtain from standard school system data. Even Type A effects are perfectly estimable only under ideal circumstances that are highly unlikely to hold in practice (for further discussion of the issues in obtaining unbiased estimates of school contributions to student learning, see McCaffrey et al. 2003; Braun, 2005a; van de Grift, 2007.). Although, these concerns might be discouraging, it should be noted that any empirically-based indicator of school performance is fallible, being subject to both variability and bias. In point of fact, value-added analysis has been more rigorously studied than other approaches such as inspection visits and the like. Consequently, when properly implemented and interpreted, a value-added analysis generates a school-level indicator that, in conjunction with other indicators, yields an informative portrait of school functioning. Indeed, because value-added estimates have a different empirical basis than most other indicators, they can be a particularly valuable addition to a school's performance review portfolio. The value-added analysis can serve as the first stage of a multi-stage process where, for example, the relationships between value-added estimates and various school characteristics are examined with the goal of identifying useful or surprising patterns. Importantly, the utility of value-added estimates is substantially greater than that of school performance measures based on the comparison of raw test scores used in some OECD member countries (OECD, 2007a), or even the results of contextualised attainment models emphasised in much decision-making concerning school performance. Our advocacy of the use of value-added measures in this report highlights the greater credibility of value-added estimates. Nonetheless, it is crucial to discuss the caveats and assumptions applicable in using value-added modelling to advance education policy objectives.

Chapter Four

Design Considerations

The design of an artefact, whether a statistical model or a house, is shaped by its intended use, the resources available and the relevant constraints. To this mix, must be added the experience of the designer with similar or related artefacts. In the context of value-added modelling, there are a number of key design factors including: data quality; data integrity and coverage; philosophy of statistical adjustment; technical complexity; transparency; and cost. Each is discussed below.

1. *Student assessment and test data quality.* Since value-added models operate on data generated by student assessments, primary consideration must be given to the nature and quality of that data. In particular, do the data adequately reflect what students know and can do with respect to the established curricular goals? That is the essence of test score validity and should be addressed in a number of ways. The four most relevant questions are: does the test provide evidence with respect to all (or, at least, all of the most important) curricular goals; do all students take the exam under comparable conditions; are the test scores sufficiently accurate to support the intended inferences; and are the test scores free of inappropriate influences and/or corruption? If the answers to these questions are all affirmative, then one can consider employing value-added modelling.

2. *Data integrity and coverage.* The procedures employed to transform the raw test data into usable data files, as well as the completeness of the data, should be carefully evaluated. Student records for two or more years are generally necessary for value-added modelling and it is not uncommon in longitudinal data files for some scores to be missing because of imperfect record matching, student absences, and in- or out-migration. Generally speaking, the greater the proportion of missing data, the weaker the credibility of the results. In addition, some value-added models employ test data from multiple subjects and/or auxiliary data derived from student characteristics (*e.g.* gender, race/ethnicity, socio-economic status). Again, the integrity and completeness of such data should be evaluated.

3. *Philosophy of adjustment.* Value-added models differ in the extent to which they incorporate adjustments for student characteristics. For some classes of models, such adjustments are the principal basis for treating the estimates as indicating the causal contributions of schools. When making adjustments, care must be exercised in the choice of characteristics, as the use of characteristics that are measured with error can also introduce bias. This might occur when adjusting for characteristics that might have been partly affected by school policies can introduce unwanted bias in the school performance estimates. Examples of such characteristics are student attitudes towards school or the average amount of weekly homework. In other classes of models, each student is employed as their own 'control' and, therefore, the models do not incorporate explicit adjustments. Instead, they either exploit the co-variation in test data gathered over multiple subjects and many years or incorporate a student 'fixed effect'. These variants will be further described below.

4. *Technical complexity.* Value-added models now range from rather simple regression models to extremely sophisticated models that require rich data bases and state-of-the-art computational procedures. In general, it could be argued that more complex models do a better job of yielding estimates of school performance that are free of the influence of confounding factors, although there is still some argument on this point. The disadvantage is that, typically, the greater the level of complexity, the greater are the staffing requirements and the longer is the time required to set up and validate the system. More complex models usually require more comprehensive data (years and subjects), so that data availability limits the complexity of the models that can be considered. In addition, the greater difficulties of communicating the workings and use of more complex models might reduce the transparency of the system and increase the problems of gaining the support of stakeholders.

5. *Transparency.* Although the notion of 'value-added' is intuitively attractive, its introduction in school settings can be controversial particularly if the motives for the introduction are viewed with suspicion among some stakeholders. If it is relatively easy to explain the workings of the model in non-technical language, many of those suspicions can be allayed. On the other hand, if the value-added model is presented as a 'black box' where inner workings are only accessible to an elite group of technocrats; obtaining general acceptance might be more difficult. Simpler models are ordinarily more transparent and consequently might be favoured for political reasons even if they are less desirable technically.

6. *Cost.* The greatest proportion of the cost is incurred in the collection of the data and the construction of a usable data base. The former is usually allocated to the instructional budget since the test scores are employed for academic purposes. Nonetheless, the construction and maintenance of an appropriate data base can be considerable, as is the cost of introducing a new system of school performance indicators, which might include outreach to (and training of) various stakeholders. The actual costs of running the model, carrying out the secondary analyses, and producing reports are relatively modest, especially after a year or two of operation. However, cost considerations and magnitudes will vary substantially across countries. The pertinent issues affecting costs and the implementation of systems that utilise value-added modelling are discussed in Part III of this report that focuses on implementation issues.

The first two issues are the essential building blocks for developing a system for value-added modelling. These are discussed below in the context of identifying key issues faced by administrators and policy makers in building an effective data base for value-added modelling. The third and fourth issues are then discussed, where statistical and methodological considerations are addressed. However, given the importance of these issues, they are also discussed in other areas of this report, particularly in Chapters Five and Six where various types of value-added models are introduced. The fifth and sixth issues listed above are treated in this report as presentation and implementation issues.

Student assessment data

This report does not dwell on the development of the assessment instruments that are used in value-added models. The focus of this report is on the development and use of value-added modelling. A large literature exists on educational assessment and the key decisions required in the development of assessment instruments. This literature describes the various methods by which general reasoning and subject-specific competencies can be assessed. This report does not evaluate this literature; however, the discussion below does address some of the decisions concerning the assessment framework that can influence the development of value-added modelling, as well as how the results are used by schools, administrators and policy makers. The student assessment frameworks in place in participating countries are also discussed in order to illustrate the various ways these issues are addressed. It is clear that most education systems have not developed a student assessment framework with the explicit objective of providing data for value-added modelling. Rather, value-added models have been developed to utilise the data generated by existing student assessments. Discussion of assessment framework design should inform policy makers and administrators in their efforts to develop assessments to enhance the utility of a system of value-added modelling.

In a number of countries, the development and implementation of a national curriculum was accompanied by the development of an assessment framework and a corresponding set of assessments. The results of these assessments could serve as the input to different types of value-added modelling. It is also possible to apply value-added modelling to the data obtained from standardised tests that are administered across multiple jurisdictions that implement different curricula. However, the development of these tests and the interpretation of the results of value-added modelling become more complex. In the design of the standardised test, there might be problems of bias when the assessment is more strongly aligned with one curriculum than another. There are also difficulties in estimating schools' contributions to student progress based on data from an assessment that is not strongly related to the curriculum that schools are either supposed to deliver or upon which they focus their resources. Interpreting the results of value-added modelling in this context can be problematic. In many countries with a federal system, the curriculum is devised at the sub-national level and therefore can differ quite substantially across regions. To avoid such difficulties, it might be prudent, therefore, for value-added modelling to be applied separately within each sub-national jurisdiction. There might also be political and institutional advantages to be gained in value-added modelling being used to monitor and inform system development at the same administrative level at which the main decision-making responsibilities reside. Naturally such considerations will vary across countries with respect to the nature of the national system, as well as the hierarchical structure of educational decision-making in those countries.

Construct validity

Test scores are the raw material of a value-added analysis and, clearly, the properties of those scores will be critical to the quality of the resulting estimated school effects. Many analyses rest on the assumption that the scores are 'good enough' – neither specifying what the term entails nor carrying out any empirical investigations into the way the scores are determined. Perhaps the assumption of adequacy is based on the fact that, in most cases, the test scores are used primarily to make decisions about students and only secondarily for school effectiveness studies. Nonetheless, it is certainly appropriate to review the desirable characteristics of test score data in the context of a value-added analysis. As the discussion presented at the start of this chapter indicated, the validity and reliability of the test for assessing academic achievement must be established. The two main threats to validity are deficiencies in construct representation and high levels of construct-irrelevant variance (Messick, 1989).

With respect to the first threat, the principal concern is with tests that are poorly designed or address only some of the learning goals or have an inappropriate topical emphasis. Typically, this occurs because of a lack of expertise among the developers of the tests and/or financial constraints that limit the types of items that can be included in the test. For example, many standardised tests comprise only multiple-choice items to minimise the cost of scoring. Consequently, some higher order learning goals might not be well tested in this format. A related concern is the degree to which the test sequence is sensitive to instruction. That is, if the tests are aligned with the changing curriculum, then there will likely be a "construct shift" as students advance to higher grades. This is perfectly appropriate for making inferences about student proficiency in each grade but can lead to bias in value-added estimates if the score scales for different years have been vertically linked. See Martineau (2006) for further discussion.

With respect to the second threat, the concern is with significant departures from standardised administration, poorly constructed or ambiguous items, and problems such as low reliability. For example, questions that require the student to provide written responses and that must be scored by human graders can contribute to unreliability because the scoring procedures are not well implemented or are poorly monitored. Fortunately, these sorts of technical problems can be resolved through training and practice. Effective implementation should assure school leaders that students' test performance is a reasonable measure of their academic standing. If not, then schools whose performance is apparently not up to standard can place the blame on the test and incorrect inferences can be drawn from the analyses, leading to sub-optimal decisions at a range of levels. Another potential difficulty is that the test results for some schools will be manipulated in an attempt to achieve a better school value-added score. This represents a particularly pernicious instance of construct-irrelevant variance. These issues can be alleviated somewhat through the structure of the framework of student assessments and their role in school accountability and school improvement programmes. The creation of incentives that might lead to such sub-optimal outcomes is discussed in Part I.

Another consideration in examining test quality is related to the question of whether and how the different assessment instruments administered in successive years are prepared. If the same (or substantially the same) form is employed each year, then its integrity is likely to be compromised over time and test performance will increase but not be accompanied by improved learning (Koretz, 2005). Such 'test score inflation' undermines the credibility of value-added analyses, particularly if its magnitude varies across schools. If different forms are created each year, then the new form must be equated with the previous form in order to maintain the comparability of the scale (Kolen and Brennan, 2004). Substantial equating error, incorporating both measurement variance and bias, also compromises value-added estimates. Finally, longitudinal value-added analyses typically employ test score scales that have been vertically linked across grades (Harris et al., 2004). Different strategies to carry out vertical linking yield score scales with different properties that, in turn, can have a substantial impact on value-added estimates (Patz, 2007).

More generally, test validity comprises both construct validity and consequential validity (Messick, 1989). The latter refers to the appropriateness of the inferences and actions taken on the basis of the scores. That the scores are of consequence is not at issue; rather, the point is whether their use can be justified given the context and the purpose. Thus, the test scores can be valid for one use but not for another. Validity is not an 'all or nothing' matter: it is a matter of degree. However, if there are serious concerns related to either the construct or consequential validity, then it might not advisable to proceed with a value-added analysis, at least until the concerns have been reasonably addressed.

Measurement error

Another characteristic of test scores is reliability, which is a measure of the replicability of the measurement process. Reliability is a dimensionless quantity (*i.e.* it is not expressed in units of measurement) that takes values between 0 and 1. High reliability (*i.e.* values close to 1) means that students would achieve very similar rankings were they to take another test that is parallel in structure and format to the test actually taken. On the other hand, if there is substantial 'noise' in the testing process, reliability is reduced. Some test features that determine reliability are aspects of the design (such as test length, item formats, etc.) and the quality of the scoring of student-produced responses. Low reliability is a threat to validity because it means that the results of the value-added analysis could have been materially different had the test administration been repeated.

Reliability is a summary indicator of one aspect of test quality. A closely related term is measurement error, which is expressed in scale score units and is employed to quantify the uncertainty associated with observed test scores. Roughly speaking, high reliability corresponds to low measurement error. There are advantages, however, to representing the replicability of test scores in terms of measurement error. For many tests it is possible to calculate the measurement error associated with each point on the reporting scale. Ordinarily, measurement error is smallest near the centre of the scale where, typically, most of the student scores are found, and it is greatest at the ends of the scale. This phenomenon is a direct result of the way the tests are designed and developed. Problems can be compounded with measuring progress in student performance over time as it might induce further measurement error in equating different student assessments (Doran and Jiang, 2006). The standard assumption in regression models is that each observed value of the criterion is drawn from a distribution with the same variance. Thus, the fact that measurement error is not uniform across the scale of measurement (termed heteroskedacticity) can be problematic when test scores are used as a criterion. Failure to account for heteroskedacticity can result in biased estimates. At this point, little is known about the relationship between the degree of departure from uniform measurement error and the resulting bias. For further discussion, see McCaffrey et al. (2003: 103).

Measurement error can also cause problems when test scores are used as control variables in a regression model. The usual assumption is that the control variables are error-free. It is well known that when test scores are used as control variables, measurement error causes a downward bias in the estimates of the corresponding regression coefficients. Relying on data from two states in the USA, Ladd and Walsh (2002) investigated the extent of this bias. The models were standard linear regression models that incorporated prior year test scores but no student characteristics. These models were employed by North Carolina and South Carolina for purposes of school evaluation. They found that the estimated effects for schools serving lower ability students (based on their prior year performance) were substantially lowered and that the estimated effects for schools serving higher ability students were substantially raised. That is, the results of the value-added analysis disadvantaged schools serving weaker students and advantaged schools serving stronger students. Further, they show how this bias could be substantially reduced if test scores from earlier years are available for use as instrumental variables. In their absence, other relevant student characteristics should be employed if they are available. This is further discussed in Chapter Six.

The distributional properties of test scores are also relevant to the implementation and interpretation of a value-added analysis. The standard assumption is that scores are distributed according to the Gaussian (normal) form, at least conditional on the other variables (student characteristics) in the model. Mild departures from this assumption are not cause for worry. However, substantial ‘floor’ or ‘ceiling’ effects could be problematic. For example, if the test in a particular grade is relatively easy for large numbers of students enrolled in a subset of schools, then the distribution of their gain scores will have a pronounced skew to the lower tail. The value-added estimates for those schools will be biased downward in comparison to what would be obtained were the test sufficiently challenging for those students.

Scaling of test scores

While the construction of student assessments and tests is not the focus of this report, the issue of scaling test scores has been considered too important not to mention. It is common for ‘raw’ test scores to be transformed to a different scale for reporting and for secondary analysis. Such transformations can make it appear as if the test scores are comparable from one year to the next. However, true comparability depends on careful implementation of the test specifications and, if necessary, score adjustment through a special process called (test) equating. Serious departures from year-to-year comparability might not be especially problematic for students if they are only being compared with others in the same cohort. However, it can be problematic for value-added analysis as it means that the distribution of gain scores varies across years (Harris et al., 2004). If school effects are obtained from the analysis of data from multiple cohorts, then this variation can introduce construct irrelevant variance.

In some settings, end-of-year tests are administered in each grade and the raw test scores from different grades are 'vertically linked' to yield a single cross-grade scale. There are a number of different procedures for carrying out the vertical linkage and each produces a cross-grade scale with different properties that can result in different estimated school effects (Patz, 2007). Although the construction of a cross-grade scale is not required for the application of many value-added models, vertically linked test scores are often used as the input file for a value-added analysis. In such situations, users should be mindful of the characteristics of the vertical scale and how it might affect the value-added model estimates. They should be wary of treating the scale as an interval scale (*i.e.* one for which score differences have the same meaning all along the scale). Though it is tempting to do so, it is rarely justified and a more conservative stance is recommended.

Assessment results reported on an ordinal scale

Heretofore, it has been assumed that test scores are reported on a scale with sufficiently many values that the scale can be treated as if it were effectively continuous. In some settings, however, final scores are reported on a coarse scale comprising as few as two ordered categories. For example, the authorities might establish two standards denoting 'competent' and 'advanced achievement'. Each standard is represented by a score, or cut-point, on the original reporting scale. Students are then classified into one of three categories ('below competent', 'competent' and 'advanced') depending on where their score falls. Although conventional value-added modelling should not be applied in such cases, it is possible, nonetheless, to carry out a value-added analysis. If there are only two categories, one could employ logistic regression or probit models in place of the usual normal-theory models. If there are more than two categories, then polytomous logistic regression models or ordered probit models can be used. See Fielding, Yang, and Goldstein (2003) for an illustration of this type of model.

Issues of validity and reliability are also relevant to ordinal scale data. If the categories are determined by a form of standard-setting procedure, then the validity of the procedure must be evaluated (Hambleton and Pitoniak, 2006). If the categories correspond to stages on a developmental scale, then the theoretical and empirical support for the scale should be evaluated. In both cases, reliability is related to the probability that a student is assigned to the appropriate category. Placement in the wrong category is a type of measurement error which can induce bias in estimation. The greater the measurement error (and the lower the reliability), the less credible are the estimates of school value-added.

In most participating countries the rationale for implementing a value-added system based on certain assessments is to focus the attention of school leaders, teachers and students on improving performance on those measures and student learning in the corresponding academic disciplines. Thus, the choice of subjects and grade levels, as well as the nature of the assessments must be made thoughtfully, as it is likely to affect the actions of all stakeholders. In particular, deficiencies in the assessments might lead to higher student scores that are not associated with desired improvements in student learning. This would be an instance of a lack of consequential validity. Decisions concerning how student performance is employed for school evaluations can alter the incentives and, therefore, the behaviour of school principals and teachers (Burgess et al., 2005). Typically, student scores are transformed or summarised into performance indicators that inform the decision-making process. A key distinction is between performance indicators that are discrete and those that are continuous. If a school is evaluated on the basis of a discrete indicator, then there is a natural incentive to focus resources on improving that indicator. For example, a value-added analysis that focuses on the proportion of children reaching or exceeding a particular reading level encourages schools to focus attention on those students who are below the literacy level but who are likely to reach that level when given adequate support. On the other hand, in this example there is little incentive for the school to improve the scores of students who are already above that level or to focus on those students who are well below the level. By contrast, a value-added analysis that focuses on a continuous indicator is more likely to encourage a more uniform allocation of resources, although it is possible that the students who appear to be best placed to make larger gains might receive greater attention. For example, it might be easier to improve the performance of high-achieving students than that of low-achieving students. Not only can this result in distortions within schools but also makes comparisons between schools more problematic. That is, schools with greater proportions of students from advantaged backgrounds (however measured) might receive higher value-added scores as their students might generally achieve greater gains. Were this the case and were teachers from schools with higher value-added scores accorded special benefits, then there would be a clear incentive for teachers to move to those schools with greater proportions of students from advantaged backgrounds.

It is possible, however, to introduce a countervailing force by employing differential weighting of score gains. For example, greater weight can be accorded to improvements at the low end of the scale in comparison to the high end. Since low-socio-economic status students are more likely to be found at the low end of the scale, such a weighting scheme can provide additional incentives for school leaders and teachers to focus on lifting the performance of these students and even to induce the most effective teachers to move to these schools. These issues are addressed in Part I, which illustrates such systems and the implications of various incentive structures.

The structure of student assessments in participating countries

A number of decisions concerning the design and use of value-added models depend on the nature of the assessment data that is available. The assessment data collected in each country is discussed below to illustrate the differences that exist across countries, as well as the strategies that can improve the data and thus enhance the policy utility of value-added analyses. In some countries, the choice of assessments that can be used for value-added analyses is essentially determined by the structure of the education system. For example, if the school system is organised into primary and secondary sectors with schools belonging to one or the other, then value-added analyses can normally only be based on assessments administered across a time-span commensurate with the time students would normally spend in either a primary or a secondary school. From the perspective of value-added analyses, it is problematic if one assessment takes place half-way through students' primary education and the second half-way through students' secondary education. Table 4.1 details the student assessments that could be used in value-added analyses in participating countries and illustrates the differences among countries in the subjects covered. It should be noted that in some countries the lack of comparability of assessments is a barrier to the implementation of value-added analyses.

Table 4.1. Student assessments in participating countries that potentially could be used for value-added modelling

Country	Year Level	Subjects
Belgium (Fl.)	Year 1-6	Mathematics, Language of instruction
	Year 1-6	Mathematics, Reading, Spelling
	Year 6 (final year of ISCED 1)	Mathematics, Reading, Nature (sub-domain of environmental studies), French, Society
	Year 8	Cross-curricular areas ('learning to learn', 'retrieval and processing of information'), Biology, French, Society
Czech Rep.*	13 (state Maturita)	Czech language, Foreign language and one of Mathematics, Social Science, Science or Technology
	Year 5,9	Czech language, Mathematics, Foreign Language, Learning skills
Denmark	Year 2, 4, 6, 7, 8,	Reading, Mathematics, English, Science
	Year 9 & 10	All compulsory subjects (assessed by teachers)
	Upper secondary	Reading, Mathematics, English, Science
England	Key stage 1: Year 2	Reading, Writing, Mathematics
	Key stage 2: Year 6	Reading, Writing, Mathematics, Science
	Key stage 3: Year 9	English, Mathematics, Science
	Key stage 4: Year 11	A wide range of subjects most of which are allowed to count towards a pupil's best 8 results
France	National exam (baccalaureate at end of upper-secondary)	Covers 15 subjects for each student
Norway	Year 5,8	National tests in Mathematics, Reading English (reading)
	Year 10	External exams (Mathematics, Norwegian or English.) All compulsory subjects (assessed by teachers)
	Year 11,12,13	Exams and teacher assessments in various subjects
Poland	Year 6 (primary school exit exam)	Cross-subject competency test
	Year 9 (lower secondary exit exam)	Humanities, Mathematics, Science
	Year 12 (Upper secondary exit exam)	Matura exam (Polish is compulsory and then assessments in a range of other subjects)
Portugal	Year 4, 9	Mathematics, Portuguese,
	Year 12	All subjects required for certification and tertiary entrance
Slovenia	Year 6	Mother tongue, Mathematics, First foreign language
	Year 9	Mother tongue, Mathematics, one mandatory school subject (decided by Ministry)
	Upper-secondary (Year 13)	Vocational: Mother tongue, either mathematics or first foreign language, two school and curriculum specific subjects
	Upper-secondary (Year 13)	General: Mother tongue, mathematics, first foreign language and two out of 30 optional subjects.

Country	Year Level	Subjects
Spain	4 (Primary), 8 (lower-secondary)	Mathematics, Language of instruction: social sciences with civic education, science, technologies of information and communication, other**
Sweden	Year 9, final grades	Assessment across 16 subjects
	Year 5, standardised test	English, Mathematics, Swedish
	Year 9, standardised test	English, Mathematics, Swedish
	Upper-secondary, final grades	Grade-point average, all subjects for each student (30-35 subjects)
	Upper-secondary standardised test	English, Mathematics, Swedish

* Data collection currently in pilot stage. The project collecting data at Year 13 will be transformed into State Maturita exam in 2010; Year 5 and 9 will not continue.

** Mathematics and language of instruction are assessed annually. Other subjects assessed on a less frequent basis.

There is considerable variation in the ages and grade/year levels at which student assessment data are collected. In considering the student assessment data that could be used for value-added analyses, the age at which students are assessed shapes the output measure through which it is possible to measure the effects of schools upon student progress. Assessments in some countries focus on primary education, while others focus upon lower and upper-secondary education. Countries such as Belgium (Flemish Community) and the Czech Republic concentrate their assessments in the earlier grades, which facilitates the use of value-added modelling in the development of the primary education sector. On the other hand, the structure of the student assessment frameworks in countries such as Norway, Poland, Portugal, Slovenia and Sweden facilitate, for the most part, the development of value-added modelling focused on the secondary education sector. In Denmark, there are assessments in both mathematics and reading in both primary and lower-secondary education and additional assessments in science and English in only lower-secondary education. The range of subjects included in the student assessment framework will reflect the priorities of the national system and will have an impact upon the use and interpretation of value-added models. If only mathematics is assessed in given years then only value-added in mathematics will be measured. If it is desired to create a more broad-based indicator of value-added, then clearly student assessments on a broader range of subjects are required. In general, students are assessed in a greater number of subjects in secondary education, particularly in upper-secondary education where the results of examinations in all subjects (*e.g.* national examinations) can be used for value-added modelling (depending upon the type of value-added model employed). At lower levels, assessments are concentrated in only a few areas. For most countries these are Mathematics, Sciences and either the national language or language of instruction (with a focus on reading and/or writing in that language).

