LITERATURE REVIEW ON
THE VALUE-ADDED MEASUREMENT IN HIGHER EDUCATION

HoonHo Kim and Diane Lalancette

TABLE OF CONTENTS

1. Introduction

1.1 Value-added measurement in the context of an AHELO feasibility study

The Organisation for Economic Co-operation and Development (OECD) conducted a feasibility study on the international Assessment of Higher Education Learning Outcomes (AHELO). AHELO emerged from a meeting, held in Athens in 2006, among OECD Education Ministers who expressed the need to develop better evidence of learning outcomes in higher education. A series of experts meetings followed in 2007 leading to the recommendation to carry out a feasibility study to assess learning outcomes.

The goal of the AHELO feasibility study was to determine whether an international assessment of higher education learning outcomes is scientifically and practically possible. Based on the recommendations that have resulted from the expert groups meetings conducted in 2007, and given its purpose and underlying motivation, the AHELO feasibility study has been designed with two key aims:

- Test the science of the assessment: whether it is possible to devise an assessment of higher education outcomes and collect contextual data that facilitates valid and reliable statements about the performance/effectiveness of learning in higher education institutions of very different types, and in countries with different cultures and languages.
- Test the practicality of implementation: whether it is possible to motivate institutions and students to take part in such an assessment and develop appropriate institutional guidelines.

The first phase in exploring the feasibility of carrying out an international assessment of higher education learning outcomes was to determine whether adequate assessment instruments can be successfully developed and administered for the purpose of the feasibility study. Three assessments were developed to examine the feasibility of capturing different types of learning outcomes. One looks at generic skills that students in all fields should be acquiring while the other two focus on skills that are specific to disciplines. Engineering and economics were chosen for this feasibility study. Along with each of these three tests, contextual information is collected from students, relevant faculty and from participating institutions' leaders. These contextual surveys were designed to identify factors that may help to explain differences in observed learning outcomes of the target population and offer insights for interpretation.

The second phase in exploring the feasibility was to implement the developed instruments and surveys in a diversity of countries, languages, and institutions to explore the feasibility of implementation. With more than 270 higher education institutions in 17 participating countries, tests were administered to students nearing the end of their Bachelor's degree programme in one, two or three of the strands while all institutions also administered contextual surveys. Data collection is now completed and the results of the study will be presented in a report on the scientific and practical feasibility of AHELO by December 2012.

A complementary phase to the feasibility study was to explore methodologies and approaches to capture value-added, or the contribution of higher education institutions to students' outcomes, irrespective of students' incoming abilities. The purpose of adding this phase, the value-added measurement strand, was to review and analyse possible methods for capturing the learning gain that can be attributed to higher education institutions' attendance. The work conducted in this strand builds upon similar work carried out at school level by the OECD (OECD 2008) to review options for value-added measurement in higher education. The intent is to bring together researchers to study methodologies with a view to providing guidance towards the development of a value-added measurement approach for a fully-fledged AHELO main study.

1.2 Purpose of this literature review

Value-added models can be used to evaluate, monitor, and improve an institution and/or other aspects of an education system. However, the use of statistical models to measure the value-added or marginal learning gain raises a number of scientific and practical issues imposing layers of complexity that, though theoretically well understood, are difficult to resolve in large scale assessments (OECD, 2008).

Understanding the characteristics and the fundamental differences between existing value-added models is essential as there are many advantages and disadvantages to each of the various models. Furthermore, accurate estimates can only be made when using the most appropriate and suitable value-added model given the data properties and the policy objectives.

This report reviews existing literature on value-added measurement approaches, methodologies, and challenges within both the K-12 (primary and secondary education) and the higher education contexts, albeit with greater emphasis on methodologies developed for the latter[1]. More concretely, it sets out the properties of different value-added models, how they are different from each other, and how they handle statistical and technical issues within their modelling procedures. This report also reviews the criteria for choosing an appropriate model in order to provide recommendations for future development.

2. Understanding value-added measurement

2.1 Definition of "value-added" and "value-added modelling"

Although in many countries, performance of educational institutions have mainly focused on student attainment measures, such as the average score on standardised test or the percentage of students in each school progressing to higher levels of education (OECD, 2008), student achievement can also be measured as growth (Teacher Advancement Program, 2012).

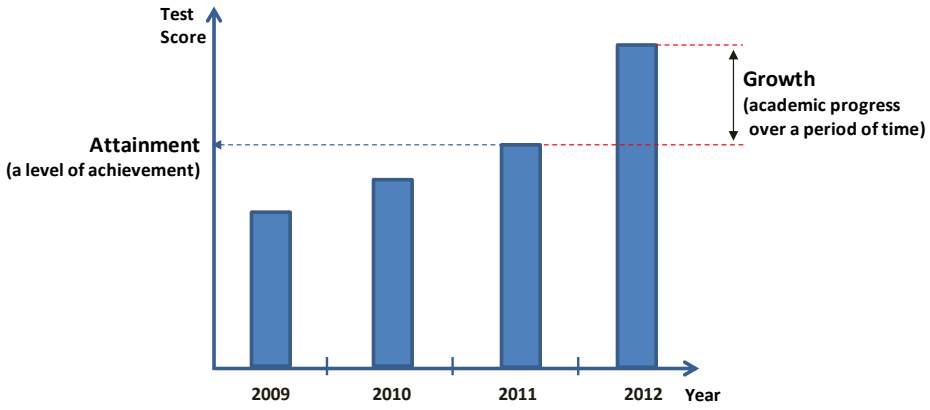

Figure 1: Attainment and growth: two different ways to measure student achievement

Attainment refers to the levels of achievement students reach at a point in time, *e.g.* on a standardised test at the end of a given school year. Academic attainment levels, usually represented by numerical scores or standards of achievement, are typically used to rate institutional performance. In contrast, growth relates to the academic gain or progress students make over a period of time (*e.g.* on a standardised test administered over several grades).

The concept of "value-added" in an education system relates to student achievement as growth in knowledge, skills, abilities, and other attributes that students have gained as a result of their experiences in an education system over time (Harvey, 2004-12). From the point of view of the educational institution, value-added could also be defined as the contribution of schools or higher education institutions (HEIs) to students' progress towards stated or prescribed education objectives over time (OECD, 2008).

"Value-added modelling" can be defined as a category of statistical models that use student achievement data over time to measure students' learning gain. As Doran and Lockwood (2006) reported, the value-added models answer research questions such as:

- what proportion of the observed variance in student achievement can be attributed to a school or teacher?
- how effective is an individual school or teacher at producing gains?
- which characteristics or institutional practices are associated with effective schools?

According to the definition of value-added modelling provided above, statistical analyses undertaken in a number of countries to monitor the performance of educational institutions cannot be considered as value-added measurement. Although many countries measure student achievement regularly, in many cases they do not focus on the changes in student achievement over time but rather on the differences in student achievement between schools in a given school year for the purpose of identifying high-achieving schools (OECD, 2008; Doran & Izumi, 2004).

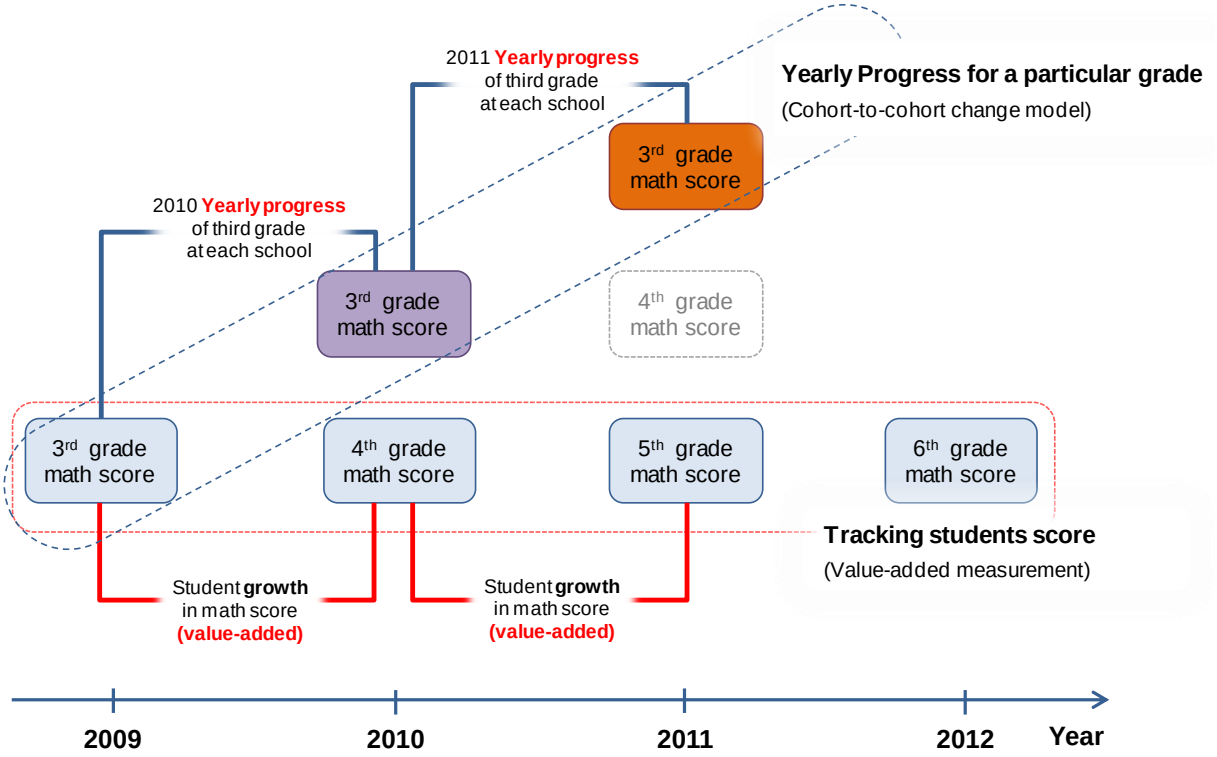

Figure 2: Comparison between yearly progress of particular grade and student growth

As shown in the upper part of Figure 2, some countries measure yearly progress of student achievement based on comparisons of test scores for a given grade in a given subject over years (*e.g.* Adequate Yearly Progress in the United States). This cohort-to-cohort change model is not considered value-added measurement as it does not measure the change in student achievement from a given grade to previous (or subsequent) grade. The cohort-to-cohort change model only refers to the changes in mean test scores for a particular grade over time, and do not reflect student academic growth by attending school over time.

2.2 Benefits of using value-added measurement

Value-added measurement provides additional indicators of institutional performance beyond student attainment levels at one point in time, which is commonly used in many countries. Positive aspects of value-added measurement can be categorized into the following two benefits:

- Value-added measurement provides a 'fairer' estimate of the contribution educational institutions make to students' academic progress as it tracks the same student over time taking into

consideration the initial achievement level of students as they begin the school year (Teacher Advancement Program, 2012; OECD, 2008; Doran & Izumi, 2004).

Value-added measurement focuses on the change in students' scores over a given time period instead of scores collected at a specific point in time (Teacher Advancement Program, 2012; OECD, 2008; Sanders, 2006; Raudenbush, 2004; Tekwe et al., 2004). It would be unfair to evaluate each institution's contribution to student achievement by only looking at attainment levels, or percentage of students meeting certain standards, as the skills and knowledge of students entering an educational institution vary greatly (Reardon & Raudenbush, 2009).

In a scenario where students enter an institution with comparatively low levels of cognitive skills, despite a significant increase in students' scores, the institution may still not be recognized as an effective institution if it does not meet a minimum success rate as the evaluation of the institution performance only takes into account the attainment level, and not the growth in student achievement.

- Value-added measurement provides a more 'accurate' estimate of the contribution educational institutions make to students' academic progress as it incorporates a set of contextual characteristics of students or institutions (Teacher Advancement Program, 2012; OECD, 2008).

  Although comparisons of raw test scores provide important information, they are poor measures of institutional performance in failing to produce results that can reflect differences in contextual characteristics such as students' socio-economic backgrounds. By evaluating only one score (*i.e.* the attainment on a standardised test at one point in time), it is difficult to identify to what extent that score was influenced by factors outside of the institution as compared with other factors that can be controlled within the institution.

  In contrast, value-added measurement can estimate the contribution of educational institutions to students' academic progress by isolating student attainment from other contributing factors such as family characteristics and socio-economic background over the course of a school year or another period of time (Teacher Advancement Program, 2012; Sanders, 2006; Braun, 2005a; Raudenbush, 2004; Tekwe et al., 2004; McCaffrey et al., 2003).

Even though fairer and greater accuracy may be obtained using value-added measurement, some difficulties remain in measuring the effects an institution might have on student achievement. Above all, value-added measurement based on the results of standardised tests can measure only part of an institution's effects. The education happening in an institution translates into accumulated knowledge, skills, customs, and ethical (or social) values, but also has effects on the way students think, feel, or act. What standardised tests usually measure refers to skills, specific facts and functions that cannot reflect the entire learning happening in an institution (Bennett, 2001; Harvey & Green, 1993). Additionally, in theory, the effects an institution might have on student education may only be revealed years later, which would require also assessing value-added later, with alumni in addition to with graduating students. In any case, the complexity of the education environment requires that interpretation of institutions' value-added scores includes various caveats for fair and correct interpretation (OECD, 2008).

3. Overview of value-added modelling used in education systems

3.1 Value-added modelling in K-12 education

Throughout the 1990s, schools were held increasingly accountable for student learning outcomes (Braun, 2005b), and many countries, especially OECD countries, are under ever more pressure to enhance schools' effectiveness and efficiency (OECD, 2008). The emphasis in K-12 education shifted from input measures, such as teacher-pupil ratio or expenditure per pupil, to output evaluations, such as determining whether students met the standards set by a state or nation. Therefore, there has recently been growing recognition of the need to develop accurate school performance measures (OECD, 2008). The assessments of student achievement at

the state-level or the national-level are now common in many countries. The results are often widely reported and used in public debate as well as for school improvement purposes.

Value-added measurement in K-12 education is rooted in a series of school effects research which began, at least in the United States, with the Coleman Report that studied the relations of schools and families to student academic attainment (OECD, 2008; Coleman et al., 1966). At first, high-achieving schools were identified by comparing the students' average test scores. Subsequent studies on school effectiveness developed the analysis models of school mean test scores at a specific point in time taking into account relevant demographic characteristics of the students, such as socio-economic background (Haveman & Wolfe, 1995) and the hierarchical structure of school systems (Aitkin & Longford, 1986; Willms & Raudenbush, 1989). These sophisticated cross-sectional models (*e.g.* education production functions) have been used to provide measures of school performance and to compare the resulting differences in school rankings (Hanushek, 2007; Burstein, 1980).

However, it was considered that such analyses on school effects did not contain the required analytic framework to be classified as value-added models because they depended on the test scores collected at a particular point in time and did not consider the differences in the initial achievement level of students between schools (OECD, 2008).

Thus there has been an increasing interest in the way to measure the performance of teachers, schools, and districts after controlling for the factors affecting student achievement such as student's entering academic ability and student composition (Hibpshman, 2004). In the mid-1980's, as a result of improvements in statistical methodology and available data, researchers began to use more advanced value-added models (Raudenbush & Bryk, 1986) making significant progress in school effect studies. Such development of value-added measurement led to the implementation of operational high-stakes teacher and school assessment systems in a number of OECD countries, including the United States (Tennessee, North Carolina, Ohio, etc.), United Kingdom, and Australia (Downes & Vindurampulle, 2007; Hibpshman, 2004).

While a number of different models have been implemented, the most commonly used, and those that have received the most attention, have been the mixed-model approach developed by William Sanders, the Tennessee Value-Added Assessment System (TVAAS) (Ballou et al., 2004; Hibpshman, 2004; Sanders & Horn, 1998, 1994) and the hierarchical linear models (HLM) introduced to reflect the multilevel (or nested) data structures and individual differences in growth curves over time in education research (Bryk & Raudenbush, 1988; Raudenbush & Bryk, 1986, 2002).

Almost all value-added models used in K-12 education employ data that track test score trajectories of individual students in one or more subjects, over one or more years (Goldstein et al., 1993; Sanders et al., 1997; Rowan et al., 2002; McCaffrey et al., 2004; Ponisciak & Bryk, 2005). Through various kinds of statistical adjustments, such student growth data can be transformed into indicators of school value-added (OECD, 2008).

Most of value-added models used in K-12 education use annual standardised test scores at the end of the school for individual students to assess student progress compared to the previous year's test scores in fundamental academic skills, and apply the results as a measure of the effectiveness of teachers and schools. In this respect, it is not surprising that value-added measurement research projects have multiplied in recent years since the annual standardised tests at state- or nation-level were administered (*e.g.* the *No Child Left Behind (NCLB) Act* of 2002 in the United States, which measures student achievement and sometimes requires teachers and schools to make annual, adequate achievement progress) (Goe et al., 2008).

