Southwest Basic Education Project (SBEP)

Analysis of the Impact of SBEP on Student Achievement

June 2012

Yang Min, Ding Yanqing and Hu Wenbin

Cambridge Education

Southwest Basic Education Project (SBEP)

Analysis of the Impact of SBEP on Student Achievement

June 2012

Yang Min, Ding Yanqing and Hu Wenbin

Cambridge Education, Demeter House, Station Road, Cambridge CB1 2RS, United Kingdom
T +44(0) 1223 578500 F +44(0) 1223 578501, W www.camb-ed.com

Issue and revision record

Revision	Date	Originator	Checker	Approver	Description
<Click here>					

Cambridge Education, Demeter House, Station Road, Cambridge CB1 2RS, United Kingdom
T +44(0) 1223 578500 F +44(0) 1223 578501, W www.camb-ed.com

Content

Executive Summary

The DFID funded Southwest Basic Education Project (SBEP) operated during 2006 – 2010 in 27 of the poorest and most remote counties in southwestern China. The project was designed to improve the Provincial and County governments' own systems or education support and development and included : stipends for poor students (especially girls and minorities) ; introduction of School Development Planning (SDP) ; teacher training on effective support, improved quality and greater relevance of schooling for disadvantaged children ; involvement of those children in their own learning ; head teacher training ; and equity training focusing on the most disadvantaged children.

The SBEP project Goal was: "Increased and equitable access to high quality basic education in all counties targeted in the Government of China's Nine Year Compulsory Education Programme in the Western Region." Its Project Purpose was: "To support the Government of China to achieve its goals in basic education, by increasing Government capacity to improve effective programmes that increase equitable access, completion and achievement for the most excluded boys and girls."

The Student Achievement Study (SAS) was designed to assess the potential impact of the project interventions on students' performance, especially the performance of disadvantaged children. It aimed to answer three questions: (i) Whether students in project counties made more progress than those in non-project counties over project period? (ii) Whether project interventions had greater impact on disadvantaged students than less disadvantaged students over project period? (iii) Could project interventions have greater impact on disadvantaged students for improved performance in some areas of knowledge than the other? All questions were related to changes of students' performance from where they started at the beginning of the SBEP project to where they ended up 2 years or 4 years after the project interventions. It was assumed that if the project interventions were effective on disadvantaged children, a greater change in performance of disadvantaged than that of not disadvantaged over the project period should be expected, and performance of children in these areas should be improved overall.

The study was a panel design of three waves, in which school achievement in Chinese and Math were examined in students in grades 3, 5, 7 and 9 in the same 520 schools at 2007 (baseline test), 2009 (middle term review, MTR) and 2011 (end of project evaluation, EoP) respectively. The baseline sample consisted of about 43,000 students from project counties and about 3,000 students from non-project counties as a control group. The control group was matched to the project intervention group by the township socio-economic status, school type, mean age of students and gender ratio. The same tests were administrated in the control group at same years.

The testing scores of three waves were equated using anchor items and applying item response theory modelling, which produced a standardized and normally distributed measure of learning ability of students. This measure was comparable over time.

The analyses were carried out from three angles. Firstly, they assessed overall progression of all students in the project sample from the baseline to the mid-term (a 2-year period), and from the baseline to the end of project (a 4-year period) in comparison to that of students in the non-project control sample (stratified by gender by grades). This analysis was based on aggregated means of the ability scores with weighting by number of students, using meta-regression modelling.

Second, it examined the impacts of the SDP interventions on performance progress made by disadvantaged children in comparison to that by not disadvantaged in the project counties. The key measures of disadvantaged were girl, minority, disabled and a composite family socio-economic scale

(SES). The performance progress was assessed over the 2 year and 4 year periods. Multilevel models were used for the analysis of the total ability score, allowing random effects among counties, schools and students with adjustment for student characteristics and school type and so on.

The third angle of analysis was to further examine possible impacts of the SDP interventions on performance differentiated by learning ability on sub-domain knowledge and skills made by disadvantaged children in the project counties. Two core sub-domains for Chinese were Recognition of Characters and Reading, and three for Math were Algebra, Space and Practice. Multivariate multilevel models were used for simultaneous analysis of sub-domain scores and for allowing random effects among schools and students with same adjustment applied in the second set of analysis. The performance change of disadvantaged students in this analysis was over a 2-year period, from the middle term to the end of project.

The key findings were:

- Despite having markedly lower average ability scores at the baseline tests, students of all grades in the project counties made significantly more progress than those in the non-project counties by the MTR and by the EoP, for both Chinese and Math, results were similar for boys and girls.

- Girls of all grades in SDP schools demonstrated a more improved performance in Chinese than girls in non-SDP schools at both mid-term and end of the project, and a small improvement in Maths at the end of project. The same patterns were found from sub-domain analysis.

- Students with low socio-economic backgrounds of all grades in the SDP schools showed a consistent pattern of greater progress in both Chinese and Maths over the project life than those in the non-SDP schools, with half the estimates of those patterns reaching statistical significance. The same consistent pattern was supported by the sub-domain analysis too.

- Minority students in SDP schools demonstrated worse progress overall for both subjects except for students in Grade 9 in Maths ; there were similar findings from the sub-domain analysis.