The frequency of assessments varies considerably across countries. It should be noted that the system of assessments in some countries do not currently permit value-added analyses as defined in this report. Our definition emphasises that a prior assessment is required to measure value-added. Moreover, the assessments have to be comparable in a manner that supports the desired inferences concerning the relationship of different factors to student progress. Countries such as England and Denmark have developed student assessment frameworks that span the primary and secondary school education sectors. In England, key stages have been identified in the progression of students through their schooling, with assessments taking place in Years 2, 6, 9 and 11. The Flemish Community of Belgium is the only example among participating countries to have annual student assessment data, if only at the primary school level. Annual testing can somewhat circumvent some of the statistical and methodological problems with value-added modelling discussed later in this report and should enhance the utility of the results.

The frequency of the assessments has an impact upon the choice of the value-added model to be used, as well as whether or not to include student background characteristics. These decisions in turn affect the interpretation of the results of the model. Decisions concerning the frequency of assessments will depend upon the nature of the curriculum and the priorities with respect to monitoring student progress at various points in their school careers. For countries preparing to develop a framework of student assessments and to utilise value-added modelling, there can be advantages to tracking progress through more frequent student assessments.

As discussed in Chapter six, increasing the number of prior attainment measures can greatly enhance the accuracy and credibility of value-added analyses. It is tempting, therefore, to encourage more frequent student assessments. There is a concern, however, that additional assessments would place an undue burden upon schools and reduce the amount of effective teaching time. That is, not only do tests take time out of the school day, but also impose organisational requirements regarding pre- and post-assessment activities. Policy makers can weigh the benefits of increasing the assessment frequency against these burdens and the financial costs. Moreover, tests can place increased pressure on students that might also have negative consequences. This is reflected in Table 4.1 which shows that in most school education systems students are currently assessed in only a few year levels and in selected subjects or learning areas.

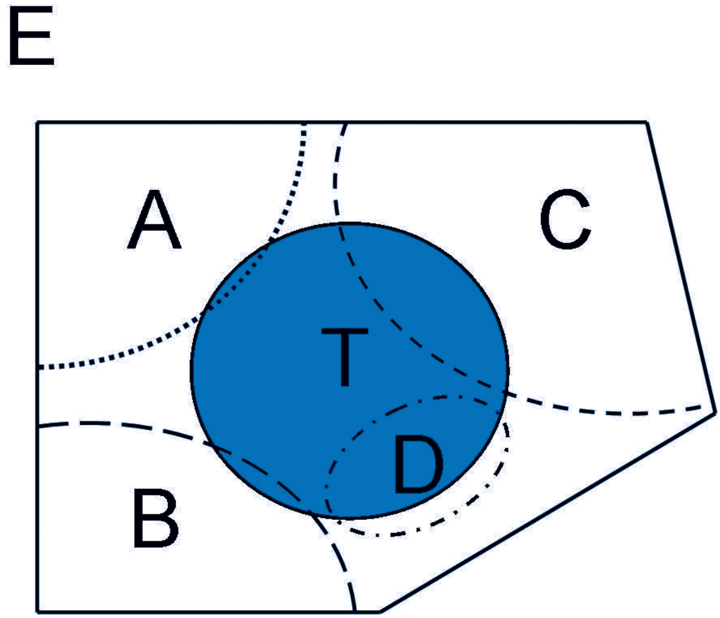
As discussed in Part I, the use of test results for high-stakes purposes can create incentives to influence student performance on these assessments in a sub-optimal manner. The practice of 'teaching to the test' is one such undesirable consequence but there are a number of documented instances where various school indicators and high-stakes tests can and have been manipulated in a manner that creates sub-optimal outcomes (Nichols & Berliner, 2005). Other problems can emerge if a school's value-added score can be more directly manipulated. Consider a scenario in which two assessments are employed to estimate schools' value-added. Suppose the first assessment occurs in Year 3 and the second assessment in Year 6. Clearly, a school's value-added increases if there is a larger positive difference between the assessments. There is an incentive, therefore, both to lift students' scores in Year 6 and to lower the scores (of those same students) in Year 3. This could be achieved by advising students not to take the Year 3 assessment as seriously as might otherwise be the case or even by encouraging them to deliberately under-perform. More radical actions could include structuring the curriculum so students are not properly prepared for the Year 3 assessment. Yet, strategies can be developed to reduce the likelihood of such sub-optimal activities. For example, the perverse incentive effect could be countered by imposing performance targets for the Year 3 assessment. More generally, schools should have an incentive to lift the performance of students in all assessments, thereby aligning their interests with those of the students. This can be achieved most simply when each assessment is both a prior and final assessment. Consider the annual assessment framework in the Flemish Community of Belgium, where each assessment (except for that in Year 1) has a dual role. Thus, the Year 3 assessment is a final performance measure in the value-added analysis between Year 2 and Year 3 (or Year 1 and Year 3) and also a prior assessment measure in the value-added analysis between Year 3 and Year 4 or some other subsequent year. This dual role mitigates the incentive to reduce performance on the Year 3 assessment. An exception would occur if policy makers place greater emphasis on the value-added measure for a specific year.

Schools can also be encouraged to lift student performance on the initial assessment by making that assessment part of general administrative procedures or school education policies or programmes. For example, student performance in the initial or prior assessment could be linked to a system of school inspections and school evaluation procedures. The assessment measures might also form part of a broader framework of school measures that are used to facilitate effective school choice. As was discussed in Part I, making these measures publicly available often creates positive incentives to lift student performance. Aside from considerations of aligning incentives, appropriate procedures should be implemented to ensure that every assessment is both fair and error-free. Test administration should be standardised and the marking of test papers should both be highly reliable and not open to tampering or manipulation at any stage of the process. This will result in greater confidence in the assessment outcomes and the value-added analysis that follows. It should also be noted that some countries utilise externally developed standardised assessments and others rely on school-level tests. A few, such as England, sometimes employ both kinds of assessment, although all the qualifications at Key Stage 4 are externally assessed. At Key Stages 2 and 3, however, data are collected from both external assessments and teacher assessments. External assessment data has been employed because it is thought to be more credible and comparable, as well as possessing superior psychometric properties. At Key Stage 1, tests were not externally marked and there have been concerns raised about robustness of the data (see Tymms and Dean, 2004). Since 2005, all Key Stage 1 (taken by seven year-old students) results are based on teacher assessments. Whilst this might introduce the potential for bias (in contrast to a standardised assessment), there is a possibility that the data are more valid since teachers draw on a broad range of evidence over a period of time rather than a single test administered on one occasion. If teacher evaluations are employed, they should be subject to external monitoring to assure comparability and validity

Philosophy of adjustment and the use of contextual characteristics

In order to obtain estimated school effects, most value-added models carry out a regression adjustment to student test scores. The intent of the adjustment is to 'level the playing field', that is, to remove from the comparisons among schools the confounding effects of systematic differences in the student populations they enrol. In doing so, the hope is that the value-added analysis will be more successful in 'isolating' the contributions of individual schools to their students' academic progress than is the case when schools are compared on the basis of student attainment alone. Although this strategy is sensible and widely used, it is important to appreciate that statistical adjustment must be carried out carefully and with due regard to possible negative consequences. With this in mind, the following paragraphs present a simplified explanation of statistical adjustment, illustrating the strengths and pitfalls of the procedure.

Figure 4.1. Graphical illustration of the process of statistical adjustment



Suppose the goal is to estimate the relative performance of a school. This is the target or parameter of interest. The circle (labelled 'T' in Figure 4.1) represents the true value of the parameter. The estimate obtained from an unadjusted comparison is represented by the four-sided figure (labelled 'E'). In this case, the estimate is too large. That is, we use the areas of the figures to indicate their magnitudes. E might be larger than T because the school's students are more advantaged than those of the average school. Since we recognise that schools are not randomly assigned to students (or vice-versa), we resort to statistical adjustment on measured student characteristics to create a more level playing field. Each adjustment is supposed to modify E and bring it closer to T. In Figure 4.1, the effect of the adjustment is represented by a figure contained in E that might or might not overlap with T.

The first adjustment (labelled 'A') reduces the area of E. The new estimate, E-A, is closer to T than is E. Note that A overlaps slightly with T, indicating that some of the adjustment removed a small portion of the true difference. However, the new estimate is still too large. Further adjustments for the next two characteristics (labelled 'B' and 'C') yield an estimate E-A-

B-C which is closer to T. In the case of C, however, there is considerable overlap with T, meaning that there has been some over-adjustment. Finally, the adjustment D has removed a good portion of T but relatively little of the part of E outside T. This means that there has been substantial over-adjustment. The resulting estimate, E-A-B-C-D, might be closer to T but it might be smaller than T rather than larger. A further adjustment similar in effect to D might yield an estimate that is poorer than previous estimates. The lesson to be drawn is that statistical adjustment must be carried out thoughtfully.

In most value-added modelling, isolating the contribution of schools requires estimating the relationship between student scores and various socio-economic and other contextual variables. Although there are measurement issues that need to be addressed in isolating the multiple impacts upon student performance, it can useful for policy makers to analyse both the extent of the relationship between student performance and specific contextual characteristics and, in some cases, analyse value-added results for particular groups of students. Analysis of this data can inform policy development in a variety of areas including school equity funding.

Importance of contextual characteristics

The OECD PISA programme does not produce value-added measures and is more closely aligned with what have been classified as contextualised-attainment models in this report. The most recent findings from PISA confirm previous evidence that students' socio-economic status is one of the largest predictors of school performance using such modelling (OECD, 2007a). These findings are consistent with the extant literature, which documents the statistical link between individual and family background variables on the one hand and youths' education on the other hand (OECD, 2007d; Haveman and Wolfe, 1995). Moreover, this link has been extended to include neighbourhood or community and peer characteristics (Ginther, Haveman, and Wolfe, 2000; Brooks-Gunn et al., 1993; Corcoran et al., 1992; Mayer, 1996). These analyses estimate the strength of the relationship between various factors and a single performance or outcome measure. These factors can include individual background characteristics and a variety of socio-economic contextual characteristics, as well as school characteristics. As discussed in the Introduction of this report, the key feature that distinguishes value-added modelling is the inclusion of a comparable prior attainment measure, thereby more accurately isolating the contribution of the school to student progress. When measures of prior attainment are included in the regression model, the incremental contribution of contextual characteristics to account for differences in student outcomes is often much reduced. Ballou, Sanders and Wright (2004) indicate that when a rich set of prior and concurrent attainment measures is available, adjustment for students' demographic characteristics has minimal impact on the estimated school effects. In addition, despite being generally in favour of including socio-economic status as a student background variable, McCaffrey et al. (2003, 2004) conclude that controlling for student-level socio-economic and demographic factors without measures of prior performance is not sufficient to remove the effects of background characteristics in all school systems, especially those systems which serve heterogeneous students. Policy makers should therefore be cautious in interpreting school performance measures from contextualised attainment models.

In the design of value-added models, policy makers and administrators must carefully consider the use of socio-economic contextual characteristics. For those more familiar with contextualised attainment models, the importance of socio-economic contextual characteristics as predictors of student attainment is well-known. Consequently, the discussion in the section above regarding the diminished role of these characteristics in value-added modelling might be somewhat surprising. Analysis of Norwegian and Portuguese data shows that the use of contextual characteristics is much more important in contextualised attainment models than in value-added models. Hægeland and Kirkebøen (2008) provide an empirical illustration of how estimates of school performance are affected by the choice of which socio-economic contextual variables are included in both contextual attainment and value-added models. The authors note that adjusting for students' prior performance and adjusting for students' socio-economic status are not mutually exclusive approaches to estimating school performance. It is also evident that the role of contextual factors can differ among countries and the type of model utilised. However, the findings of the Norwegian study concerning the influence of socio-economic status characteristics in value-added estimates were also obtained in the Portuguese longitudinal study. The analysis of the Norwegian data sheds light on the use of contextual variables in value-added models and illustrates the differences on this point with contextual attainment models. The study compared the results of four different specifications, incorporating an increasing amount of socio-economic data as control variables. The comparison of the results showed that adding socio-economic characteristics increased the amount of explained variance in student scores and reduced the dispersion of the distribution of school performance indicators in contextualised-attainment modelling. This is consistent with the literature, which finds that socio-economic characteristics are correlated with student performance and are not uniformly distributed across schools. However, their results indicate that in their value-added modelling, the effects of including additional socio-economic status variables are limited due to the presence of prior performance measures. They show that a simple value-added model that contains only basic demographic information (gender and year of birth), in addition to prior attainment measures, had much greater explanatory power than the most comprehensive contextualised attainment model. The inclusion of additional socio-economic characteristics to this value-added model had only a minor impact on the explanatory power of the model and on the estimates of school performance.

On the other hand, incorporating additional measures of prior performance had a greater impact on the predictive power of the model.

Notwithstanding the findings above, the addition of socio-economic characteristics to a value-added model might be consequential for particular schools. With regard to the Norwegian data, the largest impact for a single school with the inclusion of the full vector of socio-economic contextual characteristics in the value-added model corresponded to one-half of a standard deviation of the distribution of estimated school performance. This result underlines the importance when developing a system of value-added modelling of conducting sensitivity analysis not only of the overall model parameters but also for individual school estimates. Substantial changes in value-added estimates should stimulate further investigation as they might signal problems with the data. Ideally, these types of analyses should be carried out during the pilot stage of the implementation process.

Though the analysis of the Norwegian data is suggestive, one cannot draw general conclusions from this exercise. The consequences of including (more) socio-economic contextual variables in (contextualised) value-added models, and of including more socio-economic contextual variables in a contextualised attainment model, might vary across levels, years and countries. If socio-economic characteristics are only related to the initial level of performance and not the growth rate, then there would be no benefit in including these characteristics in value-added models. On the other hand, there would be some benefit if these characteristics were correlated with growth in student performance. In some OECD member countries the inclusion of ‘year of birth’ in the value-added model captures the effect of ‘repetition’ or grade retention, which is a phenomenon negatively correlated with socio-economic status (OECD, 2007c). It is also possible that the inclusion of ‘year of birth’ captures the effect of differential age of entry into the education system. Employing a contextualised attainment model (variance component model) to PISA 2000 data, Ferrão (2007a) shows that the ‘repetition’ explains 45% of the variability of the Portuguese students’ performance in Maths (measured by PISA). From the educational point of view, the inclusion of the variable ‘year of birth’ as covariate in the value-added model might be controversial and should be appropriately addressed by each country.

An analysis of Portuguese data (representative of Cova da Beira region) yielded similar findings to the Norwegian analysis with respect to the effect of including various socio-economic characteristics in value-added models (Ferrão, 2008). This analysis utilised data collected at the beginning and end of the academic year 2005-06 for students enrolled in the 1st, 3rd, 5th, 7th and 8th grades. The response variable was the maths score in a standardised test equated[14] with maths prior achievement (Ferrão et al., 2006). The socio-

[14] Equalisation via common items.

economic characteristics analysed include those measuring parental education and student eligibility for free school meals and books. Eligibility for free school meals is a common measure used in similar estimations that have included socio-economic contextual characteristics (see Goldstein et al., 2008; Braun. 2005a; Ballou, Sanders and Wright, 2004; McCaffrey et al., 2004; Sammons et al., 1994; Thomas and Mortimore. 1996). The focal issue was the sensitivity of school value-added estimates to different single-variable operationalisations of the construct of socio-economic status. Results showed correlations near 0.90, suggesting that the use of simple alternative proxies might yield comparable results (Ferrao, 2007a). However, it is important to note that the rankings of some schools do undergo substantial shifts over time. Although these findings are somewhat encouraging, further work should be carried out focusing on other commonly used characteristics, with attention to the use of multiple covariates.

When considering the use of socio-economic characteristics, the frequency and range of student assessments must also be taken into account. If students are frequently assessed in a number of subjects, and the number of test scores is correspondingly large, then the contribution of background variables in value-added models is greatly reduced. However, if there are less frequent assessments and there is a longer gap between student assessments then the potential contribution of background variables is greater. For example, if a student who has been assessed in Year 3 is not assessed again until Year 6 then contextual variables such as socio-economic status might be strongly correlated with the student's rate of progress over this three-year period. Leaving aside technical considerations, it might be advisable to include socio-economic characteristics in a value-added model in order to gain the confidence of stakeholders. One approach would be to present the results for different models that include none, some or all available socio-economic and other background characteristics. The importance of such an approach will depend on the intended use of the school value-added estimates. The concerns of key stakeholders might be greater if a strong school and/or teacher accountability system is being enacted than they would be if value-added estimates are being used solely for school improvement purposes.

Which socio-economic contextual characteristics?

It is useful to recall that the estimated school effects generated by value-added modelling represent the combined contributions of schools' actions and policies together with the peer effects stemming from the interactions among students and their impact on school climate, attitudes towards academics and other school-level variables. To the extent that adjustments for individual and school-level characteristics do not fully capture such peer effects, the estimated school performance measures are not unbiased estimates of schools' contributions to student learning. Note too that the interpretation of the estimated school performance measures depends on which variables are used for the adjustment. Each set of variables implicitly establishes the 'level playing-field' on which schools are compared. That is, when we state that the estimated school performance measures give us the relative ranking of schools' performance with all 'other things' being equal, it is the adjustment that determines what comprises those 'other things'. It should be borne in mind that the main purpose for including explanatory variables in the model is to reduce bias in the estimated school performance measures. To accomplish this goal, these variables must be both related to the outcome and differentially distributed among schools. The stronger the relationship and the greater the variation among schools, the more will the adjustment have its desired effect. In any event, the addition of these variables will generally increase the accuracy of prediction.

The student characteristics that are typically employed in the adjustment process include such variables as gender, race/ethnicity, and level of parental education. These characteristics are generally associated with academic achievement (OECD, 2007b; Lissitz et al., 2006). If these characteristics are unequally distributed across schools, then failing to take them into account can lead to biased estimates of schools' value-added. That is, in the absence of any adjustment, schools enrolling students with more 'favourable' characteristics, on average, will be advantaged in comparison with schools enrolling students with less 'favourable' characteristics, on average. An analysis of existing data and data collected during the pilot programme should reveal the appropriate contextual characteristics to include in the value-added modelling. In doing so, it should be recognised that the inclusion of (multiple) prior performance measures will generally weaken the relationship between current test scores and socio-economic characteristics. At the same time, the inclusion of certain characteristics in the model might be valuable for public acceptance and can have an impact on the value-added scores of individual schools.

The success of the adjustment process depends both on the appropriateness of the model as well as the scope and quality of the variables used in the adjustment. With respect to the first consideration, the adjustment is usually carried out by fitting a linear regression model. If the relationship is strongly non-linear then the model is misspecified and value-added estimates will be biased. The problem can sometimes be mitigated by introducing interactions among the predictors. For example, it might be that for certain immigrant groups there is a gender gap in performance that is different in magnitude and even in direction from that observed for the majority group. The standard linear regression model would be misspecified and the resulting value-added estimates will be biased. The bias might be particularly problematic if members of the minority group students are concentrated in certain schools, which in a number of systems might be a likely occurrence.

With respect to the second consideration, limitations in data collection usually result in only a small set of student characteristics being available for analysis. If there are unmeasured characteristics that are independently related to the outcome, then the adjustment model is misspecified and, again, the resulting estimates will be biased to some extent. Furthermore, data quality is always a concern, since poor quality can lead to increases in both the variance and bias of the estimated school effects. Inaccuracies can arise when the data are obtained from student self-reports, especially those from younger students. Parental self-report data can be problematic if the questionnaires are ambiguous or if the parents are not familiar with the language. Even administrative data drawn from school files can suffer from gross errors.

An advantage of using value-added modelling is that they permit the quantitative assessment of the magnitude of the disadvantage associated with particular characteristics (*e.g.* ethnicity, income, level of familial education) in relation to student progress, not just in relation to student attainment at a particular point in time. The patterns that emerge over time in these relationships are important for policy development. For example, do particular forms of disadvantage exist, are they sustained over the course of students' school education and does the impact of such disadvantage expand or decline over time? Moreover, careful use of the results of value-added models makes it possible to identify schools that are more successful in lifting the performance of disadvantaged students. This can lead to the dissemination of 'best practice' among schools, provided that channels are in place to facilitate such information transfers.

The analysis conducted by Hægeland and Kirkebøen (2008) demonstrated, *inter alia*, that by international standards Norway has an extensive set of student-level contextual data available for analysis. Clearly, the level of data availability differs across countries and, typically, it is data availability that constrains the contextual characteristics that can be included in various models. On the other hand, the availability of prior measures of academic achievement might lessen the need for an extensive set of contextual variables. Most countries collect some form of demographic information from students and include them in their value-added models. Table 4.2 details the range of contextual data collected and available for use in value-added modelling across participating countries. Student age, gender and a variable indicating immigrant status and/or ethnicity are the main individual demographic characteristics included across countries.

The results from a number of countries illustrate the importance of including a measure of students' age (Ray, 2006; Hægeland et al., 2005). Even when excluding mature-age students or those students repeating a grade or year level, the age of students in a given grade or year level can vary by up to a year in some systems. Age has been shown to have a statistically significant relationship to student progress and, therefore to the estimation of schools' value-added. The recording of age varies across countries and this, in part, reflects differences in data collection methods. In some countries, school enrolment data specifies students' date of birth, while in other countries, the lack of such data means that either there are other administrative data sources or that the data (exact age or age range) is obtained directly from the students themselves.

Student gender is a characteristic used in most value-added analyses across participating countries. This characteristic does not often influence schools' value-added scores as the distribution of male and female students is typically uniform (with the obvious exception of same-sex schools). However, gender might be important for more detailed analysis of value-added information that fosters school improvement initiatives. Differences in the performance of male and female students have received increasing attention in recent years as female students have achieved higher levels of performance and attainment than male students in a number of domains and in a number of attainment measures. However, the magnitude and possibly the direction of the expected effect of a gender variable might differ depending upon the measure. In some countries, performance comparisons show male students performing more strongly in subject areas such as mathematics and science and females performing more strongly in reading and writing literacy (OECD, 2007a; 2007b). Such gender disparities might not have an impact on value-added estimates. However, it might be useful to conduct the value-added analyses separately by gender for specific subjects as the results could signal the need for specific policies and programmes that seek to address such disparities.

Immigrant status and/or ethnicity are identified differently across countries and reflect differences in the ethnic mix, the policy focus, and the data available. In some countries, a single variable reflecting immigrant status can be included in their modelling. In others, specific ethnic groups or the region from which the student immigrated are included as some groups are relatively disadvantaged in comparison with the majority group. The results of a value-added analysis for specific groups of students might indicate the need to further disaggregate the student population. For example, an analysis of a single variable identifying immigrant status might yield a bi-modal distribution or a distribution of scores comprising distinct clusters. This might indicate that particular ethnic or immigrant groups are progressing at different rates and that schools' contributions to that progress also differs. There is some evidence that such patterns can persist and even grow over time (Borjas, 1995, 2001). Additional analysis might indicate which groups should be separately identified. In these situations, including a simple measure of immigrant status will not fully capture the disadvantage faced by distinct immigrant groups and will therefore not be as useful for policy initiatives. In some instances, interaction variables might prove useful, particularly if there is substantial economic heterogeneity with particular ethnic groups. To accommodate such changes, flexibility is required in both the data collection and in the information technology used to compile the data. Administrators and policy makers require this flexibility to better specify the value-added modelling and produce more useful results, as well as for *ad hoc* data collections required for specific policy objectives such as programmes aimed at specific regions or groups of students. In some countries, the language barriers to student progress are of concern, particularly when the language of instruction differs from the language spoken at home or the students' first language. These barriers are considered to be particularly important (both from an educational and a political perspective) when these students exhibit poor performance in a number of subject areas.

Table 4.2 organises contextual variables into distinct categories. This categorisation has been made for illustrative purposes and does not necessarily apply to a specific country. To assist in their modelling, most countries collect measures of student learning difficulties, level of family education, level of economic resources and welfare benefits. The latter could also be considered a measure of economic resources. Some countries also collect characteristics related to a student's family structure that have been shown to affect outcomes such as parental marital status, whether the student is being raised outside the family home, and a measure of family size (Amato and Keith, 1991). It is important to note that some characteristics are fixed and do not change over the course of students' schooling, but others characteristics might change over time. The data collection and storage systems must be flexible enough to accommodate both kinds of characteristics.

The socio-economic characteristics collected across countries concentrate on the level of parental education levels and family income. Characteristics denoting whether students and/or their families are in receipt of welfare payments such as educational and household support are also included in some countries. These can be further indications of the level of economic resources available to students and families. In the Flemish Community of Belgium, a variety of data is collected to form an index of students 'Being at Risk'. Norway also includes measures on the level of family wealth and the incidence of parental unemployment over the 10 years prior to the assessment.

Characteristics identifying students with learning difficulties are collected in most countries. The typology of learning needs differs across countries and is normally aligned with existing data collections in the education system. While not considered to be an indicator of a special learning need, data identifying if the student has repeated a grade in the school is included by a number of countries. This can be particularly important if the student is repeating the grade in which the assessment is being administered or a grade between the current assessment and the prior assessment. Estimates of the contribution of a school to student progress between the two assessments could be biased by differences in the number of years of instruction.