3.2 Value-added modelling in higher education

In recent decades, even more emphasis is being placed on accountability in higher education. This can be explained by rising tuition costs, disappointing rates of retention and graduation, employers' concerns regarding insufficient knowledge and skills that are expected in the workplace, and the emerging fundamental

questions about the value that higher education provides to students (Leveille, 2006). Where the focus of the assessment is on accountability, institutions are required to demonstrate, with evidence, conformity with an established standard of process or outcomes (Ewell, 2009). Therefore there is greater reliance on quantitative evidence such as standardised tests and surveys, as the main objective is to compare institutions and/or programmes against fixed standards of achievement.

Along with demands for external accountability of higher education, higher education institutions have been under increasing pressure from governments, policymakers, and other stakeholders as well as students to improve the quality of education and to enhance the effectiveness of higher education (Liu, 2011; Ewell, 2009). Internally, institutions also need to measure achievement and track their own progress so that they can know where they stand, correct shortcomings in teaching, and improve the quality of education (Liu, 2011; Steedle, 2011). Assessment tools could include both quantitative and qualitative evidence-gathering instruments such as standardised and faculty-designed examinations, self-report surveys (*e.g.* National Survey of Student Engagement (NSSE) in the United States), capstone projects, demonstrations, portfolios, and specially designed assignments embedded in regular courses (Ewell, 2009). Although assessment results can be used to compare achievement amongst students (normative approach), in order to improve, the tracking over time or against established institutional goals could prove more useful (criterion-referenced approach).

In response to growing demands, both externally and internally, on the quality of education, many countries and higher education institutions now focus on the assessment of student learning outcomes (Ewell, 2009; Liu, 2011; Steedle, 2011). As an example, in the United States, approximately 25% of Association of American Colleges and Universities (AACU) member institutions are now administering standardised tests of high-order skills, such as communication, critical thinking, and problem solving (Hart Research Associates, 2009).

The value-added models used in higher education differ in many ways from the models used in K-12 education as the type of data available differs significantly. Almost all value-added models used in K-12 education are developed based on longitudinal data pertaining to the same students and the same subjects over years (Ballou et al., 2004; McCaffrey et al., 2004; Tekwe et al., 2004; OECD, 2008).

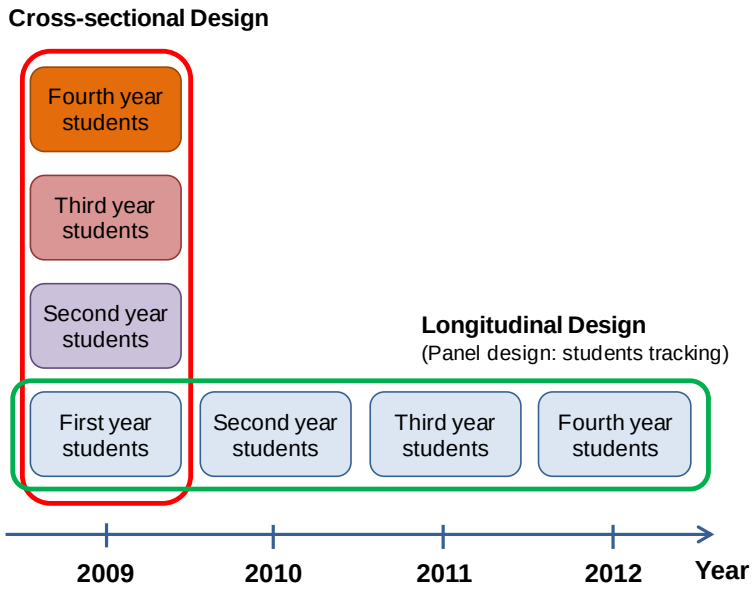

Figure 3: Cross-sectional and longitudinal design

However, such assessment conditions are rarely met in higher education. One major difference is the difficulty to track individual students in higher education due to a relatively high level of student mobility. Higher education students tend to change programmes, take leaves of absence, or even drop out of school halfway

through. For this, and other similar reasons, the longitudinal approach, with a repeated measures design often used in K-12 education, may not be logistically feasible or could be extraordinarily expensive in higher education, even when it is technically possible.

Few longitudinal studies have been conducted in higher education and only some of them include value-added modelling to assess education quality across programmes or institutions. In the United States, for example, two longitudinal studies with repeated measures design were implemented, the Lumina Longitudinal Study and the Wabash National Study.

The Lumina Longitudinal Study administered the Collegiate Learning Assessment (CLA) developed by the Council for Aid to Education (CAE) to a longitudinal cohort of entering freshmen and tested them at three different points in time. The freshmen were first tested in the fall of 2005, retested in the spring of 2006, and retested again in the spring of 2009, at which point they became graduating seniors. Nearly 50 colleges and over 11,000 students in the U.S. participated in the study. With the results from the longitudinal study, institutions could measure how well their students, as a group, perform relative to the sample of all other institutions participating in the study and admitting students of similar entering academic ability.

In addition to the use of the longitudinal approach with a repeated measures design, the Lumina Longitudinal Study also integrated a cross-sectional design in order to compare results from the two different designs. The cross-sectional design was applied to freshmen and seniors who took the test at the same time in 2006 (Klein et al., 2009). The results indicated that the score differences between the freshmen and the seniors obtained with the cross-sectional design were consistent with the score differences in the longitudinal cohort of freshmen that was retested as seniors in 2009. The report then concluded that the cross-sectional design seemed preferable, because the longitudinal data obtained with the repeated measures design did not generate more accurate results than the cross-sectional design despite the fact that it may take a lot of time and much cost to track the same students. However, as the Voluntary System of Accountability (2008) and Liu (2011) note, more empirical evidence is needed to compare results from both designs.

The Wabash National Study, another longitudinal research project implemented in the U.S., was designed to provide participating institutions with extensive evidence on teaching practices, student experiences, and institutional conditions to promote student growth across multiple outcomes (Blaich & Wise, 2011). This study measured student outcomes using 13 instruments including the Collegiate Assessment of Academic Proficiency (CAAP) from the American College Testing (ACT), along with student experiences using two experience surveys including the National Survey of Student Engagement (NSSE). Although the study started in fall 2006 with 4,501 first-year students from 19 higher education institutions, more than 49 institutions and 17,000 students from three cohorts have since participated in the study. At each institution, random samples of students were assessed three times: twice in the fall and spring of their first-year in 2006, and once again in the spring at the end of the fourth-year in 2009. Twenty-nine colleges and universities joined the new version of the Wabash Study in fall 2010, followed by the Wabash National Study 2006-2009. This study will last three years ending in the fall of 2013.

Unlike these two research studies using the longitudinal data with a repeated measures design, many other value-added models used in higher education, at least in the United States, are based on a cross-sectional design (Klein et al., 2007; Steedle, 2009, 2011; Liu, 2011), also called “contextualized attainment models (Lenkeit, 2012; Ray et al., 2009)”. In a cross-sectional design, value-added scores are calculated based on the difference between ‘observed’ mean score from raw data and ‘expected’ mean score from the linear regression (*i.e.* residuals) (Klein et al., 2007; Liu, 2011). Schools with large positive residuals are considered to be more effective. On the other hand, schools with large negative residuals are considered to be problematic and require further improvement.

It should be noted that the value-added models in higher education using a cross-sectional design require additional data such as entering academic ability scores (*e.g.* SAT or ACT scores used as a standardised test for college admissions in the United States) to control for initial status in addition to the general background of students affecting the test score, while the longitudinal design uses test scores collected over time.

Recently, an alternative approach to measure value-added measurement using the cross-section design was proposed. This alternative model employs the hierarchical linear models in estimating institutions' value-added scores, incorporating two levels of analysis, the student- and institution-level. The model allows for differentiating between factors within- and between-institutions explaining senior student achievement variance (Steedle, 2009) (for more information, see 4.2.3 HLM-based residual analysis model).

4. Illustrative value-added models

The selection of a model for value-added measurement in education will vary depending on the types of data available for analysis, even though the differences in system and policy objectives of value-added measurement surely have some effects (Liu, 2011; Ewell, 2009; Steedle, 2009). In K-12 education, in many countries, students take a standardised achievement test every year. The results are accumulated with a variety of school contextual variables and student backgrounds. Because of sufficient data available, almost all value-added models used in K-12 education are developed to analyse the longitudinal data pertaining to same students and same subjects over years (Ballou et al., 2004; McCaffrey et al., 2004; Tekwe et al., 2004; OECD, 2008), and therefore there is a wider variety of value-added models than in higher education.

In higher education, most value-added models, other than those longitudinal research projects such as the Lumina Longitudinal Study and the Wabash National Study, use the cross-sectional design where the same test is administered to incoming freshmen and graduating seniors in a given institution at the same time (Klein et al., 2007; Steedle, 2009, 2011; Liu, 2011). As a result of limitations on available data, research on institution's contribution to student achievement in higher education is greatly restricted, and therefore limits the development of value-added models that could be used in the context of higher education.

This section includes a brief review of the statistical and methodological properties of a selection of value-added models along with their advantages and disadvantages. The intent is to illustrate some properties of the models and how specific issues are handled within different modelling procedures rather than to provide an exhaustive review of all different types of value-added models. The value-added models presented here are classified by system level and types of measurement design.

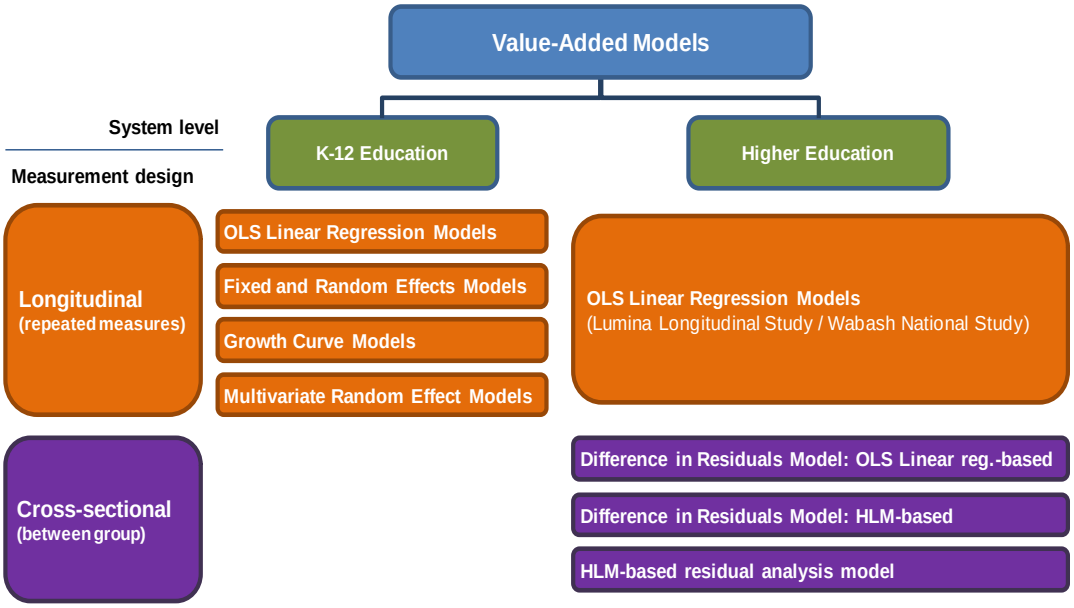

Figure 4: The structure of the value-added models in education system

For the purpose of the AHELO feasibility study which endeavours to measure student learning outcomes and the contribution of an institution to student academic growth, this report focuses on estimating school effects, rather than teacher effects.

4.1 Models used in K-12 education

Since the State of Tennessee (USA) enacted the use of Value-Added Assessment (VAA) in 1992, the use of the value-added modelling has been expanded to estimate teacher, school and district effects on student achievement in K-12 education (OECD, 2008; Sanders, 2006; Wainer, 2004; Sanders & Horn, 1998). Today, a variety of value-added models are used in K-12 education and the specificities of their analysis methods will depend on the effects (teachers, schools and districts) the researchers choose to investigate.

In this section, four general categories of value-added models are discussed: i) ordinary least squares (OLS) linear regression models, ii) fixed and random effects models, iii) growth curve models, and iv) multivariate random effects models.

4.1.1 OLS linear regression models

The OLS linear regression models are typically used in K-12 education to adjust student test scores for students' prior test scores and student or school characteristics (Jakubowski, 2008; OECD, 2008; McCaffrey et al., 2003; Ladd & Walsh, 2002). These approaches are also called "covariate adjustment models" because they specify the current score as a function of the prior score and possibly other covariates, using separate models for each year and explicitly linking students' scores to the effects of only their current school (McCaffrey et al., 2003).

In these approaches, it is assumed that the regression coefficients are the same for all schools (Gujarati & Porter, 2009). The common models specify current scores as linear functions of the covariates as follows (OECD, 2008):

$$y_{ij(2)} = \beta_0 + \beta_1 y_{ij(1)} + \beta_2 X_{ij} + \beta_3 X_j + \varepsilon_{ij} \quad (1)$$

where

$y_{ij(2)}$: the current test score of student i within school j (i = 1,…, n_j; j = 1,…, J)

$y_{ij(1)}$: the prior test score of student i within school j

β_0: the intercept (*e.g.* which is the mean for all students when all of the independent variables take on the value 0)

$\beta_1 \sim \beta_3$: the regression slope for independent variables

X_{ij}: the student characteristics

X_j: the school characteristics

ε_{ij}: the residual, known as the error term, which is assumed to be normally distributed and independent of the covariates (*i.e.* $y_{ij(1)}$, X_{ij}, and X_j).

In the equation above, each student's current test score ($y_{ij(2)}$) is specified as a function of his/her prior test score ($y_{ij(1)}$) and other covariates (X_{ij} and X_j), using separate models for each year. In other words, these models are fit separately for each year of data, and therefore only the prior test score can be used in these analyses, although the information from multiple years is available (McCaffrey et al., 2003).

The estimated value-added score for school j (VA_j) is then taken to be a mean residual of students from this school (Jakubowski, 2008; OECD, 2008):

$$VA_j = ave(y_{ij(2)} - \hat{y}_{ij(2)}) = \frac{1}{n_j}\sum_{i=1}^{n_j}\varepsilon_{ij} \quad (2)$$

where

n_j: the number of students in school j

$\hat{y}_{ij(2)}$: the estimated linear regression prediction of current test score of student i.

Thus, if students in school j achieve higher current test scores on average in comparison with students from other schools with similar covariate values, then the corresponding residuals tend to be positive, yielding a positive estimated value-added for the school (OECD, 2008).

Alternatively, we can use successive-year gain scores as a dependent variable (McCaffrey et al., 2003). The 'gain score models' specify a one-year gain score (current test score minus prior test score) separately for each year and link student gains to their current-year school effects. Specifically, the gain score models assume:

$$y_{ij(2)} - y_{ij(1)} = \beta_0 + \beta_1 X_{ij} + \beta_2 X_j + \varepsilon_{ij} \quad (3)$$

There has been a substantial and long-standing debate over the use of the OLS linear regression models. Their primary advantage is that they are simple to specify and fit using any standard statistical software (Sanders, 2006; McCaffrey et al., 2003). Given the public attention on value-added measurement and accountability in education, sometimes the method underpinning the value-added models should be explicable to people with basic statistical knowledge and proper interpretation of estimates should be understandable for all stakeholders (Jakubowski, 2008).

Another positive aspect of the OLS linear regression models is their ability to be extended naturally to models where scores from successive years are nonlinearly related, via higher-order polynomial terms (*e.g.* $y^2_{ij(1)}$ or $y^3_{ij(1)}$) (McCaffrey et al., 2003).

A major disadvantage of the OLS linear regression models is that students without the prior or current test scores are excluded from the analysis (Sanders, 2006; McCaffrey et al., 2003). As the current score is regressed on the previous score, students would be excluded from the analysis if prior test scores, or sometimes current test scores, are unavailable. If some students are excluded from the analysis due to missing test scores, the value-added measures tend to be unstable and biased when students whose score gains are missing are not selected randomly from the school. Specialised methods, such as imputation or weighting, are required to eliminate bias and allow to use all available data (for more information, see 6.1 Missing data) (McCaffrey et al., 2003).

One of the limitations of OLS linear regression models is that these models do not reflect the multilevel nature of the data structure in an education system, where students are nested within schools or programmes (Clark et al., 2010). The OLS linear regression models assume that the students are homogeneous, differing only in the levels of their independent variables, the explanatory variables (Beck, 2001). With these linear regression models, it can be expected that the independent variables included in the models explain much of difference in a student, a school, or a year. However, in reality, statistical controlling of all possible factors that might be affecting student achievement or academic growth is difficult to achieve. For example, as shown in Figure 5, achievement of students in the same school is likely to be clustered due to the influence of unmeasured school

characteristics such as a learning atmosphere. If these unmeasured differences between students attending different schools are not accounted for in the value-added model (*i.e.* unmodeled heterogeneity), this unmodeled heterogeneity would increase the error (ε_{ij}), and result in biased estimates (Clark et al., 2010; Gujarati & Porter, 2009; Jakubowski, 2008; McCaffrey et al., 2003; Beck, 2001). Therefore, it is necessary to remove this unmodeled systematic heterogeneity from the error term.

