Pai PENG, Jan HOCHWEBER, Eckhard KLIEME

Test Score or Student Progress? A Value-Added Evaluation of School Effectiveness in Urban China

Abstract Outcome-oriented evaluation of school effectiveness is often based on student test scores in certain critical examinations. This study provides another method of evaluation—value-added—which is based on student achievement progress. This paper introduces the method of estimating the value-added score of schools in multi-level models. Based on longitudinal student achievement data, two measures of school effectiveness in one local education authority in China are compared. It is found that the between-school difference in both test-score and value-added is large comparable with that of Western countries. The results of the two measures of school effectiveness are highly different. The value-added measures lack consistency across different subject areas within schools while the test score measures are highly correlated between subjects. Teachers show their preference for value-added measures over test-score measures of education quality. It is suggested that value-added measures of school effectiveness should be used as a complement to rather than a substitute for test-score measures. The shortcomings of value-added approach are also discussed.

Keywords value-added, test score, school effectiveness, school evaluation

Background

The evaluation of school effectiveness is always a challenging task in various

Pai PENG (✉)

German Institute for International Educational Research (DIPF), Frankfurt am Main 60486, Germany

E-mail: peng@dipf.de

Jan HOCHWEBER

German Institute for International Educational Research (DIPF), Frankfurt am Main 60486, Germany

Eckhard KLIEME

German Institute for International Educational Research (DIPF), Frankfurt am Main 60486, Germany

social and educational contexts. The output of a school, often measured by its average student achievement in some critical examinations (e.g., GCSE in England, *Gaokao* and *Zhongkao* in China), has been the predominant criterion for assessing school effectiveness. Although academic achievement only reflects a narrow set of educational goals¹, it is widely used in school evaluation for its availability and understandability. From the perspective of students, parents, and other stakeholders of the educational system, students' test scores are natural indicators of school effectiveness and quality. The successful implementation of PISA now makes even international comparisons possible, though mainly depending on test scores for reading, mathematics, and science (OECD, 2010). A number of factors make test scores a strong signal for school effectiveness evaluation in China, such as emphasis on academic excellence in the Confucian tradition, the highly selective high school system, and widespread school choice practices.

However, test scores at one time point cannot reveal the extent of a school's contribution, or in other words, how effectively a school promotes student learning. The deficiency of test score measures or status models mainly lies in the fact that schools and students are not randomly assigned: Some schools always have higher-achieving intakes than others. In the past two decades of educational effectiveness research, the value-added approach has been viewed as a fairer and more accurate way of evaluating the academic effectiveness of schools (Sammons, 2007; Thomas, 2010; Braun, 2005; OECD, 2008). The multi-purpose, uses and consequences of value-added measures have been discussed extensively (Thomas, 1998; Thomas & Mortimore, 1996; Braun, Chudowsky, & Koenig, 2010). Value-added models have also attracted considerable attention from practitioners, particularly in some US states, for teacher performance (Braun, 2005) and in the UK for school effectiveness (Thomas, 2010). There are also some exploratory studies conducted in high schools in China based on the value-added approach (e.g., Peng, Thomas, Yang, & Li, 2006; Ma, Peng, & Thomas, 2006; Thomas et al., 2012).

This study is one part of the research project, "An Exploratory Study of School Effectiveness in China." The project is funded by a scholarship offered by Konrad-Adenauer-Foundation (KAS) in Germany. In this study, the value-added approach is applied to evaluate the effectiveness of lower secondary schools in one urban Local Education Authority (LEA) of China. The purpose of this study is to compare value-added and test score measures of school effectiveness both theoretically and empirically. Some policy implications will also be discussed.

¹ Test scores are incomplete measures of student achievement, since scores of some core subjects only partly represent the cognitive skills of students. Also, scores cannot reveal intellectual curiosity, motivation, persistence in tackling difficult tasks, or the ability to collaborate well with others.

Research Questions

In contrast to test scores, the term value-added refers to the school residuals/school effects² in modeling student final achievements after statistically adjusting for student intake factors that are not controlled by the schools (e.g., baseline attainment, socioeconomic status). The value-added score is interpreted as a measure of the direct effect that schools have had on student achievement (McCaffrey, Lockwood, Koretz, & Hamilton, 2004). It reveals the effects of a wide range of school/classroom processes and context factors. Focusing on value-added provides an estimate of school effectiveness in addition to test scores. Many types of statistical models can be applied to estimate the value-added score of schools, but in practice there is no dominant one (Braun et al., 2010). Peng and Klieme (2012, October) listed three value-added models that have been commonly used: the linear regression model, the fixed-effects model, and the random-effects or multilevel model. They also provided different model adjustment strategies that have been adopted in practice.