Table 4.2. Contextual data collected across participating countries that potentially could be used for value-added modelling

Country	Demographic Information	Immigrant status	Student learning difficulties	Family structure	Family education	Economic resources	Welfare benefits
Belgium (Fl.)	Age, Gender, Country of birth of student and both parents, Age when immigrated	Language spoken with mother at home, migration background	Identified learning difficulties, history of repeating a grade	Student's being not raised at home (*e.g.* foster parents, institution) as constituent part of student's status of being at risk (BAR)	Maternal education qualification		Study grant, household replacement income, household depending on welfare benefit as constituent part of student's BAR status
Czech Rep.	Age, Gender, Place of birth		Students with special learning needs		Parents' highest level of education completed	Parental occupation categories	
England	Age, Gender, Ethnic group	English as a first language (student)	Student recorded as having special learning needs			Neighbourhood income deprivation (measured by postcode data)	Student entitled to Free School Meals (dependent on family income)

Table 4.2. Contextual data collected across participating countries… (cont.)

Country	Demographic Information	Immigrant status	Student learning difficulties	Family structure	Family education	Economic resources	Welfare benefits
France	Age, Gender, Place of birth	Nationality, Place of birth	Students' class, subject options			Parents' occupation (divided between 4 occupational categories), Family size,	Financial aid received,
Norway	Age, Gender, Graduation in years earlier than expected	Born outside of Norway, country/region of origin, Age of immigration		Parents' marital status, Age of parents at birth of first child, Number of siblings and half siblings, Birth order	Parents' highest level of education completed	Family income, Family wealth (based on family taxable wealth)	Incidence of parental unemployment over prior 10 years,
Poland	Age, Gender		Dyslexia				
Portugal	Age, gender	Language spoken at home	Student marks, grade repetition, Special education needs	Number of siblings	Parents' education (ISCED classification)	Parents' occupation, computer at home, internet at home	Student entitlements to support (depend on family income)

Table 4.2. Contextual data collected across participating countries… (cont.)

Country	Demographic Information	Immigrant status	Student learning difficulties	Family structure	Family education	Economic resources	Welfare benefits
Slovenia	Age, Gender		Special education needs				
Spain	Age, gender	Country of birth of student and parents, Age of immigration, Language spoken at home	Students with special learning needs, history of grade repetition	Questionnaire on family structure	Parents' education levels	Parents' occupation levels, cultural and other possessions at home	Student grants
Sweden	Age, Gender, Place of birth, Ethnic group	Immigrant background of students and parents, Year of immigration			Parents' highest level of education completed	Household income	Household social benefits

School-level data

Up to this point, the discussion has focused on adjustments for student-level characteristics. It is also possible to adjust for school-level or contextual characteristics.[15] Such characteristics might be aggregations of student variables (*e.g.* mean test scores) or those that are only defined at the school level (*e.g.* racial/ethnic composition of the school population, community socio-economic status). Although one can quite easily incorporate such variables in a model, the danger of over-adjustment remains. That is, if the contextual variable is associated with true school performance, then adjusting for that variable biases the estimates of school effects. Thus, caution is warranted when deciding whether to carry out such adjustments.

In some countries, the type of school is incorporated as a covariate although this might not extend to a distinction between government and non-government schools, as the latter are sometimes not included in the value-added analysis. Additional information might be available concerning the level of school resources and, to some extent, on school processes. Incorporating school-level covariates might be particularly useful to those interested in school development. Analyses that focus on certain types of schools or on particular groups of students (*e.g.* students with special learning needs) can prove to be more useful when both contextual and school-level variables are used to adjust student outcomes. One example is programme evaluation when programs are implemented in some schools but not in others. In some settings, it might also be possible to incorporate classroom-level data for more detailed analyses of teacher value-added. As an example, in the Flemish Community of Belgium information is collected on: the use of particular textbooks; the gender and experience of the teacher; whether there is a computer in the class; the use of computers and the Internet in lessons; and the teaching time allocated to the subject. Such analyses can be readily applied in more targeted analyses of value-added. Analyses that regress value-added estimates on school practices to ascertain if they account for a substantial amount of the variance in the value-added estimates can be effective secondary analyses and offer another option for policy makers.

Appropriate steps should be taken to ensure the integrity of all data, regardless of whether it is part of a broader administrative data collection or if it is gathered alongside other data for particular use in the value-added analysis. Ray (2006) points out that some school-level covariates are subject to manipulation by school authorities. For some models, the impact of a change in the covariate on a school’s value-added can be worked out in

[15] These adjustments are not possible with models that incorporate school fixed effects.

advance and, hence, there is an incentive to shift the value in the desired direction. For example, in the contextualised value-added modelling used in England, the higher a school's proportion of students unclassified with respect to ethnicity, the higher its value-added, all else remaining constant. Thus, it would be in the school's interest either to not find out or to not report students' ethnicity. Quite sensibly, Ray points out that the models selected should be designed to minimise such perverse incentives. Ideally, such data would be collected outside the student assessment framework and collected in a system that does not involve the school administration and so reduce the likelihood of data corruption.

Chapter Five

Illustrative Value-Added Models

This Chapter introduces a number of different value-added models to provide some examples that can be used in education systems. The objective of this Chapter is not to present a complete list or review of the different types of value-added models as this is outside the scope and purpose of this report. Rather, the types of models presented illustrate some of their differences and illustrate how specific issues are handled with different modelling procedures. The design features discussed in Chapter Four affect these models to varying degrees and each model has both advantages and disadvantages with respect to the full set of issues. Five general categories of value-added models are discussed: linear regression models; variance component models; fixed effects models; multivariate random effect response models; and some discussion of growth curve analysis. Value-added modelling can be used to estimate either annual or cumulative school effects but in a number of the models presented as examples here the school effect is measured as an annual rather than a cumulative effect.

The discussion of these types of models should also inform decisions of the choice of the most appropriate model given the methodological issues discussed in Chapter Six. It should also be noted that this report does not advocate the use of one model over another. Rather, it points out how some models can be more appropriate given the different policy objectives and the constraints under which the analyses must be carried out. Nonetheless, during the development of a system of value-added analysis, it is imperative that a variety of models be examined to evaluate their relative suitability with respect to a number of criteria.

Linear regression value-added models

This first set of models employs simple linear regression to adjust outcome test scores for some combination of student prior test scores and student or contextual characteristics. One form of the model is:

$$y_{ij(2)} = a_0 + a_1 y_{ij(1)} + b_1 X_{1ij} + \ldots + b_p X_{pij} + \varepsilon_{ij} \qquad (1)$$

where

i indexes students within schools j,

$y_{ij(2)}$ = final test score,

$y_{ij(1)}$ = prior test score,

{X} denotes a set of student and family characteristics,

a_0, a_1, b_1, … b_p denote a set of regression coefficients,

ε_{ij} denotes independent and normally distributed deviations with a common variance for all students

Denote the predicted value for student i in school j by $\hat{y}_{ij(2)}$, based on fitting equation (1) to the full data set. Then, the estimated value-added for school j is taken to be the average over its students of the fitted residuals: $ave_i\{y_{ij(2)} - \hat{y}_{ij(2)}\}$.

Thus, if students in school j achieve higher final test scores on average (in comparison with students from other schools with similar predictor values), then the corresponding residuals tend to be positive, yielding a positive estimated value-added for the school. There are many variants of the basic model. In particular, if prior year test scores are available from earlier years or other subjects, then these can be easily accommodated. See Ladd and Walsh (2002) and Jakubowski (2007) for other examples. For this method to yield consistent estimates requires that the included covariates are uncorrelated with the error term, which may include a school effect in addition to idiosyncratic errors. In addition, it does not take into account the structure of the error term that is a feature of some of the models illustrated below.

Variance component or random effect models

Another type of model comprises two regression equations: a student-level regression as in (1) above; and a school-level regression that models the variation in adjusted school intercepts obtained from the student-level regression. A technical advantage of such so-called hierarchical (or multi-level) models is that they take into account the grouping of students within schools, yielding more accurate estimates of the uncertainty to be attached to the estimates of school value-added.

A typical formulation of such models is:

$$
\begin{aligned}
& y_{ij(2)} = a_{0j} + a_1 y_{ij(1)} + b_1 X_{1ij} + ... + b_p X_{pij} + \varepsilon_{ij} \\
& a_{oj} = A + \delta_{0j} \\
& \text{where} \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (2) \\
& \varepsilon_{ij} \sim N(o, \sigma^2) \\
& \delta_{0j} \sim N(0, \tau^2).
\end{aligned}
$$

Each residual in both equations is assumed to be independent of all other residuals. The rationale for the second equation is that the adjusted school intercepts $\{a_{0j}\}$ are thought of as being randomly distributed about a grand mean (A) and the deviations from that mean are taken as estimates of school value-added. Interest centres on those schools with large deviations (positive or negative). This sort of model is employed in the 'contextual value-added' modelling that has been implemented in England, although the actual school value-added estimates are obtained through further analysis and computations. The model utilised in England is further discussed below.

These types of models are often referred to as 'random effects' models because the parameters that are intended to capture the schools' contributions to student performances are treated as random variables. Consequently, the estimated effect for a particular school is influenced by the data from all the other schools, as well as the data from the school itself. The resulting estimates are sometimes called 'shrinkage' estimates because they can usually be represented as a weighted average of the ordinary least squares estimate for the school and an estimate related to the data for all the schools. The specific combination depends both on the model and the data available. Shrinkage estimates are biased but typically have smaller mean squared error than ordinary least squares estimates.

With multi-level modelling, the residual variance is partitioned into two levels: the student (Level 1) and the school (Level 2). These are the model's 'random effects'. Within an education system, it is possible to have other levels. For example, *within* schools, students are grouped into classes, but if there is no national data on teaching groups, this level cannot be modelled. Level 1 residuals show variation in students' outcomes in relation to their schools. The Level 2 residuals show schools' outcomes in relation to the national expected results, given the included covariates. These Level 2 residuals are the school value-added scores.

A closely related model is the variance component model (see Raudenbush and Willms 1995: p.321) with a different set of level one and/or level two covariates, depending on the type of school effect (*type A* or *type B*) the analyst intends to estimate. The model is as follows:

$$y_{ij} = \mu + \beta_W \left(x_{ij} - \bar{x}_j\right) + \beta_b \bar{x}_j + u0_j + \varepsilon_{ij} \qquad (3)$$

where y_{ij} is the test score result for student i in school j; x_{ij} is the student prior achievement; $\bar{x}j$ is the school sample mean prior achievement for school j; u_{0j} is the school-level random component, also called random effect or value-added of school j, that is assumed to be normally distributed with mean of zero and variance σ^2_{u0}; and ε_{ij} is the student-level random component assumed to be identically, independently and normally distributed with mean zero and a variance σ^2_ε. Fixed parameters μ, β_w, β_b, represent, respectively, the mean of test score, the within-school regression coefficient relating the student prior achievement to the outcome test score, and the between school slope.

Antelius (2006: p.4) illustrates how a variance component model could be used to calculate value-added in upper-secondary schools in Sweden. The grades obtained when leaving compulsory comprehensive education were assumed to reflect the previous knowledge of students and educational background while the grades obtained from upper-secondary school show the level of knowledge students have achieved in the core subjects (mathematics, natural science, Swedish, English, social science, artistic activities, physical education and health and religious studies). Measures of each school are presented for a period over three years to ascertain whether or not this value changes over time (Antelius, 2006).

In Portugal, analysis of three different variance component models were considered for the region of Cova da Beira, involving a representative sample of students at the primary, elementary and lower-secondary levels of education (Vicente, 2007). A different set of predictor variables were included in each model: a null model; a Traditional Value-Added (TVA) model that included student socio-economic status and prior achievement; and in addition, a model that included other student variables such as gender, whether the student was classified as special needs, if they attended kindergarten, type of class in primary education, and grade repetition (TVA+). The correlation between value-added estimates generated from the Null and TVA models varied from 0.61 to 0.94 depending on the grade. In contrast, with the exception of scores for the 3rd grade, the values of the correlation between TVA and TVA+ estimates were equal or larger than 0.96. Ferrão and Goldstein (2008) also evaluated the impact of measurement error in those estimates.

Fixed-effects value-added models

A rather different approach employs so-called fixed-effects models. As the name implies, these models represent school contributions as fixed parameters as opposed to random effect models where the school contributions are assumed to be random variables with a common distribution. In random effects models, correlations between covariates and the random effects can introduce bias into the estimates of the school effects. That problem does not exist with fixed effects models and this, arguably, is their main advantage. On the other hand, the estimated school effects might vary considerably from year to year, since there is no use of 'shrinkage'. A simple version of such a model is given below:

$$y_{ij(2)} = a_0 + a_1 y_{ij(1)} + \sum_k b_{kij} X_{kij} + \theta_j + \varepsilon_{ij} \tag{4}$$

where

$\theta_j =$ effect of school j.

Hægeland and Kirkebøen (2008) utilise a fixed-effects model to analyse school value-added in Norway. They provided an empirical illustration of how estimates of school performance are affected by the choice of which socio-economic contextual variables are included in either contextual attainment models or value-added models. The authors note that adjusting for students' prior performance and adjusting for students' socio-economic status are not mutually exclusive approaches to estimating school performance. It is also evident that the role of contextual factors might differ among countries and the type of model utilised.

The Dallas model

A well-known model that combines the features from different classes of models is the two-stage model employed in Dallas, Texas, presented in Webster and Mendro (1997; see also Webster (2005)). The role of the first stage was to adjust the student test score variables (current scores as well as prior scores) appearing in the second stage. The adjustment was carried out for a number of relevant student characteristics. In the second stage, the adjusted current score was regressed on the adjusted prior scores in a hierarchical linear model that took into account the grouping of students within schools. Moreover, this model easily accommodated the inclusion of school-level covariates that could further enhance the statistical characteristics of the resulting estimates of schools' value-added.

Specifically, let

$$y_{ij} = b_0 + b_1 X_{1ij} + \ldots + b_p X_{pi\,j} + \varepsilon_{ij} \tag{5}$$

Where

i indexes students within schools j,

y denotes a current or prior test score outcome,

{X} denotes a set of student characteristics that include ethnicity/language proficiency, gender, student poverty level, first- and second-order interactions among these characteristics, as well as a number of indicators of neighbourhood socio-economic status,

{b} denotes a set of regression coefficients,

ε_{ij} denotes independent, normally distributed deviations with a common variance for all students.

Thus, the coefficients of equation (5) are estimated for each possible choice of y. Typically, ordinary least squares is employed. Interest, however, focuses not on the estimated coefficients, but on the residuals from the regression. For each fitted regression, the residuals are standardised. Suppose we use a ~ to denote a standardised residual.

Stage 2 employs a two-level model. Level 1 takes the following form:

$$\tilde{Z}_{ij} = c_{0j} + c_{1j}\,\tilde{P}^{1}_{ij} + c_{2j}\,\tilde{P}^{2}_{ij} + \delta_{ij} \tag{6}$$

and level 2 takes the form:

$$c_{oj} = G_{00} + \sum_{k=1}^{m} G_{0k} W_{kj} + u_{0j}$$

$$c_{1j} = G_{10} + \sum_{k=1}^{m} G_{1k} W_{kj}$$

$$c_{2j} = G_{20} + \sum_{k=1}^{m} G_{2k} W_{kj}. \tag{7}$$

In level 1:

i indexes students within schools j,

$\tilde{Z}_{ij}$ denotes a student's adjusted current test score,

$\tilde{P}^{1}_{ij}$ and $\tilde{P}^{2}_{ij}$ denote a student's adjusted prior test scores,

{c} denote a set of regression coefficients,

δ_{ij} denotes independent, normally distributed deviations with a common variance for all students.

Note that the term 'adjustment' refers to the results of carrying out the stage 1 analysis. In principle, more than two prior measures of prior achievement could be employed.

In level 2:

{W} denotes a set of m school characteristics, including various indicators of the demographic composition of the school, multiple indicators of the socio-economic status of the school community, school mobility and school crowding,

{G} denotes a matrix of regression coefficients,

u0j denotes a school-specific deviation of its intercept in the level 1 equation from the general linear regression relating school intercepts to school characteristics.

The stage 2 model, which is similar to a random-effect model, is fit using multi-level software. The estimated school effect is again a reliability-adjusted estimate of u0j. This is sometimes called an empirical Bayes estimate because it is equal to the estimate of u0j obtained from a least squares regression for that school alone shrunk toward the estimated regression plane, with the amount of shrinkage inversely proportional to the relative precision of that estimate (see Braun (2006b) for an introduction to empirical Bayes methodology). The overall performance index for a particular school is constructed as a weighted average of the estimated school effects for different courses and grades. In Dallas, the weights were determined in advance by a designated group of stakeholders, the Accountability Task Force.

In England, a simplified version of a multi-level model has been employed to facilitate effective interpretation for stakeholders. An example of such efforts is the decision not to include any explanatory variables for the random component of the model. Such a decision simplifies the model but introduces the assumption of uniformity in value-added between students within schools such that performance can be illustrated with a single value-added score. A more complex approach is to assume variation within schools so that a range of measures is produced for each school. A significant feature of multi-level modelling is the application of 'shrinkage', where the value-added scores for small schools tend to be closer to the national mean, making it less likely that extreme value-added scores will be recorded for these schools. The model can be kept relatively simple: it could, in theory, have more levels of analysis and more explanatory variables both in the 'fixed' and 'random' parts of the model.

Multivariate random effect response models

The EVAAS (Education Value-Added Assessment System) model is an example of a multivariate, longitudinal, mixed effects model; that is, test data is collected on students in multiple subjects over several grades. While the EVAAS model continues to be slightly updated over time, published versions are not yet available and a recent application takes the following form:

Let

i index students,

j index transitions,

n_i the school attended by student *i*.

Then, the bivariate model is of the form:

$$\left(y_{ij}, z_{ij}\right) = \left(\mu_j, \gamma_j\right) + \sum_{k \le j} \left(\theta_{n_i k}, \varphi_{n_i k}\right) + \left(\varepsilon_{ij}, \delta_{ij}\right); \quad (j = 1, 2, 3) \qquad (8)$$

where

y_{ij} represents the student's reading score;

z_{ij} represents the student's math score;

μ_j represents the average reading score over the whole population;

γ_j represents the average math score over the whole population;

$\theta_{n_i k}$ represents a school effect in reading;

$\varphi_{n_i k}$ represents a school effect in math; and

ε_{ij} and δ_{ij} are the random error terms in reading and math, respectively.

The parameters $\{\mu\}$ and $\{\gamma\}$ are assumed to be fixed, whereas the parameters $\{\theta\}$ and $\{\varphi\}$ are assumed to be random and jointly independent. Let $\underline{\varepsilon_i} = (\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3})$ and $\underline{\delta_i} = (\delta_{i1}, \delta_{i2}, \delta_{i3})$, then $(\underline{\varepsilon_i}, \underline{\delta_i})$ are assumed to follow a multivariate normal distribution with mean vector zero and an unstructured positive definite covariance matrix. Conditional on the other parameters in the model, $(\underline{\varepsilon_i}, \underline{\delta_i})$ are assumed to be independent across students. The joint normality assumption of the error terms is critical for multilevel modelling of this type to correct for confounding or non-random assignment.

The layered model is sometimes referred to as a *persistence model* because the school effects at one transition are carried over to succeeding transitions. Typically, the variance-covariance matrix for the student-level error components is left unstructured. It is assumed to be common to all students within the cohort but might vary across cohorts. Consequently, the number of parameters can be large and a substantial amount of data is required for accurate estimation.

It should be clear that both the data base requirements and the computational demands are very substantial. The EVAAS model is implemented on proprietary software and the model described above has been used to analyse data from more than a hundred school districts for more than a decade. It has been recently modified but there are no descriptions yet publicly available. A more complex version of the EVAAS model is employed to estimate teacher effects. School and teacher models can be, and are, run in parallel, but there is little discussion in the literature as to how the two sets of estimated effects can be used jointly.

The primary attraction of the EVAAS model is that, because it focuses on student progress across a number of assessments, it affords no obvious advantage to schools with students who enter with comparatively high test scores. Another attraction is that there is no need to discard student records that have missing data. Missing data are dealt with as a matter of course. Recent studies support the robustness of estimates obtained from EVAAS to departures from assumptions about the nature of the missing data (Lockwood and McCaffrey, 2007). An obvious distinction between the Dallas and EVAAS models is that the latter includes neither student nor school covariates. Since the Dallas model employs data from only two time points, it must rely on covariance adjustments to make comparisons between schools fairer. Furthermore, consideration of political imperatives and acceptability to stakeholders can provide additional impetus for incorporating student characteristics into the stage 1 model. On the other hand, Sanders et al. (1997) has argued that with multivariate longitudinal data, each student acts as their own 'block', and this obviates the need to incorporate such data into the model (Sanders et al., 1997; Ballou, Sanders and Wright, 2004). Although it is certainly true that simple gain scores are more weakly correlated with student characteristics than are current scores, Sanders' assertion is not a mathematical certainty and requires further investigation.

To this end, Ballou, Sanders and Wright (2004) showed how student covariates could be included in the EVAAS model for teachers without introducing bias into the estimated teacher effects (denoted as EVAAS-C.) They applied both models to data from a school district and found that the estimated teacher effects from the two models were very similar. In other words, the EVAAS estimates were robust to the inclusion of student covariates. It is an open question whether these findings generalise to other settings and to the estimation of school effects.

For some, the fact that the EVAAS does not employ student covariates is an advantage because there is no suggestion that there are different expectations for students with different backgrounds On the other hand, there might be situations in which non-statistical considerations, for example, might lead to the adoption of EVAAS-C in preference to EVAAS. It should be kept in mind that adjusting for student covariates in models less encompassing than EVAAS could bias the estimates of school performance in systematic ways. For example, if student covariates are correlated with school performance (*e.g.* higher levels of parental education are correlated with schools having more qualified teachers) then adjusting for the covariate will result in an underestimate of school performance.

Goldstein (1987) offers another example of a multivariate response model that allows for the cross-classification of students both by their Junior and Secondary schools. The results of the cross-classified model suggest that the Secondary school value-added is influenced by the particular Junior school the student attended. Another example can be found in the work of Ponisciak and Bryk (2005). Building on earlier work of the Consortium on Chicago School Research, they introduced a three-factor, cross-classified model, which they denoted HCM3. The model made use of the longitudinal records of students in a single subject. Separate analyses were conducted for each subject. Students were cross-classified by the class and school attended for each grade. As the authors point out, their 'model is a combination of two simpler models – a two-level model for student growth in achievement over time, and a two-level model for the value each school and classroom adds to student learning over time' (Ponisciak and Bryk, 2005: 44).

While the final version of the model is rather complex, the basic idea is quite simple. Each student is assumed to have a linear latent growth trajectory. The slope of that trajectory in a given year and grade is deflected, positively or negatively, by the combined effects of the classroom and school in that year. The deflection is assumed to be permanent; that is, it persists through the next assessment and beyond. Note that this model assumes that the test score scale can be treated as if it were an interval scale, an assumption that is at best an approximation.

Growth curve analysis

Some consideration should also be given to growth curve analysis that utilises longitudinal data with more than two observations of student performance to estimate the contribution of schools to students' growth in that performance. A growth (in performance) curve is depicted by a growth curve of a performance measure (or other outcome) over time. When estimating growth curves, the model smoothes over the observed measures to estimate the continuous trajectories that are believed to underlie the observations. Growth curve models assume that there is a latent growth curve that has given rise to the scores on the measurement occasions (it is for this reason that they are sometimes referred to as 'latent growth curve models'). In individual growth curve analysis, a growth curve for each subject is estimated to represent the development over time. With linear growth curves, two growth parameters are estimated, namely an initial level growth parameter (intercept or status) and a growth rate parameter (growth or slope). Both parameters vary between individuals meaning that for each individual a growth curve is estimated with a specific initial level and a specific rate of change. There is a 'base growth model' for a cohort entering in a particular grade and year:

$$E[y_{it}] = c_{0i} + c_{1i}t \qquad (9)$$

Here

i indexes students and t indexes grades,

E denotes the expectation operator,

y denotes the test score,

c_0 and c_1 denote the initial level and the slope of growth.

It is assumed that the pair (c_0, c_1) are randomly distributed over the students in the cohort. Equation (10) represents the latent growth trajectory for student i in the absence of class and school effects. Now, let v_t denote the deflection to the slope by the class and school in which the student was enrolled in grade t. Then

$$E[y_{it}] = c_{0i} + tc_{1i} + \sum_{k=1}^{t} v_k \qquad (10)$$

The last term on the right hand side, the summation, represents the cumulative contribution of the class and school effects over the t grades. The {v} (the school effects) are assumed to be random across classrooms nested within schools and independent of the student level effects.