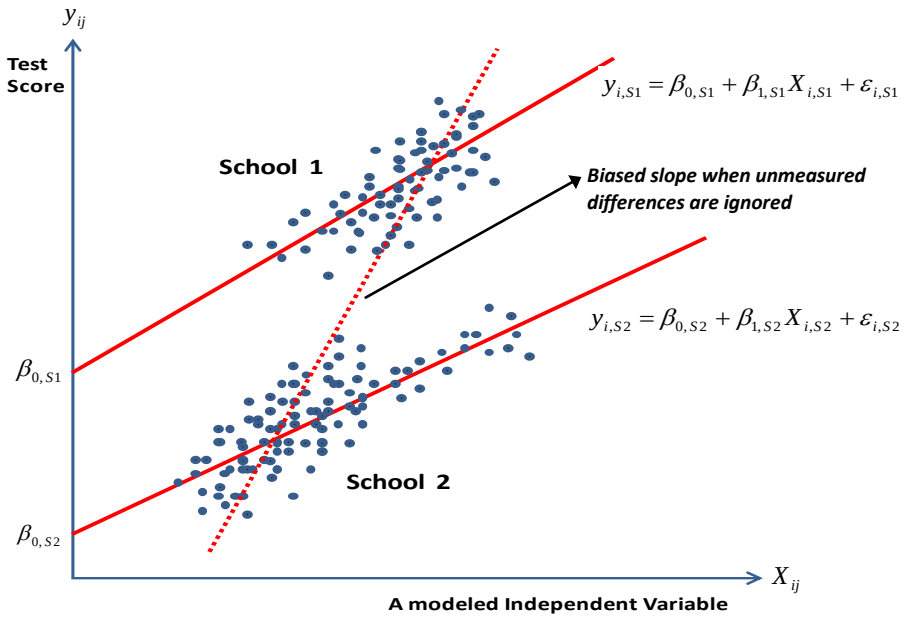

Figure 5: Bias from ignoring unmeasured differences between groups
(adapted from Gujarati & Porter, 2009, p. 596; Raudenbush & Bryk, 2002, p. 137)

4.1.2 Fixed and random effects models

A common way to handle multilevel data structure in education system is to apply multilevel regression models with a two-level nested structure in which students (at level-1) are grouped within schools (at level-2) (Clark et al., 2010; Gujarati & Porter, 2009; Raudenbush & Bryk, 2002). This two-level linear regression model for student achievement can be written as:

$$y_{ij(2)} = \beta_0 + \beta_1 y_{ij(1)} + X_{ij}^{'} \beta_2 + X_{j}^{'} \beta_3 + u_j + \varepsilon_{ij} \qquad (4)$$

where

$y_{ij(2)}$: the current test score of student i within school j (i = 1,..., n_j; j = 1,..., J)

$y_{ij(1)}$: the prior test score of student i within school j

β_0: the intercept (*e.g.* which is the mean for all students when all of the independent variables take on the value 0)

$\beta_1 \sim \beta_3$: the regression slope for independent variables

$X_{ij}^{'}$: the student characteristics

$X_{j}^{'}$: the school characteristics

u_j : the effects of a school j on students achievement

ε_{ij} : the residual at student-level.

Before conducting the multilevel regression models, researchers must decide whether to treat the school effects u_j, also known as the school residual, as fixed or random. The approaches chosen by researchers primarily depend on the types of research questions traditionally studied within each discipline (Clark et al., 2010).

Economists, for example, are more likely to focus on the effects of student and family characteristics rather than on student achievement (Todd & Wolpin, 2003), and hence prefer using fixed effects models.

In contrast, education researchers tend to use random effects models because they mainly focus on the school's contribution to the student achievement and academic growth (Townsend, 2007), and fixed effects approaches would not allow school characteristics to be modelled (for more information, see the next unit, 4.1.2.1 Fixed effects model) (Gujarati & Porter, 2009; OECD, 2008). The next two units discuss each of these models in turn.

4.1.2.1 Fixed effects models

As the name implies, these models estimate school effects on student achievement or academic growth as a fixed parameter (*i.e.* a value which has to be estimated). In other words, to reduce bias caused by not reflecting unmodeled differences between schools (*e.g.* the learning atmosphere or teacher expectations for students), fixed effects models assume that each school has its own fixed effects on student achievement (Gujarati & Porter, 2009). In contrast, random effects models assume that there is a bigger population of schools and their school effects, and each school effect on student achievement is chosen at random from the population (OECD, 2008).

There are two alternative fixed effects models: one is the fixed effects least-squares dummy variable (LSDV) model based on using dummy variables for school effects, and the other is the fixed effects within-group model in which school effects are differenced out (Clark et al., 2010; Gujarati & Porter, 2009).

The first approach, the LSDV model, takes account of unmeasured differences (*i.e.* heterogeneity) between schools by allowing each school to have its own intercept value and the dummy variables as additional predictors in an analysis of covariance (ANCOVA) model (Gujarati & Porter, 2009), which is shown in equation (5):

$$y_{ij(2)} = \beta_{01} + \beta_{02}D_2 + \beta_{03}D_3 + \cdots + \beta_{0J}D_J + \beta_1 y_{ij(1)} + X'_{ij}\beta_2 + X'_j\beta_3 + \varepsilon_{ij} \quad (5)$$

where the reference group is school 1,

$y_{ij(2)}$: the current test score of student i within school j (i = 1,…, n_j ; j = 1,…, J)

$y_{ij(1)}$: the prior test score of student i within school j

β_{01} : the effects of school 1 (*i.e.* the reference school)

β_{0j} : the difference between the effects of school j and the reference school.

D_2 : 1 for school 2 and 0 otherwise

D_3 : 1 for school 3 and 0 otherwise

$\vdots$

D_J : 1 for school J and 0 otherwise

Therefore, the sum ($\beta_{01}+\beta_{0j}$) represents the school effects for school j.

Although this LSDV model does not require assuming normal distribution of the school effects, it has some problems (Clark et al., 2010; Gujarati & Porter, 2009). First, this model could run up against the degrees of freedom problem if too many dummy variables are introduced into the model. When there are too many schools to be measured, the LSDV model sacrifices degrees of freedom as many as ($n-1$) dummy variables. Second, there is always a possibility of multicollinearity among a lot of dummy variables in the model, which might make a precise estimation of one or more parameters difficult. Third, the effects of school-level covariates are treated as nuisances and may not be able to be estimated. Because the school-specific intercepts (*i.e.* dummy variables) would absorb all differences in student achievement between schools, the school-level covariates ($X_j^{'}$) included in the model seem to have no impact on the differences in student achievement.

On the other hand, the second equivalent approach, known as the fixed effects within-group model (Gujarati & Porter, 2009), or pupil demeaned model (Clark et al., 2010), differences out the school effects by subtracting the school mean test scores ($\bar{y}_{j(1)}$, $\bar{y}_{j(2)}$) and mean covariates ($\bar{X}_j$) from individual student's test scores ($y_{ij(1)}$, $y_{ij(2)}$) and covariates (X_{ij}), and thereby, unmodeled differences between schools can also be differenced out of the model. A typical formulation of such model is:

$$y_{ij(2)}-\bar{y}_{j(2)}=\beta_1(y_{ij(1)}-\bar{y}_{j(1)})+(X_{ij}-\bar{X}_j)^{'}\beta_2+(\varepsilon_{ij}-\bar{\varepsilon}_j) \quad (6)$$

The resulting values ($y_{ij(2)}-\bar{y}_{j(2)}$) are called 'demeaned' or 'mean corrected' values (Clark et al., 2010; Gujarati & Porter, 2009). After subtracting school mean values for each student and school in the same way, researchers can run an OLS linear regression model using all demeaned values at the student-level. Finally, the value-added measure for each school can be obtained by calculating the mean residual of students in a given school, just like in equation (2) in the OLS linear regression model.

In comparison to the OLS linear regression models, the fixed effects within-group model produces more consistent estimates of the slope coefficients (Gujarati & Porter, 2009). By subtracting school mean values from individual student values, the school fixed effects (*i.e.* u_j in equation (4) or β_{0j} in equation (5)) and unmodeled differences between schools included in the error term (ε_{ij}) would be removed. As a result, this fixed effects within-group model become free from the regression assumption that the school effects (u_j) and the error term (ε_{ij}) must be uncorrelated with the student, family, and school characteristics (Clark et al., 2010; Gujarati & Porter, 2009).

However, the use of fixed effects within-group model comes with its own costs. The most important restriction is that the effects of school-level covariates ($X_j^{'}$) cannot be identified, as it is in the LSDV model, because these school characteristics would be differenced out along with school fixed effects (u_j) when doing subtraction in equation (6) (Clark et al., 2010). In addition, the estimated school effects obtained by the fixed effects within-group model may vary considerably from year to year (OECD, 2008), since there is no use of 'shrinkage' which is used in the random effects model to reduce the effects of sampling variation (for more information, see next unit, 4.1.2.2. Random Effects Models; Efron & Morris, 1973; and Copas, 1983).

4.1.2.2 Random effects models

Random effects models are also known as multilevel models, hierarchical linear models, or mixed models (Clark et al., 2010; Gujarati & Porter, 2009; Raudenbush & Bryk, 2002). At level-1, the unit of analysis is student

and each student's test score is represented as a function of a set of individual characteristics. At level-2, the unit of analysis is school. The regression coefficients at level-1 for each school are conceived as dependent variables that are hypothesized to depend on various school characteristics (OECD, 2008; Raudenbush & Bryk, 2002).

A typical formulation of such model, a random-intercept model, is:

Student-level (level-1) $$y_{ij(2)} = \beta_{0j} + \beta_{1j}(y_{ij(1)} - \bar{y}_{j(1)}) + \beta_{2j}(X_{ij} - \bar{X}_j) + \varepsilon_{ij} \quad (7)$$

School-level (level-2) $$\beta_{0j} = \gamma_{00} + \gamma_{0s}W_{sj} + u_{0j} \quad (8)$$

$$\beta_{1j} = \gamma_{10} \quad (9)$$

$$\beta_{2j} = \gamma_{20} \quad (10)$$

where

$y_{ij(2)}$: the current test score of student i within school j (i = 1,…, n_j; j = 1,…, J)

$y_{ij(1)}$: the prior test score of student i within school j

$\bar{y}_{j(1)}$: the mean prior test score for school j

X_{ij}: the student characteristics

$\bar{X}_j$: the mean of each student characteristic for school j

β_{0j}: the intercept of school j

β_1 & β_2: the level-1 regression slope for student's prior test score and characteristic

W_{sj}: the school characteristics (s denotes the number of the school characteristics)

γ_{00}: the level-2 intercept

γ_{0s}: the level-2 regression slope for school characteristics

ε_{ij}: the residual which is assumed to be normally distributed and independent of level-1 covariates

u_{0j}: the residual which is assumed to be normally distributed and independent of level-2 covariates.

As set out above, in the hierarchical linear models, the intercept and coefficients at level-1 become dependent variables at level-2. Each coefficient represents the slope for each independent variable at school j, but the meaning of the intercept (β_{0j}) at level-1 is determined by the location of the level-1 covariates: simple X_{ij}, centring around the grand mean (*i.e.* the mean of the means of several subsamples) ($X_{ij} - \bar{X}$), centring around the group mean ($X_{ij} - \bar{X}_j$) (Raudenbush & Bryk, 2002).

Even though it would be perfectly practicable to use simple X_{ij} in regression models, sometimes this may lead to nonsensical results (Raudenbush & Bryk, 2002). For example, suppose that age (X_{ij}) is the only determinant of the first grade student achievement test score (y_{ij}) in a primary school and simple X_{ij} is used without

centring (*e.g.* $y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + \varepsilon_{ij}$). Then, the intercept β_{0j} will be the expected outcome ($\hat{y}_{ij}$) for student i at school j whose age is 0. In this case, the intercept β_{0j} is meaningless because the minimum age requirement for admission in primary school is usually five or six.

If researchers are interested in the variation in β_{0j}, the simple X_{ij} needs to be transformed into the group-mean centring around ($X_{ij} - \bar{X}_j$) as in equation (7). In this case, the intercept β_{0j} becomes the unadjusted mean test score for school j (Raudenbush & Bryk, 2002). Therefore, at level-2, the mean test score for school j (β_{0j}) is represented by a set of school characteristics which is expected to affect the mean test score.

The total residual variance can be broken out into two components: the within-school (between-student) variance ($\mathrm{var}(\varepsilon_{ij})$) and the between-school variance ($\mathrm{var}(u_{0j})$) (Clark et al., 2010). As opposed to the fixed effects LSDV model, where each school has its own intercept value (*i.e.* fixed effects dummy variable coefficients) in the random effects models, the school intercept β_{0j} is thought of as randomly distributed. In the level-2 equation, the deviation (u_{0j}) from the expected test score ($\gamma_{00} + \gamma_{0s} W_{sj}$) is taken as an estimate of school value-added effects (Gujarati & Porter, 2009; OECD, 2008), which is assumed to be drawn from a normal and independent distribution (Clark et al., 2010; Raudenbush & Bryk, 2002).

These random effects models have at least two major advantages over fixed effect models. The most important being the regression coefficients. Estimates of school effects are more statistically efficient (*i.e.* having smaller mean-squared error thereby generating narrower confidence intervals) than those for fixed effects models (Clark et al., 2010; McCaffrey et al., 2003). For a better understanding, the meaning of 'shrinkage' should first be set out.

In the value-added models, estimates of the school effects (u_{0j}) are of major interest, but the accuracy of such estimates depends on the sample size for each school (Clark et al., 2010; Jakubowski, 2008). In schools having only small numbers of students, sampling variability will lead to some estimates being extremely small or extremely large relative to the true effect (Goldstein, 1997), and thereby increase the uncertainty and instability in the estimates (Raudenbush & Bryk, 2002). The random effects models introduce 'shrunken estimates' to reduce such uncertainty in estimates for small schools by shrinking the estimates of a given school toward the grand mean for all students (i.e. the mean of the means of several subsamples). A simple shrunken estimate for the school effects ($\hat{u}_{j,adj}$) is given by:

$$\hat{u}_{j,adj} = \lambda_j (\bar{y}_j - \bar{y}) \tag{11}$$

where

$\bar{y}_j$: the mean score for students in school j

$\bar{y}$: the grand mean score for all students

λ_j: the shrinkage weight factor which is less than 1

The differences between the observed ($\bar{y}_j$) and predicted mean score ($\bar{y}$) for school j is multiplied by a constant shrinkage factor (λ_j) (Clark et al., 2010; McCaffrey et al., 2003; Raudenbush & Bryk, 2002). This shrinkage weight factor can be defined as:

$$\lambda_j = \frac{\hat{\sigma}^2_{u_j}}{\hat{\sigma}^2_{u_j} + (\hat{\sigma}^2_{\varepsilon_{ij}} / n_j)} \quad (12)$$

$\hat{\sigma}^2_{u_j}$: between-school variance (the variance of the true means, β_{0j}, about the grand mean, γ_{00})

$\hat{\sigma}^2_{\varepsilon_{ij}}$: within-school variance (the variance of $\bar{y}_j$)

If the number of students in school j (n_j) is small, or the within-school variance ($\hat{\sigma}^2_{\varepsilon_{ij}}$) is large relative to the between-school variance ($\hat{\sigma}^2_{u_j}$) in equation (12), the shrinkage factor (λ_j) will be noticeably less than 1. As a result, in equation (11), an expected school effects ($\hat{u}_{j,adj}$) is shrunken towards zero.

The amount of shrinkage toward zero is determined by the number of students in a school (Clark et al., 2010; McCaffrey et al., 2003; Raudenbush & Bryk, 2002). The fewer the number of students in a given school, the greater shrinkage occurs. Hence, the extreme expected school effects due to their instability will be shrunken toward zero, and thereby the variance (*i.e.* the mean-squared error) in the estimated school effects can be reduced (OECD, 2008; Lindley & Smith, 1972). Statistically, such shrinkage estimates, on average, tend to be closer to parameters than any unbiased estimators, and provide a stable indicator for evaluating individual school performance (Jakubowski, 2008; Raudenbush & Bryk, 2002; Beck, 2001).

These random effects models can also estimate coefficients of school-level covariates (Gujarati & Porter, 2009). As seen earlier, the fixed effects models cannot estimate the impact of the school-level covariates on student achievement because of the use of the school dummy variables and subtraction method. On the other hand, the random effects models include school-level covariates, such as teacher-to-student ratio and mean socio-economic status, to explore the extent to which between-school differences can be explained by such school characteristics (Clark et al., 2010). If level-1 coefficients, β_{1j} and β_{2j}, are assumed as varying randomly over schools in equations (9) and (10), the random-intercept model can be extended to the random-coefficients model, in which the prior test score ($y_{ij(1)}$) and the students characteristics (X_{ij}) are permitted to vary across schools, and the differential school effects on the coefficients of student-level covariates can be obtained (Raudenbush & Bryk, 2002).