The authors consider the following four questions in detail:

- (1) How large is the between-school difference in school effectiveness based on test score and value-added (covariate-adjusted test score)?
- (2) To what extent are test score and value-added measures of school effectiveness different?
- (3) Are test score and value-added measures of school effectiveness consistent across different subjects?
- (4) Which measure of effectiveness is more sensible from the perspective of teachers?

Methodology

Sample and Data

The data collection was conducted in one LEA of urban China (Hongshan district in Wuhan city). This is mainly because (1) national representative longitudinal data on student achievement is not available; and (2) the existing study in China (e.g., Thomas et al., 2012) indicates that the school effectiveness evaluation in a single LEA might be more appropriate than that in the whole country³. There are 25 middle schools (Grades 7 to 9) in this LEA. One cohort of students (aged

² In this study, school effectiveness, school effects, and school quality are used interchangeably.

³ Their study was conducted in three LEAs and the results (parameter estimates) were quite different across LEAs.

15/16, Grade 9 in 2011/12 school year, as called “2012 cohort” later) in each of these schools is included in the sample. The following variables are used in the analysis.

Student Achievement Measures

At the end of Grade 9, all students in China take the High School Entrance Examination (HSEE). The student’s total scores in this exam determine the kind of high school in which he/she can be enrolled (key high school, ordinary high school or vocational school). The HSEE test scores provide an estimate of students’ final achievements in lower secondary schooling. At the end of the 2011/12 school year, 4,490 students took the HSEE in this LEA (exam information is shown in Table 1).

Table 1 High School Entrance Exam in Wuhan, 2012

Subject	Type	Full mark	Time (minutes)	Date
Chinese		120	150	20 June 2012
Mathematics	Paper-based (Closed-book)	120	120	20 June 2012
English		120	120	21 June 2012
Physics and chemistry		130	120	21 June 2012
Politics and history and society	(Open-book)	80	90	20 June 2012
Physical education and health	/	30	/	8 to 20 April 2012

Source: Wuhan Office of Admission and Examination. (2012). 2012 年武汉市高中阶段学校考试招生问答 [2012 Wuhan HSEE Admission Questions Manual]. Retrieved May 16, 2013, from <http://www.whzkb.com/html/2012-4-28/3284.html>; Wuhan City Office of Admission and Examination. (2012). 市教育局关于印发2012年武汉市初中毕业升学体育考试实施方案的通知 [Wuhan City Education Bureau Notice on the 2012 Issuance of Junior and Middle School P.E. Examination Procedures]. Retrieved May 16, 2013, from <http://www.whzkb.com/html/2012-4-10/3262.html>

According to the 2006 *Compulsory Education Law*, middle schools cannot select students on the basis of any type of test scores (Standing Committee of the National People’s Congress, 2006, Art. 12). They are required to abandon ability sorting and accept any students who finish primary schools in their districts. Although there is an exit exam at the end of primary education, this exam lacks difficulty and discrimination. Therefore, in many LEAs there is no proper prior attainment measure of students, which is a necessity in estimating the value-added of middle schools. The Hongshan LEA may be an exception. All new entrants of middle schools are required to take a diagnostic test (DT). This test aims to check the initial ability of students and how they grasped the basic knowledge and competences learned in primary schools. The test was conducted

shortly after the students were enrolled. The DT score as the student baseline attainment was used. Of the 2012 cohort, 4,053 students took part in this test in 2009. The student achievement data was linked according to school and student identifier.

Both the HSEE and the DT are strictly designed to meet the National Curriculum Standards. They are organized and graded by external experts. One difference between them is that the diagnostic test is for new Grade 7 students only in the Hongshan LEA and the HSEE is for Grade 9 students in the whole of Wuhan city. The descriptive statistics of the total test scores along with scores for each subject are shown in Table 2. The correlation coefficient between the DT and the HSEE score is 0.87 (total score), 0.69 (Chinese), 0.80 (mathematics), and 0.81 (English). In the analysis, both test scores were standardized (mean = 0, standard deviation = 1).

Table 2 Descriptive Statistics of Student Achievement

Variable	Mean	Std. Dev.	Min.	Max.
DT scores (prior attainment)				
Chinese	68.75	13.92	2.00	98.00
Mathematics	60.30	23.53	2.00	100.00
English	67.63	22.83	8.00	100.00
Total	196.68	54.41	20.00	298.00
HSEE scores (final achievement)				
Chinese	79.46	13.46	3.00	111.00
Mathematics	77.57	18.31	6.00	117.00
English	74.84	20.25	8.00	115.00
Total	231.87	47.38	24.00	337.00

Note. Total = Chinese + mathematics + English; Full DT scores are 100 for each subject; Full HSEE scores are 120 for each subject. The sample includes students who have both DT and HSEE scores.