- There was some weak evidence showing impacts of SDP interventions on improved performance of disabled students to the end of the project though without reaching statistical significance. Positive effects were found among disabled students of Grades 3, 7 and 9, with more estimates of those effects reaching significance level for Chinese sub-domains than for Maths

- The SDP interventions did not show sub-domain differentiated impacts on performance of disadvantaged children.

To conclude, the study has provided evidence to support the positive effects of the SBEP interventions on disadvantaged children in the poorest regions in China. Efficacy of such interventions in the education system appeared measurable by progress of students' achievement or learning ability.

1. Introduction

1.1 Background of the SBEP

From 2006 to 2010, The China-UK Southwest Basic Education Project (SBEP) was carried out in the 27 remotest national poverty counties in the four provinces of Yunnan, Sichuan, Guangxi and Guizhou. The Southwest Basic Education Project (SBEP) supported the achievement of the Chinese Government's target of Nine Year Compulsory Education. It did so by increasing government capacity to implement effective programmes that increased equitable access, improved completion rates and resulted in greater achievement for the most disadvantaged girls and boys. The project directly benefited 1,668,000 children in 27 of the poorest counties in Yunnan, Sichuan, Guizhou and Guangxi provinces (see the map below).

The main rationale for the launch of the project in 2006 was that, even though China had made great progress towards universal primary education in the previous 25 years (official net enrolment rates had increased from 93% in 1980 to 99% in 2005), children from poor families, often from ethnic minorities, faced substantial barriers in accessing quality basic education. Some scholars believe that "the effects of this developmental tendency are the worst for children in poor rural areas and for female children, as inequalities in the distribution of opportunities have already serious affected their opportunities for personal development and upward mobility. Inequality in educational opportunities also leads to the polarization of society, enhances people's sense of inequity, and negatively affects social integration" (Liu, etc. 2009; Li, Chunlin 2012).

The Government turned its attention to access for marginal groups in the poorest counties, with the aim of ensuring that all counties achieve Nine Year Compulsory Education by 2010. The Government was also paying increasing attention to improving the quality of basic education (MOE, 2010). This project aimed to

support the Government in addressing these two challenges by focusing on three main problems: low enrolment and retention rates, particularly at junior middle school level; poor quality of education; and weak education management.

One of the main outcomes the project expected to achieve was that the quality and relevance of schooling for approximately 1.6 million disadvantaged children in 27 counties would be improved. The total budget at the time of project design and approval was £27.0 million, reduced for DFID budget reasons to £23.6 million at the time of the mid-term review in 2009. Each province provided 10% of the amount as counterpart funding which was targeted and used for student assistance. The project was officially launched in November 2006 and was completed by the end of March 2011, thus lasting about five years.

1.2 Intervention activities and expected outputs

Through the integrated implementation of a series of intervention activities, the project was expected to improve equitable access, education quality and education management in project counties. The intervention activities and expected outcomes are as follows (focusing on 5 outputs):

- (Output 2) By providing training for approximately 77,000 teachers on how to use a participatory teaching approach and by supporting effective support systems for teachers' professional development, to achieve the goal of improved quality and relevance of schooling for the disadvantaged children.

- (Output 3) By supporting 1,400 schools in poor townships to carry out school development planning (SDP), by improving the leadership and management capacity of head teachers, and supporting the reform and upgrading of the school inspection system, to achieve the goal of improved systems of school management which promote the interests of the most disadvantaged girls and boys.

- (Output 4) By improving government education management information systems through designing an integrated student-based database, to achieve the goal of improved capacity of monitoring and evaluation systems to orient policy and practice in favour of the most disadvantaged boys and girls.

- (Output 5) By supporting the institutional analyses of the education system at national, provincial and county levels and relevant capacity building activities, to achieve the goal of improved capacity of government education systems to better meet the needs of the most disadvantaged girls and boys.

It can be seen from the outputs, which can be reviewed for effectiveness,that apart from indicators for educational equity, the academic performance of students, especially the performance of disadvantaged students, is also central to the study of the impact of SBEP. The research into the academic performance of students (and disadvantaged students in particular) has to do with the impact evaluation of 4 outputs (Outputs 2 - 5) out of the total of five outputs of the project.

Underpinning the inputs described above were convictions, expressed in the Project Memorandum and Logframe, that improvements in equitable access, quality and management could have impacts on learning achievement.

Globally, there is growing evidence that leadership and management, particularly through a school based management approach, is critical for effective teaching to take place (Khattri, Barrera-Osorio). Related to this is evidence to support the need to improve teacher quality and professionalism as the quality of the teaching can have a significant impact on learning outcomes (vid. Alexander, Barber).

1.3 The Student Achievement Study (SAS)

The purpose of the SAS is to explore whether there have been positive changes in the academic achievements of students in project-counties as a result of SBEP interventions. As the SBEP interventions were emphasised and implemented particularly in the poorest areas for most disadvantaged student groups, examining project effects on performance of disadvantaged students is a focus of the SAS study. This study is purely quantitative, which complements the qualitative study in the End-of-Project (EoP) evaluation and specifically addresses the changes in achievement as measured by tests on students in project and non-project schools, at three points over the life span of the project (from 2006 to 2010). The study aimed at answering three questions: (i) Whether students in project counties made more progress than those in non-project counties over project period (ii) Whether project interventions had a greater impact on disadvantaged students than on less disadvantaged students over project period? (iii) Whether project interventions have had greater impact on disadvantaged students for improved performance in some areas of knowledge than in others.