The use of such random effects models comes at the cost of an important additional assumption: both residual ε_{ij} and u_{0j} must be uncorrelated with the student, family, and school characteristics (Clark et al., 2010). However, in practice, unobserved school characteristics that influence student achievement (*e.g.* teacher quality, student motivation, or learning atmosphere) can be correlated with student and school characteristics that are included in the model (Gujarati & Porter, 2009; Jakubowski, 2008; Clark et al., 2010).

In addition, random effects models can introduce bias, although shrunken estimates minimise the mean-squared error between observed and estimated school effects (Raudenbush & Bryk, 2002). As seen above, when some schools have a small number of students or when the within-school variance is large relative to the between-school variance, the shrunken estimates would not provide accurate estimates for those schools (Clark et al., 2010), although more precise estimate could be obtained by reducing variance of the estimated school effects. In that case, if they are either highly effective or extremely ineffective, the shrunken estimates would be far below or far above the true school effects, respectively.

4.1.3 Growth curve models

The growth curve models refer to types of approaches that analyse trajectories of students over time to estimate the school contribution to student academic growth in achievement test score (Bollen et al., 2004).

As such, they apply to longitudinal data, where the same students are repeatedly observed. A minimum of three time points is required for proper estimation, and four or five time points are preferable in order to estimate more complex models involving trajectories following quadratic or cubic trends, as illustrated in Figure 6 (Raudenbush & Bryk, 2002; Curran & Muthén, 1999), while the value-added models set out above use only two consecutive annual test scores. In practice, the objects of study, especially in the social and behavioural sciences, might grow, decline, or follow other patterns rather than linear growth (Bollen et al., 2004).

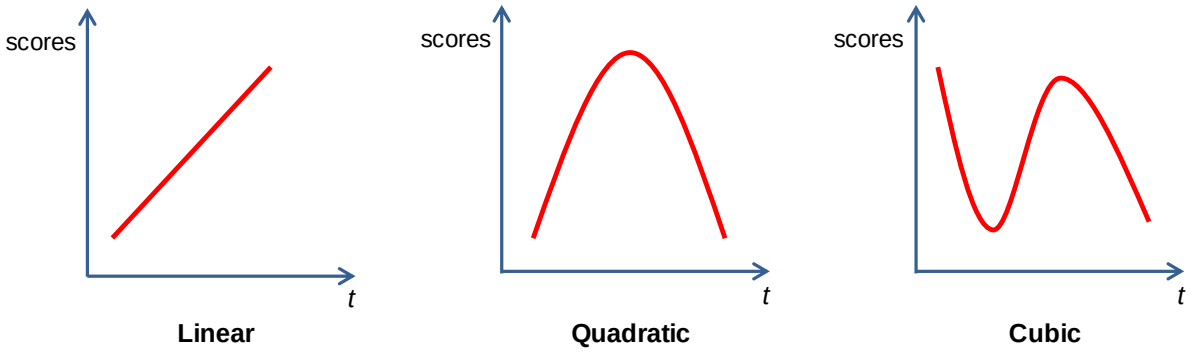

Figure 6: Types of the relationship between instructional time and student achievement

There are two different growth curve models in social and behavioural sciences. One treats growth curve models as a special case of hierarchical linear models (HLM) (Ponisciak & Bryk, 2005; Goldschmidt et al., 2004; Raudenbush & Bryk, 2002), and the other treats them as a special case of structural equation models (SEM) (Bollen & Curran, 2006; Hox & Stoel, 2005). Although both growth curve models allow analysis at multiple levels of hierarchical structured data and estimate the effects of independent variables on student initial status and growth rate (also, if the sample size is large enough, both models produce, theoretically and practically, identical results) (Hox, 2002), the estimation methods of both models are fundamentally different (Shin, 2007; Curran, 2003; Kline, 1998).

The HLM growth curve models do not require that all students be measured at the same time point, but they require a large sample size and are more sensitive to the sample size than the SEM growth curve models (Shin, 2007; Kline, 1998). On the other hand, the SEM growth curve models allow for measurement error in independent variables of change and provide a number of model fit indices but require that each student has the same number and spacing of time points. With the SEM growth curve models, student and school data collected periodically over time could not be analysed together with those measured less regularly.

This advantage of the HLM growth curve models approach over the SEM growth curve models approach may be significant in research on school effect on student achievement and academic growth, as many reasons such as school schedules or special needs can make it difficult to carry out follow-up tests and surveys regularly (Kline, 1998).

For these reasons, in this section, the growth curve models are explained from the HLM perspective, which views the multiple observations on each student as nested within students, like the students are nested within their schools (for more information about SEM, see Hox & Stoel, 2005; Curran, 2003; Hox, 2002; Kaplan, 2000; and Kline, 1998).

The growth curve models encompass a wide array of applications from individual growth curve modelling to programme evaluation and school performance modelling (Goldschmidt et al., 2004). The structure and components for a growth curve model will depend on researchers' primary interest. Individual student's

academic growth and each school's contribution to it are of major interest in education research (Raudenbush & Bryk, 2002). In this section, a three-level hierarchical growth model is introduced to present both value-added measures.

At level-1, each student's development is represented by an individual growth trajectory that depends on a unique set of parameters which become the dependent variables in a level-2 model. At level-2, the dependent variables (*i.e.* initial status and student's growth rate ij) are represented as a function of a set of individual characteristics, and the variation in growth parameters among students within a school is captured. In the same way, at level-3, dependent variables (*i.e.* mean initial status and mean growth rate within school j) are represented as a function of a set of school characteristics, and the variation among schools can be obtained.

Then, the three-level hierarchical growth model is:

Repeated-observations model (level-1) $$y_{tij} = \pi_{0ij} + \pi_{1ij} AY_{tij} + \varepsilon_{tij} \quad (13)$$

Student-level model (level-2) $$\pi_{0ij} = \beta_{00j} + \beta_{01j} X_{ij} + r_{0ij} \quad (14)$$

$$\pi_{1ij} = \beta_{10j} + r_{1ij} \quad (15)$$

School-level model (level-3) $$\beta_{00j} = \gamma_{000} + \gamma_{001} S_j + u_{00j} \quad (16)$$

$$\beta_{01j} = \gamma_{010} \quad (17)$$

$$\beta_{10j} = \gamma_{100} + \gamma_{101} S_j + u_{10j} \quad (18)$$

where

y_{tij}: the academic achievement at time t of student i in school j

AY_{tij}: the academic year (time)

π_{0ij}: the initial status of student ij

π_{1ij}: the growth rate for student ij during the academic year

ε_{tij}: the residual which is assumed to be normally distributed and independent of level-1 covariates

β_{00j}: the intercept (i.e. the mean initial status in school j when X_{ij} is zero)

X_{ij}: the student characteristics

β_{01j}: the effects of student characteristics X on individual initial status, which is fixed for all schools at γ_{010}

β_{10j}: the intercept (i.e. the growth rate in school j)

r_{0ij} & r_{1ij}: the residuals which are assumed to be normally distributed and independent of level-2 covariates

γ_{000}: the intercept (i.e. the initial status of student ij when X and S_j are zero)

S_j: the school characteristics

γ_{001}: the effects of school characteristics S on the mean initial status in school j

γ_{100}: the intercept (i.e. the growth rate of student ij when S_j is zero)

u_{00j} & u_{10j}: the residuals which are assumed to be normally distributed and independent of level-3 covariates.

In the hierarchical growth model above, the substantive interest is the decomposition of the variance in π_{0ij} and π_{1ij} as well as β_{00j} and β_{10j} into their within- and between-schools components (Raudenbush & Bryk, 2002; Kaplan, 2000). If the results show that the variance component of r_{0ij} and r_{1ij} are statistically significant, there are significant differences in the initial status and academic growth rate between students within schools. Similarly, if the variance of u_{00j} and u_{10j} are statistically significant, there are significant differences in mean initial status and mean academic growth rate between schools.

One advantage of these growth curve models is their congruency with the reality faced by many schools, where students often start out at different levels and grow at different rates (Heck, 2006). These models can produce the correlation of the growth parameters, such as initial status and growth rate, as well as their relation with time-varying and time-invariant covariates (Raudenbush & Bryk, 2002; Kaplan, 2000). For example, if initial status and growth rate are negatively correlated, it can be said that students who have lower achievement scores at the entry stage tend to gain at a somewhat faster rate.

In addition, these models can use data from all students, even those with partially complete records (McCaffrey et al., 2004; Little & Rubin, 1987), as opposed to other value-added models discussed earlier that can use data only from the students participating in both tests which are administered at two consecutive points in time, unless imputation or another missing data method is applied. Therefore, the growth curve models tend to be robust to missing data (McCaffrey et al., 2004).

The main weakness of these growth curve models is that they rely heavily on the quality of the longitudinal data set, which is greatly affected by student mobility or grade repetition (OECD, 2008). Furthermore, because of repeated surveys, there would be a great deal of measurement error in the test scores, thereby the precision and accuracy of estimation in such trajectory models could be negatively affected by such repeated measurements over time (Schmitz & Raymond, 2008).

4.1.4 Multivariate random effects model

The multivariate random effects model has been used in Tennessee since 1993, and adopted by a number of other school districts across the United States (Wainer, 2004; Sanders & Horn, 1998). This model, also known as the Educational Value-Added Assessment System (EVAAS) model, was developed by Sanders and Horn, and forms the basis of the Tennessee Value-Added Assessment System (TVAAS) (Sanders & Horn, 1998; Tekwe et al., 2004). This EVAAS model was introduced to measure teacher, school, or district effects on student achievement and academic growth by tracking the progress of students against themselves over the course of their studies with their assignment to various teachers' classes, schools, or districts (Sanders, 2006; Sanders et al., 1997; Sanders & Horn, 1998, 1994). Considering the objectives of this study, this section focuses on school effects on student achievement and academic growth over time.

The EVAAS model allows school effects to accumulate over time. This model focuses not only on how well a student does in a given subject, grade, and year at the school the student is currently attending, but also on the accumulated knowledge and skills acquired in the previous school the student attended as well as in the previous school year. Because of this accumulation of school effects on student achievement, the EVAAS model is often called the “layered” model (Wright et al., 2010).

Statistically speaking, this model can also be defined as a multivariate, longitudinal, and mixed effects model because it measures score changes in multiple subjects over time with fixed and random effects. For the analysis, details of each school and its students should also be collected annually from multiple subjects across several grades (Sanders, 2006; Sanders et al., 1997; Sanders & Horn, 1998, 1994).

The simplest form of the EVAAS model focusing on school effects is (Sanders et al., 1997; Sanders & Horn, 1994): the student achievement is represented by a vertically linked series of a standardised achievement test scores which is administered annually in one or more subjects. The sequence of test scores of a student who was first tested in 2010 in the third grade, for example, is assumed to satisfy the following equations:

$$Y_{10}^{3} = b_{10}^{3} + u_{10}^{3} + e_{10}^{3} \tag{19}$$

$$Y_{11}^{4} = b_{11}^{4} + u_{10}^{3} + u_{11}^{4} + e_{11}^{4} \tag{20}$$

$$Y_{12}^{5} = b_{12}^{5} + u_{10}^{3} + u_{11}^{4} + u_{12}^{5} + e_{12}^{5} \tag{21}$$

$$\vdots$$

where

Y_t^k : the test score in year t and grade k

b_t^k : the district mean test score in year t and grade k

u_t^k : the contribution of the grade k school to the year t test score

e_t^k : the influence of student specific factors on the test score in year t and grade k .

The statistical specification of the layered equations for estimating school effects is (Sanders et al., 1997; Sanders & Horn, 1994; Tekwe et al., 2004):

$$y_{ijst} = \mu_{st} + \sum_{l=1}^{t}\sum_{k=1}^{J} P_{ijkl} u_{ksl} + \varepsilon_{ijst} \tag{22}$$

where

y_{ijst} : the test score on the s^{th} subject at time t for the i^{th} student in the j^{th} school (*e.g.* s=1, 2; t=1, 2; i=1, 2,…, n_i ; j=1, 2,…, J)

μ_{st} : the population mean parameter for the s^{th} subject test score at time t

P_{ijkl} : the proportion of academic year time spent by the i^{th} student, who was in the j^{th} school at time 2 test, in the k^{th} school during the year prior to the test at time l ($1 \le l \le t$)

u_{ksl} : the random school effects of the k^{th} school on subject s test score at time 1

ε_{ijst} : the random within-school error for the i^{th} student in j^{th} school for the s^{th} subject at time t .

The EVAAS model assumes that the random effects u_{ksl} and ε_{ijst} are independent and normally distributed with mean zero, Var(u_{ksl}) = σ_{sl}^2, Cov(u_{ksl}, $u_{k's'l'}$) = 0 for all $k \neq k'$, $s \neq s'$, or $l \neq l'$, and Cov(ε_{ijst}, $\varepsilon_{i'j's't'}$) = 0 for all $(i, j) \neq (i', j')$. However, the covariance matrix of ε_{ijst} is unstructured, requires no assumption and constraint

regarding the error terms, and allows each variance and covariance to be completely different and to have no relation to the others. In addition, while classical regression analysis assumes that the errors should be uncorrelated with each other and the predictors should be linearly independent, this EVAAS model allows for the intra-student correlations among test scores at different times, and reflects the fact that student's test scores at different times are actually highly correlated with each other (OECD, 2008; Gujarati & Porter, 2009; Tekwe et al., 2004).

A unique and attractive aspect of the EVAAS model is the total school effects on student academic growth can be divided according to the proportion of time spent in each school (Tekwe et al., 2004). A school's contribution to student academic growth on a standardised test in one year is approximately proportional to the time enrolled during the given year. If one student, for example, attended school k' for the entire year in year 1 and the first half of the year prior to the test in year 2, and the second half in school k, then the total school effects for the s^{th} subject in year 2 will be $u_{k's1} + 0.5u_{k's2} + 0.5u_{ks2}$, which is reflects the proportion of the academic year spent in each school. In this case, $u_{k's1} + 0.5u_{k's2}$ is the amount of effects of school k' on this student's academic growth over the last two years, and $0.5u_{ks2}$ is for school k.

This model also allows for the use of incomplete data (Ballou et al., 2004; Sanders & Horn, 1994). The equations above pertain to the same students and the same subjects, and the same school effects enter more than one equation, as achievement in the later year is "layered" on top of earlier achievements (Sanders & Horn, 1994). Therefore it is possible to estimate the school effects even if all data for a given school during the year in question are missing (Ballou et al., 2004; Sanders & Horn, 1994). Even though Y_{11}^4 is missing, for example, u_{11}^4 would still appear in the equation for Y_{12}^5, provided the latter is not also missing. Therefore, a missing record, in the example above, can be obtained from the imputation of unobserved scores using observed scores. By minimising the loss of data due to missing observations, the EVAAS model can reduce the sample selection bias and consequently provide more precise estimates and narrower confidence interval (Lockwood & McCaffrey, 2007).

In addition, compared with other value-added models, the EVAAS model is highly parsimonious and efficient (Tekwe et al., 2004). This model does not require controlling for either incoming academic ability or other covariates such as socio-economic status, demographic characteristics, or other factors that influence student achievement and academic growth as the model assumes student scores in prior years adequately reflect student characteristics (Wiley, 2006).

However, some researchers criticize the parsimony of the EVAAS model. They are concerned that the omission of covariates that could contribute to student achievement can lead to biased estimates when students are systematically different from each other and stratified by those covariates (McCaffrey et al., 2004). Without the covariates, the model cannot account for the possible influence of such factors on student achievement and academic growth (Ballou et al., 2004; McCaffrey et al., 2004).

To address this critique, Ballou, Sanders, and Wright (2004) introduced student background, such as eligibility for free and reduced-price lunch, race, and gender, into the original EVAAS model for estimating teacher effects using the TVAAS data set. Researchers found that controlling for student-level covariates makes very little difference to teacher effects, and reconfirmed that statistical adjustment for student and school characteristics was unnecessary in the EVAAS model for estimating teacher effects as well as school effects.

However, the debate over the issue of controlling for student and school characteristics is not yet concluded (McCaffrey et al., 2004). Above all, the findings of the research conducted by Ballou, Sanders, and Wright (2004) cannot be generalised to other assessment settings because researchers only used the data from the Tennessee Value-added Assessment System. In addition, the research only focused on the teacher effects while excluding school-level covariates. Therefore, the impact of student- and school-level covariates on the estimation of school effects has still not been clarified.

Another disadvantage of the EVAAS model is that school effects at one point in time can be affected by the earlier school effects: if the school effects in previous years are relatively high, this year's school effects would be lower than its true value (OECD, 2008; Wiley, 2006). The EVAAS model assumes that the school effects on student achievement persist in and can be carried over to all succeeding years. However, this assumption turns out to be problematic since in reality, school effects actually have diminished over time and may not affect student's future growth (McCaffrey et al., 2003).

4.2 Models used in higher education

The value-added models used in higher education differ in many ways from those in K-12 education due to different contexts surrounding the assessments and data availability from the assessments conducted in higher education. While almost all value-added models used in K-12 education are developed based on longitudinal data pertaining to the same students and the same subjects over years (Ballou et al., 2004; McCaffrey et al., 2004; Tekwe et al., 2004; OECD, 2008), such a longitudinal approach with a repeated measures design is rare in higher education and most value-added models employ a cross-sectional design testing entering freshmen and graduating seniors at the same time (Liu et al., 2012; Liu, 2011; Steedle, 2009, 2011; Klein et al., 2007).