It should be noted that there exists sample attrition over the course of the three years. For students with DT scores, 20.1% do not have HSEE scores; for students with HSEE scores, 9.7% do not have DT scores. One possible reason is that there is a large number of pupils from rural migrant workers' families in this LEA. These students often transfer to another city (or back home to the countryside) if their parents change jobs and move. Also, the correlations between missing data rate (defined as percentage of students who have DT score but no HSEE score in a school) and school mean DT score is -0.43 , showing lower achieving schools have a slightly higher percentage of sample attrition. Furthermore, students

without HSEE scores have achieved less than students with HSEE scores in the DT within each school. These factors may lead to an underestimation of the value-added of low achieving schools because students with lower-than-average achievement are likely to have higher than average progress⁴.

Student Background Variables

Student background information was collected from the questionnaire survey, which was conducted when the students were in Grade 9. The data collection was conducted with the help of the Research Center of Hongshan Education Authority, who issued an approval letter for the survey and informed principals of the 25 schools about this survey, asking them to provide assistance. Ten research assistants were recruited to conduct the survey after training. The collected questionnaires were delivered to two graduate students for data entry. More details of the survey can be found in Peng (2013, forthcoming).

In the analysis three background variables were used: gender, age in months and socioeconomic status (SES). Of the sample, 54.6% were boys and their HSEE total scores are statistically lower than girls. In total, 99.3% of students were born between 1995 and 1998. The age gap seems quite large for one cohort of students. One possible reason is the large number of students from migrant workers' families. These students sometimes repeated a grade on transferring from the countryside to urban schools. The correlation of age and HSEE total score is -0.20 . Father's education level is used as a proxy of the student SES. It is a categorical variable ranging from 1 (primary education or lower) to 5 (graduate education or higher). In the analysis it is treated as a continuous variable. The correlation of SES and HSEE total score is 0.39 . Note that the sample size was reduced from 4,053 to 3,568 because of missing data on these background variables.

School-Level Variables

School-level variables are not often included in the value-added models since they may be related to the estimated school effects. However, for the purpose of evaluation, a school should not be held accountable for its context, which it cannot control. In this study, two school aggregate variables were used to describe the school context, namely the mean student DT scores and mean SES. Both of the variables are centered on the grand mean in the analysis.

⁴ An indicator variable for students who performed above-average was included in the DT in the value-added models. It was found these students make less progress than students with below-average DT score.

Teacher Attitudes

To investigate teachers’ opinions regarding the two measures of effectiveness, the authors conducted a teacher survey in this LEA. The questionnaires were randomly allocated to teachers of the 2012 cohort when the student survey was conducted. A total of 168 teacher questionnaires were collected (with a response rate of 88%). Teachers were asked to give a point score to alternative evaluation criteria, as shown in Table 3.

Table 3 Item Design for Teachers’ Attitudes towards Alternative Evaluating Methods

“To what extent do you think it is reasonable to use the following methods of evaluating education quality?”	
Items	Point score
1. Test score (e.g., HSEE) of the students in my class;	1 (least reasonable)
2. Student progress relative to their baseline attainment at the entry of school;	2
3. Student progress in my class relative to other classes with the similar intake in the school;	3
4. Student progress in my class relative to other classes with the similar intake in the whole district.	4
	5
	6
	7 (most reasonable)

Methods

To answer questions 1 to 3, the authors apply multi-level modeling to estimate the value-added school effectiveness. Multilevel modeling, which was specially designed to analyze data with nesting or hierarchical structure, has been frequently employed in education effectiveness studies. This is because student achievement in the same class or school is more alike than the achievement from different classes or schools. This “clustering” or “nesting” feature of the data violates one assumption of the single-level multiple regression: The measured units are independent and the residuals of units are uncorrelated. A special advantage of multilevel modeling is that it enables the researchers to investigate the nature of between-group variance (e.g., between-school variance), and the effects of group-level predictors (e.g., school type, leadership) on individual outcomes. For simplicity’s sake, only the random intercept was included in the model. The simplest form of value-added models is shown in model (1)

$$Y_{ij2} = \alpha_0 + \alpha_1 Y_{ij1} + \zeta_j + \varepsilon_{ij} \tag{1}$$

where $\zeta_j \sim N(0, \tau^2)$ and $\varepsilon_{ij} \sim N(0, \sigma^2)$. Y_{ij2} represents the final achievement of student i in school j . Y_{ij1} represents the baseline achievement at the entry of school and α_1 is the regression coefficient. The fixed intercept α_0 is the intercept for the average school. The random intercept ζ_j is the individual school’s

value-added effect on student learning. ε_{ij} represents the student residual. Model (1) only includes student prior attainment and sometimes called the “progress-only” specification.