Additional complexity is introduced by taking into account the realities of working school systems. For example, secular changes might take place in the system and affect all the students who entered the system in a given year and are enrolled in a particular grade. It is assumed that such changes shift the mean for that grade/year cohort. In addition, a random effect is introduced for each school to account for selection effects due to students not being randomly assigned to schools. The model can also be expanded to accommodate changes in the class and school effects over time. For further details, consult Ponisciak and Bryk (2005). The cited reference contains an extended analysis of data from the Chicago Public Schools system, as well as a comparison of the HCM3 results with those of simpler models. A closely related model, utilising latent variable regression, has been proposed by Choi and Seltzer (2005). See also the review by Choi, Goldschmidt and Yamashiro (2005).

As growth curve models are a type of multi-level model (measurements nested within students), it is straightforward to include an extra level, such as the school-level (students are nested in schools), in order to estimate school residuals. These school residuals reflect the relative contribution of a school to their students' status and growth over time and, thus, can be used as value-added scores of schools. Growth models are intuitively appealing and can be considered in education systems that have a large number of observations of student performance (growth curve modelling is not suited to situations where only two measures of student performance are available). The models rely heavily on the quality of the longitudinal data set and issues such as student mobility and grade repetition must be considered (these issues are discussed in more detail in Chapter Six).

Conclusion

This chapter has provided some key examples of value-added models and discussed their statistical properties, illustrating advantages and disadvantages of their use in specific circumstances. Each model has different data requirements and therefore each has different costs associated with its implementation. Different models can also be suited to particular policy and analytic objectives so it is impossible to state, *a priori*, that there is a 'true' or 'best' model across education systems. Instead, analysis needs to be undertaken of how each model can be used to meet the required objectives and meet the desired statistical criteria during the implementation stage of the system of value-added modelling.

Chapter Six further discusses the criteria that should further an understanding of the statistical operating characteristics of different value-added models so that policy makers and administrators can make informed choices in their selection of a model when implementing a system of value-added modelling.

Chapter Six

Model Choice: Statistical and Methodological Issues

The objective of this Chapter is to assist administrators and policy makers in their decision-making regarding the appropriate value-added model to be used in their education system. The decision to employ value-added modelling and, if so, which model in particular, involves many factors, both technical and non-technical. Some key design issues were touched upon in Chapters Four and Five. The focus of this chapter is on statistical and methodological considerations which are important because their explication reveals both the strengths and limitations of the different models in various contexts. Even judged by purely technical criteria, there are few, if any, cases where there is a single 'best model' that can be implemented in every situation. Although technical analyses are rarely definitive, they do contribute to informed decision-making. Moreover, if a value-added model is implemented, then an appreciation of the strengths and weaknesses of the model reduces the risk of improper interpretations and inappropriate use of the estimated school value-added scores.

There are three main statistical issues to be considered. First is the variance of the estimates, including their inter-temporal stability, which can be a particularly complex problem because of the difficulty in disentangling true changes in school performance from various sources of noise. The second issue is bias and robustness to departures from underlying assumptions. Finally, there is the question of the degree of similarity between the value-added estimates produced by the different models. Part III of this report includes a discussion of how such criteria can be practically applied in choosing the most appropriate model in the pilot stage of the implementation process. The material in this report should enable policy makers to utilise the appropriate estimation and garner the confidence of stakeholders in the use of the value-added estimation.

Before proceeding with the main task of the chapter, it is worthwhile to recollect the reason we are grappling with this set of complex issues. From a policy point of view, the capacity to identify both unusually effective and ineffective schools is extremely important. Such data-based indicators can be used in conjunction with other indicators for various purposes, including evaluation, improvement or the provision of information to the public. It is intuitively plausible that it is possible to employ longitudinal test score data (in the aggregate) to make credible judgments about school quality. However, it is quite challenging to build a proper evaluation system.

The application of a value-added model to a particular data set is intended to yield estimates of the contributions made by schools to student progress. The objective is to try to isolate the contribution of the school itself (its personnel, policies and resources) to student learning. In other words, the use of such models is intended to emulate (to the greatest extent possible) the situation of a randomised experiment. This is challenging and the statistical criteria to be discussed serve as the basis for deciding how close to achieving this goal one can come with a particular model in a specific setting. The preferred model will vary between education systems because of differences in objectives, the samples and contextual data used, and the nature of student assessments. From a practical point of view, model choice should not be made without extensive pilot testing, analysis and consultation with various stakeholders. These considerations are discussed further in Part III.

Statistical Criterion: Variance and inter-temporal stability

Typically, the application of a value-added model produces a set of estimated school effects, along with estimates of the variances of those estimates. The (estimated) variance of a school effect is a measure of the uncertainty that is attached to that estimate. Generally speaking, the amount of variance is largely determined by the particular value-added model used and the amount of data available, especially the number of observations that can be obtained from the school. Variance estimates are important, not least because they provide a counterweight to the natural inclination to over-interpret small differences between school effects. They can also be used to construct confidence intervals around the estimated school effects.

Obviously, one would prefer that the variances to be as small as possible, leading to short confidence intervals. When the confidence intervals are small in comparison with the spread among the estimated school performance measures then ‘extreme’ schools can be easily identified. That is, schools with true effects that are substantially higher (or lower) than average, will typically be associated with estimates that are relatively accurate and judged to be statistically significantly different from the average. Accordingly, substantial effort is expended in trying to reduce the level of the variances of the school performance estimates. This usually involves obtaining more relevant data (*e.g.* longer test score sequences or test data in multiple subjects) as well as selecting a model that makes more efficient use of the data at hand.

A key element in choosing an appropriate value-added model is the stability of results over time. If schools’ value-added scores fluctuate substantially and, more importantly, in an apparent random manner, then it is difficult to be confident that accurate estimates of the contribution of a school to growth in student performance are being obtained. A reduction in confidence might have serious repercussions for various stakeholders in the education system, particularly those that might feel the brunt of a punitive school accountability system. Stability of school results should therefore be analysed in the development of value-added modelling and in the regular monitoring of the system. However, given that some changes in schools' value-added scores are expected and desired over time, there are difficulties in determining if instability is due to real changes in school performance or just chance fluctuations.

Year-on-year correlations of schools' value-added estimates depend on school size, the type of model used, the number of contextual variables included, the number of years between prior attainment and outcomes and the coverage of the comparison (all schools in the country or some subset). When school effects are calculated annually, it is not unusual to find that many fluctuate rather widely. Kane and Staiger (2002) observed this phenomenon in North Carolina. Some schools will appear to be unusual on the basis of changes in the data that are used in the value-added model, but for some schools it is hard to say whether a rise or fall in value-added looks 'genuine'. More detailed value-added data (*e.g.* from models for subjects or subgroups within a school) can be used to establish whether the changes are plausible.

As an example, analysis was undertaken of English data of the stability of schools' value-added and contextualised value-added scores compared with the stability of schools' raw results (Ray, 2007). Table 6.1 shows the average absolute change in each of the measures and the standard deviation of these changes. These statistics all are presented in the same units: Key Stage 4 points. Raw results increased between 2005 and 2006, whereas value-added and contextualised value-added scores changed little on average because they are relative measures. Importantly, the standard deviations of these changes are of a similar size. The results here show that although value-added and contextualised value-added are more variable than raw scores in relative terms (*e.g.* as measured by correlations between 2005 and 2006), stability isn't necessarily lower for value-added in absolute terms. In fact, stability in this case is slightly higher for both value-added and contextualised value-added scores than for raw results, with the value-added estimation producing the most stable measure.

Table 6.1. Absolute changes in Contextualised Value-Added (CVA), Value-Added (VA) and raw results (APS): Summary Statistics, Key Stage 4, 2005-2006 (U.K.).

	Mean change	Standard deviation of changes	25th Percentile change	Median change	75th Percentile change
Change in raw APS	5.4	14.9	-4.1	4.9	14.2
Change in VA	-0.1	12.3	-7.9	-0.4	7.3
Change in CVA	-0.3	13.4	-8.1	-0.4	7.5

Source: Ray, A. (2007)

Three factors other than variation in true school performance that affect the stability of value-added scores over time are: changes in the assessment instrument being utilised; changes in the accompanying data (usually the contextual data); and the greater volatility in the results for smaller schools. Test score characteristics can vary from year to year because of insufficient control in development, problems in equating test forms, or even planned changes. Similarly, there can be changes in the number, meaning and quality of the variables used for adjustment. A common remedy that is recommended in this report is to use three-year moving averages for schools' reported value-added scores. This tends to smooth out random fluctuations and should provide more stable measures. The cost of this procedure is that it can make it more difficult to identify true changes in schools' effectiveness. Three-year moving averages can be applied to the results of any value-added model. In particular, recall that the so-called random effects models exhibit an important characteristic; namely, that schools' value-added estimates are 'shrunk' toward the overall average of zero, with the amount of shrinkage inversely related to the relative amount of information available from the school. Thus, estimates for small schools tend to experience a great deal of shrinkage, which contributes to stability but, again, makes it more difficult to identify schools that are significantly different from the average. In a sense, this is a version of the familiar trade-off between Type I and Type II errors. It should be noted, however, that views differ on the appropriateness of using shrunken residuals in the context of a system for providing value-added scores schools (Kreft and De Leeuw, 1998: 52).

Changes in tests might increase or decrease the numbers passing or getting higher grades. This could create instability for school indicators if the models rely on vertical equating to produce growth scores or 'progression' statistics.[16] Even with value-added scores that simply compare schools

[16] An example in England is a simple statistic currently being considered (though not yet in use): the number of pupils in a school who progress two National Curriculum levels or more within a Key Stage.

against each other and produce estimates centred round the average, there would be a problem of instability if changes in the tests favoured some schools more than others. For example, if pass rates rise in a vocational subject that is part of the value-added output measure and this subject is taken mainly by students in particular schools, these schools could end up with higher value-added scores than in the previous year.

A related issue is the robustness of value-added results to different data. For example, suppose that there are two different tests in the same subject, each given over a number of years. If the same value-added model is applied to each data set, how similar are the results? Sass and Harris (2007) carried out such a study using data from Florida in the course of estimating teacher effects and obtained qualitatively different results. This result is not surprising as the tests were built using different frameworks and had different psychometric characteristics. Nonetheless, this finding serves as a reminder that the nature and quality of the test data can and should have a material effect on the output of the analysis. Further work in this direction can be found in Fielding et al. (2003) and Lockwood et al. (2007).

When the value-added model includes contextual data, discontinuities can also lead to instability. For example, in England, a particular Local Authority changing its policy on entitlement to Free School Meals might affect contextualised value-added scores in its schools during that year. In comparing the stability of contextualised value-added scores with raw scores, Thomas et al. (2007) illustrated that correlations based on raw scores are considerably higher. Value-added scores were found to be less stable than raw results because the latter are regularly subject to factors that the value-added scores have factored out. For example, a school's results might be relatively low over time because it usually has an intake with low prior attainment and high levels of deprivation; if the value-added scores measure residual variation in outcomes after taking these factors into account, then there is a greater possibility of instability of scores. However, it should be noted that despite this instability, the value-added results are likely to be a more equitable measure of this school's effectiveness.

Estimates for small schools will be subject to greater sampling variability. Plots of year-on-year differences in school effects against school sample sizes display a characteristic pattern with greater dispersion associated with smaller sample sizes and negligible dispersion associated with larger sample sizes. More generally, since estimated school effects are deviations from an overall average, a school's result also depends on the (adjusted) test score gains in other schools. These too, can vary across years. In most education systems, smaller schools are more common in the primary school sector than in the secondary school sector. Accordingly, the value-added estimates of primary schools are more likely to exhibit greater relative instability, making it more difficult to isolate persistent 'underperformers'. Ray (2007) investigated the number of primary schools that might plausibly be labelled as underperforming on the basis of data accumulated over three years in England. Of the 16 200 primary schools examined, relatively few (424 primary schools) had a value-added estimate more than one standard deviation below the average for three consecutive years. This was not calculated using the contextualised value-added scores but was based on the median method (so without any shrinkage). In order to increase the membership of the group qualifying as underperforming on the basis of having 'low' value-added in each of the three years, the definition of 'low' would have to be made less restrictive (*e.g.* 0.75 standard deviations below average in all three years). Clearly, one could set a criterion based on three-year averages in order to smooth out some of the instability. Other options would be to exclude schools below a certain size along with general warnings to the user about the accuracy of assessing annual changes in value-added scores. Smoothing across years and/or excluding small schools involves a trade-off between having estimated school effects that are less affected by random variation and discovering true changes in school effects at a later period. In discussion within the expert group formed for the development of this report, it was generally considered that schools with annual cohorts of less than 20-30 students were more prone to produce less stable results. However, it was recognised that school size can vary considerably across countries and that practical considerations need to be included in any decisions concerning removing schools from the sampling or analysis. Additional investigation of the stability of schools' value-added results should guide judgments about their inclusion in the sample.

Statistical Criterion: Bias

The utility of a value-added model also depends on the amount of bias in the estimates it produces. Bias is a measure of essential inaccuracy. An estimator is biased if its average value over many replications of a study does not tend towards its 'true' value. Typically, bias is not reduced by simply adding more data of the kind that has already been included in the model. In this respect, bias is fundamentally different from variance because ordinarily, the latter can be reduced by increasing the amount of data available for analysis.

Bias is also more difficult to quantify and to ameliorate than is variance because, in a sense, it lies 'outside' the model. For example, suppose it is common in some districts for students to attend private tutoring sessions in preparation for examinations. If these sessions are well designed, the students will advance academically and, presumably, this will be reflected in their performance on the test. However, if the test scores are used for a value-added analysis, the schools these students attended will appear to be more successful than they really are, resulting in a distorted or 'biased' picture of their relative performance. In this example, the bias enters into the estimation of school effects because of an omitted variable (attending private tutoring) creating a correlation between the school variables and the error term. While the calculation of a variance is based on assuming the model is correct, bias usually arises when the assumptions underlying the model are not satisfied. The assumptions might relate to the nature of the data (such as the omissions of relevant variables), the structure of the model, or both. So, while variance estimates for school effects are generated as a matter of course by most value-added models, estimates of bias are never produced. Approximations to the bias can sometimes be calculated analytically. More often, they are obtained through simulations in which departures from the assumptions are systematically explored.

Estimated school effects will be biased to the extent that there is systematic under- or over-adjustment (see discussion in Chapter 4). The student-level data available for analysis rarely fully represents those aspects of the student's background that are related to academic achievement. For example, the level of parental education is usually considered as a proxy for general socio-economic status. However, a fully specified model for socio-economic status usually would also include parental occupation(s), family income and further inter-generational transfers. Evidently, the level of parental education alone does not do justice to the concept of socio-economic status. It is likely, therefore, that a model incorporating the level of parental education alone results in under-adjustment. That is, the estimated effects of schools with higher socio-economic status populations are biased upward, while the estimated effects of schools with lower socio-economic status populations are biased downward.

Unfortunately, there are myriad ways for bias to confound estimates of school performance. Consider, for example, the situation in which student mobility varies among schools. In schools with highly mobile student populations, substantial school resources might be directed toward transient students, only for it to be the case that they either have left before the test has been administered or have not spent sufficient time in the school to be counted. This difficulty is compounded by the effect of the changes in class composition on the non-transient students. Thus, some amount of the school's efforts is not reflected in the data for the model and could result in a lower estimate of the school's performance. If mobility rates are greater in schools serving more disadvantaged populations and with fewer resources overall, then these schools' estimates could be biased downward. These and other similar scenarios suggest that great care should be exercised in comparing schools with very different mobility patterns.

Measurement error is also a potential source of bias. It is well-known that the theorems of classical regression theory assume that the explanatory variables in the model are measured without error. In the present case, both prior test scores and contextual variables might contain substantial amounts of noise, with the consequence that the estimates of the regression coefficients used for adjustment are biased toward zero. Ladd and Walsh

(2002) show that the use of a single prior test score can lead to value-added estimates with poor operating characteristics. They suggest using twice-lagged test scores (*i.e.* scores from two years earlier) as an instrument for the prior year test scores. There is lack of consensus, however, as to whether the twice-lagged score fully meets the requirements for an instrumental variable.

Statistical Criterion: Mean Squared Error

In practice, assumptions are never completely satisfied and no model is perfectly appropriate. Thus, bias might always be present. The issue is the direction of the bias and its magnitude (both absolutely and in relation to the magnitude of the variance). Bias is often a greater concern than variance, not least because it is a more subtle danger to the utility of the estimates produced by a value-added model. Traditionally, statisticians judge an estimator on the basis of a measure of total error, called the mean squared error (MSE). A convenient expression for the MSE is:

$$\text{MSE} = \text{Variance} + (\text{Bias})^2$$

Thus, some models accept a small amount of bias in order to reduce the variance sufficiently to yield a smaller MSE. This is the strategy of value-added models that model school contributions as random effects. They yield estimated school effects that are shrunk toward the average (introducing bias) but the variances of the estimates are substantially reduced in comparison to those not based on sharing data across schools. The former usually have a lower MSE than the latter. An alternative approach to dealing with adjustment concerns is to employ models in which both students and schools are treated as fixed effects. This eliminates the problem of correlated errors and the like. However, when the numbers of students and schools is large, there are computational issues with large numbers of students and schools that can lead to greater uncertainty with the school value-added estimates that need to be addressed because of the large number of parameters to be estimated. Fixed-effects estimates are consistent but can be quite variable because there is no ‘borrowing of information’ across schools, as is the case with random effects models. There is a trade-off between the bias and variance found in random-effects as opposed to fixed-effects models. Lockwood and McCaffrey (2007) have investigated the statistical properties of random effects models. They demonstrate that, with sufficient data on prior attainment, the bias introduced by correlation between student specific errors and (random) school effects is small enough to be ignored. The models yield estimates that are shrunk towards the mean which induces some bias but also reduced variance. These models are generally preferred due to the resultant lower MSE. However, one should always be aware of the trade-off that is present when using random effect models, since borrowing of information produces estimates that are less variable (*i.e.* more precise) at the cost of a bias.

Missing data

To this point, the report has considered three statistical criteria with the assumption that the database employed in the analysis is complete. In practice, however, that positive circumstance is rarely obtained, in part because value-added models are so greedy for data. They require student records of test performance in one or more subjects for two or more years. Many require student characteristics and other contextual data as well. In most settings, some student records will be incomplete. Of course, most worrisome is the situation in which enrolled students are entirely absent from the database. It is essential, therefore to conduct a number of data quality evaluations before proceeding to the analysis. These issues are treated more fully in Part III.

A substantial amount of missing data, especially test score data, is a cause for concern, with respect to considerations of both variance and bias, especially the latter. Now, it is certainly the case that there are legitimate reasons for test score data to be missing. These include the student leaving the school or area/region or taking another form of the assessment (especially in a system with explicit educational tracks). On the other hand, the student might have been absent on the day of the test with no opportunity for a make-up session. The question then devolves to asking whether the characteristics of the students with such missing data are consistent with the assumptions of the model – a question that is now addressed.

To begin with, consider first the situation in which the value-added model requires test scores from two successive occasions, as well as some student characteristics. If all student records contain the prior score but some are missing the current score, then something must be done to ameliorate the situation. One possibility is to simply delete those records with missing data and carry out the analysis on a set of complete records. Unfortunately, this is likely to produce biased estimates unless the missing data are missing at random. The assumption that missing data are missing completely at random means that the distribution of missing scores is the same as the distribution of observed scores (McCaffrey et al., 2003: p. 82). This assumption is unlikely to hold in school systems. It does not hold, for example, if students with unfavourable characteristics (*i.e.* characteristics that are associated with smaller gains) are more likely to be missing test scores, other things being equal. This would be particularly important for differences in retention rates in both post-compulsory schooling and in different subjects. In that case, schools with higher proportions of such students and, typically, higher proportions of deleted records, will be advantaged in the analysis. This is a form of bias.

More complex models (*e.g.* EVAAS) are able to accommodate both complete and incomplete records. The incomplete records will not introduce bias if the missing data is missing at random. The assumption that missing data is missing at random is a weaker assumption than missing *completely* at random. This means that, conditional on the student characteristics and test scores included in the model, the distribution of the missing scores is assumed to be the same as the distribution of observed scores, *e.g.* within a group of students with the same characteristics and test scores in the model, the missing scores are not systematically different from the non-missing scores. In other words, the process generating the pattern of missing values and the test score outcomes are independent of one another (Rubin, 1976; Little and Rubin, 1987).

Even the weaker missing at random assumption can fail in many ways. It fails, for example, if for a fixed set of student characteristics, weaker students (*i.e.* those with more shallow test score trajectories) are more likely to be absent on the day of testing. They might be absent because they choose to do so or they might even be encouraged to do so. Of course, the missing at random assumption is unlikely to be fully satisfied. The question then is how robust are the estimated school effects to departures from the missing at random assumption. A recent study (McCaffrey et al., 2004) suggests that, under certain conditions for some models, there is a fair degree of robustness. In other words, the bias in the estimates introduced by the missing data is relatively small.

This good news should be interpreted cautiously. First, the robustness is partly due to the extensive data employed by these models. That is, the effect of the departure from the missing-at-random assumption is mitigated by the contributions of the extensive information employed by the model. Second, missing data leads to greater variance in the estimates in comparison with what would be obtained with complete data. So substantial amounts of missing data will reduce the utility of the estimates if, for example, the main goal is to identify schools that are significantly different from the average. If truly less effective schools are more likely to have incomplete databases, then with random effects models, their value-added estimates will experience greater shrinkage and it will be more difficult to distinguish them statistically from the average.

Model choice in value-added analysis

In implementing a value-added model it is advisable, where possible, to compare the characteristics of the school value-added estimates from different model specifications. From a practical point of view, the most important issue is to what extent different value-added models yield generally similar results, *i.e.* whether the choice of model makes any difference empirically. Jakubowski (2007) undertook a comparative study, using data from Poland and Slovenia, to compare different value-added models with respect to the stability of the results. These models have been often used in value-added research and some of them have been implemented operationally. They are not described here as they are treated in the literature on multi-level (hierarchical linear or mixed) models and value-added methods for school assessment (see Goldstein, 1997, 1999; Raudenbush and Bryk, 2002; Snijders and Bosker, 1999).

In both countries the data included individual student scores from exams conducted at the end of primary school and at the end of secondary school. However, the age of the students and subjects that were examined differed. It is important to note that the two countries differ substantially with respect to population size, the organisation of schools, and many social and economic characteristics. The first model was a simple linear regression model, with regression residuals used to calculate schools' value-added. The second model was a linear regression fixed effects model. The third model was a random effects model, with school effects assumed to be independently and normally distributed. The fourth model considered was a random slope (or random coefficient) model where not only the intercepts (school effects) but also the intake score slopes were assumed to be randomly distributed and allowed to vary between schools.

The key finding was that the correlations among different sets of value-added estimates were very high (Jakubowski, 2007). Therefore, from a practical viewpoint it was judged that simpler models were preferable to more complicated ones in conditions where simplicity and accessibility are more important for policy makers than theoretical optimality. The random slope model also provided very similar estimates to the simpler models. Allowing for variation in intake score slopes did not produce significantly different results alone. This does not mean that model choice is an irrelevant question nor does it mean that simpler models should always be preferred and will always produce similar results. Rather, it illustrates that different value-added estimates might not produce substantially different results and that these differences should be tested and analysed. Comparing estimates of different value-added models with respect to some set of pre-determined criteria and objectives should allow a suitable model to be identified. However, in reviewing such comparisons general correlations might not be as important as the consistency of schools' value-added scores at either end of the distribution. In comparing different models, it should be recognised that there are costs and benefits associated with different models and that while more complex models might yield superior statistical properties, such as some robustness against missing data and selection bias, they might also be more costly in terms of transparency and, particularly for some countries with poor centralised data collections, data requirements.

There have been a number of other relevant studies. Gray et al. (1995) calculated value-added scores for a group of secondary schools between 1990 and 1991 and between 1991 and 1992 and found strong correlations of between 0.94 and 0.96. The authors consider that their findings, along with earlier research suggest "that there is a good deal of stability in schools' effectiveness from year-to-year" (p.97). In their more recent study of 63 secondary schools in Lancashire, Thomas, Peng and Gray (2007) found correlations in contextualised value-added for adjacent years in the range 0.80 to 0.89. Comparative analyses have also been conducted by Ponisciak and Bryk (2005), who found modest correlations among methods. In the USA, Tekwe et al. (2004) carried out a study comparing estimated school effects for four models employing data for grades 3, 4 and 5 from a Florida school district with 22 elementary schools. The models ranged from the simple to the complex. Correlations among the model estimates typically exceeded 0.90, except those involving a complex multi-level model where they exceeded 0.70. The authors concluded that there does not appear to be any substantial advantage gained from using more complex models rather than a simple change score model. In response to the analysis of Tekwe et al. (2004), Wright (2004) carried out a simulation employing a factorial design for the different parameters: number of students; gain patterns; and the degree to which missing values might have biased schools' value-added scores. He compared a simple gain score model with two more complex, longitudinal models. Using a MSE criterion, he concluded that the more complex models are to be preferred in view of their lower MSE in those cells of the design that are more likely to represent real-world data. It is also possible that the typical size of the estimated standard errors attached to the estimated school performance measures can be different across models. Therefore, one method might be preferred because a greater number of schools can be accurately distinguished from the average. However the question of whether stability is 'reasonable' depends critically on how the value-added scores are to be used and how notions like 'underperformance' are defined. The results described above are consistent with empirical work on the EVAAS model.