1.4 Design of SAS

An ideal design of the study was to follow mixed cohorts of students over the project period from the Baseline to the end of the project, i.e. the student cohort starting from Grade 3 at the Baseline would be in Grade 5 in the mid-term of the project, and in Grade 7 at the end of the project. Similarly the student cohort starting from Grade 5 at the Baseline would be in Grade 7 in the mid-term and in Grade 9 at the end of the project. Each student in the two cohorts would have 3 test results during the project period for observing changes over time. However, in reality such tracking at individual level was extremely difficult because students moved between schools, in particular from primary school after Grade 5 to middle school for Grade 7. There was no system in those counties to track students once they moved to different schools. Given the large scale of the project and its limited resources, the SAS adapted panel design to track students at the grade level.

This meant that the same grade in the same school but different students would be followed and tested 3 times over the project period. However, with this design the differences in student achievement over time could come from several sources: (a) the project interventions, (b) the background of different students at different time of testing and (c) other projects' interventions and exposure of students to the rapid socio-economic changes of society as time went by. Although the effects of (b) could be adjusted based on information of student surveys in statistical analysis, there is no way to separate the project effects (a) from the effect (c) unless a control group that had no exposure to the project but was otherwise similar to the students in the project county was set up.

Therefore a control group also in panel design was added in the study for comparison at the average level. The group contained students from non-project counties of the project provinces matched in socio-economic status at both township and school levels, and by age range, percentage of girls, but no exposure to the SBEP project interventions. Two hundred students in each of the Grades 3, 5, 7 and 9 for each province were given the same tests as the project students at the Baseline, mid-term and the end of the project tests.

Assuming a moderate 0.3 SD unit of improvement in the mean test marks by standardised z-score and unbalanced sample with the intervention group 5 times larger than the control group, to detect such change with 90% statistical power at a significant level 5%, 141 students in the control group and 705 in the intervention group would be sufficient in a well-controlled experimental setting. However, as the study is observational with many factors in student background potentially having effects on the exam score and not

controllable, the study required a minimum of 200 students in the control group of each grade at each phase of testing and 2% of student population in the project intervention group of at each phase. A stratified random sampling method was applied at the Baseline test to draw the school samples. It was advised that the same county and schools sampled in the Baseline test would be used for the MTR and EoP tests during the project period. Table 1 presents the actual numbers of schools and students drawn from project counties for the Baseline test.

Table 1: Percentage of Sampled Schools and Students in Project Counties

	Schools			Students		
School	Total Number of Schools in Project Counties	Total Number of Sampled Schools	%	Total Number of Students in Project Counties	Total Number of Students in Sampled Schools	%
Primary	7,927	405	5.1	1,167,000	34,967	2.3
Middle	483	115	23.8	501,000	9,258	1.8

This design implied a sample of above 30,000 students in the project counties and at least 800 or more students in the non-project counties being tested at the Baseline test (2007), mid-term review (MTR, 2009) and end of project evaluation (EoP, 2011) respectively. It was assumed that a much larger sample size than what a well-controlled randomized trial would require should give enough power to detect moderate changes in students' mean test scores over time that could be attributable to project intervention.

As a concern of comparability of the two groups of students, Table 2 below presents 3 key student characteristics between project and non-project district samples for the Baseline test in 2007.

Table 2: Comparison of student characteristics between project and non-project samples at the 2007 test

Grade	Girl:Boy		Minority: Han		Mean age in years	
	Project	Non-proj	Project	Non-proj	Project	Non-proj
3	0.97	1.18	0.92	2.85	10.3	9.7
5	0.92	0.89	0.88	4.77	12.3	12.1
7	0.87	0.90	1.36	3.13	13.6	13.8
9	0.88	0.82	1.33	2.59	15.8	15.9

The project and non-project samples were comparable in their distributions of students' gender and age overall, but there were large differences in the distribution of minority students. This was because the non-project sample was matched by school and township socio-economic status. Only schools with a high concentration of minority students in those counties could be close to the similar social and economic status (SES, please refer to 3.1 for more details) status for the match. Future advanced data analysis should be able to take such difference into consideration.

2. Data Collection

2.1 Administration of student tests

The test papers were developed by local education experts following the national curriculum of the test subjects, Chinese and Mathematics. The project management offices (PMO) at national, provincial and county levels were responsible for the organization of the tests as embedded in the student assessment routines in the education system, including appointment of experts, resourcing and other logistic arrangements for the activity. The National Support Team (NST) of the project was responsible for working with the national project management office (NPMO) to ensure that all the technical preparations of the test (including surveys of students, teachers and schools) were ready. The international and national consultants provided technical support in the development of test items and the piloting of them, as well as manuals to guide the tests in schools and sampling schools. Each county PMO had a team to conduct the test in the sampled schools as instructed by the manual. The same project management system administrated student tests in 2007 and 2009. Since the project officially finished in March 2012, the NPMO was dismissed. For the final wave of testing in 2011, the NST worked directly with the provincial and county PMOs as well as county experts to conduct the EOP test.