$$Y_{ij2} = \alpha_0 + \alpha_1 Y_{ij1} + \sum_p \beta_{pij} X_{pij} + \zeta_j + \varepsilon_{ij} \quad (2)$$

$$Y_{ij2} = \alpha_0 + \alpha_1 Y_{ij1} + \sum_p \beta_{pij} X_{pij} + \sum_q \beta_{qj} Z_{qj} + \zeta_j + \varepsilon_{ij} \quad (3)$$

The Model (2) is, called “Type-A” specification. It includes a vector of p variables of student background (X_{pij}) besides prior achievement. In this study, they are student gender, age, and SES. The Model (3) adds a vector of q variables of school context factors (Z_{qj}), including the mean prior achievement and the mean SES. It is called the “Type-B” specification. The Type-A effect is the effect parents generally consider when choosing a school and The Type-B effect is the one school officials consider when evaluating the performance of those who work in schools (Raudenbush & Willms, 1995). Each school’s value-added score (ζ_j) can be estimated by empirical Bayes prediction. For example, the estimate of Type-B effect is shown in the formula (4), which was provided by Raudenbush and Bryk (2002, p. 154)

$$\begin{aligned} \zeta_j^{EB} &= \frac{\tau^2}{\tau^2 + \sigma^2 / n_j} \hat{\zeta}_j \\ &= \frac{\tau^2}{\tau^2 + \sigma^2 / n_j} (\bar{Y}_{2j} - \hat{\alpha}_0 - \hat{\alpha}_1 Y_{j1} - \sum_p \hat{\beta}_{pj} X_{pj} - \sum_q \hat{\beta}_{qj} Z_{qj}) \end{aligned} \quad (4)$$

where ζ_j^{EB} is the empirical Bayes estimate of the value-added, n_j is the number of students in school j , $\hat{\zeta}_j$ (the term in parenthesis) is the OLS estimator of the effects for each school. Other things being equal, schools with fewer students will have stronger shrinkage. For the ranking and comparison purpose, an approximate 95% confidence interval of the value-added is calculated: $(\zeta_j^{EB} \pm 1.96 * S.E._j)$.

To answer the question 4, T-tests were employed to check the difference in teachers’ opinion between alternative evaluation criteria, particularly between the test score and value-added (item 1 and 2).

Findings

Firstly the authors used a two-level (school, student) empty model to decompose the variation of HSEE total score into two parts. Then a step by step process of the student baseline achievement, student background characteristics and

school-level characteristics was used to build value-added models of various specifications. The results are shown in the Table 4.

Table 4 Estimates of Multi-level Models of HSEE Total Score

	Test-Score Model		Value-Added Model	
	empty model (1)	progress-only (2)	Type-A (3)	Type-B (4)
fixed part				
DT score		0.916*** (0.008)	0.886*** (0.009)	0.887*** (0.009)
SES			0.028** (0.009)	0.028** (0.009)
Boy			-0.040** (0.015)	-0.039** (0.015)
Age in months			-0.003*** (0.001)	-0.003*** (0.001)
Mean DT score				0.447** (0.133)
Mean SES				0.383** (0.144)
Constant	-0.192* (0.092)	0.002 (0.052)	0.609*** (0.156)	0.6479*** (0.153)
Random part				
School level	0.212*** (0.064)	0.066*** (0.020)	0.061*** (0.018)	0.043*** (0.014)
Student level	0.818*** (0.018)	0.198*** (0.004)	0.182*** (0.004)	0.182*** (0.004)
ICC	0.206	0.250	0.250	0.192
log likelihood	-5964.3	-2521.3	-2089.1	-2086.9
N of students	4490	4053	3568	3568

Note. a) * Standard errors in parentheses. * $p<0.05$, ** $p<0.01$, *** $p<0.001$; b) In the Type B model, student-level DT score and SES are centered around the school mean; School-level DT score and SES are centered around the grand mean.