The similarity of schools' value-added scores using different models illustrates that the choices faced by policy makers and administrators are not simply choices between good and bad models. In general, most models will produce similar results if the data used is the same across models, the test data is reliable, and particularly if multiple prior attainment measures are incorporated into the estimation process. It appears, though, that more complex models, given the limitations of the data available, can provide greater accuracy and also appear to be less sensitive to departures from the underlying assumptions. Models can be complex in different ways. One model might introduce complexity by including multiple assessment scores on multiple subjects such as in the EVAAS model. Another model might take into account a variety of additional factors affecting performance scores (Ponisciak and Bryk, 2005). The increased level of complexity in either of these models (or any complex model) is only beneficial if it captures meaningful patterns or sources of noise in the data. The disadvantage lies in the greater level of complexity and the need for more data so that the parameters of the model can be well estimated. This trade-off needs to be analysed in the pilot stage of the implementation of a system of value-added modelling, including an assessment of the extent that additional data is required for more complex modelling.

In the recommendation to the UK Government concerning the implementation of value-added modelling, Fitz-Gibbon (1997: 38) found that "the value-added indicators produced by the simple procedure of comparing students' performance directly with the performance of similar students, regardless of the school attended, and then summing the value-added scores (residual scores) gave indicators that correlated so highly with indicators from more complex models that the simple methods could be recommended". Given the advantages of communicating simpler models to stakeholders, such a finding lends itself to the adoption of more simple value-added estimations. These could then be supported with more complex models both for internal analysis and to monitor the results of the simpler model.

An additional issue that can be analysed is the differences in modelling of different structures of student assessment scores. Fielding, Yang and Goldstein (2003) compared value-added estimates based on a multi-level model for point scores and a multi-level model for ordered categories. The models were applied to a large database of the General Certificate of Education Advanced Level examination in England and Wales. For both kinds of models, the covariates were: student prior achievement; gender; age; school; type of funding and admission policy; and, examination board. It was shown that the correlation coefficients and rank correlations between the institution residual estimates and value-added estimates from each pair of models were larger than 0.96. However, if it is true that an individual school's value-added estimates can differ substantially among models then the choice of the most appropriate value-added model is an important one. Therefore, in comparing the impact of different models, the identification of single schools for which there are significant differences should be undertaken. In addition, it should be emphasised that consistency of findings does not necessarily imply that bias or measurement error do not exist.

Conclusion

A school's estimated contribution to student learning can alter with the specific value-added model employed. Differences in specifications can derive from a number of factors such as the range of test data used (*i.e.* the number of years and the number of subjects), the treatment of missing data and the kinds of adjustments employed. With these differences, each value-added model brings advantages and disadvantages that must be considered in light of the context in which they are used and the nature of the data available. In general, the more complex models have greater data requirements, are more difficult to implement and evaluate, and pose greater challenges in trying to communicate their logic to different stakeholders, including the public at large. A natural question then arises, "Is it worthwhile using more complex models?" With greater complexity come additional costs, particularly if additional data must be collected for the more complex models (which is often the case). The advantages of this increased complexity, such as reduced variance, need to be weighed against the costs. Among policy makers there is an understandable preference for simpler value-added models that are easier (and cheaper) to implement and more amenable to effective communication with stakeholders. However, if simpler models result in more misspecification then the school performance estimates will be biased and costs will be larger in the long-run. These costs and benefits will differ between education systems and can be analysed during the pilot phase of the implementation process to illuminate the extent of the trade-offs.

Given the particular characteristics of each education system, the objectives of the system of value-added modelling and the type of student assessment upon which it is based, it is not possible to identify a single value-added model that is suitable to all education systems. Instead, different models should be analysed for their fit with each system. The discussion of the issues in this chapter that should be analysed to inform decisions of model choice has included:

- The variance in each value-added model should be analysed to evaluate the suitability of particular models. The estimated standard errors attached to the estimated school effects can differ across models. One method might be preferred because smaller standard errors mean that a greater number of schools can be accurately distinguished from the average or classified as reaching some pre-defined target. Analyses comparing value-added models against this criterion might be conducted in the implementation stage. For example, pilot data can be tested to identify the most appropriate model by minimising variance to produce more interpretable results.

- The use of socio-economic contextual data and the roles that different data components play in a value-added analysis as all value-added models involve some sort of adjustment to the sequence of raw test scores attached to each student. Although the need for adjustment flows naturally from the rationale behind value-added modelling, it must be done carefully or it will produce estimates that can be quite misleading. Analyses should be conducted to assess the impact of the inclusion of socio-economic characteristics upon schools' value-added scores and aspects of the overall value-added model (*e.g.* the predictive power of the model and the standard errors associated with school estimates).

- The potential bias in the model needs to be analysed and the potential for how it can be reduced tested during the pilot phase of implementation. While the extent of bias in estimations is not straightforward to analyse, approximations can be made and simulations run to assess potential bias. The potential of missing data can be explored and the inclusion or exclusion of specific variables in the model might highlight specific problems. Comparisons with actual raw test scores further illustrate potential bias in the estimations.

- The assumptions concerning missing data made in the specification of value-added modelling can be compared with the pattern of missing data evident in the sample and the estimates of the effects of missing data can be calculated. Procedures can also be implemented to reduce the frequency of missing data in the implementation of student assessments and other data collections (*e.g.* creating (dis)incentives for (low) high levels of student participation).

- Small sample size is an issue given the greater levels of uncertainty usually surrounding estimating school value-added with small sample sizes and the reduced stability of these schools' value-added scores. Estimates of value-added for small schools can be tested and recommendations made for both the analysis and presentation of school results. In general, participating countries considered cohorts with fewer than 20-30 students produced school value-added estimates that led to problematic interpretation of results.

- Stability of schools' value-added scores and how this is affected by the classification of school performance and the choice of value-added models. Analyses such as those presented in this report can be undertaken to ascertain the degree of stability of school scores and whether it can be minimised. In such analyses, it is important to consider not only the overall level of stability (or lack thereof) but changes in individual school scores. Analysis can then be conducted of the causes of such instability and to identify whether particular schools are more susceptible to instability their school results.

Given the need for straightforward value-added models that can be effectively communicated to stakeholders, the analysis outlined above should compare the results with relatively more simple and more complex value-added models and an assessment made of the differences. If there are few significant differences between these models then it might be appropriate to use the simpler value-added models to present results to the public and to some other stakeholders. This would facilitate effective communication and ease the use of value-added information to advance specific policy purposes. The presentation of the results of simpler models would then need to be supported by extensive on-going internal analysis that compared these results with those obtained from more complex value-added models. Comparative analysis would ensure that the simpler models produced estimates that were accurate and did not unfairly affect specific schools or school groups. As the model is developed over time, such analysis would need to be continually undertaken. This would be particularly important in instances where data availability and requirements change over time.

If such a decision is made to employ two levels of modelling then it requires a set of actions to ameliorate any discrepancy in the results between the simpler and more complex models. As shown in this Chapter, such discrepancies might not necessarily be common to a large number of schools. Moreover, during the implementation phase, the choice of the specific model that is used and presented to stakeholders should be based upon analysis that illustrates that such discrepancies have been minimised. But it is important that there is a pre-determined set of criteria for assessing the validity of differing results, particularly if value-added results are to be used for school accountability purposes. Such criteria should identify the source of the difference in a school's results and then enable an identification of the more accurate measure of a school's performance. If value-added information is used for school improvement purposes, then such procedures can provide further valuable information. In some instances, they could be incorporated into the system of school improvement. A discrepancy in a school's results might trigger an expanded data collection that helps to identify the source of the discrepancy. Regardless of the actions for individual schools, the analysis of discrepancies in results between more simple and more complex value-added models should then feed into the ongoing development of the system of value-added modelling. This should help to reduce the number and size of discrepancies between simple and complex models over time. It might be prudent to initiate value-added analyses through simpler models, with more complex models being reserved for research and introduced perhaps at a later stage when all the technical issues have been satisfactorily resolved.

Part III

Implementation of a System of Value-Added Modelling

Introduction

Regardless of the nature of the statistical and methodological underpinnings of the value-added modelling, the impact upon policies, practices and outcomes can be negligible or even negative if an effective implementation is not undertaken. This belief was evident in a number of countries involved in the development of this project and led to more detailed analysis of the methods for implementing a system of value-added modelling. Part III of this report builds on the discussion presented in Part I and Part II to provide a guide for the implementation of a system of value-added modelling in education systems. Such a guide is not a definitive list, nor will each aspect be applicable to all education systems. Rather, it builds on the knowledge gained both in various education systems and from the expert group who have experience in implementing systems of value-added modelling in various education systems.

A number of issues need to be addressed in order to implement a system of value-added modelling effectively. These follow the issues already raised in this report and are presented here under the following implementation themes: setting policy objectives and school performance measures; choosing an appropriate value-added model; development of an effective database; running an effective pilot programme; monitoring the results of value-added analyses; developing a communication and stakeholder engagement strategy and commensurate training programmes; and, presentation and use of value-added information. To complement Part III, a list is provided at the end of Part I that it is hoped will provide practitioners with a short checklist of the main issues in the implementation of a system of value-added modelling.

Chapter Seven

Setting Policy Objectives and Choosing the Appropriate Value-Added Model

Value-added analysis can be used to advance a number of policy and programme objectives. These were discussed in detail in Part I of this report and need not be repeated in detail here. The implementation of a system of value-added modelling to further specific policy objectives requires a number of key decisions to be made and steps to be enacted. These derive from the three main policy objectives discussed in Part I of this report: school improvement; school accountability; and school choice.

School improvement efforts can be greatly assisted with the use of value-added information, particularly in systems that enable schools to use value-added results to develop and monitor school improvement initiatives. The key features affecting implementation efforts centre on the use of value-added information to support and advance systems of data-based decision-making that can empower schools and other decision-makers to analyse variation in school and student performance. This can inform decisions to better allocate resources, identify areas of best-practice and those in need of improvement to develop an ongoing school-improvement system.

School accountability can be informed through the use of schools' value-added scores to hold schools to account for their performance. Accountability can take numerous forms with links to school funding, specific interventions for low-performing schools, or consequences for the remuneration of administrators, school principals and teachers. More implicit accountability systems can also be developed to increase the focus on schools' results without explicit links to resources, autonomy or remuneration. The first step in implementing a system of value-added modelling for school accountability purposes is to consider the current school accountability arrangements and how changes might affect stakeholders. A key component of the successful engagement of stakeholders is to provide clarity on the objectives and operations of a system of value-added modelling. In regard to school accountability arrangements, key questions arise about the use of rewards and sanctions and the level at which they will be applied. This report has focused exclusively upon school-level value-added measures but value-added models have also been used to advance individual teacher accountability (Braun 2005b; McCaffrey et al., 2004, McCaffrey et al., 2003) and it is important to explicitly make such a delineation given the potential impact upon key stakeholders and the development of specific value-added estimations.

School choice can assist the development of school education systems by allowing parents and families to choose the school that best suits their needs. Through this, schools are encouraged to develop the education they offer to meet the needs of parents and families. The benefits of a system that facilitates school choice rests on the assumption that parents and families have the required information to distinguish schools. Value-added measures are invaluable as they provide greatly improved measures of school performance compared with, for example, raw test scores. These improved measures should enable better decision-making and therefore improve the matching of schools with parents and families' needs. In turn, this should provide schools with better information as they seek to develop the education they offer to attract students and families to their school. If the advancement of school choice is a key objective for the implementation of a system of value-added modelling then it can be beneficial to conduct a review of the extent to which parents and families are actually able to choose between schools when making their education decisions. In some countries, legislative and regulatory requirements restrict school choice while in others, institutional, geographical and resource constraints restrict the choices families can make (OECD, 2006). In such circumstances, additional information might have a reduced benefit in lifting school choice. A review of these circumstances should provide important context for decisions concerning the use of value-added information.

A key question in the implementation of a system of value-added modelling is whether schools' value-added results will be published and in what form. Clearly, the publication of results is required to expand school choice in an education system. Part I of this report provides numerous examples of how school results can be presented to suit particular aims. It is beneficial to detail the presentation of results early in the implementation process. It can assist in both the development of specific value-added models, the use of value-added results to categorise school performance, and it can affect a number of facets of the development of school improvement and accountability systems. The decision of how to present school value-added results should be tested and then further developed in the pilot stage of the implementation process and it is crucial in effectively engaging key stakeholders in the process (NASBE, 2005).

While declaring the objectives could be considered a pre-requisite for the development of any policy or programme under a system of good governance, explicitly stating these objectives shapes decisions such as identifying the appropriate value-added model, the form of publication of schools' value-added scores, and a communication strategy that gains the support of key stakeholders. If value-added information is to be used in evaluating school performance and shaping school improvement initiatives, it is important to consider how that information will be incorporated into the existing system of school evaluation to increase its effectiveness. In most OECD member countries, the current system of school evaluation utilises school inspectorates (or a similar institution) and/or school self-evaluations (OECD, 2007a). As described in Part I, a number of methods can be utilised to increase both the efficiency and effectiveness of school evaluations. For example, a system can be developed whereby value-added results trigger specific school evaluations. This can increase efficiency through targeting lower-performing schools or groups of students considered to be at-risk, and increase information flows when mechanisms are established to allow high-performing schools to share their best practice.

Determining the variable upon which to measure value-added

After explicitly defining the objectives of the development of value-added modelling it is necessary to specify the measure(s) upon which schools' performance will be gauged. This requires identifying the appropriate student assessment instruments and the dependent variable(s) to be used in the value-added modelling. The construction of this variable should be directly related to the objectives of developing the system of value-added modelling. For example, if the objective is for students' to attain minimum literacy and numeracy levels then the assessment instruments and the appropriate variable can be identified to measure schools' value-added performance in lifting students above these levels.

A value-added model could focus on various aspects of schools' performance. Decisions regarding this focus affect the type of model that can be employed and also the policy and programme actions stemming from the use of the value-added model. Decisions regarding the subject areas and grades or year levels in which student assessments will be used for value-added modelling are particularly important as they delineate the aspects of a school upon which performance is measured. These decisions therefore define what is meant by a school when estimating schools' value-added scores to promote school accountability, school choice or school improvement. If students are assessed in only mathematics and the language of instruction then the definition of a school is those aspects of a school that contribute to performance in those measures in the grade or year level in which the assessment takes place and, depending on the structure of the school system, the grades or year levels leading up to the assessment. It could be argued that judging school performance on assessments of students' numeracy at a specific grade places a disproportionate emphasis on a school's mathematics teachers in that grade. This might be an intentional policy decision but these issues need to be considered and explicitly addressed. The breadth of testing of students varies considerably across OECD member countries. In general, at lower levels of education, only key learning areas such as numeracy and literacy are tested. In the later years of secondary education, a greater number of subjects are often tested but these are sometimes not tested using standardised assessment instruments. Such difficulties can be overcome for modelling purposes but they should be recognised in the development of a system of value-added modelling. In systems that use value-added results for internal school improvement purposes, multiple value-added measures can significantly add to the explanatory power of the analysis of school performance and greatly assist decision-making. Such decision-making would benefit from a range of data specifying performance in different subject areas that is supported with student-level contextual data. Conversely, systems focused on improving school accountability or school choice might require a focus on a single performance measure.

Categorical and continuous measures

Given the choice of assessments in particular subject areas, a further issue that needs to be addressed is how performance will be measured or categorised. The measurement of student performance can be a continuous measure that identifies student performance across a range of scores (notwithstanding the ceiling effects of student assessment instruments) or it could be a categorical or dichotomous measure. Student assessment instruments can also be devised to better delineate students achieving pre-determined levels. It might be preferred to specify particular levels of performance that categorise students according to measures of, for example, low, medium and high ability. Schools' value-added scores would therefore measure the contribution of the school to these pre-determined categories. Dichotomous measures can be appealing if the objective is to measure the performance of schools in lifting students to or above a single ability or performance measure. Common examples would include minimum literacy and numeracy skills at given grade or year levels. Such measures can be the focus of specific student assessments or can be extrapolated from continuous measures. This provides schools with the incentive to focus on this aspect of performance which can be viewed as a positive consequence. It also has a potentially negative consequence if such a focus comes at the expense of students at other levels of performance (Fitz-Gibbon and Tymms, 2002). The decision to focus on specific measures should be aligned with the policy objectives of the development of value-added modelling and feed into programme development.

A focus on specific performance levels provides incentives for school principals and teachers to reach these particular levels and can also create a focus on particular students or subjects. For systems that do not want to emphasise a specific measure, a continuous variable that measures student performance and, through this, school value-added, might be the most appropriate. This would enable schools and other administrators to analyse a larger distribution of data to develop and monitor school performance and specific programmes and policies. It would also provide a more even distribution of incentives within schools rather than a focus on a specific skill level. In some instances, continuous measures can be developed that can then be grouped into pre-defined categories or minimum standards. This can be advantageous if the appropriate student assessment instruments can be developed.

Decisions concerning the development of student assessments for value-added modelling can be strongly influenced by the existing structure of student assessments, which might already be well established in an education system. It should be decided whether additional assessments should be developed to complement the existing framework. An additional complexity might lie in ensuring that the new assessments do not disrupt the objectives of the education system. Existing assessments can often be a determinant of student progression through their education and any additional assessments could disrupt education leading to these assessments. Instead, if new forms of assessment are developed then both forms of assessments should complement each other.

The structure of the dependent variable impacts upon decisions concerning model choice as it can determine the type of models from which to choose. If the dependent variable in the value-added model is a dichotomous variable (or will be reconstructed in such a manner for particular applications) then this needs to be identified at an early stage given the ramifications for model choice. Dichotomous dependent variables have different modelling requirements than continuous dependent variables. Such models were more fully discussed in Part II of this report.

Identifying the appropriate value-added model to best address policy objectives

Given the policy objectives driving the development of a system of value-added modelling, it is possible to establish the key stages in a process through which an appropriate value-added model is chosen for the main implementation. This process begins with identifying the main factors that will affect the choice of the model, such as how the model will be applied and the results interpreted to achieve policy objectives and, connected to this, the structure of the student performance measure (dependent variable) upon which value-added will be estimated. Each value-added model has advantages and disadvantages that must be considered in the context of the overall objectives and use of value-added information. The second stage of the process is identifying the statistical and methodological criteria to choose the most appropriate value-added model. This will be based upon the results of estimations of different value-added models upon either the pilot data or pre-existing data from student assessments already in place within education systems.

The specification of policy and analytic objectives establishes a framework with which to assess the validity of different value-added models. The use of value-added modelling to advance school accountability, school improvement, or school choice places specific requirements on the value-added modelling and the requirements of addressing various statistical and methodological issues. A key distinction is whether the modelling is to be used internally or also to be published. This will guide decisions such as how to address the instability of school scores and measurement error with smaller schools, and also provide answers to larger questions about the additional analysis that could be conducted with more complex modelling to analyse specific schools, students, or education programmes. It is also important to realise that when choosing between distinct value-added models, analysis should be conducted of the potential impact for schools of using these models. For example, if low-performing schools are to be categorised as such, then the differences in such categorisation (particularly over years, if possible, with the available data) with different models should be analysed to identify the different impacts upon schools and how such differences would be addressed in the live implementation.

Numerous statistical and methodological criteria should be identified. Part II of this report identified a number of these issues and it is possible to decide on the preferred model against such criteria. The over-arching policy objectives should be kept in mind when choosing such criteria. For example, greater emphasis might be placed upon being able to significantly separate the performance of different schools or to minimise the instability of school scores across years. Decisions could be made to exclude particular schools (*e.g.* small schools or those serving students with special learning needs) from the main analysis to achieve the 'best fit' for the chosen model. Such decisions would benefit from clearly specified policy objectives and how value-added information would be used as a basis for action (*e.g.* in particular education programmes).

The key criteria emphasised in Part II of this report can be established during the implementation phased which would then be tested during the pilot phase so that a clear decision can be made on the most appropriate value-added model. Such criteria could focus upon:

- *The amount of variance and bias in distinct models.* Different models will produce differences in the estimated standard errors attached to each school's value-added score. This has consequences for being able to make statistically significant distinctions between schools' performance which might be a key policy objective of the modelling. This will be of particular importance if schools' value-added scores are to be published and if scores will be categorised based upon statistically significant differences. One model might be preferred because smaller standard errors mean that a greater number of schools can be accurately distinguished from the average or classified as reaching some predefined target.

- *The use of socio-economic contextual data in different value-added models.* Some models include few contextual characteristics while some contextualised value-added models include a large number of socio-economic measures. The number and frequency of current and prior attainment measures can affect the explanatory power of including such characteristics and this can be tested in the pilot phase of the implementation process. The impact upon incentives should also be considered as well as how such model adjustments affect the actions stemming from schools' value-added scores. The inclusion of socio-economic characteristics can also affect the standard errors associated with school estimates and how the model stands against the underlying assumptions.

- *Missing data and how it is accounted for in the modelling.* As discussed in Part II, some value-added models are better equipped to account for missing data. In other models, an impact will be evident upon the predictive power of the model and the level of variance and bias in schools' value-added scores. A decision will need to be made about the exclusion of some variables but procedures can also be developed in the implementation phase to reduce the pattern of missing data by creating (dis)incentives for (low) high student participation.

- *How the results of smaller schools change in different models.* Small sample sizes in smaller schools often produce less precise and reliable measures that are also less stable over subsequent years. Models that 'shrink' smaller schools' value-added results to the mean can produce more useable results but there are clear problems with this level of intervention in the data. In general, participating countries considered cohorts with fewer than 20-30 students produced value-added estimates that led to problematic interpretation of results. This problem should be analysed during the pilot phase of implementation.

- *Changes in schools' value-added scores over time.* The stability of school scores over time might also be analysed and the impact upon particular schools measured. This would be related to the size of the variance and potential bias in the model. If the stability of school scores is considered to be too low, then standards can be imposed to minimise any negative impacts. For example, if instability is concentrated in particular schools then these could be removed from

> the main analysis. For such schools, additional estimations can be applied and, depending upon the main policy objectives, separate accountability or improvement initiatives introduced. Standards could also be applied to remove schools with an abnormally large change over a number of years. This could be applied as a proportion of the change in all or similar schools' scores. The use of a three-year moving average in the measurement of value-added would smooth changes over time. In addition, they might provide the opportunity to conduct further analysis of schools with abnormal changes in a single year value-added score.

These issues can all be assessed while the models are under consideration in the pilot stage. Such analysis also provides policy makers an opportunity to analyse the impact of applying different standards to the use of data such as the inclusion of missing data and schools with smaller sample sizes. Differences in such standards would have different impacts under different value-added models. To increase transparency, such criteria can be weighted to guide later decision-making. Decisions on these issues will not be clear choices as some models might be superior against some criteria but inferior against others. The decisions will require judgements to be made of the performance of each model under the chosen criteria. When difficulties arise, it is worth considering analysing differences in value-added scores between the two models and estimating the impact of such differences under the prescribed policy objectives (*e.g.* schools being identified as low-performing).

After specifying the key characteristics of what is required from value-added modelling, analysis can be undertaken on either existing student assessment data or the data obtained in the pilot stage of the development of a value-added modelling system. This analysis can assess the appropriateness of distinct value-added models to meet the objectives of the system and to address pre-determined statistical and methodological criteria. The results of this analysis should present the advantages and disadvantages of distinct value-added models and, from these, recommend a preferred model. Most importantly, it should identify the implications of the model choice upon the use and application of schools' value-added scores and the prescribed policy and programme objectives. This will highlight the impact for particular types of schools but should also identify the extent to which distinct models could meet the prescribed policy objectives. To achieve these ends, it is important when assessing the appropriateness of distinct value-added models to analyse not just the overall model (*e.g.* goodness of fit) but also the impact of different models upon individual schools.

Development of an effective database

This section discusses the key aspects of developing a database that supports the efficient development and administration of a system of value-added modelling. Given the discussion of measurement error and model misspecification in Part II of this report, the quality of the data used should be considered and, if possible, improved upon in the developmental stage. This requirement affects the key issue of the scope of the dataset that can provide opportunities to build more comprehensive data systems to analyse value-added and broader aspects of the school education system. However, the broadening of the database should not be accompanied by a reduction in the quality of the data. The discussion presented below of the development of an integrated database to assist in decision-making and policy development should be considered in the context of the current data collected in each education system and the costs of developing an effective database given the importance of data quality.