The county PMO team was responsible for the quality of the data as well as for administering tests at school level. They were required to visit all schools at the time when implementing the tests, and to ensure there was no missing data from student exam papers. The team collected all the test papers and answer sheets, including the unused ones, marked the papers, and then sent the test papers and scored answer sheets back to NPMO for further data cleaning and data entry

2.2 Data entry and error checking procedure

All data were entered into the data base by scanning. This method certainly saved an enormous amount of time and cost, by cutting down the resources required to enter data and by eliminating human error in manual data entry.

Data was cleaned using systematic approaches. Taking EoP data as an example, data cleaning went through 5 major procedures with 14 steps to check for consistency, data reading errors, duplicated IDs and matching data from different types of questionnaires. The systematic approach assured a high degree of data cleanliness.

2.3 An overview of school and student samples collected in three phases

Table 3 shows the number of school samples of three phases (Baseline, MTR and EoP). The project counties showed a stable school sample, while a decreased school sample from non-project counties was observed over the project period. Table 4 shows the number of students in the samples from project and non-project counties by gender. These figures were the final data used in the analysis for the study report.

The numbers suggest that overall sample size as designed by the study was achieved, with total students from project counties around 30,000, and those from non-project counties between 1,300 and 2,251, well above the minimum 800 at each test. However, a much reduced number of schools and students in the Grades 7 and 9 from non-project counties at the EoP phase may indicate that caution is required in explaining results.

Table 3: Number of schools in the final study sample

Phase	Project counties	Non-project counties
Baseline	499	45
Mid-term	473	33
EoP	487	12

Table 4: Number of students in the final study sample (line 1 for Chinese, line 2 for Maths)

Grade	Baseline (t1)		Mid term (t2)		EOP (t3)	
	Project	Non-project	Project	Non-project	Project	Non-project
Grade 3	12388	673	12341	531	10572	435
	13050	675	12517	528	10816	431
Grade 5	12009	598	11665	562	10602	656
	11937	655	11461	560	10713	642
Grade 7	3140	691	2974	415	3904	133
	3200	691	3058	437	3769	133
Grade 9	2782	627	2699	439	3252	123
	3329	625	2654	433	3108	103
Total	30309	2589	29679	1947	28330	1347
	31516	2646	29690	1958	28406	1309

2.4 Limitations of the data

In preparing for data analysis, a number of problems in the data were identified that will require caution when interpreting findings.

First of all, as a consequence of small student samples from non-project counties, only a few schools were included in the study. Although the designed sample size in total was more than sufficient to detect moderate changes between project and non-project students in exam tests, they might not be a representative sample of non-project schools. The exam results of students in absolute value would not be representative of the population. However, as the SAS was only interested in relative changes in students' achievement over time and differences in such relative changes between project and non-project students, the problem of non-representativeness should not be a major concern.

Secondly, data collection did not strictly follow the study design of school/grade panel which caused a loss of the panel feature to some degree. This was particularly the case for the EOP data collected from the non-project counties. Some provinces drew different non-project schools at different phases of testing. This was less a problem with project counties. School panel information in Table 5 shows a somewhat imbalanced panel sample of the project counties with about two thirds of schools being followed up at least twice or three times and one third with only one data collection. Multi-level model analysis can provide robust and efficient estimates of data from imbalanced design (Goldstein 2010). However, the violation of the study design implies that the mean changes of students' achievement scores over time in non-project students will come from more sources of variation than those of project students, which include differences between testing times, between schools and between students, hence large sampling errors in the change scores. Moderate effects due to project intervention may not be detectable by statistical testing due to large sample errors embedded in the non-project comparator. The authors shall pay attention to the trend of changes between project and non-project students, and discuss it explicitly with caution.

Table 5: Frequency of repeated tests by schools

Number of tests	Project counties	Non-project counties
Grade 3		
1	106 (25.0)	31(96.9)
2	98 (23.1)	1(3.1)
3	220 (51.9)	0
Total	**424 (100.0)**	**32 (100.0)**
Grade 5		
1	104 (26.7)	30 (96.8)
2	127 (32.6)	1(3.2)
3	158 (40.6)	0
Total	**389 (100.0)**	**31 (100.0)**
Grade 7		
1	42 (39.3)	11 (100.0)
2	29 (27.1)	0
3	36 (33.6)	0
Total	**107 (100.0)**	**11 (100.0)**
Grade 9		
1	34 (32.4)	10 (100.0)
2	30 (28.6)	0
3	41 (39.0)	0
Total	**105 (100.0)**	**10(100.0)**

Thirdly, there were possible selection biases in non-project schools in favour of high performers. Table 6 shows overall mean test scores of students at Grades 3 and 5 in Chinese testing for project and non-project counties by three phases of tests. While an increased mean score over time is observed in both project and non-project samples, such change in the latter at the EoP phase was dramatically large. The same pattern was found in Maths mean scores for the two grades too. Further enquiry in the EoP data revealed that the non-project county selected in Yunnan province came from the Xishan District of Kunming City, the capital of the province where schools had much better education quality with better resources than the SBEP project counties.