The parameter estimates of all models are briefly introduced. As shown in the fixed part of the value-added model, baseline achievement (DT score) is the best predictor of student final achievement in the HSEE. One standard deviation change of baseline achievement can predict at least 0.88 standard deviation changes of the

HSEE total score. The effect of individual SES is significantly positive after student baseline achievement is controlled for⁵. The school mean SES and mean achievement can also positively predict the HSEE score. This implies that students with similar baseline achievement and SES benefit more from the schools with higher mean SES and higher mean baseline achievement. The authors find that boys make less progress than girls, and older students achieve less than younger students in the same cohort. The reason may be that older students have repeated a grade because of previous underachievement or self-selection by their parents⁶.

Question 1: How Large Is the Between-School Difference in Effectiveness based on Test-Score and Value-Added?

The size of between-school difference can be indicated by the intra-class correlation (ICC), which measures the proportion of the total variance that is accounted for by the higher-level units (Snijders & Bosker, 2012). From the random part of the empty model it is found that 20.6% of HSEE total score variance is situated at the school level. After controlling for student baseline attainment, variance at the school and student level are reduced dramatically. Student DT score can explain 69% of the between-school variance and 76% of the within-school variance in HSEE total score. Other student-level background variables (gender, age and SES) also reduce the HSEE variance at both levels. The authors also found that the residual ICCs (0.250) in progress-only and Type-A models are higher than that of the empty model⁷. This finding may be unique in the Chinese educational context because the residual ICCs often range between 0.05 and 0.20 and they are often smaller than raw ICCs in western countries (Hedges & Hedberg, 2007, as cited in Snijders & Bosker, 2012). This implies that schools are more different in the value-added measure of effectiveness than in test-score measures in this education authority. It should be borne in mind that schools are even more heterogeneous in student achievement at the entry into school, for the raw ICC of DT total score is 0.295⁸. This means

⁵ In an early version of this paper, a composite SES indicator generated from several indices (parental education, books at home, occupational status, and family wealth) was used. The effect of SES was not significant in the early analysis so we used a more transparent measure of SES in this version: father's education level.

⁶ Parents tend to send their children to school later than required if they think their children are not ready or not clever enough for the schooling. These older students in the same cohort may make less progress than their schoolmates.

⁷ The ICC of Type-B model cannot be directly compared with those of model (1), (2), and (3) because the predictors are scaled differently.

⁸ The authors collected student achievement data at the end of each school year. We found ICC reduced gradually from Grade 7 to Grade 9, for both total score and individual subject score.

that some schools have made such great progress in student achievement during three years that they have reduced the achievement gap between schools in this district. Both the raw and residual ICC found in this study are larger than those found in developed countries but lower than those found in developing countries (Scheerens & Bosker, 1997; Scheerens, 2001).

Question 2: To What Extent Are Test-Score and Value-Added Estimates of School Effectiveness Different?

The value-added measure of school effectiveness can be predicted using the equation (4). The authors’ estimate each school’s value-added based on the Type-B model. The result is shown in Fig. 1. School ID 18 has the highest value-added and school ID 23 has the lowest. A 95% confidence interval of each school’s value-added is also shown. Value-added scores of nine schools are statistically different from zero. These schools are called outliers in the value-added evaluation. The other 16 schools are not statistically different from the average⁹.

Fig. 1 Value-Added Score and Ranking of 25 Schools

The test score measure of school effectiveness can be divided into two types: a/ average HSEE score of schools and b/ the percentage of students above a

⁹ The authors also estimate the value-added in the progress-only model and the Type-A model. The ranking is similar to that in the Type-B model. Nonetheless, there are 11 outliers in the progress-only model and 13 outliers in the Type-A model.

cut-off score. The latter is considered because only high-achieving students above a cut-off score in the HSEE can be accepted by key high schools in China. In 2012, the cut-off score was 458 points in total score of HSEE. The authors list the mean total score of the three subjects, the percentage of students above the cut-off score and the value-added scores of all 25 schools in Table 5.

Table 5 Different Measures of Effectiveness and the Ranking of 25 Schools

School ID	Mean score	Percentage above the cut-off score	Value-added score	Rank of mean score	Rank of percentage above cut-off score	Rank of value-added
1	220.48	0.32	0.19	15	13	5
2	208.40	0.19	-0.32	19	20	24
3	231.63	0.43	-0.05	11	7	16
4	247.02	0.46	0.01	3	6	12
5	263.72	0.67	0.03	1	1	10
6	243.94	0.47	-0.07	5	5	18
7	216.94	0.22	0.14	17	18	6
8	244.54	0.47	-0.08	4	4	19
9	207.76	0.19	-0.05	20	19	15
10	186.07	0.11	-0.14	24	22	20
11	192.32	0.13	-0.17	22	21	21
12	223.58	0.26	0.20	14	16	4
13	200.04	0.06	0.05	21	25	9
14	212.57	0.35	-0.05	18	10	17
15	224.25	0.31	0.07	12	14	7
16	239.01	0.48	0.06	7	3	8
17	232.90	0.33	0.01	8	12	13
18	232.32	0.43	0.47	9	8	1
19	191.41	0.10	-0.24	23	23	23
20	217.05	0.23	0.02	16	17	11
21	223.67	0.29	0.30	13	15	2
22	254.78	0.58	-0.04	2	2	14
23	174.46	0.09	-0.39	25	24	25
24	231.91	0.33	0.25	10	11	3
25	239.81	0.42	-0.19	6	9	22