As value-added estimates can be a powerful force for change, it is critical that the database be constructed and maintained with utmost care to prevent errors or omissions from contaminating the results. The quality of the data used in value-added modelling has a clear impact upon the confidence with which interpretations of school performance can be made. The development of data systems varies across countries for a variety of reasons. The development of an effective student-level database has been crucial to the effectiveness of the system of value-added modelling in England. In 1997, the development of better student-level data was highlighted and a unique student identifier that would help data to be matched throughout the school system was introduced in 1999. Another key development was the move to an annual *student-level* census of schools in 2002, which collected the background characteristic data that schools recorded for administrative purposes. To increase the breadth and effectiveness of analysis, this data then needed to be incorporated into a single data system that allowed users to analyse schools' value-added results in conjunction with various contextual and school-level data. There might also be efficiency gains of consolidating data sources into a single comprehensive data system.

The first step in developing the required high-quality database is to identify the data that will be used for value-added modelling. For policy makers who wish to develop a system of value-added modelling to facilitate decision-making for school improvement and policy development purposes, it can be beneficial to develop a comprehensive database that extends beyond the minimum data requirements for value-added modelling. A key decision to be made at this step is whether the benefits of a more comprehensive data system outweigh the development and maintenance costs. Such a system could include complementary data from various sources, but for those systems that do not wish to supplement their basic student assessment data, then resources can be concentrated in ensuring a high-quality database is developed and maintained to produce high-quality value-added estimates. If a more comprehensive data system is required, then the question of what information should be collected needs to then be addressed. Four main types of data could be collected for inclusion in value-added analysis and to further policy development. These could be used for a variety of school improvement purposes that are more fully discussed in Part I of this report. The four main types of data are:

Student assessment data that encompasses all student assessment scores to be used for value-added modelling. This would include all prior and current student assessment scores cross-referenced using student identifiers. It would also include any composite measures of combined assessment scores (*e.g.* an average of scores across different subjects) and specific measures considered to be of importance for policy purposes (*e.g.* minimum literacy requirements). Additional indicators or variables that could be calculated are performance targets, or school or student scores that might be used to trigger specific actions. As the database is developed over time, it might be useful to track students to identify additional education and labour market outcomes. This is necessary for analysis of schools' value-added measured against such outcomes as the percentage of students progressing to post-secondary education and to analyse school data against other socio-economic outcomes.

Student-level contextual information that includes all individual (*e.g.* students' age), family and other characteristics that are considered necessary for analysis in the (contextualised) value-added model. The choice of such characteristics has been discussed in Chapter Six, and should be driven by two objectives. The first objective is the use of these contextual characteristics in value-added modelling, particularly more extensive contextualised value-added modelling. These can be important characteristics with which to capture the effect of factors outside the control of the school that affect student progress. However, for some value-added models they are not required as they add little to the predictive power of the model and have little impact upon school results. The second objective is the use of characteristics to investigate value-added in particular schools or specific groups of students. For example, particular interest might exist in the value-added for students from poorer socio-economic backgrounds or particular immigrant groups. The analysis of these sub-groups requires the commensurate student-level contextual data.

To measure the contribution of schools and other factors to student progress requires a database that can identify and accurately delineate student-level data. This requires students being identified normally with some form of an identification number or code that is distinguishable on student assessment data and on all other student-level contextual information. Student identification numbers are required to identify students and to track students as they both enter and leave schools. The issue of student mobility is an issue that must be addressed in value-added modelling both for the missing values it can create in the dataset and the problems of attributing growth in student performance to different schools. To enable the accurate analysis of this issue, an information system must exist that properly tracks student mobility between schools, particularly mobility between the pre-determined student assessment periods that feed into value-added analyses. In some countries this is a more difficult task than others. Some countries, such as Norway and Denmark, utilise existing administrative information systems that systematically assign all students an identification number and enable effective tracking of students. The establishment of such systems can be costly and resource intensive. Additional complexities can be encountered if multiple jurisdictions and institutions are involved. In Poland, tracking students was first attempted through data held by the National Examination Boards. However, the required student-level data only existed in the data collected by Regional Boards. A process was then undertaken to match the data held by the various Regional Boards that was hampered by the lack of student identification numbers (only data for the name, gender and date of birth were held). This was considered to be a costly and resource-intensive process but one that was a necessary pre-requisite for the development of a system of value-added modelling. As such, it also led to changes in the management of data systems such as the introduction of student identification numbers.

School-level information considered necessary in a number of systems includes data on school sector and school type and data indicating if the school is located in specific regions. School size (as measured by the number of students) should be able to be identified given the instability often associated with value-added scores for smaller schools. It might also be considered beneficial to collect information that identifies key programme and policy information that facilitates analysis of their relationships to value-added scores. This data can provide a key ingredient in overall quality control in the school education system and will facilitate the development and monitoring of specific programmes and policies aimed at increasing school improvement. This could be done at the school, district or regional level depending upon the nature of the programme. For example, Goldhaber and Brewer (2000) analysed the relationship between teacher certification and teacher-level value-added scores. In England, analysis has been conducted internally on specific programmes such as the specialist schools programmes that provide additional funding and an expanded curriculum in particular areas. The design of the value-added model and the information supporting the model enabled a performance measure of the impact of these programmes to be developed.

School-level contextual information can also be collected that has the aim, similar to the focus of student-level contextual information, of includeing characteristics in a contextualised value-added model that ‘level the playing field’ for comparative analysis of schools’ values-added scores.

School-level information might be used instead of student-level information if the data cannot be collected at the student level or if it is easier to be collected at the school level. This might be the case if school-level administrative data already exists that sufficiently measures the required contextual factors. However, care should be taken to ensure the reliability of such data. In some systems, various socio-economic measures are used in administrative data as part of programmes to provide additional resources to disadvantaged schools. These measures might not necessarily adequately measure the factors that need to be captured in order to isolate school effects in value-added models, particularly if they provide crude measures of socio-economic status. Less accurate measures can also be less effective in providing data that facilitates analysis of particular groups of students and school-level measures can negate the potential for analysis of within-school differences. An additional problem with administrative data is the potential for bias. In some education systems, school-level administrative data on the socio-economic status or learning disadvantage in the school is provided by school principals or administrators. If school principals provide these measures with the knowledge that it might affect either the school's value-added score or the level of resources the school receives then the provision of such data should be considered as susceptible to bias. These problems have been evident in a number of education systems and can create difficulties in the interpretation of contextualised value-added models.

School evaluation information and reports that provide further evaluative information concerning school performance can aid the interpretation of value-added scores, the use of value-added models for programme development, and lead to developments that improve the system of school evaluations. As discussed throughout this report, value-added scores do not provide a complete picture of schools' performance. Greater confidence can be placed in interpretations of, and actions stemming from, value-added scores if additional evaluative information is obtained. If it is part of a comprehensive data system, linking schools' value-added information with evaluative information from school inspectorates and school-self evaluations can provide a valuable resource for the development of school improvement initiatives. Additional school-level information would enable more detailed analysis of high- and low-performing schools. In addition, there are efficiency gains in enabling the institutions and actors who evaluate schools and school programmes to analyse schools' value-added information. This facilitates the targeting of school evaluations to pertinent areas and enables an output-based, rather than input-based, school evaluation. This also aids the functioning of school inspectorates as it permits the analysis of the recommendations and judgements made by inspectors and how these relate to schools' value-added scores. This can greatly facilitate the quality control monitoring of, and within, school inspectorates.

The linking of schools' value-added information to other evaluative information can also be considered in light of the use of value-added data to improve school choice. The publication of schools' value-added scores can be beneficial to parents and families as it informs their decisions about which school best suits their needs. Given the variety of needs and requirements placed on schools by parents and families, it might be considered appropriate to provide further evaluative information to better facilitate school choice. This could be presented in a format similar to the School Performance Tables available in England or the school evaluative information now publicly available in the Flemish Community of Belgium.

While the creation of a flexible database and data collection methods creates the potential to greatly facilitate the use of value-added models for ongoing policy development, it is beneficial if the required student-level data is identified in the initial development period. An important step in identifying the required data is to ensure an agreed set of core definitions of all variables that would be collected. In some countries privacy laws might restrict the use of contextual data. In Poland, privacy laws have prevented the extensive use of socio-economic status in their value-added modelling and in Slovenia, signed agreements are required from parents before socio-economic data can be obtained from students. Central to the issue of identifying data requirements is the articulation of the objectives and the specific actions linked to value-added modelling. This facilitates the identification of the key characteristics and information that needs to be collected and to identify in advance whether this data is to be used internally, will extend to use by schools and other education stakeholders, or if it is to be used by the general public. Once these issues have been addressed, and a broad strategy agreed upon to develop a data system, then it is possible to review the existing data systems and the capabilities of the resources invested in them. This would include a consideration of practical issues such as the software currently used and quality control issues such as ensuring common standards in data collections. It is possible to then determine if further data is necessary, if new data collections methods need to be implemented, and if new information systems infrastructure needs to be established.

Pilot programme for the system of value-added modelling

The objective of the pilot programme is to assess and further develop different aspects of the system of value-added modelling. This includes: operation and implementation issues; decisions concerning student assessments and the choice of the specific value-added model; the development of stakeholder communication and engagement strategies; and, to assess how schools' value-added scores and other information are interpreted and best utilised to meet stated policy objectives. These issues have been discussed throughout this report and need to be assessed during the pilot programme. The pilot programme should therefore not be viewed as solely a test of the specific value-added model to be used in an education system. The discussion of these issues in this report has been informed by the results of pilot programmes implemented in participating countries.

A pilot programme is often conducted on a sub-set of schools and can be considered to be a trial run before the live implementation. It should be treated in the same manner as the live implementation of a system of value-added modelling to create a realistic and valid assessment. The method with which a sub-set of schools is selected or asked to join the pilot programme will vary between countries but it is important that a sample of schools participate that can properly inform a live implementation. This requires obtaining a sample of schools that is representative of the broader school population and can be effectively engaged in the process of assessing the implementation of value-added modelling. To encourage effective engagement in real-life pilot studies, some education systems have emphasised that the pilot study was not to be used as a tool of school accountability. In selecting a sub-set of schools it is worth considering that schools might feel less inclined to participate in a study that subjects them to additional accountability and performance measurement.

In cases where a representative sample cannot feasibly be obtained, it can be beneficial to ensure that a sufficient number of schools from different sectors and regions are included in the pilot programme as this will enable a better analysis of whether there are specific factors in, for example, a particular region, that need to be accounted for in the live implementation. Specific factors might be found that require a change to a specific variable in the value-added model (*e.g.* variables measuring school sector or the proportion of students with special learning needs or from disadvantaged backgrounds) but there are also a number of implementation issues that will need to be considered. For example, stakeholder communication and engagement strategies might need to be modified for schools in regional or rural areas.

All aspects relating to the assessment of students, the use of information systems to compile datasets, and running value-added estimations should be conducted as though it is the live implementation. If the structure of student assessments already exists, then it would be pertinent to utilise such data to assess the reliability of the information systems being used and the modelling of the value-added estimations. This would provide an assessment of any capacity constraints in the information system to be utilised. It would also allow a more complete judgement to be made on the appropriateness of the choice of the value-added model.

As discussed above, it is not appropriate to make an *a priori* decision on which specific model to implement in an education system. The pilot stage should be considered as the time to assess the most appropriate value-added model to be used in the live implementation. Such an assessment should be made against a set of pre-determined criteria as discussed above. For a pilot programme to be optimally useful, a number of years of data will be needed to ascertain how the stability of schools' scores differ between different models. In some education systems, the structure of student assessments existed well before the implementation of a system of value-added modelling. Assessment data from multiple years could therefore be used to inform model choice. In education systems where such a framework does not exist, the final decision of the most appropriate model might extend into the initial implementation of the value-added modelling of the broader school population. This could extend the period of the analysis of schools' value-added scores over subsequent assessments which might be important if there is found to be excessive instability in specific school scores. It might therefore be prudent to postpone the use of value-added scores for school accountability purposes given the greater uncertainty in the estimations. Depending on the extent of the instability and the ability to isolate it in a particular sub-set of schools, this might also be viewed as part of the broader ongoing development of the value-added modelling. Incremental changes to the value-added specification are to be expected as any analysis undertaken to ascertain how the model can be improved should be considered to be part of an ongoing process.

The pilot programme provides an excellent opportunity to further develop a stakeholder engagement and communication strategy. The engagement process can begin with the opportunity to recruit schools into the pilot programme and provide them with the opportunity to contribute to the objectives of the overall system of value-added modelling. School principals, teachers and other staff can contribute to: assessing and further developing responses to operation and implementation issues; the effective use of schools' value-added information, particularly at the school level; and, the communication and engagement strategy. Further input could be garnered from participating staff concerning the framework of student assessments, the collection of complementary data, particularly data collected at the school level, and the development of the most appropriate information system. It is considered an important part of a sampling procedure that the level of inconvenience and work imposed upon the sampling unit (in this case, the school) is minimised. Feedback during the pilot stage could greatly increase operational efficiency and reduce the impact upon schools' normal operations.

An important element of the pilot programme, with respect to operational procedures, is ensuring accurate data collection procedures. If additional information is to be gathered from schools then the appropriate piloting and questionnaire development needs to be undertaken. If administrative data is to be used then this should also be checked with schools to ensure the accuracy and completeness of data. Quality control monitoring of the data and data collection should be part of the live implementation but the monitoring procedures can be developed and assessed during the pilot programme. While the choice of the value-added model requires statistical expertise and is not ideally suited to the input of all stakeholders, it can be beneficial to gain input on the use of data for the development of a contextualised value-added model. Stakeholders would be able to advise on the need to include specific factors that affect student performance in the value-added modelling that might also affect the actions stemming from schools' value-added scores.

As schools would be the main target of a communication strategy, school principals, teachers and other staff can provide essential input into developing effective communication with schools and other stakeholders. Such input could influence the objectives of the strategy but the pilot programme also provides an opportunity to assess the value of specific information and guidance materials (*e.g.* on the use of the information system to analyse schools' value-added information) and seminars and workshops that can be developed for schools. This would extend beyond the correct interpretation of schools' value-added scores to the use of information systems containing school- and student-level value-added information to monitor school performance and develop commensurate school improvement programmes.

The pilot programme provides an important opportunity to develop effective training programmes and also to engage school principals and teachers in the use of value-added information for school improvement. Such engagement should be an important step in gaining the support of stakeholders for the live implementation of the system of value-added modelling. School principals and teachers can provide valuable input into how to best interpret and present value-added information. This would include the presentation of value-added information, including the ranking of specific scores, and the use of other evaluative information. The value of various training programmes can also be assessed to better align both the focus and delivery of training. In some countries, a key aspect was to engage with school principals and teachers to address their concerns about school scores that they perceived as unrealistic. The benefits from such dialogue required additional stakeholder training that extended beyond information sessions to develop analytical capabilities within schools. Follow-up assessments of the value of such training can ascertain if particular aspects of value-added modelling or the interpretation of value-added information could be enhanced.

The development of student assessment instruments has not been a focus of this report. Nevertheless, the pilot programme should be used to further assess the suitability of the assessment instruments. Standardised test instruments are the end result of a lengthy process of design and development, shaped by a multitude of goals and constraints (Braun, 2000). In assessing the validity of the assessment instruments, both substantive and technical issues need to be addressed. For example, further analysis could include the degree of articulation between the actual test content and the content standards that are supposed to be implemented by the school. This and other issues should be analysed to ensure the reliability of assessment instruments before the live implementation of the system of value-added modelling.

Given the objectives of a pilot programme, it is to be expected that problems will be encountered. Plans should be put in place to document and then address such problems. This is a central step in meeting the objective of the pilot programme of further developing the system of value-added modelling. Problems encountered in the pilot programme could therefore be viewed as an opportunity rather than a failure, and should be incorporated into a quality control system that operates throughout the life of the system of value-added modelling. An effective quality control system should ensure that high quality procedures are maintained and that issues are addressed to ensure continual improvement. Such procedures should monitor aspects of the system such as the framework of student assessments, the model employed to estimate value-added, the interpretation of schools' value-added scores, and the accuracy of the data used in the system. Any issues that need to be addressed in the pilot programme should serve as examples of the issues that need to be monitored once the system is operational. Such monitoring should aim to ensure that schools' value-added scores are accurate estimates of school performance.

Chapter Eight

Further Development and Use of Value-Added Modelling

The effectiveness of a system that uses school performance measures as a basis for actions rests on the confidence stakeholders have in the reliability of the measures of performance over time. Effective quality control monitoring of the results of the modelling and the data that feeds such analysis is therefore central to the effective use of a system of value-added modelling. The discussion presented here focuses on the importance of monitoring schools' value-added results over time, emphasising that such monitoring needs to focus on changes in individual school results as these are the focus of stakeholders and efforts to lift performance. Given the need to minimise unstable variation in schools' value-added results, the discussion emphasises the need to calculate and present a three-year moving average of each school's value-added score as the central or published school performance indicator. A discussion is then presented of how systems can develop successful communication and stakeholder engagement strategies and the training of stakeholders, particularly teachers and school principals, connected with such strategies. The chapter concludes with a discussion of how the pilot phase of the implementation process can feed decision-making regarding the publication of schools' value-added scores.

The credibility of any statistical system rests, in the first instance, on the integrity of the data and the operations performed on that data. Thus, developing and implementing effective quality control procedures at every stage of the process is an essential aspect of a value-added analysis. It is evident that both test data and covariates should be carefully checked and edited before analysis. This can involve identifying out-of-bound or unusual values, as well as unexpected distributional characteristics. Comparisons with prior year data can sometimes be helpful. Patterns in missing data might also call for analysis and consequent actions devised and implemented. As an example of a specific monitoring initiative, a sample of schools could be selected following each data collection and further analysed to ensure that the data is accurate and can be correctly interpreted. In particular, substantial changes in the number of students excluded from testing (*e.g.* because of disabilities) or the number of students absent on the day of testing should be highlighted as this could signal the presence of bias in the estimated school effects. In some countries, schools face strong disincentives if students miss the specified assessments.

Changes in schools' value-added scores will tend to be seen as indicating changes in school performance, even though this might not be justified on statistical grounds. Less stable value-added scores could lead directly or indirectly to incorrect inferences or actions, and their potential usefulness could be obscured by an impression of inaccuracy. Ideally, school performance indicators would be relatively stable but retain the ability to rise and fall in response to real changes in school performance. This ideal situation is unlikely to be achieved on all occasions. It is therefore necessary to analyse changes in value-added results extensively during the pilot programme and conduct analyses of changes over time once the system is established. In education systems that can analyse existing data, the opportunity exists to further test model specification and assess the stability of school scores over time to feed into decisions of model choice and the appropriateness of the student assessments and the data used in the model.

An analysis of the stability of school results was conducted in some participating countries. Some instability of school scores is to be expected across all value-added models and some instability, of course, is desired. Greater instability was evident in some education systems and this might reflect the lower quality of the examination systems in those countries. The student assessments upon which value-added is being measured should therefore be examined if the instability of school scores is considered excessive. The stability of school scores depends not only on the definition of abnormal or excessive instability but also on the categorisation of schools by their level of performance. It was also found that instability of scores differs with school size, the type of model used, the number of contextual variables included, the number of years between prior and current attainment, and the coverage of the value-added comparison (all schools in the country or a subset). These findings illustrate the benefits of conducting further analysis on schools with large changes in school scores or seemingly random changes over a number of years.

The additional analysis of schools with less stable scores over time can be difficult given the complex objective of distilling observed changes into what might be termed 'persistent' and 'transient' components. The former refers to stable changes in true performance and the latter refers to all other factors. The transient component of instability can be dampened to some degree by incorporating more data (*i.e.* more prior years and more subjects) and by averaging the results over successive cohorts. However, more detailed analysis of the data can reveal the source of instability in school scores. For example, modelling of particular subjects or subgroups within a school can be used to establish whether the changes in scores look plausible.

It might also indicate that changes in the data used or in specific student assessments have resulted in variation in school scores. This can assist in an analysis that estimates differences between persistent and transitory school-level effects. If there are specific known issues, such as a change in the classification of contextual data, this could be flagged in publications and drawn to the attention of school inspectors and other users of the data. For presentation of the results, instability of school results can be viewed as another argument for presenting confidence intervals around point estimates. Associating a confidence interval with each school's value-added score can lessen the likelihood of misinterpretation.

It has been considered advisable in a number of education systems to present point estimates with confidence intervals, with the advice that overlapping confidence intervals indicate that the corresponding point estimates are not statistically significantly different. When many such comparisons are made, there is a high risk of making too many Type I errors. This danger can be mitigated by employing the techniques of simultaneous inference, the best known of which are termed 'Bonferroni methods'. Newer techniques, such as those based on the False Discovery Rate approach (Benjamini and Hochberg, 2000), are becoming increasingly popular. For general audiences, graphical presentations can be very effective. The so-called caterpillar plot can be particularly effective. The estimated effects are ordered by rank along the X-axis and by magnitude along the Y-axis. In addition, a confidence interval for each effect is placed vertically and centred on the point estimate. In addition, it must be explained that while confidence intervals can better illustrate statistically significant differences in school results they are not a panacea and do not capture the uncertainty due to bias and other secular changes. Some potential sources of bias can be incorporated into the model, as is the case with the model of Ponisciak and Bryk (2005) discussed in Part II.

In practice, schools displaying unusually large changes should be studied carefully. If the instability is substantial and thought to be primarily due to transient factors, then consequences that are directly linked to a school's value-added estimate should be avoided. In these cases, triangulation utilising additional evidence (*e.g.* from school inspections) is beneficial, particularly if changes in school scores lead to actions such as strong sanctions or rewards. If the results are to be used internally, then appropriate cautions can be attached. On the other hand, if the results are to be made public, then guidelines should be put in place to determine whether the results should be suppressed. The guidelines should take account of school sample size as well as other factors. Across participating countries, schools with smaller sampled cohorts were found to have much larger instability of results across years. One possibility open to policy makers is therefore not to report results from schools that do not meet minimal sample size requirements, or for schools for which the length of the confidence interval associated with a difference exceeds a pre-determined threshold. These two results are often related and the question of how to deal with small schools must be taken into consideration by administrators and policy makers. The expert group in this project considered interpreting value-added results for schools with less than 20-30 students in a cohort to be problematic, but it is recognised that school size can differ substantially between countries. However, it is possible to group smaller schools to obtain larger samples that, statistically speaking at least, can be better interpreted. Problems can arise however in how to interpret the results of groups of smaller schools if there is no *a priori* rationale for such a grouping. In some OECD member countries, it is possible to group smaller schools that exist under specific geographic regions and administrative units. Value-added results can therefore be analysed to ascertain performance measures for these regions or administrative units. Such measures can be particularly useful for policy analysis if these administrative units have distinct education programmes, the impact of which can then be informed by the value-added measures. However, the interpretation of such results should be made with caution given that differences might exist between schools that will make the interpretation of a single score for a group of heterogeneous schools problematic. This could be particularly important in systems that have greater levels of school autonomy and therefore a potentially greater divergence in education policies and programmes.

Instability in school value-added scores might not only arise through changes in school performance or through problems in the value-added estimates. Schools' scores can also be affected by changes in the value-added model used to estimate school performance. Over time, it is inevitable that changes will be made to the model, the data, or both, in response to continual inspection of the value-added analyses or to external demands. It is valuable to regularly confirm that the model is still appropriate for current policy purposes and to consider the implications of changes in available data. While such changes should be minimised so that they do not overly negate the comparability of results over time, it is only natural to assume that statistical estimations will be slightly altered and improved upon with the further development of the system. Such changes might also come from changes in policy that wish to focus on different aspects of school performance or to focus on more extensive contextualised value-added estimations. Such changes should be tested to ascertain the impact on all schools' value-added scores (not just the overall model) and it is important that they are discussed with stakeholders to ensure that the interpretation of value-added remains constant over time.

In England, changes have been kept to a minimum but there was a difference between the 2005 and 2006 secondary school contextualised value-added model specification and between the 2006 and 2007 primary school models.[17] If value-added scores are being compared or averaged over time to judge school performance, it is clearly important that any changes in the underlying model need to be taken into consideration. With some changes, it might be possible to calculate value-added scores from both the old and new bases, but in other cases, such as the inclusion of a new piece of data, this is not possible. The effect for earlier years could be estimated based on the change for the most recent year, although this would not necessarily provide a robust estimate for the earlier years. Where it is possible to recalculate earlier years on the new basis or, conversely, to obtain an estimate for the new year on the old basis, two sets of figures can be given, and a trend or average can be calculated on a consistent basis. However, this does not remove the difficulties for a school whose earlier value-added score would have been different if calculated on the new basis, especially if actions were undertaken and the school incurred specific consequences based on the results obtained under the earlier model. There are also important judgements to be made regarding how significant a change in the model has to be in order to warrant the calculation and dissemination of revised earlier figures or new figures on the old basis. These judgements will depend on the number of schools affected, the size of the impact, and the resources required in calculating the alternative figures.