In the analysis of relative changes over time, such selection bias will clearly favour students from non-project counties. If analysis showed that students from non-project counties were found making greater progress over time than those from project counties based on the data, the result would not be reliable due to such selection bias. However, if the analysis showed a greater progress from students of project counties than those from non-project counties over time, some degree of positive project impact which would be larger than that observed in the data if the selection bias was not in favor of the latter group could be reported

Table 6: Test scores of students in project and non-project counties in three phases (for Chinese test)

Period	Grade	Project counties		Non-project counties	
		Sample size	Mean score	Sample size	Mean score
Baseline	Grade3	12388	45.74	612	46.64
	Grade5	12009	46.19	489	48.11
MTR	Grade3	12341	47.95	531	47.23
	Grade5	11665	50.32	562	51.68
EoP	Grade3	10572	54.89	435	70.26
	Grade5	10602	54.65	656	61.72

3. Methods

3.1 Key variables

Exam scores

At the Baseline, MTR and EoP stage, students were tested on two subjects: Maths and Chinese. The test contents were developed by the project M&E expert team, and tests on the same subject in the three phases had received equalising disposal to make the test scores comparable. For both Maths and Chinese tests, students were given a total score and scores in sub-domains. Three sub-domains of Chinese included recognition of Chinese characters, reading and ancient Chinese, and only the 1st two for Grades 3 and 5 students. The three sub-domains of Maths included arithmetic, space and practice.

First of all are the anchor items. The parameter estimation in the MTR and the EOP evaluations added the task of equating with the tests at the Baseline stage[1]. The method of equating with common test items was employed in the design thus the quality of the anchor items (compared with other test items) had a bigger impact on the parameter estimation errors. In the mid-term evaluation, before the equating, there was an independent estimation of the anchor items, which obtained the difficulty and discrimination indexes (of the anchor items). These indexes were compared with the same set of parameters in the Baseline to identify whether there were linear relationships. Those anchor items which inferred a linear relationship were put back to the Baseline or mid-term test as normal test items for estimation. The items with a discrimination degree of less than 0.2 were deleted.

In the equating process in the mid-term evaluation, it was found that the process of designing anchor items was not well regulated. Some anchor items had been changed (mainly in the subject of Chinese in Grades 5, 7 and 9). Those changed items were treated as non-anchor items for parameter estimations.

An Item Response Theory model was used to perform the equating, which produced a new score θ for each student to measure learning ability of the student. The new variable followed Standard Normal distribution as observed for each grade from each test time and by subject areas. To get rid of the negative value of the standardised variable and for an easy interpretation of results, the authors further performed a linear transformation of the ability score θ by using the following formula:

$$T=10*\theta+50$$

The same procedure was used for test scores of the three phases: Baseline, MTR and EoP. The T score is the outcome variable in this study and is used for all analyses unless another form of the score is mentioned in a particular analysis. The ability score or test score or exam score is used in the report exchangeable.

[1] The equating of the test-scores in two tests is needed in order to assess the effect of SBEP intervention by comparing the differences between students' test scores in the Baseline investigation and those in the midterm review. Equating compares the difficulty indexes of the two tests by using the same measurement via certain transformation method, for example, if the unit of *jin* needs to be converted to the unit of kilogram, the conversion relationship is like this: 2 *jin* equal 1 kilo. After the conversion, the unit of kilogram is used as the unified unit of measurement which can be used to compare the weight of different objects.

Project indicator

This is a dichotomous variable, coded 1 for students from project districts and 0 from non-project districts. It is used to differentiate change patterns of students' test scores between the two groups. This variable is used to differentiate the overall impact of the SBEP project interventions on students' learning ability from effects of other projects that may apply to the SBEP project schools as well as to other non SBEP project schools.

Measure for intensity of project interventions

Within project counties, the amount of project interventions varied around the key component: School Development Plan (SDP). The SDP initiative aimed at improving school management in general, and was implemented by increasing the capacity of schools to prioritise the development needs through reinforcement of school planning and participation of stakeholders, such as community and students, in the process of school planning and implementation of the plans. Roughly one third of the schools (mainly the schools in the poorest townships) in the project counties implemented the SDP, and became 'SDP schools'. The rest of schools as a 'non SDP' group received project interventions less comprehensive or less intensive in terms of head teacher and teacher training in general.

The main consideration behind the principle of choosing the schools from the poorest townships in the counties to do SDP was that this arrangement would make it possible to give more support to the most needy schools through giving extra funds to these schools for implementation of SDP. The SDP could also be used as an effective way of making parents who did not send their children to school understand the value of schooling, by involving them fully in the process of developing and implementing SDP for schools their children were supposed to be attending.

Table 7 below shows some key differences between SDP and non-SDP schools in receiving project interventions. It was assumed that if project interventions were effective on students' learning, one would find more improvement over time in exam scores among students from SDP schools than from those from non-SDP schools.

Table 7: Differences of SDP and non-SDP schools in project interventions

Project intervention	SDP schools	Non-SDP schools
% of students received SBEP aid (Grade 7)	18.6	14.4
% of students received SBEP aid (Grade 9)	25.3	12.7
% of teachers received SBEP training (primary school)	79.7	70.5
% of teachers received SBEP training (lower secondary school)	56.1	53.9

Measures for disadvantaged groups

One of the main outputs of the project (Output 1) was to improve equality in education for disadvantaged groups in poor counties. It was important to assess whether project interventions had impacted on students with disadvantages. To track the participation and performance of disadvantaged children, both in absolute and relation terms to other groups, the project developed the "Social and Economic Status (SES) Index" for each student based on family assets, housing condition, agriculture incomes, family size and household wares and wealth. The index is ranged from 0-12, with a lower score being indicative of higher SES status or less poverty.