This section presents three different cross-sectional approaches for value-added measurement used in higher education. Although their detailed calculations of value-added scores vary, the three models are similar in their use of the test scores of freshmen and seniors who take the test at the same time (*i.e.* cross-sectional design), and the fact that they are based on regression residuals (*i.e.* differences between estimated scores produced by linear regression model and actual observed scores) rather than simple differences between freshmen and seniors actual scores (Steedle, 2010).

4.2.1 Difference in residuals model: OLS linear regression-based approach

The OLS linear regression models intend to capture whether students' mean academic growth between entering freshmen and graduating seniors in a given institution is near or above what is typically observed at institutions admitting students of similar entering academic ability (*i.e.* the 'expected' test scores or overall mean test scores) (Steedle, 2010, 2011).

To measure the 'expected' test scores, this model carries out regressions of the current mean test scores on the mean entering academic ability scores (*e.g.* mean SAT scores) for freshmen and seniors, respectively. The important thing is that the unit of analysis is institutions instead of students, and thereby the dependent variable is institution' current mean test score. A typical formulation of the OLS linear regression model is:

$$\bar{y}_j = \beta_0 + \beta_1 \overline{SAT}_j + \varepsilon_j \tag{23}$$

where

$\bar{y}_j$: the current mean test score of institution j (j = 1,..., J)

$\overline{SAT}_j$: the mean entering academic ability score of institution j

ε_j : the residual which is assumed to be normally distributed and independent of the covariates.

Each OLS linear regression equation produces its own residuals, which are the differences between 'expected' test scores produced by the regression model and actual 'observed' mean test scores. As the OLS linear regression models for freshmen and seniors are conducted separately, residuals are also obtained for freshmen and seniors, respectively.

Value-added score for institution j (VA_j) is:

$$VA_j = [\bar{y}_{j,se} - E(\bar{y}_{se})] - [\bar{y}_{j,fr} - E(\bar{y}_{fr})] \quad (24)$$

where

$\bar{y}_{j,se}$: the observed senior mean test score at institution j

$\bar{y}_{j,fr}$: the observed freshman mean test score at institution j

$E(\bar{y}_{se})$: the expected senior mean test score

$E(\bar{y}_{fr})$: the expected freshman mean test score.

$E(\bar{y}_{se})$ represents expected mean test scores for seniors attending institutions admitting students of similar entering academic ability, and the difference between the observed mean test score ($\bar{y}_j$) and the expected test score ($E(\bar{y})$) represents residual. Therefore, in equation (24), value-added score for institution j can be obtained by subtracting the freshman residual from the senior residual.

Equation 24 above can also be rearranged as follows:

$$VA_j = [\bar{y}_{j,se} - \bar{y}_{j,fr}] - [E(\bar{y}_{se}) - E(\bar{y}_{fr})] \quad (25)$$

Finally, the value-added score for institution j can be defined as the difference between the institution's observed freshman-senior mean difference (*i.e.* observed value-added score) and expected freshman-senior mean difference (*i.e.* expected value-added score) (Klein et al., 2007; Steedle, 2009, 2010, 2011; Liu, 2008, 2011).

As illustrated in Figure 7, each institution has two value-added scores, the 'observed' value-added score and the 'expected' value-added score produced by the OLS linear regression model on institution's mean entering academic ability score. When the observed value-added score for a given institution j exceeds the expected value-added score (*i.e.* overall mean for institutions admitting students of similar entering academic ability), it can be said that the institution j has relatively 'high' value-added. In other words, students attending institution j appear to have 'grown' more in academic competences than students at other institutions after controlling for the entering academic ability (Steedle, 2009).

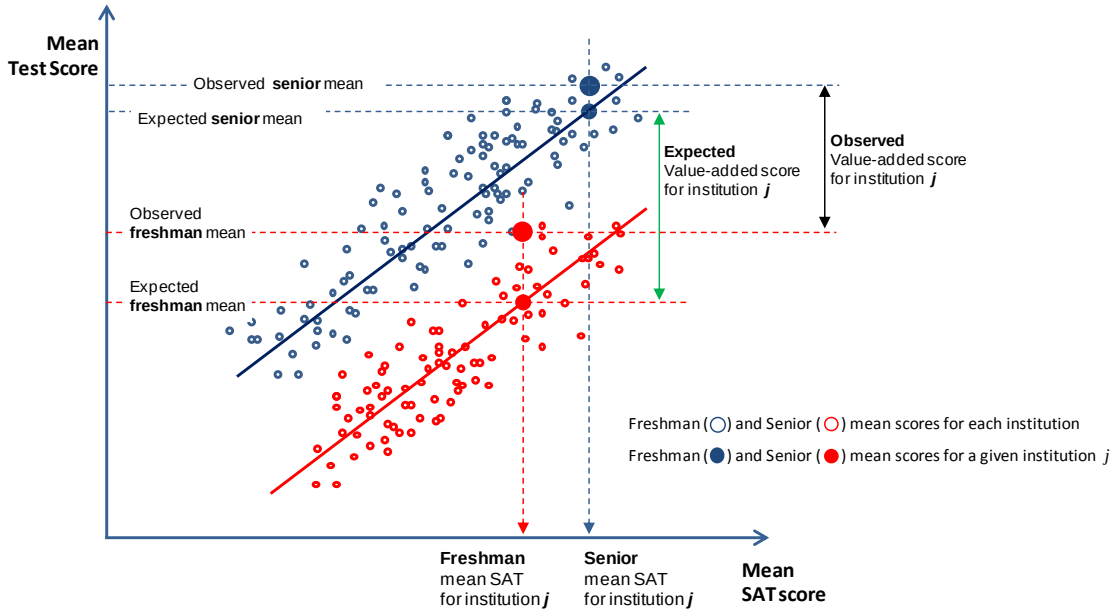

Figure 7: Value-added score estimation approach
of the OLS linear regression-based difference in residuals model
(adapted from Steedle, 2010)

This model is easy to implement and its results are fairly straightforward because it depends on the OLS linear regression model and a simple subtraction method. In addition, using the cross-sectional design for value-added measurement is less costly and more feasible to implement than the longitudinal design (Liu, 2008, 2011; Steedle, 2009, 2011; Klein et al., 2007).

However, there are some potential problems with this value-added model. First, this model relies on appropriate standardised test scores reflecting students' entering academic ability. As seen above, in this model, value-added scores are obtained by taking differences between the 'expected' mean test scores produced by the OLS linear regression model on institution's mean entering academic ability scores and the actual 'observed' mean scores.

A more significant problem is that the use of such a value-added model based on the differences between freshman and senior residuals is faced with a dilemma of an assumption of a linear relationship between the mean current test scores and the entering academic ability scores (Traub, 1967; Cronbach & Furby, 1970; Klein et al., 2007). If the mean test scores and the entering academic ability scores are not linearly related to each other, the assumption underlying this model is substantially violated, and thereby biased estimates are produced. However, if these two variables are highly linearly related to each other (*i.e.* substantially correlated with each other), the reliability of residuals (*i.e.* the consistency of residuals produced under consistent conditions) is fairly low and tends to decrease as the correlation between two variables increases (Traub, 1967; Pike, 1992; Banta & Pike, 2007).

4.2.2 Difference in residuals model: HLM-based approach

Like the OLS linear regression-based model above, this model computes value-added scores of each institution based on the differences of residuals which are obtained by subtracting the freshman residuals from the senior residuals.

However, this model differs from the OLS linear regression model in that it takes a multilevel approach (Liu, 2011). This model uses two-level HLM in calculating freshman and senior residuals. Given that students are nested within institutions and that student achievement can be affected by various institutional characteristics, the value-added modelling needs to reflect the hierarchical data structure, and to consider the influence of institutional characteristics in estimation of each institution's contribution to student academic achievement (for more information on the HLM, see 4.1.2.2 Random Effects Models). In addition, student is the unit of analysis, while the OLS linear regression models conduct the analysis at the institution level using each institution's mean scores. This helps utilize full information at the student level and present more accurate relationship between current test scores and entering test score.

As set out in previous hierarchical linear models earlier, at level-1, the unit of analysis is students, and each student's test score is represented as a function of the student's entering academic ability scores (*e.g.* SAT scores). At level-2, the unit of analysis is institutions. The level-1 regression coefficients for each institution are conceived as dependent variables that are assumed to depend on institutional characteristics (OECD, 2008; Raudenbush & Bryk, 2002). This multilevel model is conducted for freshman and senior students separately, using entering academic ability scores (*e.g.* SAT scores).

A simple version of such a model is given below:

Level-1 (student) $$y_{ij} = \beta_{0j} + \beta_{1j}(SAT_{ij} - \overline{SAT}_j) + \varepsilon_{ij} \quad (26)$$

Level-2 (institution) $$\beta_{0j} = \gamma_{00} + \gamma_{0s} W_{sj} + u_{0j} \quad (27)$$

$$\beta_{1j} = \gamma_{10} \quad (28)$$

where

y_{ij}: the current test score of student i within institution j (i = 1,…, n_j; j = 1,…, J)

SAT_{ij}: the entering academic ability score of student i within institution j

$\overline{SAT}_j$: the mean SAT scores at institution j

β_{0j}: the level-1 intercept (equal to the mean current test score at institution j)

β_{1j}: the level-1 regression slope for student's entering academic ability score

ε_{ij}: the residual which is assumed to be normally distributed and independent of level-1 covariates

W_{sj}: the institution characteristics (s = 1, …. , S)

γ_{00}: the level-2 intercept

γ_{0s}: the level-2 regression slope for school characteristics

u_{0j}: the residual which is assumed to be normally distributed and independent of level-2 covariates.

The total residual variance can be broken out into two components: the within-institution (between-students) variance ε_{ij} and the between-institution variance u_{0j}.

At level-1, if the entering academic ability score (SAT_{ij}) is equal to the mean SAT score at institution j ($\overline{SAT}_j$), the expected test score becomes β_{0j}. Therefore, it can be said that the β_{0j} represents the mean test score at institution j.

At level-2, the institution-level intercept β_{0j} is thought of as being randomly distributed and the deviation (u_{0j}) from the grand mean after controlling for the institutional characteristics ($\gamma_{00} + \gamma_{0s} W_{sj}$) is taken as the estimate of freshman (or senior) residuals. From equation (27), the differences between these two residuals (u_{0j}) for freshmen and seniors represent the value-added score for institution j.

This model applies the multilevel modelling to reflect the nested data structure in higher education, whereby more precise estimates to calculate the school effects (*e.g.* freshman and senior residuals) are obtained. Moreover, the HLM approach provides the standard error of residuals for freshmen and seniors, respectively. It can be used as an estimate of precision of residuals to compute the confidence interval for freshman and senior residuals (Steedle, 2011).

However, value-added scores for each institution should be interpreted with caution, as the HLM-based difference in the residuals model exhibits many of the shortcomings common to the OLS linear regression-based approach. This HLM-based model also requires a variable reflecting students' entering academic ability (*e.g.* SAT scores), and assumes a linear relationship between student current test scores and entering academic ability scores.

Furthermore, this model, like the OLS linear regression-based model above, uses the difference in residuals between freshmen and seniors in order to produce value-added scores for each institution (Liu, 2011). Hence, this HLM-based difference in residuals model also has the problem of the reliability of residuals, just like the OLS linear regression-based model (for more information, see 4.2.1 Difference in residuals model: OLS linear regression-based approach).

In addition, the potential problems with the random effects models should also be considered in advance (for more information, see 4.1.2.2 Random Effects Models). This HLM-based difference in residuals model based on random effects models should satisfy the assumption that both residual ε_{ij} and u_{0j} must be uncorrelated with the student and institution characteristics (Clark et al., 2010). In addition, bias can be introduced to the estimated institution effects by the use of shrunken estimates, although more precise estimate could be obtained by reducing variance of the estimated institution effects (Raudenbush & Bryk, 2002).

4.2.3 HLM-based residual analysis model

As shown in Figure 8, the two previous value-added models used in higher education compute value-added scores for each institution based on the difference between freshman and senior residuals, which are obtained by subtracting 'expected' test score from 'observed' mean test score. In other words, these two models focus on the amount of student academic growth at each institution between entering freshmen and graduating seniors, and compare the amount for each institution to the estimated overall growth for all institutions admitting students of similar entering academic ability.

The alternative model, the HLM-based residual analysis model, compares the senior mean test scores for each institution instead of the difference in scores between freshman and senior. As shown in Steedle (2009, 2010, 2011), this model produces value-added scores based on the degree to which observed senior mean test scores exceed or fall below expectations after controlling for entering students' academic ability (*e.g.* the entering academic ability of seniors, the senior mean SAT scores, and freshman mean test scores). For example, if seniors in a particular institution perform better on an achievement test compared with other seniors having similar entering academic ability in a typical institution, it can be said that former seniors have

grown more in their academic ability than expected and that their institution has provided greater value-added education for its students.

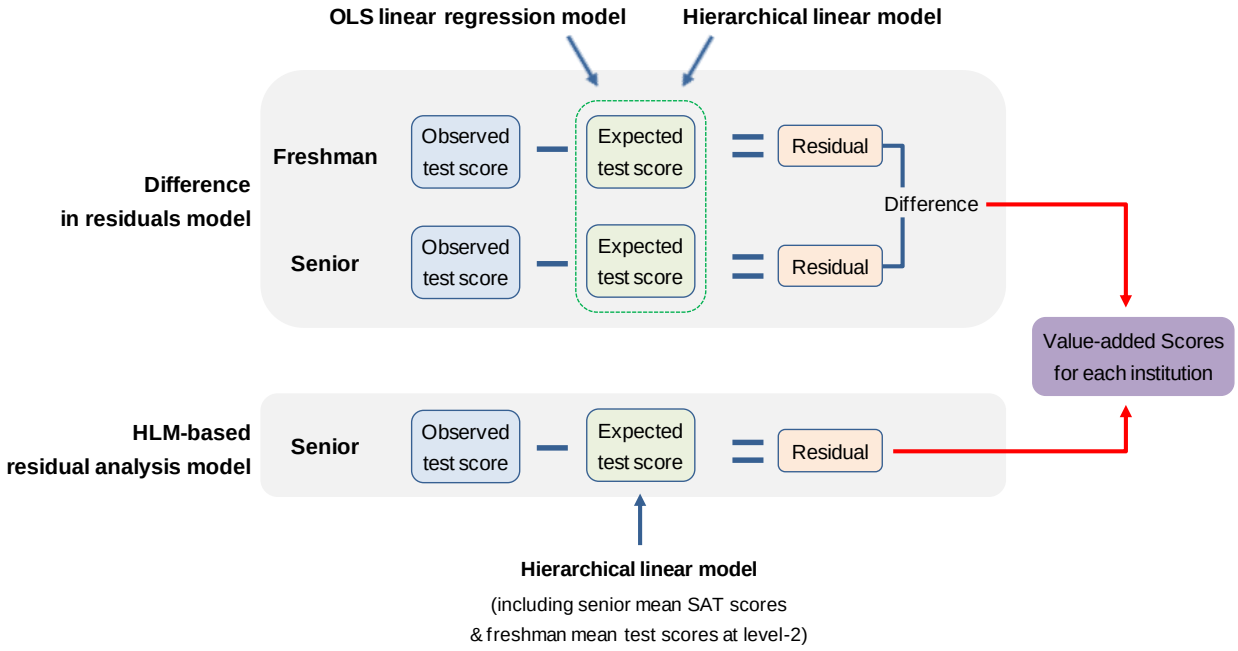

Figure 8: Differences between three value-added models used in higher education

To produce the institutional effects on academic achievement of senior students, this approach incorporates two levels of analysis (Steedle, 2010). At level-1, the unit of analysis is students and the senior current test score is represented as a function of the entering student's academic ability score (*e.g.* SAT score). At level-2, on the other hand, the unit of analysis is institutions and the level-1 regression coefficients for each institution are conceived as dependent variables that are hypothesized to depend on the senior mean entering ability score and the freshman mean current test score at each institution.

The two-level hierarchical linear model takes the following form:

Level-1 (student) $$y_{ij,se} = \beta_{0j} + \beta_{1j}(SAT_{ij,se} - \overline{SAT}_{j,se}) + \varepsilon_{ij} \quad (29)$$

Level-2 (institution) $$\beta_{0j} = \gamma_{00} + \gamma_{01}\overline{SAT}_{j,se} + \gamma_{02}\bar{y}_{j,fr} + u_{0j} \quad (30)$$

$$\beta_{1j} = \gamma_{10} \quad (31)$$

where

$y_{ij,se}$: the current test score of senor student i within institution j

$SAT_{ij,se}$: the entering academic ability score of senor student i within institution j

$\overline{SAT}_{j,se}$: the senior mean entering academic ability score for institution j

β_{0j}: the level-1 intercept (equal to the senior mean current test score at institution j)

β_{1j}: the level-1 regression slope for student's entering academic ability score

ε_{ij}: the residual which is assumed to be normally distributed and independent of level-1 covariates

$\bar{y}_{j,fr}$: the freshman mean current test score for institution j

γ_{00}: the level-2 intercept

γ_{01} & γ_{02}: the level-2 regression slope for school characteristics

u_{0j}: the residual which is assumed to be normally distributed and independent of level-2 covariates.