Note. Value-added estimated from the Type-B models of HSEE total score.

Obviously mean score and percent above cut-off score are perfectly correlated

(Pearson’s $r = 0.96$, $p < 0.001$). The ranking of schools between the two test-score measures are also highly correlated as expected (Spearman’s $\rho = 0.93$, $p < 0.001$). The value-added measure is neither correlated with the mean score (Pearson’s $r = 0.29$, $p > 0.05$, Spearman’s $\rho = 0.30$, $p > 0.05$) nor correlated with the percentage above the cut-off score (Pearson’s $r = 0.23$, $p > 0.05$, Spearman’s $\rho = 0.23$, $p > 0.05$). This means that test score and value-added measures of school effectiveness are highly different. In Table 5, it is shown that schools with high HSEE scores and a high percentage of students above the cut-off score do not necessarily have high value-added (e.g., school ID 8 ranks the fourth in test score measure but it is at the bottom in the value-added score), and vice versa (e.g., school ID 13). This pattern can also be seen in Fig. 2. Schools in the first quadrant have above average test scores and value-added scores and those in the third quadrant are below average in both measures. Schools in the second and fourth quadrants are inconsistent between the two measures of school effectiveness. In practice, schools at the top are more popular with parents because their students are high-achieving. But some of them (in the second quadrant) may make less contribution to their students’ progress than schools in the first quadrant.

Fig. 2 Scatter of HSEE Score and Value-added Score of 25 Schools

Question 3: Are Both Test-Score and Value-Added Measures of School Effectiveness Consistent Across Different Subject Areas?

The authors applied the Type-B model to estimate schools’ value-added in each subject. The parameter estimates are not reported here as the result is not of

direct interest to this study. In Table 6, the correlations of mean test scores and value-added scores in each subject are presented. It is found that mean test scores of schools are highly correlated across subjects ($r = 0.90\text{--}0.95$), which means that schools have higher scores in one subject tend to have higher scores in other subjects. Value-added measures of teaching quality are moderately correlated between subjects ($r = 0.54\text{--}0.77$). This finding demonstrates that value-added measures of school effectiveness are far less consistent across subject areas than test score measures. The finding is in line with what Thomas and Mortimore (1996) reported. They found that value-added measures for GCSE English and mathematics have a correlation of 0.46 in England. This result implies that there may be a subject department effect within these schools. The total picture of school effectiveness should be separated into different subject areas.

Table 6 Correlation Matrix of Test-Score and Value-Added for Each Subject

	Test-score Chinese	Test-score mathematics	Test-score English	Value-added Chinese	Value-added mathematics
Test-score mathematics	0.95				
Test-score English	0.92	0.90			
Value-added Chinese	0.26	0.23	0.05		
Value-added mathematics	0.21	0.30	0.08	0.77	
Value-added English	0.14	0.08	0.19	0.57	0.54

Note. Value-added is estimated from the Type-B models of HSEE subject score.

Question 4: Which Measure of Education Quality Is More Sensible for Evaluation from the Perspective of Teachers?

The authors examined teacher opinions about alternative evaluation methods. Average scores that teachers gave to items 1 and 2 are compared in Table 3. From teachers’ perspective, it is more reasonable to evaluate education quality by students’ progress with respect to baseline achievement than by the average test score of examinations at the end of the school year or the HSEE ($5.1 > 4.1$, $df = 167$, $t = 8.15$, $p < .000$). This means that teachers of the 2012 cohort of students preferred the value-added measure to the test-score measure of education quality. With respect to value-added measures, the authors compared the average score that teachers gave to items 3 and 4 in Table 3, finding that teachers also showed their preference for being compared with all classes in the LEA rather than just being compared with the classes in the same school ($4.8 > 4.2$, $df = 167$, $t = 4.90$, $p < .000$). This finding is not surprising because most schools have fewer than four classes in the same cohort.