Use of results of a three-year moving average

Given the potential for the excessive instability of some schools' value-added results over subsequent years, it is clear why the expert group considered it beneficial for actions stemming from schools' value-added results to be based on a three-year moving average of scores. It is considered that there is a need for caution in the interpretation of data from just one or two years. The question therefore arises of how to utilise interim data when establishing a system of value-added modelling and how to ensure timely responses that might not be highlighted as quickly in analyses of three-year moving averages. Given the difficulties in interpreting a single year of results, it could be appropriate that actions with potentially large consequences for schools (and teachers and principals) be tempered or postponed until data from further years are available and a three-year average can be obtained or the results are supported by other information. Defining exactly what small and large consequences are is a subjective calculation and one that cannot be defined precisely in this report given the breadth of policy actions and the degree to which the relevant parameters can differ between countries and school systems. From a policy perspective, a distinction could be made between actions that are more aligned with school accountability and those that are more aligned with school improvement. Actions within school

[17] These changes took place between the pilot analysis for Performance Tables and the national publication of contextualised value-added scores for all schools.

accountability systems could have potentially large negative consequences (as perceived within schools) compared with the use of value-added scores for internal school improvement purposes, but this is not true for all actions and interventions in these systems.

In waiting for three years' of data to obtain a precise value-added score it is recognised that it is inefficient not to use the data in some manner and that this delay can be harmful to students if actions are not taken to lift low-performing schools. Value-added scores indicating low performance could trigger further analysis of existing data and school processes. Such analysis would be aimed at identifying additional indicators of low performance to make a more complete assessment upon which appropriate remedial actions could be based. Further data could be collected and analysed (although there are obviously resource constraints in collecting that data) that could include an analysis of student performance such as raw test scores, student retention and pass rates, and further analysis of student in-take data and other administrative data such as student mobility. This might yield a further indication of either changes within the school (such as changes in student composition) or changes in that performance of students that might confirm (or contradict) the single-year value-added result. Further analysis might be undertaken of additional school indicators. Teacher turnover rates might be a sign of a problem at a school or that the changes might have occurred with an influx of new teachers. A change of school principal might have resulted in changes either to school programmes or to the organisation of the school that could be important in the context of the result of the value-added model. Information on school processes would also be valuable to support the information of the result of the one-year value-added model. Information from value-added models and information on school processes are complementary rather than substitutes and the combination of multiple indicators can provide greater confidence in the decisions to undertake specific actions.

It might also be advisable not to publish the results of value-added models until a three-year moving average can be obtained. Greater instability of schools' scores in these early years can lead to problems if these results are published. Stakeholders can quickly lose confidence in a system with such instability particularly if the publication of school results is a new venture in an education system. Therefore, in the initial years, it is considered that there are benefits to begin with a process focusing upon school improvement measures and, if desired, this can build into a system that has stronger actions based on the results of the value-added model including the publication of results. Alternatively, results can be published on an interim basis and additional information can be used to support one to two years' of value-added results.

Communication and stakeholder engagement

Numerous stakeholders can benefit from a system that utilises value-added models. With these benefits, it is recognised that if value-added results are used as a basis for action, such action could have a negative impact upon particular individuals and organisations (*e.g.* the placing of sanctions upon schools). This potential can create an unwelcome response to the introduction of new systems that measure, among other things, the performance of individuals or organisations. This reaction might be particularly apparent if value-added modelling is introduced as part of a broader school or teacher accountability programme. Given these potential problems, it would be pertinent to engage stakeholders in the development, implementation, and continued use of value-added modelling. Effective engagement could be achieved through an extensive communication strategy that complements extensive training programmes. Such efforts would recognise and facilitate the development of schools as effective learning organisations, and are discussed below.

Teachers, school principals and other staff within schools are the main stakeholders whose work would be affected by the implementation of a system of value-added modelling. Perceptions of mistrust, increased pressure, frustration and a fear of a loss of autonomy are common reactions to the implementation of a system that monitors performance (Saunders, 2000). Although many reform efforts must confront these problems, those that rely on value-added analyses can face some specific difficulties. First, value-added modelling can be seen as particularly unmerited as the models are sufficiently complex so as to appear opaque to many stakeholders. Second, the information generated is at the school level but any real improvement depends on changes at both the school- and teacher-level. Thus, one challenge is how to generate and present information that can be understood by teachers. Another is to build capacity so that teachers, school principals and other school staff can use the information effectively. Building capacity will involve increased mentoring and training of teachers, school principals and other stakeholders (Saunders, 2000). It will also require investment in central office staff and analytic resources. A communication and stakeholder-engagement strategy should focus on the greater accuracy inherent in value-added modelling of measures of school performance. This has been shown to be a significant benefit, with stakeholders coming to favour value-added modelling as it provides a more accurate, and thus fairer, measure of school performance than other indicators that have been used in some education systems (Dudley, 1999). For example, Fitz-Gibbon (1997) highlighted the favourable views of head teachers in England to the introduction of value-added modelling and, as detailed below, Jakubowski (2007) found teachers also favoured the use of value-added modelling to measure school performance. An important benefit of the effective engagement of key stakeholders is that it should reduce the possibility of behaviour that can potentially bias the data used in value-added modelling. As discussed in Part I, a number of systems can suffer from adverse behaviour that can bias the student assessment and school-level data collected and also create incentives for sub-optimal teacher and school behaviour. Overcoming these problems requires teachers and school principals to trust the system to be fair and to reflect true school performance. It must also be made clear that this effort requires a long-term commitment that might alter both the relations between the central authority and schools and the dynamics within schools.

Successful communication strategies in a number of education systems have involved the engagement of stakeholders in a number of facets of the implementation of the system. Such systems have moved beyond merely communicating the details of the value-added model being developed to encouraging stakeholders to utilise value-added information for their own benefit. Effective engagement involves multi-channelled communication in the development and operation of value-added models and the system that utilises schools' value-added scores as a basis for actions (Saunders, 2000). This is particularly important if value-added models will be utilised for a system of school-improvement measures that requires the interpretation of school results and the formulation of actions stemming from such interpretations at the school level.

Effective communication encompasses each stage of the process. Each stage needs to be effectively communicated to stakeholders and initiatives implemented to engage and garner their support. This includes: the objectives and rationale of the system; the development and choice of the value-added model to be used; the implementation of the system, particularly the system of student assessments; and the use of value-added information by different stakeholders. These strategies have been integral to the success of value-added modelling in school education systems in various participating countries and are discussed below.

As with the development of the overall system, the objectives of introducing a system of value-added modelling need to be clearly articulated to stakeholders. The main elements of key policy objectives have been discussed in Part I of this report and need not be repeated here, but it is important to identify and carefully consider the impact upon school principals, teachers and other school staff. There are benefits to clearly articulating how value-added scores are going to be used to measure school performance. Of particular importance to stakeholders might be the unit of analysis in the value-added modelling and how the results will be used and presented. The unit of analysis might vary to focus on regions, administrative units, schools, and teachers. This report has focused on value-added at the school-level but the unit of analysis should be explicitly addressed, including a discussion of whether schools will be explicitly identified in any published materials.

Providing schools' value-added scores to the general public is central to the objective of promoting school choice. Regardless of the intention of the publication of schools' value-added results, teachers, school principals and other stakeholders can perceive this as a form of school accountability. As has been discussed in Part I, the publication of results can create negative perceptions among schools and fuel suspicion of the motives for the introduction of a system of value-added modelling. The development of a communication strategy that addresses these needs can be constructive. In some education systems, school visits and promotional materials have been used to convey how school value-added results can be presented. These have often complemented education and training initiatives aimed at increasing understanding of value-added modelling and of the use of such information. The communication strategy can include producing publications for schools and information sessions with explicit examples of how schools' value-added scores can be published, including illustrative tables and diagrams. This would also explain how to interpret such tables and diagrams, particularly the statistical interpretation of schools' value-added scores and, if relevant, confidence intervals and how they can be used to classify significant differences in school performance. Again, including teachers and school principals in the decision-making of how to present school results and other information (*e.g.* in a school profile) can be an effective engagement strategy and improve the overall quality of the system.

In developing this system, most Governments will develop a media strategy for the release of school value-added measures and an explanation of how they should be interpreted. Value-added data can be complex and multi-dimensional and a simple ranking of schools can be misleading if the rankings are not aligned with the specified policy objectives and practices. Steps need to be taken to ensure that the presentation of value-added information in the media does not negate the positive aspects of the development of the system. In a number of participating countries it was considered that even if the objective was not to convert schools' value-added scores into a ranking of schools, then this would be done by the media. Further unintended presentation of results could manifest itself as a media focus on raw test scores. It might be considered prudent to produce information on school and student performance that includes raw test scores, value-added scores and contextual value-added scores to provide a more complete picture for internal analysis and also to better facilitate school choice for parents and families. While it is difficult to control the media story, steps can be taken to both educate the media in how to interpret value-added information and to provide explicit statements about what can and cannot be interpreted from value-added scores and other information. In addition, it is possible to emphasise particular aspects of the performance measures. For example, in conjunction with the different information presented, a single school ranking could be produced based on schools' contextual value-added scores if it is considered that this is the more accurate measure.

It might be advantageous to develop a media strategy in conjunction with teachers, school principals and other stakeholders, as schools are often the targets of media stories. Moreover, education stakeholders can be effective in communicating a common message of how to interpret value-added measures and it could be assumed that information gained from school principals, teachers and other stakeholders can inform the development of a more comprehensive media strategy. Such a strategy might facilitate a smooth implementation of the system and negate the likelihood of misleading media stories negating the advantages of the introduction of a system of value-added modelling. It is important in any organisational context that individuals feel empowered in their workplace, particularly one undertaking organisational change initiatives such as the introduction of a system of performance measurement (O'Day, 2006). Empowering school principals, teachers and other school staff with not only a greater understanding of value-added modelling but training in how to interpret and analyse value-added information for school improvement purposes can facilitate the effective implementation of a system of value-added modelling. In addition, training in how to analyse the data, develop school programmes and monitor student progress might alleviate suspicion and illustrate the tangible benefits to stakeholders. It could also be advantageous to allocate resources for particular school improvement actions stemming from value-added modelling. To emphasise the use of systems of value-added modelling for school improvement, a given sum of resources could be allocated for use by schools that analyse value-added information to develop specific programmes aimed at lifting student performance. This could act as both an incentive to undertake such analyses but also to emphasise to stakeholders that the system is being implemented for school improvement purposes and is not purely an additional layer of bureaucracy or school accountability. It would also emphasise the belief in data-based decision-making in lifting performance throughout the school education sector.

Development of a training programme

The close inspection of school-related data as a foundational exercise in school development is a relatively new phenomenon. Many educators are not well-trained in measurement or statistics and some might not feel confident in the interpretation of value-added information. Consequently, the introduction of school performance indicators based on value-added analyses must be carefully designed and implemented with training considered to be a key requirement in the introduction of new quantitative performance measures (Yang et al., 1999).

Training programmes have proven to be effective in England with the analysis of value-added results by school principals and school improvement partners. Such empowerment requires effective communication and training strategies so that school principals and teachers can more effectively utilise value-added information for school improvement purposes. A system is more likely to be supported if more tangible benefits can be articulated to, and utilised by schools. To this end, it might prove beneficial to engage school principals, teachers and other stakeholders in the development of the system through which analysis of value-added information is conducted at the school-level. This might encompass decisions of which information is to be collected and included in the modelling, and what sort of analysis is particularly beneficial to schools. The engagement of these stakeholders might also help to develop a user-friendly interface for the information system and software to aid analysis conducted at the school-level. Conversely, a system that involves little use of data at the school-level could be viewed as being imposed upon schools following a top-down management approach (Wikeley, 1998).

Efforts to bring any recalcitrant stakeholders into the system would require careful planning that reflected an understanding of how value-added information can affect school principals and teachers. The experience with such training in participating countries has shown a need for communicating fundamental statistical information of how schools' value-added results are estimated. While this training could be considered to cover only the basics of value-added modelling (the training is not intended to equip stakeholders with the skills to conduct their own high-level modelling), positive feedback was received from individuals working with teachers on how to interpret schools' value-added scores and issues such as confidence intervals and the calculation of statistically significant differences between schools' results. This training can extend to discussions of the stability of school scores over subsequent years and the impact of such instability upon the use of schools' value-added scores to attain stated policy objectives. Training should also enable the analysis of student-level data that both illustrate the variation in student performance across a school and also of particular groups of students. This would enable schools to identify value-added scores in different subjects and in different age groups and enable analysis of particular groups of students delineated by, for example, socio-economic status, gender, ethnicity, or family status. Schools with such analytical capabilities should be able to better identify those students that have poorer performance measures, to devise appropriate measures to lift their performance and to monitor the impact of such measures. This should also facilitate extensive school-level organisational learning of the effectiveness of various approaches as schools benefit from effective data-based decision-making and schools and teachers seek to improve their methods based on an accurate understanding of school performance.

Some participating countries that developed training programmes reported that in undertaking the training it can become apparent to individuals that value-added modelling provides more accurate data of school performance than analysis of raw scores. In gaining an understanding of how to interpret value-added information and what schools' value-added scores actually measure, the benefits of such analysis became clearer to stakeholders. Many stakeholders in England welcomed the introduction of value-added modelling as it was perceived as being a much fairer performance measure than analysis of raw test scores. Training would further emphasise such benefits and therefore increase the likelihood of stakeholder acceptance of value-added modelling.

The discussion of value-added models would need to delineate contextualised value-added models and the interpretation of schools' value-added scores and coefficients for the included contextual variables. This would include discussion of whether contextualised value-added modelling would be utilised, the rationale behind such a decision, and the testing of the model in the pilot phase. Depending on the structure of student assessments, the predictive power of value-added models might not be greatly enhanced by the use of contextual characteristics. Yet, they might be important for the purposes of policy development and in effective stakeholder engagement. The use of contextualised value-added models might alleviate the concerns that models are simply measuring student intake rather than school performance. While models that use higher numbers of repeated measures can overcome the need for measures of student background characteristics, this is less easily conveyed to stakeholders who might be less versed in statistical analysis. It can also be advantageous to consult stakeholders on the inclusion of contextual data to feed into a contextualised value-added model and the additional school-level information that would complement such data. Relevant stakeholders have considerable experience with the student and school characteristics that can affect student performance and they can provide valuable insight into how such data could add to policy development. In addition, this is a further opportunity to include stakeholders in the developmental process and engage their support for the use of such modelling.

Developing effective training programmes and implementing effective communication strategies can be resource-intensive activities. Fortunately, these challenges can bring concomitant rewards. Estimated school effects, when accompanied by other contextual and comparative information, can provide a useful starting point for conversations both among and within schools. By breaking down the results by various student characteristics, a fairly detailed picture can be painted of the strengths and weaknesses of the school's programmes. Such analyses are conducted regularly in England and in some education systems in the USA, such as in Dallas, Texas and in a number of districts in Tennessee (Braun, 2005a). Developing more effective data-based decision-making with the use of value-added information encompasses the development of more extensive information systems within schools. It needs to be recognised that the development of effective information systems in complex organisations such as schools requires more than just analytical training and capabilities (O'Day, 2002). It can be beneficial to emphasise communication and effective collaboration in schools to ensure that decision-making concerning the development and monitoring of school programmes is effective across an entire school and is not confined to senior management. If it is considered beneficial to place a greater emphasis upon a school-wide approach to data-based decision-making, then training to promote peer collaboration and the development of school programmes by teams of teachers could be promoted.

Box 8.1. Training programmes in Poland

An extensive training programme was implemented in Poland alongside the introduction of a system of value-added modelling. Conducted in 2006, a tiered structure was established whereby teacher trainers were educated centrally and then trained teachers in local training centres. The objectives of the training centred upon:

– the interpretation of value-added scores;

– illustrating how value-added methods could be used to assess student's progress to facilitate school improvement programmes; and

– creating a group of teachers, school principals, inspectors, and teacher advisors able to teach others and promote valid use of value-added information.

The training programme consisted of a combination of lectures, exercise classes, and open floor sessions. The opening lectures introduced the idea of value-added assessment of schools and explained the theoretical issues in value-added modelling. It was believed that such an approach, even if too demanding, would eliminate the feeling that a small group of experts was imposing methods that were not transparent to the public.

The lectures were followed by classroom exercises in small groups. All attendees received tables with lower secondary school exam predicted scores from a regression on primary school scores. Additionally, coefficients for dummy variables estimated in the model were also presented (*e.g.* gender, dyslexic students). Teachers were then able to calculate regression residuals (through subtracting the actual scores of each student from the predicted scores). Using residuals, participants calculated school's value-added as the average of residuals for students in a particular school. In addition, teachers were taught how to calculate confidence intervals for the mean of residuals that were then used to compare schools. It was explained that such an approach could be only used as a heuristic tool and was not fully valid from the statistical point of view. It was emphasised that value-added assessment done in this simple way could be a helpful tool to check if there were any significant differences in school performance and to create preliminary hypotheses which could be then interpreted by school personnel who had greater knowledge about a school, teachers and students.

It should be noted that this simple value-added model was preferred over more complicated models, because it could be used by schools internally and was relatively simple to explain. The virtue of this simple model was further explored during the training. Experts showed teachers and school principals the way they could calculate value-added scores for defined groups within schools, showing how value-added scores could be compared between girls and boys or between classes. These simple exercises were done using an Excel spreadsheet which is commonly used in schools in Poland.

A lecture was then given which summarised the advantages and disadvantages of value-added assessment in Poland based on research conducted to test the external validity of value-added approach. The lecture was followed by an open floor session where attendees were able to ask questions and experts were able to clarify misunderstandings and explain some technicalities. Finally, a short survey was conducted among participants who evaluated the training and the introduction of a system of value-added modelling more generally.

Participants benefited not only from the three-day training but also received materials which could be used to train other teachers. Materials were printed as a booklet containing a technical description of the value-added model implemented in Poland as well as all the exercises that had been developed and taught during the training. Additionally, exercises were implemented in Excel and given to participants on CD to make further training easier. The seminar was followed by a five-hour training session conducted in the following month in each of the 50 regional and local teacher training centres. In addition, representatives of school inspectorates (Kuratoria) participated in this training and additional special training sessions were designed where value-added models were presented and discussed as a potential tool for monitoring of teaching quality in lower secondary schools. Finally, in 2007 a 'value-added calculator' website was launched and information on how to use this new Internet tool was released and incorporated into teacher training programmes in local centres. Many local teacher training centres have since responded to growing interest in value-added by incorporating value-added courses into their training programmes.

Presentation and use of value-added information

A school's value-added score will be a number that reflects its performance relative to other schools. The interpretation of this score requires an evaluative assessment that should be used as a basis for actions that advance stated policy objectives. Numerous examples were given in

Part I of this report of how value-added information can be presented both for internal use and for public consumption. These need not be repeated here, and the discussion is kept to the issues pertinent to the implementation phase. These centre on the assessment of the appropriate method to publish value-added information, their use internally and within schools, and how they will result in specific actions.

The publication of school results should be aligned with the desired policy objectives. There can be benefits in developing such publications with the pilot data and receiving feedback on these publications from relevant stakeholders. This feedback can assist in the overall development of the publications themselves but also highlight areas that stakeholders consider to be particularly sensitive. This can inform decisions on the publication of value-added scores in the live implementation. The use of value-added information within schools and for internal policy development requires training, the development of pertinent software, and judgements to be made concerning which information should be available for analysis and in what form. All can be informed through the pilot phase with analysis by relevant stakeholders creating feedback that should then inform planning for the live implementation.

Guidelines for the interpretation of value-added scores should be established to assist the development of appropriate actions and interpretations made by stakeholders. In a number of countries, this has focused upon classifying results as indicators of specific performance categories (*e.g.* low- and high-performing schools). Such guidelines should be developed and then assessed through interaction with relevant stakeholders during the pilot phase and throughout the implementation. Explicitly identifying how value-added scores will be interpreted and used to trigger specific actions increases the level of transparency and internal efficiency. Stakeholders need to know these actions in order to have confidence in the system and also to design appropriate measures to lift performance. The actors and institutions (*e.g.* inspectors, ministries, departments, and schools) that implement the pre-determined actions can also better plan and develop interventions to lift school performance. For example, a school classified as low-performing might trigger a school inspection and a period of more intense evaluation. If the criteria for this classification and resultant action is clearly defined, procedures can be put in place that allow schools and school inspectorates (or an appropriate institution) to prepare and better develop an evaluative framework that efficiently responds to the classification, thus undertaking an analysis of the value-added data to implement a school evaluation that addresses the needs of each specific school. This would allow targeted strategies to be more efficiently developed and equip school inspectors, school principals and teachers with greater information to first address and to then increase school performance.

In the pilot phase and the early implementation of the system of value-added modelling, it is possible to analyse either the pilot data or pre-existing student assessment data to gauge the impact of value-added scores and resultant actions. For example, analysis can be undertaken of the proportion of schools that would receive specified rewards and sanctions, that would receive a school inspection, that would be placed on probation, and would be classified as high- and low-performing. This analysis can inform decisions of where the 'trigger points' should be situated in the distribution of school value-added scores and also the resource implications of the proposed point in regard to subsequent actions such as school inspections and specific rewards and sanctions.

Bibliography

Aitkin, M. and N. T. Longford. (1986). Statistical Modelling Issues in School Effectiveness Studies. *Royal Statistical Society*, Series A, 149 (1), 1-43.

Amato, P. and B. Keith. (1991). Parental Divorce and Adult Well-Being: A Meta-Analysis. *Journal of Marriage and Family,* 53 (1), 43-58.

Antelius, J. (2006). *Value-Added Modelling in Sweden: A Background Report for the OECD Project on the Development of Value-added Models in Education Systems.* Skolverket.

Atkinson Review. (2005). *Final Report: Measurement of Government Output and Productivity for the National Accounts.* Palgrave McMillan.

Ballou, D. (2001). Pay for Performance in Public and Private Schools. *Economics of Education Review,* February, 51-61.

Ballou, D., W. Sanders and P. Wright. (2004). Controlling for Student Background in Value-Added Assessment of Teachers. *Journal of Educational and Behavioral Statistics,* 29.

BBC News (2007), *Guide to the secondary tables*, BBC News website, http://news.bbc.co.uk/1/hi/education/7176947.stm, November.

BBC News (2008), BBC News website, http://news.bbc.co.uk/1/shared/bsp/hi/education/07/school_tables/secondary_schools/html/320_4075.stm, 10 January.

Becker, G. (1964). *Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education.* New York: Columbia University Press.

Benjamini, Y. and Y. Hochberg. (2000). The Adaptive Control of the False Discovery Rate in Multiple Hypotheses Testing. *Journal of Behavioural Education Statistics,* 25, 60-83.

Betebenner, D. (2007). *Growth as a Description of Process.* Unpublished manuscript.

Bethell, G. (2005). *Value-Added Indicators of School Performance: The English Experience Anglia Assessment.* Battisford, Suffolk, England: Unpublished report.

Borjas, G. (1995). Ethnicity, Neighborhoods, and Human-Capital Externalities. *American Economic Review,* 85, 365-90.

Borjas, G. (2001). Long-Run Convergence of Ethnic Skill Differentials, Revisited. *Demography,* 38 (3), 357-61.

Bourque, M. L. (2005). The History of No Child Left Behind. In R. Phelps (ed.), *Defending Standardized Testing* (pp. 227-254). Hillsdale, NJ: Lawrence Erlbaum Associates.

Braun, H. I. (2000). A Post-Modern View of the Problem of Language Assessment. In A. J. (ed.), *Studies in Language Testing 9: Fairness and Validation in Language Assessment. Selected Papers from the 19th Language Testing Research Colloquium* (pp. 263-272). Cambridge: University of Cambridge, Local Examinations Syndicate.

Braun, H. I. (2005a). Value-Added Modelling: What Does Due Diligence Require? In R. Lissitz, *Value Added Models in Education: Theory and Applications.* Maple Grove, Minnesota: JAM Press.

Braun, H.I. (2005b). *Using Student Progress to Evaluate Teachers: A Primer on Value-Added Models.* Policy Information Perspective. ETS.

Braun, H.I. (2006a). *Background Paper: The use of value-added models for school improvement.* Paris: OECD.

Braun, H. I. (2006b). Empirical Bayes. In J. G. (eds.), *Complementary Methods for Research in Education.* Washington, DC.: American Educational Research Association.

Braun, H. I., Y. Qu and C. S. Trapani. (2008). *Robustness of Value-added Analysis of School Effectiveness.* ETS RR-08-22. Princeton, NJ: Educational Testing Service.

Brooks-Gunn, J., G. Duncan, P. Klebanov and N. Sealand. (1993). Do Neighborhoods Influence Child and Adolescent Development? *American Journal of Sociology*, 99, 353-93.

Bryk, A., Y. Thum, J. Easton and S. Luppescu. (1998). *Academic Productivity of Chicago Public Elementary Schools, Technical Report.* Chicago, Il.: The Consortium on Chicago School Research.

Burgess, S., C. Propper, H. Slater and D. Wilson. (2005). *Who Wins and Who Loses from School Accountability? The Distribution of Educational Gain in English Secondary Schools.* CMPO, The University of Bristol: CMPO Working Paper Series NO. 05/128.