The u_{0j} reflects the difference between observed senior mean test score (β_{0j}) and expected senior test score ($\gamma_{00} + \gamma_{01}\overline{SAT}_{j,se} + \gamma_{02}\bar{y}_{j,fr}$), that is to say value-added score for institution j.

Steedle (2011), initiator of the HLM-based residuals analysis model, found that the HLM-based residuals analysis model and the OLS-based difference in residuals model produced similar results, but the former increased reliability and year-to-year consistency of value-added scores for each institution compared to the OLS-based difference in residuals model.

Moreover, the HLM-based residuals analysis model fits the nested data structure in education system by using the hierarchical linear modelling, whereby more accurate institution effects on the student academic achievement are produced (Steedle, 2011; Clark et al., 2010). When the hierarchical structure of data is ignored, the institution effects tend to be under estimated (Steedle, 2011; Raudenbush & Bryk, 2002). Steedle (2009, 2010, 2011) also indicates that this approach can provide an estimate of value-added score precision for each institution that can be used to compute a unique confidence interval for each institution's value-added score.

Even though this third model, the HLM-based residuals analysis , increases the reliability of the institution effects compared with the OLS-based difference in residuals model, it is unlikely to be adequate for using value-added scores to make high-stakes decisions, such as decisions about funding for higher education institutions (Steedle, 2011). As seen in Traub (1967) and Pike (1992), the reliability issue is attributed to the analysis method for this HLM-based residual analysis model using the residuals between the expected scores produced by the regression model and the observed actual scores. In practice, this model does not depend on the difference scores between freshmen and seniors, although it still uses the residuals between observed senior mean test scores and expected senior scores to produce the institution effects. As Steedle (2011) points out, a larger sample size would help increase the reliability of residuals even though it is important to find other ways to obtain substantially higher reliability as there is a limit to the increasing sample size.

In addition, the HLM-based residuals analysis model includes the senior mean SAT score and the freshman mean current test score at level-2 to control for its effects on the senior mean current test score for institution j (β_{0j}). As Steedle (2011) illustrated, these two variables accounted for considerable variation in the mean current test scores. However, problems will arise in estimating regression coefficients at level-2 as the senior mean SAT scores and the freshman mean current test scores are highly correlated (Gujarati & Porter, 2009). Although this multicollinearity problem does not reduce the predictive power or reliability of the model as a whole, the regression model at level-2 may not produce valid coefficients. Therefore, the coefficients of these two variables at level-2 should be interpreted with caution.

Like the previous two differences in residuals models, the HLM-based residual analysis model also needs an appropriate standardised test score to control for the initial academic achievement level of students as they

begin the school year (*e.g.* SAT score). Thus, if there is not an appropriate indicator reflecting entering students' academic ability, this model cannot be used.

In addition, as it is the case for the HLM-based difference in residuals model, the potential problems with the random effects models should also be considered in advance (for more information, see 4.1.2.2 Random Effects Models).

5. Model choice: mean–variance–complexity trade-off

Despite researchers' efforts to develop alternative value-added models to address weaknesses of the existing ones, no single value-added model has proven superior to any others. There are still numerous open questions about accuracy and precision of estimates derived from each value-added model (Hibpshman, 2004). Therefore, this report does not advocate the use of one value-added model over others, but rather provides criteria which can be practically applied in selecting the most appropriate value-added model for a given data set in the context of higher education.

In order to select an appropriate value-added model from several candidates, a bias-variance-complexity trade-off framework is proposed by researchers (Hastie et al., 2011; Yu et al., 2006; Geman et al., 1992). This framework is based on the bias-variance decomposition of the mean squared error (MSE) of the value-added estimate.

In statistics, the estimation error refers to the difference between the unknown value of a parameter (θ) and its estimator ($\hat{\theta}$).

$$\varepsilon = \hat{\theta} - \theta \tag{32}$$

In general, a good estimator ($\hat{\theta}$) should be close to the parameter (θ). Therefore, the degree of closeness is usually measured by the mean of the squared estimation error (ε). This value is called the mean squared error (MSE) of the estimator ($\hat{\theta}$) which means the expected value of the squared difference between estimator ($\hat{\theta}$) and its parameter (θ).

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] \tag{33}$$

The MSE is used to determine the statistical significance of an estimator. If MSE is zero, then it means the estimator ($\hat{\theta}$) predicts observations of the parameter (θ) with perfect accuracy, but is practically never possible (Gujarati & Porter, 2009). In addition, values of MSE can also be used for comparing two or more statistical models using their MSEs as a measure of how well they explain a given set of observations. Therefore, minimizing MSE is a key criterion for selecting one model in a competing set of statistical models.

Statistically, the MSE can be decomposed into the sum of variance and squared bias of the estimator ($\hat{\theta}$) (Hastie et al., 2011; Gujarati & Porter, 2009). In equation (34), bias refers to the difference between the average of estimates ($\overline{\theta}$) (*i.e.* estimator $\hat{\theta}$'s expected value) and the true value of the parameter being estimated (θ), and variance indicates how far, on average, the collection of estimates ($\hat{\theta}$) are from the expected value of the estimates ($\overline{\theta}$).

$$\begin{aligned} MSE(\hat{\theta}) &= (\overline{\theta} - \theta)^2 + E[(\hat{\theta} - \overline{\theta})^2] \\ &= Bias(\hat{\theta})^2 + Var(\hat{\theta}) \end{aligned} \tag{34}$$

Mathematically, it is very common that there may be perceived to be a bias-variance trade-off, such that a small increase in bias can be traded for a larger decrease in variance, resulting in a more desirable estimator

overall (Gujarati & Porter, 2009). For that reason, unbiased estimator may not produce estimates with the smallest total variance, where estimators with the smallest total variance may produce biased estimates.

On the other hand, if model complexity is considered together, bias decreases as model complexity (*i.e.* which can be measured by the number of parameters or degrees of freedom) increases, whereas variance increases with model complexity (Geman et al., 1992). When the model becomes simpler, bias increases and variance decreases. Given this trade-off relationship among bias, variance, and model complexity, the most appropriate analysis model can be found at the lowest point of the total error (*i.e.* the MSE) (for more information, see Hastie et al., 2011; and Geman et al., 1992).

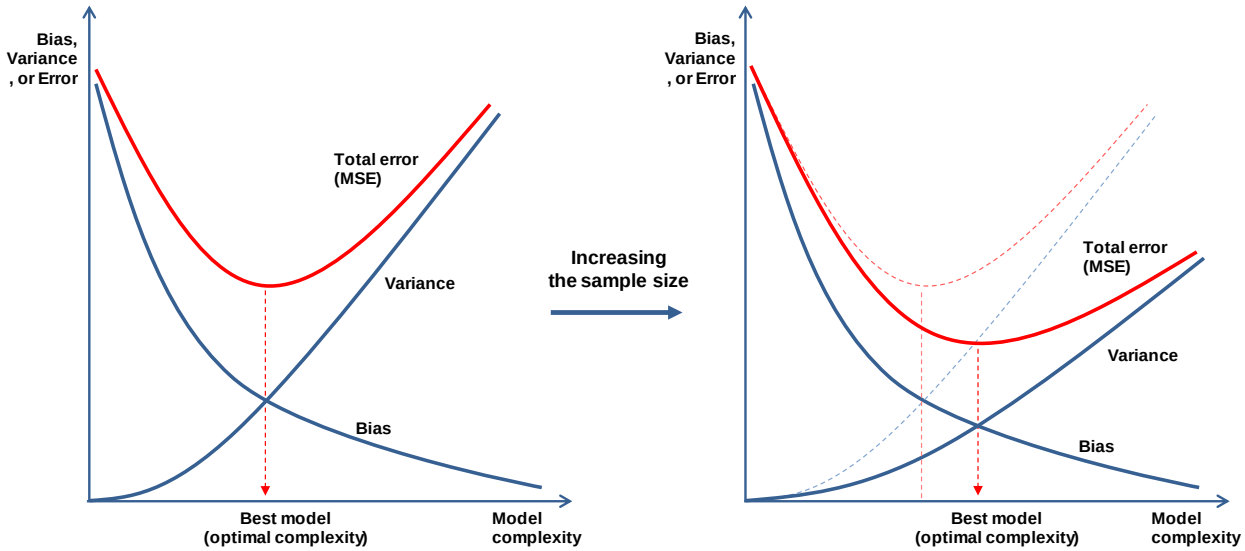

Figure 9: The relationship among bias, variance, and complexity

(adapted from Hastie et al., 2011; Yu et al., 2006; and Geman et al., 1992)

As shown in Figure 9, when the sample size is increased without changing the analysis model, the slope of variance is decreased and the optimal complexity point moves to the right. This means that variance is reduced by increasing the sample size, and so a slightly more complex model minimizes the expected total error, the MSE (Yu et al., 2006).

This bias-variance-complexity trade-off framework can be used to select the most appropriate value-added model for a given data set from various models. Yu et al. (2006) introduced a model selection criterion function to find the most appropriate analysis model for a given data set from various models.

$$\text{Model selection criterion} = f(\text{bias, variance, complexity}) \tag{35}$$

$$= [(n+d)/(n-d)] \times [(\bar{\theta} - \theta)^2 + E[(\hat{\theta} - \bar{\theta})^2]]$$

where

n : sample size

d : degrees of freedom

After conducting a series of analyses with various value-added models using the same data set, researchers produce each model's selection criterion value with the results, and then compare the selection criterion values to find the most appropriate value-added model for a given data set and education context. Finally, using equation 35, a good model can be selected by the following rule: the smaller the selection criterion value, the better the model would be (Yu et al., 2006).

6. Model improvement

Although an appropriate value-added model is selected, the model is no guarantee of accurate and unbiased results because there are still many other bias and error factors that may influence the results outside the model, such as missing data, student mobility, model misspecification, and fluctuations in value-added scores across years (Gujarati & Porter, 2009; OECD, 2008; Schmitz & Raymond, 2008; Abbasi, 2000). In order for the value-added models to produce accurate and unbiased estimates, it is necessary to reduce bias and improve reliability of the estimated institution effects.

6.1 Missing data

Missing data is a result of the failure to obtain a complete response from all students (or institutions) included in a survey such as students not taking the test or not providing their background information. Whatever the reason, a substantial missing response rate can make survey results unrepresentative of the population by distorting the estimates in one direction, and result in biased estimates (Schmitz & Raymond, 2008; Abbasi, 2000).

One simplified approach to deal with missing data is to exclude an entire student record from the data set if any single value in test scores, student characteristics, or other contextual data is missing (*i.e.* listwise deletion, also known as complete-case analysis). However, this listwise deletion approach may lead to biased estimates due to possible systematic differences between students with missing values and completely observed students, and also create larger standard errors due to reduced sample size (Gelman & Hill, 2007). On the other hand, in some cases, students are dropped only on analyses involving variables that have missing values (*i.e.* pairwise deletion, also known as available-case analysis). This pairwise deletion approach also poses difficulties in interpreting the results as each analysis involving different variables will possibly be based on different subsets of the data and thus will not necessarily be consistent with each other, although this approach would retain more of the data than listwise deletion approach (Gelman & Hill, 2007).

Other than excluding students with missing values, another approach to reduce bias arising from non-response or missing data is to impute missing values (Gujarati & Porter, 2009; Gelman & Hill, 2007; Abbasi, 2000; Montaquila & Ponikowski, 1995). Each missing data is replaced with at least one imputed response and the full sample size is maintained. A number of different types of imputation methods have been developed and used in statistical analysis, such as deductive imputation, mean imputation, random imputation, overall mean imputation within classes, hot-deck imputation, cold-deck imputation, flexible matching imputation, ratio imputation, predicted regression imputation, random regression imputation, and distance function matching (Kalton & Kasprzyk, 1982; Seastrom, 2002). For more detailed information about specific imputation methods and their relative advantages and disadvantages, see the following sources: Little & Rubin, 1987; Kalton & Kasprzyk, 1982; Hu et al., 2000; and Seastrom, 2002.

Most researchers examining imputations for non-response warn that imputations may have both positive and negative effects on estimates. Kalton & Kasprzyk (1982) pointed out three positive aspects of imputations. First, imputations may reduce bias of estimates arising from missing responses. Second, by filling out missing data with plausible values, the analysis can be conducted as if the data set is complete, and therefore complex analysis methods are not required even in the presence of missing values. Third, imputations produce consistent results across analyses as researchers do not need to apply an incomplete data set and work with the same set of complete cases.

However, imputations also have negative impacts. Imputation methods do not necessarily lead to a bias reduction in estimation when compared with analyses using incomplete data set (Kalton & Kasprzyk, 1982). Rather, imputations can introduce more bias depending on the imputation method or the properties of estimates of interest (Kalton & Kasprzyk, 1982). Imputations may also distort the relationship between variables (*e.g.* an attenuation of relationship with other variables) (Brick & Kalton, 1996). For these reasons, Kalton & Kasprzyk (1982) cautioned against the danger of researchers treating the completed data obtained from imputations as if all the data were actual responses, and overstating the accuracy and precision of the estimates. They emphasised that researchers working with a data set containing imputed values should proceed with caution, and be aware of imputations for the variables in their analyses as well as the details of the possible impact of imputations on the estimates.

A third approach in dealing with missing data is weighting. Weight adjustment, also known as statistical benchmarking, can be used in statistical analyses to ensure that demographic composition of sample is consistent with the population, whereby bias arising from the missing information can also be reduced (Abbasi, 2000). The achieved sample may not accurately represent the population due to non-random sampling or differing response rates across subgroups of the population. Therefore, weighting begins by breaking the population into benchmarking subgroups with shared common characteristics, such as gender, race, and socio-economic status. For example, suppose the ratio of female to male students in a given institution is 50:50. If the achieved student sample is composed of about 30% females and 70% males, it would not accurately represent the entire institution and cause bias in estimation. To obtain more accurate and precise estimates, there needs to be an adjustment of the weights of the respondents used in the analysis, so that the sampling distribution corresponds to its total population. In the sample institution above, the female weight would be increased while the male weight reduced after adjusting the weights.

6.2 Response rate and student motivation

One issue associated with conducting low-stakes learning outcomes assessments in a higher education context for which participation is not mandatory for students, is the low response rate. A low response rate will increase sampling bias if the non-response is unequal among participants. Generally, students' response rate is viewed as an important indicator of the quality of the assessment and high response rates are more likely to provide more accurate assessment results (Rea & Parker, 1997).

One additional issue occurs when the test results have little impact on students' academic standing or graduation. In such cases, the students' lack of motivation to perform well on the tests could seriously threaten the validity of the test scores and interpretation accuracy of the test results (Banta, 2008; Wise & DeMars, 2005, 2010; Haladyna & Downing, 2004). For example, motivated students could outperform unmotivated students (Liu, 2012; Cole & Osterlind, 2008; Wise & DeMars, 2005; O'Neil et al., 1995/1996), and therefore, institutions with more motivated students may appear to have produced more value-added scores in the learning outcomes assessments (Liu, 2012).

To avoid such problems, institutions must develop appropriate mechanisms, use a variety of incentives to recruit students and enhance participation rate, as well as to motivate students to perform well on the tests. Despite careful random selection of students, there is no guarantee that the selected students will be equally motivated to make their best effort in responding to the test. Strategies to motivate students include setting up specific assessment days, mandating students to take the test, increasing the stakes of the tests by making the test scores contribute to the course grades, providing extra monetary compensation for higher performance, and providing feedback to students to maximize their effort in taking low-stakes learning outcomes assessments (Braun et al., 2011; Duckworth et al, 2011; O'Neil et al., 2005; Baumert & Demmrich, 2001; O'Neil et al., 1995/1996; Wolf & Smith, 1995). Other strategies focus on a range of incentives to students (e.g. cash rewards, gift certificates, curse credits, bookstore coupons, and campus copy cards) in exchange for participation (Liu, 2012).

Researchers especially have proposed various solutions to examine student motivation in taking low-stakes tests and eliminate the impact of low performance motivation on test results. One possibility is to use self-report surveys to measure student motivation in taking low-stakes tests (Liu, 2012), such as the Student Opinion Survey (SOS) capturing students' reported effort and their perception of the importance of the test. The other option is to examine response time effort for computer-based tests, being provided with as much time as is necessary to finish all the questions, to determine student motivation (Liu, 2012; Wise & Kong, 2005).

6.3 Student mobility

The student mobility also causes bias especially in higher education where student mobility is relatively high (OECD, 2008). Higher education students tend to change programmes, take a leave of absence, or even drop out of school halfway through, which results in difficulties keeping track of students for years.

