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.

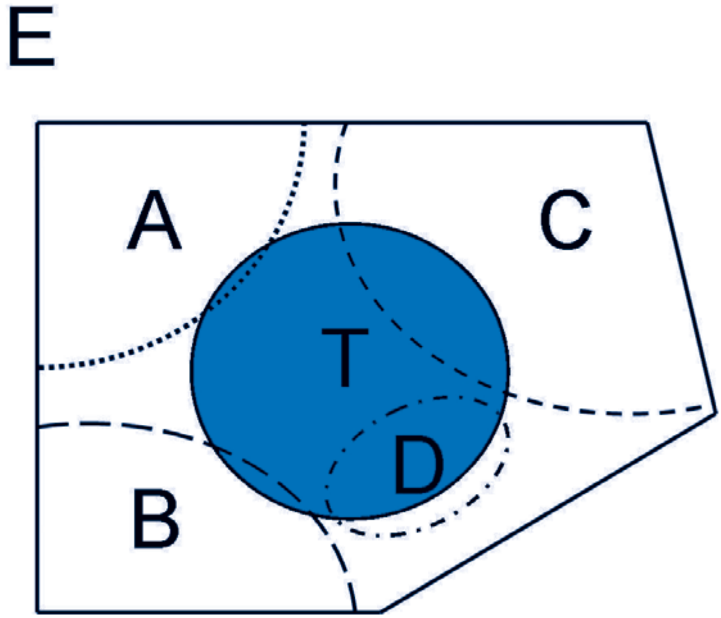
As discussed in Part I, the use of test results for high-stakes purposes can create incentives to influence student performance on these assessments in a sub-optimal manner. The practice of 'teaching to the test' is one such undesirable consequence but there are a number of documented instances where 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:

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

where y_{ij} is the test score result for student i in school j; x_{ij} is the student prior achievement; $\overline{x}j$ 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:

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

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

$$y_{ij} = b_0 + b_1 X_{1ij} + ... + b_p X_{pi\,j} + \varepsilon_{ij} \qquad (5)$$

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:

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

and level 2 takes the form:

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

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

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

In level 1:

i indexes students within schools j,

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

$\tilde{P}_{ij}^{1}$ and $\tilde{P}_{ij}^{2}$ 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:

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

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 $\underline{\varepsilon_i} = \left(\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3}\right)$ and $\underline{\delta_i} = \left(\delta_{i1}, \delta_{i2}, \delta_{i3}\right)$, then $\left(\underline{\varepsilon_i}, \underline{\delta_i}\right)$ 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, $\left(\underline{\varepsilon_i}, \underline{\delta_i}\right)$ 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:

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

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.

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.

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.

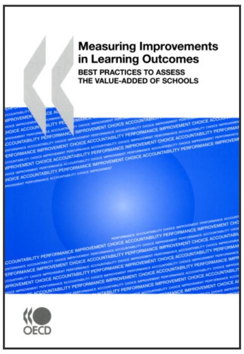
Zvoch, K. and J. Stevens. (2006). Successive Student Cohorts and Lonigtudinal Growth Models: An Investigation of Elementary School Mathematics Performance. *Education Policy Analysis Archives,* 14 (2).

Table of Contents

List of Figures

List of Tables

List of Boxes


Measuring Improvements in Learning Outcomes
BEST PRACTICES TO ASSESS THE VALUE-ADDED OF SCHOOLS
OECD

From:

Measuring Improvements in Learning Outcomes

Best Practices to Assess the Value-Added of Schools

Access the complete publication at:

https://doi.org/10.1787/9789264050259-en

Please cite this chapter as:

OECD (2008), "The Design of Value-Added Models", in *Measuring Improvements in Learning Outcomes: Best Practices to Assess the Value-Added of Schools*, OECD Publishing, Paris.

DOI: https://doi.org/10.1787/9789264050259-3-en