Conclusions and Implications

The two measures of school effectiveness in one urban district in the Mainland of China are summarized as follows, with results as shown below:

1) There are significant variations across schools, in both the test-score and value-added measure of school effectiveness.

2) Student baseline attainment, together with other student background variables can explain a large amount of variation in student HSEE achievement.

3) Value-added measures were not found to be significantly related with the test-score measure of school effectiveness. Some schools with “better” intakes have lower value-added than those with lower achieving students, and vice versa.

4) The test-score measure of school effectiveness is highly consistent across subject areas, while value-added measures lack consistency across different subjects. A school which has above-average value-added in one subject does not necessarily have an above-average value-added score in other subjects.

5) From the perspective of teachers, value-added measures based on student progress are more reasonable indicators of education quality than test-score measures.

Evaluating school effectiveness based on value-added is not new in the UK and the US. Some of the empirical findings in this study are not new, either. The literature accumulated on value-added over the last two decades is extensive (e.g., Clotfelter & Ladd, 1996; Ladd & Walsh, 2002; OECD, 2008). However, value-added evaluation is a relatively new practice in the Mainland of China. The authors argue that the value-added measure of school effectiveness is particularly important in Chinese basic education. This is mainly because achievement level measured in mean test scores is a strong signal of education quality on which parents rely to make school choice decisions. It is also a major criterion for governments to evaluate schools. This has led to the Mathew effect of school development. The gap between schools has become larger and larger because more resources were given to schools with higher student achievement in HSEE and other critical examinations. “Balanced development” of compulsory education schools is one of the mid- and long-term aims between 2010 and 2020 in the recent agenda of educational reforms¹⁰. This education policy aims to reduce the large gap between schools, particularly within LEAs (counties and districts). As former Prime Minister Wen Jiabao said, “currently test scores and the rate of admission into high schools and colleges are practically the sole criterion to evaluate students, teachers, and schools. The pressure is great for

¹⁰ See Central People’s Government of the People’s Republic of China. (2010). 国家中长期教育改革发展规划纲要 (2010–2020 年) [National agenda for mid- and long-term educational reform and development (2010–2020)]. Retrieved May 13, 2013, from http://www.gov.cn/jrzg/2010-07/29/content_1667143.htm

students, as well as teachers and schools. A scientific and diversified education evaluation system should be built as soon as possible" (Central People's Government of the People's Republic of China, 2012, Art. 2.2). Furthermore, the surveyed teachers tend to prefer the student-progress measures over level of achievement as the evaluation criterion of education quality. The authors believe that the value-added approach can help the realization of the policy aim if the value-added score of schools are publicized in a proper and transparent way.

Although value-added is a fairer way of evaluating schools than the test-score league tables, its measurement is more complex. The basic requirements are: a longitudinal database that tracks individual students accurately and links them to their schools; legitimate reasons for missing data; and expert staff to run or monitor the value-added analyses (Braun et al., 2010). These requirements can only be met in a limited number of LEAs in China. Also, various editions of value-added models can be adopted in practice. It is necessary to consider which model is more suitable in China and adapt this approach to meet the particular need of each LEA. It is necessary to disseminate more knowledge about value-added evaluation in China.

As Sammons (1999) pointed out, value-added is a necessary but not sufficient condition for the development of "good" schools. Our findings show that value-added is quite different from test-score as a measure of school effectiveness. Also, value-added is sensitive to the choice of baseline achievement. Schools and teachers might "cheat" in the baseline achievement test at the entry of schooling to give students a "low" starting achievement, in order to get higher value-added. As Braun et al. stated (2010), "even if we are attracted to value-added, and certainly value-added has many advantages over status systems, we are not ready to give up on status" (p. 61). Consequently, value-added score is not a proper substitute for test score in evaluating school effectiveness. It should be used as a complement in combination with the test score measures.

The lack of consistency of value-added scores across different subjects implies that school effectiveness or quality should not be evaluated only by an overall value-added score. This finding is of vital importance to self-evaluation of schools. Schools can identify more effective subject departments or even teachers. If value-added measure is used for high-stakes purposes (e.g., rewarding schools with high value-added or punishing schools with low value-added), overall value-added based on total score may lead to unintended consequences, such as the "free rider" problem. The ineffectiveness of some departments may be masked by effective teachers of other subject areas within the same schools. Lots of research findings in school effectiveness research imply that the concept of secondary school effectiveness needs to be qualified by the term "school and departmental effectiveness" (Sammons, 2007; Sammons, Thomas, & Mortimore,

1997). Thus a more elaborate value-added evaluating system should be built, in order to identify the multidimensionality of school effectiveness.