The rate of student mobility varies among institutions. When there is greater student mobility in one institution than in others, the estimated institution effects produced by the value-added models can be biased. For example, some students may leave the institution right before the test administration or may not have spent sufficient time in the institution to be included in the analysis. Therefore, an amount of the institution's efforts on student achievement may not be reflected in the results. Furthermore, if an institution continues to have an intake of students with low capabilities or high achieving students continue to leave the institution as a result of student mobility, this institution's value-added score would be lower than its true value and results in downward bias.

Therefore, it is necessary to examine how the student mobility rate influences the estimation of institution effects on student achievement in the value-added models, and find an appropriate way to reflect the level of student mobility of each institution in the model.

6.4 Model misspecification

Model misspecification should also be examined. When omitted variables are one or more important causal factors that should be in the model, the estimated institution effects can be distorted or biased (*i.e.* omitted variable bias), although the analysis model used is satisfactory in itself (OECD, 2008).

In statistics, the classical linear regression model based on ordinary least squares can provide the best, linear, and unbiased estimates when the model fulfils the assumptions. One of the most important assumptions is that the error term should be uncorrelated with the independent variables. The independent variables are usually expected to explain much of differences on an observation or a unit.

However, if some variables explaining the heterogeneity between observations or units are not included in the model, the unexplained heterogeneity goes into the error term and then the included independent variables must be correlated with the error term (Gujarati & Porter, 2009). For example, one value-added model includes only the levels of parental education attainment to control for the influence of socio-economic factors of institution effects on student achievement. There could be, however, other socio-economic factors, such as parents' occupations, family income, and further inter-generational relationships that could influence student achievement. In that case, the value-added model may result in under-adjustment, which means that the estimates could be biased (Gujarati & Porter, 2009; OECD, 2008).

A response to omitted variable bias might be to include every possible variable in the model. In this case, including irrelevant variables possibly increases the standard errors of other variables and then leads to inaccurate confidence intervals. Moreover, including too many variables that measure the same concept can lead to multicollinearity issues.

Consequently, variable selection and model specification may be a complex problem. Rosenbaum (1999) and Clarke (2005) suggested starting with a study on broad theories in relation to the research field which can help

make predictions on relationships across a variety of fields and factors. It allows a far more comprehensive and elaborated research design and helps in selecting an appropriate analysis model.

In addition, it is essential for researchers to think carefully about the relationships between student achievement and the factors affecting it. Researchers must ensure that these relationships are correctly modelled. They also need to take into account the omission of important variables from the selected value-added model and consider how to address the resulting associated issues.

6.5 Fluctuations in value-added scores across years

One of the most important issues in value-added modelling is the stability of the institution effects across years (Steedle, 2011; OECD, 2008). The institution effects on student achievement, of course, can vary from year to year, but it can cause some problems if the institution effects have changed radically. If the estimated institution effects fluctuate substantially from year to year, it is hardly convincing that accurate estimates of the institution effects on student achievement have been produced.

Fluctuations in value-added scores over time could be due to the actual changes in institution effects on student achievement. In many cases, however, fluctuations can be caused by other factors such as the change of the assessment instruments, the drastic change in each institution's administrative or financial situations, and/or the change in student composition (OECD, 2008; Ray, 2007). If the tests being used to assess student achievement are constructed using different frameworks and have different psychometric characteristics compared with the tests used to assess entering academic ability, qualitatively as well as quantitatively different outcomes could be obtained, even though the same value-added model is employed (Lockwood et al., 2007; Sass & Harris, 2007). For these reasons, it is essential to examine the year-to-year consistency of a value-added score for each institution after selecting the value-added model. Subsequently, if there are radical and substantial changes in value-added scores across years, careful considerations should be made to any important factors affecting the year-to-year consistency of value-added scores which should be included in the value-added model but which are not yet included.

In addition, the smaller the institution, the more the sampling variability increases. In general, the smaller the institution's sample size, the greater and unstable year-to-year differences in the institution effects can be observed (OECD, 2008). Therefore, the use of a three-year average as each institution's value-added scores or the exclusion of institutions below a certain size (*e.g.* institutions with annual cohorts of less than 20-30 students) from the sample are suggested to reduce the year-to-year fluctuation in institution effects (OECD, 2008; Ballou, 2005).

7. Conclusion

This report presented various value-added models used in K-12 education and higher education for a better understanding of value-added measurement. The report presented the key features of each model, its strengths and weaknesses. In addition, the criteria which can be practically applied in selecting the most appropriate value-added model for a given data set and education context in higher education, and various measurement issues for improving selected value-added models were also discussed.

As seen above, value-added measurement can provide policy makers and prospective students with evidence of student learning in educational institutions for external accountability purposes. It can also be used internally by institutions to inform discussions on ways to improve general education programmes or the general intellectual skills of their students (Steedle et al., 2010). The results of value-added measurement can help institutions identify their own strengths and weaknesses in their service provisions and learn more about achieving learning outcomes by benchmarking against other institutions admitting students of similar entering academic ability.

The estimated institution effects on student academic growth (*i.e.* value-added scores) can vary depending on the type of the value-added model selected and its specifications (Steedle, 2011; Steedle et al., 2010; OECD,

2008; Banta & Pike, 2007; Klein et al., 2007). The selection of the appropriate value-added model may be guided by the advantages and disadvantages of each value-added model as they relate to:

- the statistical and methodological issues,
- the properties of the data available (*i.e.* the points in time where the data is collected and the number of observations at a time),
- the complexity of modelling,
- the difficulties of interpretation,
- the costs and resources needed for implementation, and
- the policy goals for a value-added measurement (*e.g.* accountability or improvement).

In addition, even when the same analysis method is used, the resulting estimates may also differ considerably depending on the model specification such as the functional form (e.g. linear function, polynomial function, or log function) and variables included in the model (Steedle et al., 2010; Gujarati & Porter, 2009; Raudenbush & Bryk, 2002). Therefore, it is important to make sure that the model is correctly specified for a given data set, although no one can be confident that all the relevant independent variables and their relations with the dependent variable have been completely identified (Raudenbush & Bryk, 2002).

The complexity of developing or selecting the appropriate value-added model clearly indicates the results of value-added measurement should not be considered as the only source of indicators for making high-stakes decisions (Klein et al., 2007) despite careful consideration of methodological issues (*i.e.* technical constraints and available resources) as well as political issues (*i.e.* the policy objective, use of results, and impact of implementation of value-added measurement). In some instances, the institution identified as 'best' based on a value-added assessment may not be regarded as 'best' with respect to other criteria, because the value-added model gives greater weight to standardised test scores and quantified information than to other indicators (Braun et al., 2010). Therefore, it is essential that the results of value-added measurement be used with other qualitative as well as quantitative indicators, such as institutional portfolios involving references to education context, academic performance, faculty and student retention rates, and institutional efforts and best practices to improve education quality.

REFERENCES

Abbasi, Z. (2000). *Reducing measurement error in informal sector surveys.* Australian Bureau of Statistics. Retrieved July 28, 2012, from: http://mospi.nic.in/informal_paper_17.htm

Aitkin, M., & Longford, N. T. (1986). Statistical modelling issues in school effectiveness studies. *Royal Statistical Society, 149*(1), 1-43.

Ballou, D. (2005). Value-added assessment: Lessons from Tennessee. In R.W. Lissitz (Ed.), *Value-added models in education: Theory and application* (pp. 272-297). Maple Grove, MN: JAM Press.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for students background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics, 29*(1), 37–66.

Banta, T. (2008). Trying to clothe the emperor. *Assessment Update*, *20*, 3-4, 16-17.

Banta, T. W., & Pike, G. R. (2007). Revisiting the blind alley of value added. *Assessment Update. 19*(1), 1-2, 14-15. San Francisco: Wiley Periodicals Inc.

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education*, *16*, 441-462.

Beck, N. (2001). Time series cross section data: What have we learned in the past few years? *Annual Review of Political Science*, 4, 271-293.

Bennett, D. C. (2001). Assessing quality in higher education. *Liberal Education, 87*(2). Retrieved July 21, 2012, from http://www.aacu.org/liberaleducation/le-sp01/le-sp01bennett2.cfm

Blaich, C. F., & Wise, K. S. (2011). *From gathering to using assessment results: Lessons from the Wabash National Study* (NILOA Occasional Paper No.8). Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes.

Bollen, K. A., Christ, S. L., & Hipp, J. R. (2004). Growth curve model. In M. Lewis-Beck, A. Bryman, & T. F. Liao (Eds.), *The Sage encyclopedia of social science research methods* (pp. G35-G38). Thousand Oaks, CA: Sage.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. New York: Wiley.

Braun, H. I. (2005a). *Using student progress to evaluate teachers: A primer on value-added models*. Princeton, NJ: Educational Testing Service, Policy Information Center. Retrieved July 17, 2012, from http://www.ets.org/Media/Research/pdf/PICVAM.pdf

Braun, H. I. (2005b). Value-added modeling: What does due diligence require? In R. W. Lissitz (Ed.), *Value added models in education: Theory and applications* (pp. 19-39). Maple Grove: JAM Press.

Braun, H., Chudowsky, N., & Koenig, J. (Eds.) (2010). *Getting value out of value-added: Report of a workshop*. Washington, DC: National Academies Press. Retrieved July 29, 2012, from http://www.nap.edu/catalog.php?record_id=12820

Braun, H., Kirsch, I., & Yamamoto, K. (2011). An experimental study of the effects of monetary incentives on performance on the 12th grade NAEP reading assessment. *Teachers College Record*, *113*, 2309-2344.

Brick, J. M., & Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research, 5*, 215-238.

Bryk, A. S., & Raudenbush, S. W. (1988). Heterogeneity of variance in experimental studies: A challenge to conventional interpretations. *Psychological Bulletin, 104*(3), 396-404.

Burstein, L. (1980). The analysis of multi-level data in educational research and evaluation. *Review of Research in Education*, 8, 158-233.

Clark, P., Crawford, C., Steele, F., & Vignoles, A. (2010). The choice between fixed and random effects models: some considerations for educational research. *DoQSS Working Papers, 10-10. D*epartment of Quantitative Social Science - Institute of Education, University of London.

Clarke, K. A. (2005). The phantom menace: Omitted variable bias in econometric research. *Conflict management and peace science, 22*, 341-352.

Cole, J. S., & Osterlind, S. J. (2008). Investigating differences between low- and high-stakes test performance on a general education exam. *The Journal of General Education*, *57*, 119-130.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, F., Mood, A. M., Weinfeld, F. D., et al. (1966). *Equality of educational opportunity.* Washington, DC: National Center for Educational Statistics.

Copas, J. B. (1983). Regression, prediction, and shrinkage. *Journal of the Royal Statistical Society, Series B (Methodological), 45*(3), 311–354.

Cronbach, L. J., & Furby, L. (1970). How should we measure "change"- or should we? *Psychological Bulletin*, 74, 68-70.

Curran, P. J. (2003). Have multilevel models been structural equation models all along ? *Multivariate Behavioral Research, 38*, 529-569.

Curran, P. J., & Muthén, B. O. (1999). The application of latent curve analysis to testing developmental theories in intervention research. *American Journal of Community Psychology, 27*, 567-595.

Doran, H. C., & Lockwood, J. R. (2006). Fitting value-added models in R. *Journal of Educational and Behavioral Statistics, 31*(2), 205-230.

Doran, H. C., & Izumi, L. T. (2004). *Putting education to the test: A value-added model for california.* San Francisco: Pacific Research Institute.

Downes, D., & Vindurampulle, O. (2007). *Value-added measures for school improvement*. Melbourne: Department of Education and Early Childhood Development.

Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R., & Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. *Proceedings of the National Academy of Sciences*, *108*, 7716-7720.

Efron, B., & Morris, C. (1973). Stein's estimation rule and its competitors : An Empirical Bayes approach. *Journal of the American Statistical Association, 68*(1), 117-130.

Ray, A., McComack, T., & Evans, H. (2009). *Value-added in English schools*. Education Finance and Policy, 4(4), 415-438. Retrieved March 31, 2013, from http://libra.msra.cn/Publication/5336508/value-added-in-english-schoolsEwell, P. T. (2009). *Assessment, accountability, and improvement: Re-visiting the tension* (NILOA Occasional Paper No.1). Urbana, IL: University of Illinois and Indiana University, National Institute of Learning Outcomes Assessment.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation, 4*(1), 1-58.

Goe, L., Bell, C., & Little, O. (2008). A*pproaches to evaluating teacher effectiveness: A research synthesis.* Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved July 28, 2012, from: http://www.tqsource.org/publications/EvaluatingTeachEffectiveness.pdf

Goldschmidt, P., Choi, K., & Martinez, F. (2004). *Using hierarchical growth models to monitor school performance over time: Comparing NCE to scale score result*s. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles.

Goldstein, H. (1997). Methods in school effectiveness research. *School Effectiveness and School Improvement, 8*(4), 369-395.

Goldstein, H., Rasbash, J., Yang, M., Woodhouse, G., Pan, H., Nuttall, D., & Thomas, S. (1993). A multilevel analysis of school examination results. *Oxford Review of Education, 19*(4), 425-433.

Gujarati, D. N., & Porter, D. C. (2009). *Basic econometrics* (5th Ed.). Boston: McGraw-Hill.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, *23*, 17-27.

Hanushek, E. A. (2007). Education production functions. In S. N. Durlauf, & L. E. Blume (eds.), *The New Palgrave Dictionary of Economics*. Basingstoke: Palgrave Macmillan. Retrieved March 31, 2013, from http://hanushek.stanford.edu/publications/resources-efficiency

Hart Research Associates. (2009). *Learning and assessment: Trends in undergraduate education – A survey among members of the Association of American Colleges and Universities.* Washington, DC: Hart Research Associates.

Harvey, L. (2004-12). Analytic q*uality glossary, quality research international.* Retrieved on July 3, 2012, from http://www.qualityresearchinternational.com/glossary/

Harvey, L., & Green, D. (1993). Defining quality. *Assessment and Evaluation in Higher Education, 18*(1), 9–34.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2011). *The elements of statistical learning: Data mining, inference, and prediction* (5th Ed.). New York: Springer. Retrieved on August 31, 2012, from http://www-stat.stanford.edu/~tibs/ElemStatLearn/

Haveman, R., & Wolfe, B. (1995). The determinants of children's attainments: A review of methods and findings. *Journal of Economic Literature, 33*, 1829-1878.

Heck, R. H. (2006). Assessing school achievement progress: Comparing alternative approaches. *Educational Administration Quarterly, 43*, 667-699.

Hibpshman, T. (2004). *A review of value-added models*. Frankfort, KY: Kentucky Education Professional Standards Board.

Hox, J. (2002). *Multilevel analysis*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Hox, J., & Stoel, R. D. (2005). Multilevel and SEM approaches to growth curve modeling. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 3, pp. 1296–1305). London: Wiley.

Hu, M., Salvucci, S. M., & Cohen, M. P. (2000). Evaluation of some popular imputation algorithms. *2000 Proceedings of the Section on Survey Research Methods.* American Statistical Association.

Jakubowski, M. (2008). Implementing value-added models of school assessment. *EUI Working Papers RSCAS 2008/06, European University Institute*.

Kalton, G., & Kasprzyk, D. (1982). Imputing for missing survey responses. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 22-31.

Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions*. Thousand Oaks, CA: Sage Publications.

Klein, S., Benjamin, R., Shavelson, R., & Bolus, R. (2007). The Collegiate learning assessment: Facts and fantasies. *Evaluation Review, 31*(5), 415-39.

Klein, S., Steedle, J., & Kugelmass, H. (2009). *CLA Lumina Longitudinal Study summary findings*. New York: Council for Aid to Education.

Kline, R. B. (1998). *Principles and practice of structural equation modelling.* New York: Guilford Press. Retrieved on August 31, 2012, from http://psychology.concordia.ca/fac/kline/Supplemental/latent_d.html

Ladd, H. F., & Walsh, R. P. (2002). Implementing value-added measures of school effectiveness: Getting the incentives right. *Economics of Education Review,* 21, 1-17.

Lenkeit, J. (2012). Effectiveness measures for cross-sectional studies: a comparison of value-added models and contextualised attainment models. School Effectiveness and School Improvement, 24(1), 39-63.

Leveille, D. E. (2006). *Accountability in higher education: A public agenda for trust and cultural change*. Berkeley, CA: Center for Studies in Higher Education, University of California, Berkeley.

Lindley, D. V., & Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society (Series B), 34*, 1-41.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: John Wiley.

Liu, O. L. (2008). *Measuring learning outcomes in higher education using the measure of academic proficiency and progress (MAPP)* (ETS RR-08-47). Princeton, NJ: ETS.

Liu, O. L. (2011). Value-added assessment in higher education: A comparison of two methods. *Higher Education, 61*, 445-461.

Liu, O. L. (2012). Outcomes assessment in higher education: Challenges and future research in the context of voluntary system of accountability. *Education Measurement: Issues and Practice, 30(3)*, 2-9.

Liu, O. L, Bridgemen, B., & Adler, R. M. (2012). Measuring learning outcomes in higher education: Motivation matters. *Educational Researcher*, 41(9), 352-362.