Other indicator variables for disadvantaged groups are minority, girl, disabled, boarder, orphan or being from a single-parent family. In addition, being a village school (not a complete school with 6 grades and rural location) is also a measure at school level for disadvantage.

Other co-variates

In an observational study, there are factors which are bound to be different between comparative groups which need to be taken into account in data analysis in order to disentangle true effects of 'treatment', or interventions in this case. Below are some of them:

- parent expectation and student's self-expectation of the education level (in EOP data, there was only self-expectation)
- number of siblings
- student's age
- home-to-school distance
- whether have received SBEP aid or other kinds of aid
- whether mandarin is spoken at home
- whether parents work out in the cities
- school type : complete or not
- head teacher's educational background (whether is a BA degree holder or above)
- head teacher has a qualification certificate or not
- head teacher's gender
- the number of years being a head teacher, school SES (the mean value of students' SES in the school)
- percentage of boarders
- percentage of ethnic minority students
- percentage of teachers with required educational background
- percentage of teachers trained at the county level and above in the last two years.

A small number of variables had different definitions in the Baseline and MTR investigations. For example, in the case of head teacher's educational background, three dummy variables labeling 4 levels of educational attainments were used in the Baseline and MTR investigation, but only one dummy variable was used (by designating those who are not a BA degree holder or above as one single category) in the EOP review.

3.2 Analytic strategy

The three research questions mentioned above were formulated into following analytic approaches.

For Q1, a simple descriptive analysis was used to present means of exam scores by grade, by gender, by project phase and by project indicator. To further estimate and test differences in mean changes of exam scores from the Baseline to the later phases of the project between students from project and non-project counties, weighted 2-level regression analysis for means was used. The main purpose of the simple descriptive analyses was to make an overall assessment of any possible impact of the project intervention of SBEP. This analyses was carried out from two angles: first, by comparing the test scores of students in the same grade in EoP, MTR and the Baseline investigations within the project counties, to find out whether there were changes in students' test scores over time (from Baseline and on, by the time EOP review was conducted); secondly, to investigate if there was more (or less) progress in the students' test scores in project counties relative to those of the non-project counties over time from Baseline and MTR to EoP. Following the direction of descriptive analysis, a two-level regression analysis for aggregated mean scores was used to estimate and test change patterns between project and non-project students with much efficiency (Goldstein, Yang at al 2000). MLwiN (Rasbash at al 2010) was used for the modeling analysis. A simple form of the two-level model is presented in Model 1. For simplicity, subscripts of variables were ignored and all interaction terms represent a set of varied estimates according to the number of categories of each variable.

$$\begin{aligned}\bar{Y} &= \beta_0 + \beta_1 proj + \beta_2 time + \beta_3 grade + \beta_4 (proj)(time) + \\ &+ \beta_5 (grade)(time) + \beta_6 (proj)(grade) + \beta_7 (proj)(grade)(time) + \\ &+ U + E\end{aligned} \qquad (1)$$

In this model, the dependent variable is the mean score by grade by time and by project indicator. Treating every mean as a cluster with students nested within each cluster, a 2-level modeling method, in which the random effects between clusters are captured by the term U in the model, was used. The term E refers to sampling errors at student level, is embedded in the standard error of the mean score, i.e. standard deviance (SD) divided by square rooted sample size and is used as a weighting factor in this model. In this model, variable time has three categories for the Baseline, MTR and EoP tests, and the grade has four categories.

The overall mean difference at the Baseline between project and non-project students is estimated by coefficient β_1, such difference in the mean change over time is by β_4. Overall difference between project and non-project students by grade level is estimated by β_6, and difference in changes over time by grade is estimated by β_7 which includes interaction terms between grade, time and the project indicator, reflecting mean difference between project and non-project students in their change scores from one time point to the next by grade. The models are weighted by number of students listed in Table 2. More details of the modeling technique can be found in Goldstein, Yang et al (2000).

For Q2 the team used 3-level models to examine possible impact of project intervention by the intensity measure, SDP indicator, on the performance of disadvantaged students in project schools. The three-level model was used to reflect the hierarchical structure in the data: students at level 1, schools at level 2 and counties at level 3. It is assumed that schools with SDP had more exposure to the project intervention than schools that were not involved in SDP during the project life. The basic model that tests overall difference

in the change of students' performance over time between SDP and non SDP schools can be expressed in Model 2 below.

$$y = \beta_0 + \beta_1 SDP + \beta_2 time + \beta_3 (SDP)(time) + \beta_4 (X) + \beta_5 (X)(time) + \\ + \beta_6 (SDP)(time)(X) + \alpha(Z) + V + U + E \qquad (2)$$

In this model, variables indicated by X are measures for disadvantaged groups, such as girl, minority, disabled and SES. The variables indicated by Z are possible confounding or differences among students to be adjusted for such as age, non mandarin speaking, long distance to school, multiple siblings, and so on. Overall difference between SDP and non-SDP groups is estimated by β_1, time effect by β_2, interaction between SDP and time by β_3, that between disadvantaged groups and time by β_5. Since the SBEP interventions targeted most needy groups with possibly poorest achievement, negative estimates on β_1, and β_4 could be anticipated for poor performance in the SDP and disadvantaged groups at the Baseline. However, the parameter β_6 can pick up relative changes among disadvantaged students with SDP interventions over time in comparison with others. Therefore significant tests on parameters β_6 of positive sign suggest that more progress was made by disadvantaged students from SDP schools than those from non-SDP schools, hence an evidence of project effects via SDP intervention. Such effects have taken into account random effects among schools by the term U and random effects among counties by the term V. The term E is sampling error among students.