Burstein, L. (1980). The Analysis of Multi-Level Data in Educational Research and Evaluation. *Review of Research in Education*, 158-233.

Caldwell, B. (2002). Autonomy and Self-managment: Concepts and Evidence. In T. Bush and L. Bell, *The Principles and Practice of Educational Management* (pp. 34-48). London: Paul Chapman.

Caldwell, B. and J. Spinks. (1998). *Beyond the Self-Managing School.* London: Falmer Press.

Carlsson, G. (1958). *Social Mobility and Class Structure.* Lund, Sweden: Gleerup.

Choi, K. and M. Seltzer. (2005). *Modelling Heterogeneity in Relationships Between Initial Status and Rates of Change: Latent Variable Regression in a Three-Level Hierarchical Model.* March. Los Angeles, California: National Center for Research on Evaluation, Standards and Student Testing/UCLA.

Choi, K., P. Goldschmidt and K.Yamashiro. (2005). Exploring Models of School Performance: From Theory to Practice. In J. H. (eds.), *Yearbook for the National Society for the Study of Education,* 104 (2), Malden, Massachusetts: Blackwell.

Coleman, J. (1966). *Equality of Educational Opportunity.* Washington D.C.: U.S. Department of Health, Education, and Welfare.

Corcoran, M., R. Gordon, D. Laren and G. Solon. (1992). The Association Between Men’s Economic Status and Their Family and Community Origins. *Journal of Human Resources,* 27 (4), 575-601.

Department for Children, Schools and Families, United Kingdom (2008), high school performance tables website, www.dcsf.gov.uk/cgi-bin/performancetables/dfe1x1_05.pl?School=8464016&Mode=Z&Type, accessed 2 October 2008.

Dixit, A. (2002). Incentives and Organisations in the Public Sector: An Interpretive Review. *Journal of Human Resources,* 37 (4), 696-727.

Doeringer, P. and M. Piore. (1985). *Internal Labour Markets and Manpower Analysis.* New York: Armonk.

Doran, H. C. and L. T. Izumi. (2004). *Putting Education to the Test: A Value-Added Model for California.* San Francisco: Pacific Research Institute.

Doran, H. and J.Cohen. (2005). The Confounding Effects of Linking Bias on Gains Estimated from Value-Added Models. In R. Lissitz, *Value-Added Models in Education: Theory and Applications.* Maple Grove, MN: JAM Press.

Doran, H. and T. Jiang. (2006). The Impact of Linking Error in Longitudinal Analysis: An Emprical Demonstration. In R. Lissitz, *Longitudinal and Value-Added Models of Student performance* (pp. 210-229). Maple Grove, MN: JAM Press.

Dorans, N., M. Pommerich and P. Holland. (2007). *Linking and Aligning Scores and Scales (Statistics for Social and Behavioral Sciences).* New York: Springer.

Dudley, P. (1999). Using Data to Drive Up Standards: Statistics or Psychology? In C. Conner (ed.), *Assessment in Action in the Primary School.* London: Falmer Press.

Dyer, H., R. Linn and M. Patton. (1969). A Comparison of Four Methods of Obtaining Discrepancy Measures Based on Observed and Predicted School System Means on Achievement Tests. *American Educational Research Journal*, 6, 591-606.

Eurostat. (2001). *Handbook on Price and Volume Measures in National Accounts.* Luxembourg: European Communities.

Ferrão, M.E., P. Costa, V. Dias and M. Dias. (2006). Medição da competência dos alunos do ensino básico em Matemática: 3EMat, uma proposta. [Measuring math skills of students in compulsory education: 3EMat, a proposal]. *Actas da XI Conferência Internacional de Avaliação Psicológica. [Proceedings of the XI International Conference on Psychological Evaluation].* Braga, Portugal.

Ferrão, M. (2007a). *Sensitivity of VAM Specifications: Measuring Socio-Economic Status: A Background Report for the OECD Project on the Development of Value-added Models in Education Systems*. Warsaw.

Ferrão, M. (2008). Sensitivity of Value-Added Model Specifications: Measuring Socio-Economic Status. *Revista de Educación.*

Ferrão, M.E., Goldstein, H. (2008). Adjusting for Measurement Error in the Value Added Model: Evidence from Portugal. *Quality and Quantity.*

Fielding, A., M.Yang and H.Goldstein. (2003). Multilevel Ordinal Models for Examination Grades. *Statistical Modelling* (3), 127-153.

Figlio, D. and L. Kenny. (2006). Individual Teacher Incentives and Student Performance. *NBER Working Paper 12627.*

Fitz-Gibbon, C. (1997). *The Value Added National Project Final Report: Feasibility Studies for a National System of Value-Added Indicators.* London: School Curriculum and Assessment Authority.

Fitz-Gibbon, C. and P.Tymms. (2002). Technical and Ethical Issues in Indicator Systems: Doing Things Right and Doing Wrong Things. *Education Policy Analysis Archives, 10* (6).

Friedman, T. (2005). *The World is Flat: A Brief History of the 21st Century.* New York: Farrar, Strauss and Giroux.

Ginther, D., R. Haveman and B.Wolfe. (2000). Neighborhood Attributes as Determinants of Children's Outcomes: How Robust are the Relationships? *Journal of Human Resources,* 35 (4), 603-42.

Glass, D. (1954). *Social Mobility in Britain.* London: Routledge & Paul.

Glenn, C. and de J. Groof. (2005). *Balancing Freedom, Autonomy and Accountability in Education.* Nijmegan NL: Wolf Legal Publishers.

Goldhaber, D. and D. Brewer. (2000). Does Teacher Certification Matter? High School Teacher Certification Status and Student Achievement. *Educational Evaluation and Policy Analysis,* 22 (2), 129-145.

Goldstein, H. (1987). Multilevel Covariance Component Models. *Biometrika*, 74, 430-431.

Goldstein, H., J. Rasbash, M. Yang, G. Woodhouse, H. Pan, D. Nuttall and S. Thomas. (1993). A Multilevel Analysis of School Examination Results. *Oxford Review of Education*, 19 (4), 425-433.

Goldstein, H. and D. J. Spiegelhalter. (1996). League Tables and their Limitations: Statistical Issues in Comparison of Institutional Performance. *Journal of Royal Statistical Society,* Series A, Part 3, 385-443.

Goldstein, H. (1986). Multilevel Mixed Linear Model Analysis Using Iterative Generalised Least Squares. *Biometrika*, 73, 43-56.

Goldstein, H. (1997). Methods in School Effectiveness Research. *School Effectiveness and School Improvement*, 8, 369-95.

Goldstein, H., D. Kounali and A. Robinson. (2008). Modelling Measurement Errors and Category Mis-classifications in Multilevel Models. Accepted for publication.

Gorard, S., J. Fitz, and C. Taylor. (2001). School Choice Impacts: What Do We Know? *Educational Researcher,* 30 (7), 18-23.

Gray, J., D. Jesson, H. Goldstein, K. Hedger and J. Rasbash. (1995). A Multilevel Analysis of School Improvement: Changes in Schools' Performance Over Time. *School Effectiveness and School Improvement*, 6 (2), 97-114.

Hægeland, T. (2006). *School Performance Indicators in Norway: A Background Report for the OECD Project on the Development of Value-added Models in Education Systems.*

Hægeland, T., L. Kirkebøen, O. Raaum and K.Salvanes. (2005). *School performance indicators for Oslo, Reports 2005/36.* Statistics Norway.

Hægeland, T. and L. Kirkebøen. (2008). School Performance and Value-Added Indicators – What is the Importance of Controlling for Socioeconomic Background?: A Background Report for the OECD Project on the Development of Value-added Models in Education Systems.

Hambleton, R. K. and M. J. Pitoniak. (2006). Setting Performance Standards. In R. Brennan, *Educational measurement (4th ed.)* (pp. 433-470). Washington D.C.: American Council on Education.

Haney, W. and Raczek, A. (1993) *Surmounting outcomes accountability in education*. Washington, DC: U.S. Congress Office of Technology Assessment.

Hanushek, E. A. and M. E. Raymond. (2004). The Effect of School Accountability Systems on the Level and Distribution of Student Achievement. *Journal of the European Economic Association*, 2, 406-415.

Harris, D., A. Hendrickson, Y. Tong, S-H. Shin and C-Y Shyu. (2004). Vertical Scales and the Measurement of Growth. *Paper presented at the 2004 annual meeting of the National Council on Measurement in Education*, April. San Diego, CA.

Haveman, R. and B.Wolfe. (1995). The Determinants of Children's Attainments: A Review of Methods and Findings. *Journal of Economic Literature*, 33, 1829-1878.

Hill, R., B. Gong, S. Marion and C. DePascale (2005). Using Value Tables to Explicitly Value Student Growth. http://www.nciea.org/cgi-bin/pubspage.cgi?sortby=pub_date, accessed 10 January 2006.

Hoxby, C. (2003). *The Economics of School Choice, National Bureau of Economic Research Conference Report.* University of Chicago Press.

IGE. (2001). *Avaliação Integrada das escolas. Relatório Nacional. Ano lectivo 1999-2000.* Inspecção Geral da Educação, Ministério da Educação.

Jacob, B. (2002). *Accountability, Incentives and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools.* Cambridge, MA.: NBER Working Paper No. 8968.

Jakubowski, M. (2007). Volatility of Value-Added Estimates of School Effectiveness: A Comparative Study of Poland and Slovenia. *Paper presented to the Robert Shurman Centre for Advanced Studies, European University.* Florence.

Jakubowski, M. (2008). Implementing Value-Added Models of School Assessment. *RSCAS Working Papers 2008/06, European University Institute*.

Kane, T.J. and D.O. Staiger. (2002). Volatility in School Test Scores: Implications for Test-Based Accountability Systems. In D. R. (Ed.), *Brookings Papers on Education Policy* (pp. 235-269). Washington, DC: Brookings Institution.

Kohn, A. (2000). *The Case Against Standardized Testing: Raising the Scores, Ruining the Schools.* Portsmouth, NH: Heineman.

Kolen, M. and R. Brennan. (2004). *Test Equating, Scaling and Linking: Methods and Practices.* New York, NY: Springer Science and Business Media.

Koretz, D. (2005). Alignment, High Stakes, and the Inflation of Test Scores. In J. L. Herman and E. H. Haertel (ed.), *Uses and Misuses of Data for Educational Accountability and Improvement* (pp. 99-118). Malden, MA: NSSE.

Kreft, I. and J. De Leeuw. (1998). *Introducing Multilevel Modelling.* London, Thousand Oaks and New Delhi: Sage Publications.

Ladd, H. F. and R. P. Walsh. (2002). Implementing Value-Added Measures of School Effectiveness: Getting the Incentives Right. *Economics of Education Review,* 21, 1-17.

Lavy, V. (2002). Evaluating the Effects of Teachers' Group Performance Incentives on Pupil Achievement. *Journal of Political Economy,* 110, 1286-1317.

Lazear, E.P. (2000). The Future of Personnel Economics. *The Economic Journal,* 110, 467, F611-F639.

Levacic, R. (2001). An Analysis of Competition and its Impact on Secondary School Examination Performance in England. *Occassional Paper No. 34, September*. National Centre for the Study of Privatisation in Education, Teachers College, Columbia University.

Linn, R. L. (2005). Conflicting demands of "No Child Left Behind" and state systems: Mixed messages about school performance, *Education Policy Analysis Archives,*13(33).

Linn, R. L. (2004). *Rethinking the No Child Left Behind accountability system*. Washington, DC. Available online at http://www.ctredpol.org: Paper presented at the Center for Education Policy Forum.

Martineau, J. A. (2006). Distorting value added: The use of longitudinal, vertically scaled student achievement data for growth-based, value-added accountability. *Journal of Educational and Behavioral Statistics,* 31(1), 35-62.

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

Lissitz, R., H. Doran, W. Schafer and J.Willhoft. (2006). Growth Modelling, Value-Added Modelling and Linking: An Introduction. In R. Lissitz, *Longitudinal and Value-Added Models of Student Performance* (pp. 1-46). Mapple Grove, MN: JAM Press.

Little, R. J. A. and D. B. Rubin. (1987). *Statistical Analysis with Missing Data.* New York: Wiley.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L., Stecher, B., Le, V., and Martinez, F. (2007). The Sensitivity of Value-Added Teacher Effect Estimates to Different Mathematics Achievement Measures. *Journal of Educational Measurement,* 44(1), 45-65.

Lockwood, J.R., and D.F. McCaffrey. (2007). Controlling for Individual Level Heterogeneity in Longitudinal Models, with Applications to Student Achievement. *Electronic Journal of Statistics*, 1, 223-252.

Lucas, R. (1988). On the Mechanics of Economic Development. *Journal of Monetary Economics,* 22 (1), 3-42.

Madaus, G., P.W. Airasian and T. Kellaghan. (1980). *School Effectiveness: A Reassessment of the Evidence*. New York: McGraw-Hill.

Mante, B. and G. O'Brien. (2002). Efficiency Measurement of Australian Public Sector Organisations: The Case of State Secondary Schools in Victoria. *Journal of Educational Administration*, 30 (7), 274-91.

Mayer, C. (1996). Does Location Matter? *New England Economic Review,* May/June, 26-40.

McCaffrey, D. F., Lockwood, J. R., Mariano, L. T. and C. Setodji, (2005). Challenges for value-added assessment of teacher effects. In R. Lissitz (Ed.) *Value added models in education: Theory and practice*. Maple Grove, MN: JAM Press.

McCaffrey, D. F., J. R. Lockwood, D. M. Koretz and L. S. Hamilton. (2003). *Evaluating Value-Added Models for Teacher Accountability.* Santa Monica, CA: The RAND Corporation.

McCaffrey, D. M., J. R. Lockwood, D. Koretz, T. A. Louis and L. Hamilton. (2004). Models for Value-Added Modelling of Teacher Effects. *Journal of Educational and Behavioral Statistics,* 29 (1), 67-101.

McCall, M. S., Kingsbury, G. G. and A. Olson. (2004). *Individual Growth and School Success.* Lake Oswego, OR: Northwest Evaluation Association.

McKewen, N. (1995). Accountability in Education in Canada. *Canadian Journal of Education,* 20 (1).

Messick, S. (1989). Validity. In R. Linn. (Ed.), *Educational Measurement.* Washington, DC: American Council on Education.

Meyer, R. (1997). Value-Added Indicators of School Performance: A Primer. *Economics of Education Review,* 16 (3), 283-301.

Ministry of National Education, Higher Education and Research, Direction de l'évaluation, de la performance et de la prospective. (2006). *Lycée Performance Indicators: 2005 general, technological and vocational baccalauréats: A Background Report for the OECD Project on the Development of Value-added Models in Education Systems.*

NASBE. (2005). *Evaluating Value-Added: Findings and Recommendations from the NASBE Study Group on Value-Added Assessments.* Alexandria, VA: National Association of State Boards of Education.

Nichols, S.L. and Berliner, D.C. *The Inevitable Corruption of Indicators of Educators through High-stakes testing,*Tempe, AZ: Education Policy Reserarch Unit, Arizona State University.

O'Day, J. (2002). Complexity, Accountability, and School Improvement. *Harvard Educational Review,* 72, (3), 293-329.

Odden, A. and Busch, C. (1998). *Financing Schools for High Performance.* San Francisco: Jossey-Bass.

OECD. (1994). *The OECD Jobs Strategy: Evidence and Explanations.* Paris: OECD.

OECD. (1996). *Lifelong Learning for All.* Paris: OECD.

OECD. (2001). *The New Economy: Beyond the Hype*. Paris: OECD.

OECD. (2004). *Learning for Tomorrow's World: First Results from PISA 2003*. Paris: OECD.

OECD. (2005). *Teachers Matter: Attracting, Developing and Retaining Effective Teachers.* Paris: OECD.

OECD. (2006). *Demand Sensitive Schooling? Evidence and Issues.* Paris: OECD.

OECD. (2007a). *Education at a Glance.* Paris: OECD.

OECD. (2007b). *Learning for Tomorrow*. Paris: OECD.

OECD. (2007c). *No More Failures: Ten Steps to Equity in Education.* Paris: OECD.

OECD. (2007d). *PISA 2006: Science Competencies for Tomorrow's World.* Paris: OECD.

OECD. (2008). *Going for Growth.* Paris: OECD.

Patz, R. (2007). *Vertical Scaling in Standards-Based Educational Assessment and Accountability Systems.* Washington D.C.: The Council of Chief State School Officers.

Ponisciak, P. M. and A. S. Bryk. (2005). Value-Added Analysis of the Chicago Public Schools: An Application of Hierarchical Models. In R. L. (Ed.), *Value Added Models in Education: Theory and Applications.* Maple Grove, MN: JAM Press.

Raudenbush, S. and J.D. Willms. (1995). The Estimation of School Effects. *Journal of Educational and Behavioral Statistics*, 20, 307-335.

Raudenbush, S. and A. Bryk. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods (2nd Edition).* Newbury Park, CA: Sage Publications.

Raudenbush, S. W. (2004). *Schooling, Statistics, and Poverty: Can We Measure School Improvement?* Princeton, NJ: Educational Testing Service.

Ray, A. (2006). *School Value Added Measures in England: A Background Report for the OECD Project on the Development of Value-Added Models in Education Systems*, www.dcsf.gov.uk/rsgateway/DB/RRP/u015013/index.shtml.

Ray, A. (2007). *The Volatility of Value-Added Scores: A Background Report for the OECD Project on the Development of Value-Added Models in Education Systems*, unpublished.

Reel, M. (2006), presentation given at the ETS National Forum on State Assessment and Student Achievement, Education Testing Service, Princeton, 13-15 September.

Romer, P. (1994). Endogenous Economic Growth, *Journal of Economic Perspectives*, 8 (1), 3-22.

Rowan, B., R. Correnti and R. J. Miller (2002). What Large-Scale Survey Research Tells Us About Teacher Effects on Student Achievement: Insights from the Prospects Study of Elementary Schools. *Teacher College Record*, 104, 1525-1567.

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, 63, 581-592.

Rubin, D., E. Stuart and E. Zanutto. (2004). A Potential Outcomes View of Value-Added Assessment in Education. *Journal of Educational and Behavioural Statistics*, 103-116.

Ryska, R. (2006). *Value-added Modelling in the Czech Republic: A Background Report for the OECD Project on the Devlopment of Value-added Models in Education Systems.*

Sammons, P. T. (1997). *Forging Links: Effective Schools and Effective Departments.* Paul Chapman Publishing Lda.

Sammons, P., S. Thomas, P. Mortimore, C. Owen and H. Pennell. (1994). *Assessing School Effectiveness: Developing Measures to put School Performance in Context.* London: Office for Standards in Education.

Sanders, W., A. Saxton, and B. Horn. (1997). The Tennessee Value-Added Assessment System: A Quantitative Outcomes-Based Approach to Educational Assessment. In J. M. (Ed.), *Grading Teachers, Grading Schools: Is Student Achievement a Valid Evaluational Measure?* (pp. 137-162). Thousand Oaks, CA: Corwin Press, Inc.

Sass, T., and D. Harris. (2007). *The Effects of NBPTS-Certified Teachers on Student Achievement.* CALDER Working Paper No. 4.

Saunders, L. (2000). Understanding Schools Use of 'Value Added' Data: The Psychology and Sociology of Numbers. *Research Papers in Education,* 15 (3), 241-58.

SCAA. (1994). *Value Added Performance Indicators for Schools.* London: School Curriculum and Assessment Authority.

Senge, P. (2000). *Schools that Learn: A Fifth Discipline Fieldbook for Educators, Parents, and Everyone Who Cares About Education.* New York, NY: Doubleday.

Snijders, T.A.B., and R.J. Bosker. (1999). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modelling.* Londen: Sage.

Taylor, J. and N.A. Nguyen. (2006). An Analysis of the Value Added by Secondary Schools in England: Is the Value Added Indicator of Any Value? *Oxford Bulletin of Economics and Statistics,* 68(2), 203-224.

Tekwe, C., R. Carter, C. Ma, J. Algina, M. Lucas and J. Roth. (2004). An Empirical Comparison of Statistical Models for Value-Added Assessment of School Performance. *Journal of Educational and Behavioral Statistics,* 29 (1), 11-36.

Thomas, S. and Mortimore, P. (1996). Comparison of Value-Added Models for Secondary School Effectiveness. *Research Papers in Education,* 11 (1), 5-33.

Thomas, S., Peng, W-J. and Gray, J. (2007). Value Added Trends in English Secondary School Performance Over Ten Years. *Oxford Review of Education,* 33 (3), in press.

Tymms, P. and C. Dean. (2004). *'Value Added in the Primary School League Tables', A Report for the National Association of Head Teachers.* May. Durham: CEM Centre, University of Durham.

van de Grift, W. (2007). *Reliability and Validity in Measuring the Added Value of Schools: A Background Report for the OECD Project on the Development of Value-Added Models in Education Systems.*

Vicente, P. (2007). O plano amostral do projecto 3EM. Actas do XIV Congresso Anual da Sociedade Portuguesa de Estatística. In M. N. Ferrão, *Proceedings of the XIV Annual Conference of the Portuguese Statistical Society.* Lisboa: SPE, Accepted for publication.

Vignoles, A., R. Levacic, J. Walker, S. Machin and D. Reynolds. (2000). *The Relationship Between Resource Allocation and Pupil Attainment: A Review.* London: Centre for the Economics of Education, London School of Economics.

Webster, W. J. (2005). The Dallas School-Level Accountability Model: The Marriage of Status and Value-Added Approaches. In R. L. (ed.), *Value added models in education: Theory and Applications.* Maple Grove, MN: JAM Press.

Webster, W. and R. Mendro. (1997). The Dallas Value-Added Accountability System. In J. M. (ed.), *Grading Teachers, Grading Schools: Is Student Achievement a Valid Evaluation Measure?* (pp. 81-99). Thousand Oaks, CA: Corwin Press.

Wikeley, F. (1998). Dissemination of Research as a Tool for School Improvement. *School Leadership and Management,* 18 (1), 59-73.

Willms, J.,and Raudenbush, S. (1989, 26(3)). A longitudinal hierarchical linear model for estimating school effects and their stability. *Journal of Educational Measurement*, 209-232.

Wilson, D. (2004). Which Ranking? The Impact of a ‘Value-Added’ Measure of Secondary School Performance. *Public Money and Management.* January. 37-45.

Wright, S., W. Sanders and J. Rivers. (2006). Measurement of Academic Growth of Individual Students toward Variable and Meaningful Academic Standards. In R. Lissitz, *Longitudinal and Value-Added Models of Student Performance* (pp. 385-406). Maple Grove, MN: JAM Press.

Yang, M., H. Goldstein, T. Rath and N. Hill. (1999). The Use of Assessment Data for School Improvement Purposes. *Oxford Review of Education,* 25 (4), 469-83.

Zvoch, K. and J. Stevens. (2006). Successive Student Cohorts and Lonigtudinal Growth Models: An Investigation of Elementary School Mathematics Performance. *Education Policy Analysis Archives,* 14 (2).

OECD PUBLICATIONS, 2, rue André-Pascal, 75775 PARIS CEDEX 16
PRINTED IN FRANCE
(96 2008 10 1 P) ISBN 978-92-64-05022-8 – 56449 2008

Measuring Improvements in Learning Outcomes

BEST PRACTICES TO ASSESS THE VALUE-ADDED OF SCHOOLS

With education systems in all OECD countries coming under increasing pressure to enhance their effectiveness and efficiency, there is a growing recognition of the need for accurate school performance measures. But how can we measure their performance in an accurate way? Raw test scores and their ranking tend to reflect students' socio-economic status. Value-added modelling is different and focuses upon progress in student performance. It refers to a class of statistical models that estimate the contributions of schools to student progress in stated or prescribed education objectives (*e.g.* cognitive achievement) measured at at least two points in time.

Value-added estimates are a significant improvement upon measures of school performance currently used in most education systems across OECD countries. They provide a fundamentally more accurate and valuable quantitative basis for school improvement planning, policy development and for enacting effective school accountability arrangements. Without an accurate performance measure, equitable outcomes and efficient policy responses can be compromised as resources are not directed to where they are most needed. Policies and practices cannot be improved if it is not known what has proven to be effective. This is where value-added modelling plays an essential role. It provides a more accurate measure of school performance, overcoming many of the problems that plague other measures, which can be biased against schools serving more socio-economically disadvantaged students.

This groundbreaking report is essential reading for anyone interested in school performance.

The full text of this book is available on line via this link:
www.sourceoecd.org/educationt/9789264050228

Those with access to all OECD books on line should use this link:
www.sourceoecd.org/9789264050228

SourceOECD is the OECD online library of books, periodicals and statistical databases. For more information about this award-winning service and free trials, ask your librarian, or write to us at **SourceOECD@oecd.org**.

OECD*publishing*
www.oecd.org/publishing

ISBN 978-92-64-05022-8
96 2008 10 1 P

9 789264 050228