Lockwood, J. R., & McCaffrey, D. F. (2007). Controlling for individual level heterogeneity in longitudinal models, with applications to student achievement. *Electronic Journal of Statistics, 1*, 223-252.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: RAND Corporation.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., Hamilton, L., & Kirby, S. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1), 67-102.

Montaquila, J. M., & Ponikowski, C. H. (1995). An evaluation of alternative imputation methods. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 281-286.

O'Neil, H. F., Abedi, J., Miyoshi, J., & Mastergeorge, A. (2005). Monetary incentives for low-stakes tests. *Educational Assessment*, *10*, 185-208.

O'Neil, H. F., Sugrue, B., & Baker, E. L. (1995/1996). Effects of motivational interventions on the National Assessment of Educational Progress mathematics performance. *Educational Assessment*, *3*, 135-157.

Organisation for Economic Co-operation and Development. (2008). *Measuring improvements in learning outcomes: Best practices to assess the value-added of schools.* Paris: OECD.

Pike, G. R. (1992). Lies, Damn Lies, and Statistics Revisited: A comparison of three methods of representing change. *Research in Higher Education, 33*, 71–84.

Ponisciak, P. M., & Bryk, A. S. (2005). Value-added analysis of the Chicago Public Schools: An application of hierarchical models. In R. Lissitz. (Ed.), *Value added models in education: Theory and applications.* Maple Grove, MN: JAM Press.

Reardon, S. F., & Raudenbush, S. W. (2009). Assumptions of value-added models for estimating school effects. *Education Finance and Policy, 4*(4), 492-519.

Raudenbush, S. W. (2004). *Schooling, statistics, and poverty: Can we measure school improvement?* Princeton, NJ: Educational Testing Service.

Raudenbush, S. W., & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of Education, 59,* 1-17.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd Ed). Thousand Oaks, CA: Sage Publications.

Rea, L. M. and Parker, R. A. (1997). *Designing and Conducting Survey Research: A Comprehensive Guide*. San Francisco: Jossey-Bass.

Rindfleisch, A., Malter, A. J., Ganesan, S., & Moorman, C. (2008). Cross-sectional versus longitudinal survey research: Concepts, findings, and guidelines. *Journal of Marketing Research, 45*, 261-279.

Rosenbaum, P. R. (1999). Choice as an alternative to control in observational studies. *Statistical Science, 1*4, 259–304.

Rowan, B., Correnti, R., & Miller, R. J. (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.

Sanders, W. L. (2006). *Comparisons among various educational assessment value-added models*. The Power of Two-National Value-Added Conference, Columbus, OH. Retrieved on July 17, 2012, from http://www.sas.com/govedu/edu/services/vaconferencepaper.pdf

Sanders, W. L., & Horn, S. P. (1994). The Tennessee Value-added Assessment System (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education, 8*, 299-311.

Sanders, W. L., & Horn, S. P. (1998). Research findings from the Tennessee Value-added Assessment System (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education, 12*(3), 247-256.

Sanders, W. L., Saxton, A. M., & Horn, S. P. (1997). The Tennessee Value-added Assessment System: A quantitative, outcomes-based approach to educational assessment. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 137-162). Thousand Oaks, CA: Corwin.

Schmitz, D. D., & Raymond, K. J. (2008). *The utility of the cumulative effects model in a statewide analysis of student achievement.* Paper presented at the American Educational Research Association Annual Meeting. New York.

Seastrom, M. M. (2002). *2002 Statistical standards and guidelines*. Washington, DC: U.S. Department of Education, National Center for Education Statistics, NCES 2003-601. Retrieved on August 08, 2012, from http://nces.ed.gov/statprog/2002/stdtoc.asp

Shin, T. (2007). Comparison of three growth modeling techniques in the multilevel analysis of longitudinal academic achievement scores: Latent growth modeling, hierarchical linear modeling, and longitudinal profile analysis via multidimensional scaling. *Asia Pacific Education Review, 8*(2), 262-275.

Steedle, J. T. (2009). *Advancing institutional value-added score estimation*. New York: Council for Aid to Education.

Steedle, J. T. (2010). *Improving the reliability and interpretability of value-added scores for post-secondary institutional assessment programs.* Paper presented at the Annual Meeting of the American Educational Research Association, Denver, CO.

Steedle, J. T. (2011). Selecting value-added models for postsecondary institutional assessment. *Assessment & Evaluation in Higher Education*, DOI:10.1080/02602938.2011.560720.

Steedle, J., Kugelmass, H., & Nemeth, A. (2010). What do they measure? Comparing three learning outcomes assessments. *Change 42*(4), 33–7.

Teacher Advancement Program. (2012). *Understanding value-added analysis of student achievement*. Retrieved on July 17, 2012, from http://www.tapsystem.org/policyresearch/policyresearch.taf?page=valueadded

Tekwe, C. D., Carter, R. L., Ma, C., Algina, J., Lucas, M., Roth, J., Ariet, M., Fisher, T., & Resnick, M. B. (2004). An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics, 29*(1), 11-36.

Todd, P., & Wolpin, K. I. (2003), On the specification and estimation of the production function for cognitive achievement. *Economic Journal, 113*(485), 3-33.

Townsend, T. (2007). School effectiveness and improvement in the twenty-first century: Reframing for the future. In T. Townsend (Ed.), *International handbook of school effectiveness and improvement* (pp. 933-962). Dordrecht, The Netherlands: Springer.

Traub, R. E. (1967). A note on the reliability of residual change scores. *Journal of Educational Measurement, 4*, 253–256.

Voluntary System of Accountability. (2008). *Background on learning outcomes measures*. Retrieved May 18, 2009, from http://www.voluntarysystem.org/index.cfm?page=about_cp

Wainer, H. (2004). Introduction to the value-added assessment. *Journal of Educational and Behavioral Statistics, 29*(1), 1-3.

Wiley, E. W. (2006). *A practioner's guide to value added assessment.* Educational Policy Studies Laboratory Research Monograph. Tempe, AZ: Arizona State University. Retrieved on August 12, 2012, from http://nepc.colorado.edu/files/Wiley_APractitionersGuide.pdf

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

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*(1), 1-17.

Wise, S. L., & DeMars, C. E. (2010). Examinee non-effort and the validity of program assessment results. *Educational Assessment*, *15*, 27-41.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163-183.

Wolf, L. F., & Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education*, *8*, 227-242.

Wright, S. P., White, J. T., Sanders, W. L., & Rivers, J. C. (2010) *SAS EVAAS statistical models.* SAS EVAAS Technical Report. Cary, NC: SAS Institute, Inc. Retrieved from

http://www.sas.com/resources/asset/SAS-EVAAS-Statistical-Models.pdf

Yu, L., Lai, K. K., Wang, S., & Huang, W. (2006). A bias-variance-complexity trade-off framework for complex system modeling. In M. Gavrilova et.al. (Eds.), *ICCSA 2006* (pp. 518-527). Retrieved on August 31, 2012, from http://www.springerlink.com/content/mw10522133566686/fulltext.pdf

NOTES

[1] This literature review focuses on the concept of value-added as it relates to learning gains, and does not address the value-added in terms of economic gains. (See Rodgers, 2007 for a possible methodology for developing a performance indicator based on the economic value added to graduates.)

Appendix: Comparison of selected value-added models used in K-12 education and higher education

Models		Key Features	Strengths	Weaknesses	References
K-12 Education	OLS linear regression models (covariate adjustment models)	Adjust student test score for some combination of student prior test score and student or contextual characteristics Assume that the regression coefficients are the same for all schools The value-added score for each school is taken to be a mean residual of students from a given school	Are simple to specify and fit using any standard statistical software Can be extended to models where scores from successive years are nonlinearly related, via higher-order polynomial terms	Students missing either the prior year or current year test score are excluded, and then value-added scores are unstable and biased when it can't be assumed that students whose score gains are missing are not selected at random Do not reflect the multilevel nature of data structure in education (*i.e.* unmodeled systematic heterogeneity needs to be removed from the error term)	Gujarati & Porter (2009) OECD (2008) Jakubowski (2008) McCaffrey et al. (2003) Sanders (2006) Beck (2001)
	Fixed effects models (LSDV model)	Assume that each school has its own fixed effects on student achievement Take account of unmeasured differences between schools by allowing each school to have its own dummy variables as additional predictors in an analysis of covariance mode1	Reflect the multilevel nature of the data structure in education, where students are nested within schools Require no assumptions about the school effects (*c.f.* random effects models assume that school effects are drawn from a normal distribution)	Run up against the degrees of freedom problem when too many dummy variables are introduced into the model Possible multicollinearity among a lot of dummy variables School-level covariates seem to have no impact on differences in student achievement, because school-specific intercepts may absorb all differences in	Clark et al. (2010) Gujarati & Porter (2009)
	Fixed effects models (within-group model)	When school mean scores and mean covariates are subtracted from individual student scores and covariates, unmodeled differences between schools	Reflect the multilevel nature of the data structure in education, where students are nested within schools By subtracting school mean values from	The effects of school-level covariates cannot be identified because these school characteristics would be differenced out when doing subtraction	Clark et al. (2010) Gujarati & Porter (2009)

Models		Key Features	Strengths	Weaknesses	References
		can also be differenced out of the model Then, an OLS regression is conducted using all demeaned values at the student-level, just like in the OLS linear regression model above	individual student values, it can produce more consistent estimates of slope coefficients than the OLS linear regression models	method The estimated school effects may vary considerably from year to year, since there is no use of 'shrinkage' used in random effects models to reduce effects of sampling variation	
	Random effects models (multilevel models, hierarchical linear models, or mixed models)	Assumption 1: there is a bigger population of schools and their value-added scores, and each value-added score is chosen at random from the population In two-level models, each student's score is represented as a function of individual characteristics at level-1 At level-2, regression coefficients at level-1 are conceived as dependent variables and represented as a function of school characteristics Assumption 2: residuals at both level (*i.e.* student- and school-level) must be uncorrelated with covariates In the level-2 equation, the deviation from the expected test score is taken as value-added score for each school	Reflect the multilevel nature of the data structure in education, where students are nested within schools The regression coefficients and value-added scores for each school are more statistically efficient by using the shrinkage estimates (*i.e.* having smaller mean-squared error, and thereby generating narrower confidence intervals for estimates) Can estimate coefficients of school-level covariates, unlike the fixed effects models	The assumption that residuals must be uncorrelated with covariates cannot be satisfied when important school characteristics affecting student achievement (*e.g.* student motivation) are not included in the model Can introduce bias because the shrunken estimates would be far below or above the true school effects if schools have a small number of students or the within-school variance is large relative to the between-school variance	Clark et al. (2010) Gujarati & Porter (2009) Raudenbush & Bryk (2002) Copas (1983) Efron & Morris (1973) Lindley & Smith, 1972
	Growth curve models	Analyse trajectories of students over time (at least three time points) to estimate the school contribution to	Reflect the reality that students start out at different levels and grow at different	Rely heavily on the quality of the longitudinal data set, which is greatly affected by student mobility or grade	Shin (2007) Ponisciak &

Models		Key Features	Strengths	Weaknesses	References
		student academic growth In a three-level growth model, each student's development is represented by an individual growth trajectory at level-1 At level-2, the dependent variables (*i.e.* initial status and growth rate of each student) are represented as a function of a set of individual characteristics At level-3, dependent variables (*i.e.* mean initial status and mean growth rate within each school) are represented as a function of a set of school characteristics	rates Can produce the correlation of the growth parameters, such as initial status and growth rate, as well as their relation with time-varying and time-invariant covariates Can use data from all students, even those with partially complete records	repetition May have a great deal of measurement errors due to repeated measures over time, and thereby the precision and accuracy of estimation could be negatively affected by repeated measurements	Bryk (2005) Bollen et al. (2004) Goldschmidt et al. (2004) Raudenbush & Bryk (2002) Curran & Muthén (1999)
	Multivariate random effects models (EVAAS model)	Assume that the school effects on student achievement persist in and can be carried over to all succeeding years Focus not only on how well a student does in a given subject, grade, and year in a school where the student is currently attending, but also on the accumulated knowledge and skills acquired in previous school For the analysis, detailed identification about each school and its students should also be collected from multiple	Total school effects on the student academic growth is partitioned according to the proportion of time spent in each school Allow for the use of incomplete data, therefore it can reduce the sample selection bias, and consequently provide more precise estimates and narrower confidence intervals Is highly parsimonious and efficient compared to other value-added models because it does not require controlling for either incoming academic ability or	The omission of covariates can lead to biased estimates when students are systematically different from each other and stratified by the covariates Cannot account for the possible influence of covariates on student achievement and academic growth School effects at one point in time can be affected by the earlier school effects In reality, school effects actually have diminished over time and may not affect student's future growth, unlike the	Wright et al. (2010) Lockwood & McCaffrey (2007) Sanders (2006) Ballou et al. (2004) Tekwe et al. (2004) Sanders & Horn (1998,

Models		Key Features	Strengths	Weaknesses	References
		subjects across several grades annually Student achievement is represented by a vertically linked series of a standardized achievement test which is administered annually in one or more subjects	other covariates	assumption underlying these models Require a longitudinal data pertaining to the same students and the same subjects over years	1994) Sanders et al. (1997)
Models		**Key Features**	**Strengths**	**Weaknesses**	**References**
Higher Education	Difference in residuals models (OLS linear regression-based)	Intend to capture whether students' academic growth between entering freshmen and graduating seniors in a given institution is near or above what is typically observed at institutions admitting students of similar entering academic ability To measure the expected test scores, these models carry out regressions of current test scores on entering academic ability scores for freshmen and seniors, respectively Value-added score for each institution can be obtained by subtracting the freshman residual from the senior residual (can also be defined as the difference between institution's observed residuals and expected residuals)	Are easy to implement and its results are fairly straightforward to interpret Using the cross-sectional design for value-added measurement is less costly and more feasible to implement than the longitudinal design	Require an appropriate standardised test score reflecting student's entering academic ability (*e.g.* SAT scores) Are faced with a dilemma of an assumption of a linear relationship between the mean current test scores and the entering academic ability scores: - if both scores are not linearly related to each other, the assumption underlying this model is substantially violated, and thereby biased estimates are produced. - if both scores are highly linearly related to each other, the reliability of residuals is fairly low and tends to decrease as correlation between both increases	Steedle (2009, 2010, 2011) Klein et al. (2007)

Models		Key Features	Strengths	Weaknesses	References
	Difference in residuals models (HLM-based)	Compute value-added scores of each institution based on the difference between freshman residual and senior residual, just like the OLS-base model Use two-level HLM in calculating freshman and senior residuals, respectively - at level-1, each student's test score is represented as a function of student's entering academic ability score - at level-2, the level-1 coefficients for each institution are conceived as dependent variables and assumed to depend on institutional characteristics (*i.e.* the deviation from the grand mean after controlling for the institutional characteristics is taken as the freshman or senior residual)	Produce more accurate and precise estimates to calculate school effects (*i.e.* freshman and senior residuals) than OLS linear regression-based difference in residuals model by: - applying the multilevel model for reflecting the nested data structure in higher education - using shrunken estimates of residuals produced by the HLM Provide the standard error of residuals for freshmen and seniors respectively, which can be used as an estimate of precision of residuals to compute the confidence interval for freshman and senior residuals	Have exactly the same shortcomings the OLS linear regression-based difference in residuals models have - Require an appropriate standardised test score reflecting student's entering academic ability - the reliability of residuals is fairly low and tends to decrease as the correlation between the mean current test scores and the entering academic ability scores increases Have the potential problems with the random effects models: - the assumption that residuals must be uncorrelated with covariates should be satisfied - can introduce bias because of the shrunken estimates of residuals	Liu (2008, 2011) Steedle (2011)
	HLM-based residual analysis models	Compare senior mean test scores for each institution instead of difference in scores between freshman and senior Produce value-added scores based on the degree to which observed senior mean test scores exceed or fall below expectations after controlling for student's entering academic ability: - at level-1, the senior test score is represented as a function of student's	Increase the reliability and year-to-year consistency of value-added scores for each institution compared to the OLS-based difference in residuals model by: - applying the multilevel model for reflecting the nested data structure in higher education - using shrunken estimates of residuals produced by the HLM Provide an estimate of value-added	Even though, these models increase the reliability of the institution effects compared to the OLS-based difference in residuals model, it is unlikely to be adequate for using value-added scores to make high-stakes decisions Although multicollinearity between senior mean SAT and freshman mean test scores at level-2 does not reduce predictive power or reliability of the	Steedle (2009, 2010, 2011)

Models		Key Features	Strengths	Weaknesses	References
		entering academic ability score - at level-2, the level-1 coefficients for each institution are conceived as dependent variables and assumed to depend on senior mean entering ability score and freshman mean test score (*i.e.* the residual represents the value-added score for each institution)	score precision for each institution, which can be used to compute a unique confidence interval for each institution's value-added score	model, the regression model at level-2 may not produce valid coefficients Have exactly the same shortcomings the OLS linear regression-based difference in residuals models have Have the potential problems with the random effects models	