The analysis was carried out for each grade in Chinese and Maths separately, for students in project counties only. The authors used StataSE 11 for this analysis.

To answer the research question 3, the team needed to assess project impact on disadvantaged students in relation to the sub-domain scores of Chinese and Maths respectively. Two core sub-domains in Chinese tests were Recognition of characters and Reading for students of all grades and a third domain Ancient Chinese for grades 5, 7 and 9. Three core sub-domains in Maths tests were Algebra, Space and Practice. Test scores or ability scores of students on these sub-domains are strongly correlated. It would be interesting to explore project interventions on specific domain or skills. For example, teacher training in classroom teaching methods and new teaching materials may be more effective on students Reading Chinese than on Recognition of characters. The same analytic strategy for Q2 was used, but fitting multivariate models of Model 2 for two advantages of the jointed analysis of sub-domain scores: (1) direct comparison in effects of project interventions among sub domain scores; (2) improved efficiency of estimates by taking into account correlations between those scores.

A simple form of the models can be expressed below

$$y_1 = \beta_0 + \beta_1 SDP + \beta_2 time + \beta_3 (SDP)(time) + \beta_4 (X) + \beta_5 (X)(time) + \\ + \beta_6 (SDP)(time)(X) + \beta_7 (Z) + V_1 + U_1 + E_1$$

$$y_2 = \alpha_0 + \alpha_1 SDP + \alpha_2 time + \alpha_3 (SDP)(time) + \alpha_4 (X) + \alpha_5 (X)(time) + \\ + \alpha_6 (SDP)(time)(X) + \alpha_7 (Z) + V_2 + U_2 + E_2 \qquad (3)$$

$$y_3 = \gamma_0 + \gamma_1 SDP + \gamma_2 time + \gamma_3 (SDP)(time) + \gamma_4 (X) + \gamma_5 (X)(time) + \\ + \gamma_6 (SDP)(time)(X) + \gamma_7 (Z) + V_3 + U_3 + E_3$$

The set of three models is for three sub-domain scores. They are marginal models, and will be fitted simultaneously. The dependent variables Ys are assumed from Multivariate Normal Distribution with variance-covariance structure at each of the levels in the data. Technical details of the model can be found in Goldstein (2010).

The team was interested in parameter estimates of β_6, α_6 and γ_6, and differences between them in order to establish evidence of the project impact on disadvantaged students in specific knowledge or skills in the two main subjects, Chinese and Maths.

4. Results

4.1 Students' overall performance in description

Tables 8 and 9 present the mean scores of students in project and non-project counties in three phases of the project for Chinese and Maths respectively. Several important observations can be made based on average test scores in the tables.

First, students' average scores were increasing over time, which was more steady and apparent for students in project counties than for those in non-project counties.

Secondly, students in project counties scored markedly lower on average than those in non-project counties at the Baseline, and the gap decreased at the MTR, but reversed at the EoP for students of Grades 7 and 9. This seems to suggest that the achievement level of the students in project counties had improved more than for those in non-project counties.

Thirdly, the pattern in gender difference was as expected, with girls doing better than boys in Chinese and the other way round in Maths. However such gender difference was somewhat smaller among students in project counties that those in non-project counties.

Finally, the standard deviation of student average scores in primary schools or lower grades was larger than that in secondary schools. This may suggest that differences in teaching quality in the primary schools are still significant and issues of equity and imbalances remain, with no observable difference between project and non-project counties.

Table 8: Mean (SD) of ability scores by grade by project phase by gender and by project indicator (Chinese)

	Project			Non-project		
	Baseline	Mid term	EoP	Baseline	Mid term	EoP
Girl						
Grade 3	46.4(11.3)	48.6(12.5)	55.7(14.1)	52.1(10.1)	49.5(10.4)	73.0(13.3)
Grade 5	47.5(10.8)	50.8(10.6)	54.8(12.0)	57.7(11.3)	54.4(9.2)	61.3(15.9)
Grade 7	45.4(8.3)	47.9(8.1)	50.2(9.9)	49.0(7.8)	49.8(8.4)	52.7(9.4)
Grade 9*	45.0(9.8)	47.5 (9.1)	54.3(10.9)	50.9(6.2)	50.1 (9.2)	55.3(13.9)
Boy						
Grade 3	46.5(11.4)	47.5(12.5)	54.6(14.0)	50.7(10.4)	46.3(11.7)	67.8(14.1)
Grade 5	47.1(10.3)	50.2(10.3)	53.9(12.2)	52.7(12.4)	49.6(10.5)	62.3(15.1)
Grade 7	45.6(8.8)	46.5(8.7)	49.4(10.0)	48.6(7.1)	46.3(9.6)	48.2(9.4)
Grade 9*	45.4(9.5)	49.6(8.7)	53.0(11.9)	51.3(5.4)	49.9(9.8)	54.3(11.7)
All						
Grade 3	46.5(11.4)	47.9 (12.6)	54.9 (14.2)	51.4(10.3)	47.2 (11.9)	70.3 (14.2)
Grade 5	47.3(10.6)	50.3 (10.7)	54.7 (12.0)	55.3(11.9)	51.7 (10.3)	61.7 (15.5)
Grade 7	45.3(8.6)	47.2 (8.4)	50.2 (10.0)	48.8(7.5)	47.9 (9.2)	50.4 (10.2)
Grade 9	45.2(9.6)	48.6 (8.9)	53.8 (11.8)	51.1(5.8)	49.9 (9.5)	55.0 (12.5)