It should be noted that this study was conducted for one cohort of students in one of nearly 3,000 LEAs in China. The conclusions in this paper should be treated with caution. For example, the narrowing gap in achievement between schools over the three years may not be true in other LEAs. Our future research agenda includes: adding more cohorts of students to look at stability of the value-added measure, conducting value-added analysis in more LEAs, and considering how to integrate the value-added indicators into the existing evaluation system in China.

References

- Braun, H. I. (2005). *Using student progress to evaluate teachers: A primer on value-added models*. Princeton, NJ: ETS.
- Braun, H., Chudowsky, N., & Koenig, J. (Eds.). (2010). *Getting value out of value added: Report of a workshop*. Washington, DC: National Academies Press.
- Central People's Government of the People's Republic of China. (2012). 温家宝在全国教师工作暨“两基”总结表彰会上讲话 [Wen Jiabao's promotional speech regarding the “Two Basics” project on nine years' compulsory education and universal literacy]. Retrieved May 17, 2013, from http://www.gov.cn/lhdh/2012-09/09/content_2220253.htm
- Clotfelter, C. T., & Ladd, H. F. (1996). Recognizing and rewarding success in public schools. In H. F. Ladd (Ed.), *Holding schools accountable: Performance-based reform in education* (pp. 23–64). Washington, DC: Brookings Institution.
- Ladd, H. F., & Walsh, R. P. (2002). Implementing value-added measures of school effectiveness: Getting the incentives right. *Economics of Education Review*, 21(1), 1–17. doi: 10.1016/S0272-7757(00)00039-X
- Ma, X. Q., Peng, W. J., & Thomas, S. (2006). 学校效能的增值评价: 对河北省保定市普通高中学校的实证研究 [School effectiveness with value-added method: Case study of senior high schools of Baoding Hebei Province]. *教育研究 [Educational Research]*, (10), 77–84.
- McCaffrey, D. F., Lockwood, J., Koretz, D. M., & Hamilton, L. S. (2004). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: RAND Corporation.
- OECD. (2008). *Measuring improvements in learning outcomes: Best practices to assess the value-added of schools*. Paris, France: OECD Publishing.
- OECD. (2010). *PISA 2009 results: What students know and can do? Student performance in reading, mathematics and science* (Vol. 1). Paris, France: OECD Publishing.
- Peng, P., & Klieme, E. (2012, October). One-year value-added school effects from various models and their inter-temporal variability: Evidence from China. Paper presented at the UMD-INVALSI Conference: Improving Education through Accountability and Evaluation Lessons from around the World. Rome, Italy.
- Peng, P. (2013, forthcoming). *An exploratory study of school effectiveness in China* (Unpublished doctoral dissertation), University of Frankfurt. Frankfurt am Main, Germany.
- Peng, W. J., Thomas, S. M., Yang, X., & Li, J. (2006). Developing school evaluation methods to improve the quality of schooling in china: A pilot “value added” study. *Assessment in*

- Education: Principles, Policy and Practice*, 13(2), 135–154. doi: 10.1080/09695940600843252
- Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
- Raudenbush, S. W., & Willms, J. D. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics*, 20(4), 307–335.
- Sammons, P. (1999). *School effectiveness: Coming of age in the twenty-first century*. Lisse, The Netherlands: Swets & Zeitlinger.
- Sammons, P., Thomas, S., & Mortimore, P. (1997). *Forging links: Effective schools and effective departments*. London, England: Paul Chapman.
- Sammons, P. (2007). *School effectiveness and equity: Making connections*. Reading, England: CfBT Education Trust.
- Scheerens, J. (2001). Monitoring school effectiveness in developing countries. *School Effectiveness and School Improvement*, 12(4), 359–384. doi: 10.1076/sesi.12.4.359.3447
- Scheerens, J., & Bosker, R. (1997). *The foundations of educational effectiveness*. Oxford, England: Pergamon.
- Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). London, England: Sage.
- Standing Committee of the National People's Congress. (2006). Compulsory Education Law of the People's Republic of China. Retrieved May 17, 2013, from <http://www.chinaeducer.com/en/cedu/cel.php>
- Thomas, S. (1998). Value-added measures of school effectiveness in the United Kingdom. *Prospects*, 28(1), 91–108. doi: 10.1007/BF02737782
- Thomas, S. (2010). Assessment and the evaluation of institutional effectiveness. In P. L. Peterson, E. L. Baker, & B. McGaw (Eds.), *International encyclopedia of education* (3rd ed., vol. 1, p. 172–180). London, England: Elsevier.