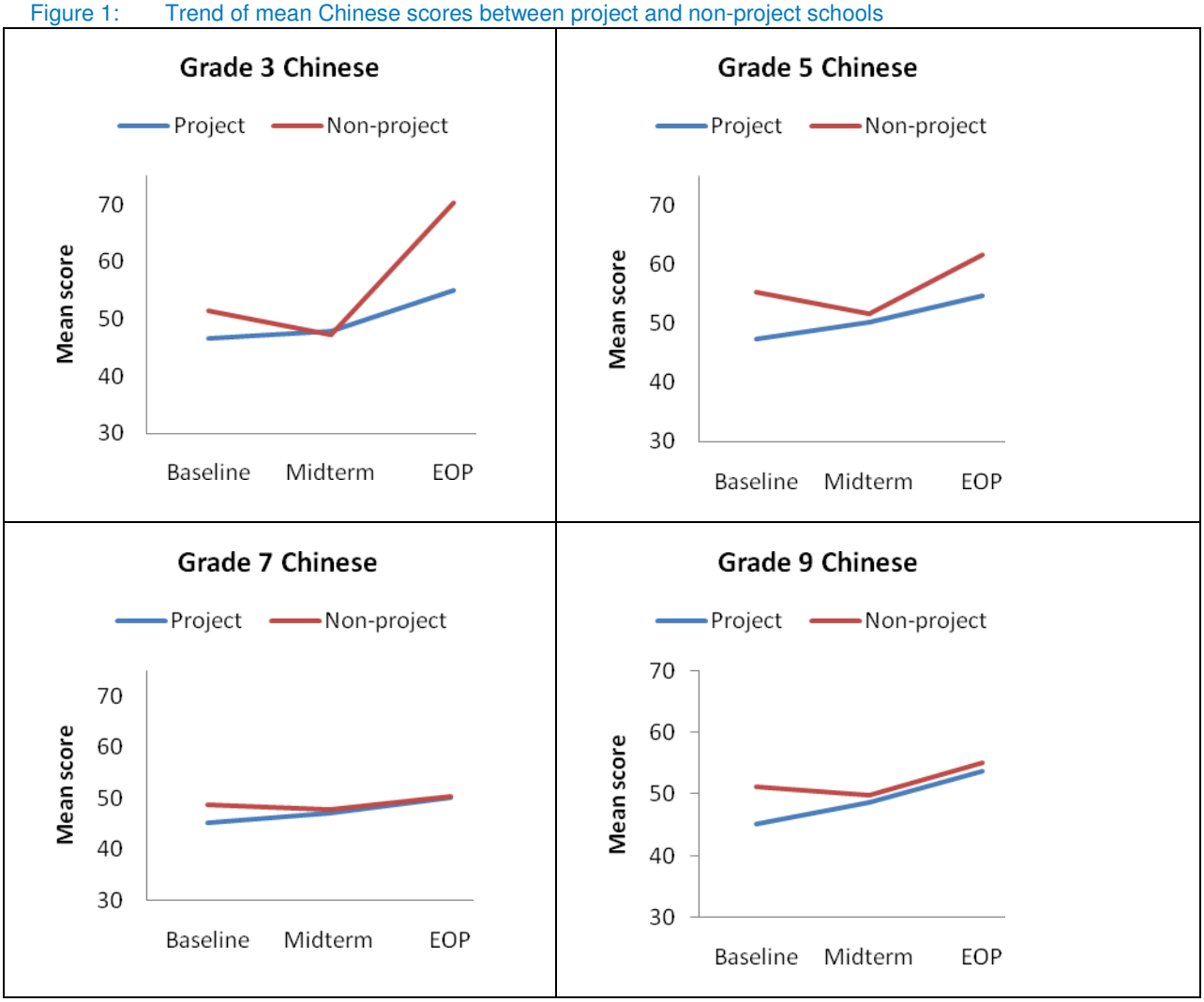
Table 9: Mean (SD) of ability scores by grade by project phase by gender and by project indicator (Math)

	Project			Non-project		
	Baseline	**Mid term**	**EoP**	**Baseline**	**Mid term**	**EoP**
Girl						
Grade 3	41.7(13.0)	48.2(14.2)	53.0(13.5)	44.9(13.7)	48.5(11.9)	62.5(11.3)
Grade 5	46.2(9.1)	47.7(10.9)	55.9(14.6)	50.1(9.4)	48.9(8.7)	58.9(16.4)
Grade 7	49.8(10.8)	50.8(8.9)	53.3(11.1)	55.7(9.9)	52.7(9.1)	52.5(8.8)
Grade 9	48.7(10.0)	51.1(9.6)	50.4(9.8)	53.5(7.9)	51.8(11.4)	49.8(8.6)
Boy						
Grade 3	42.3(12.4)	48.7(14.3)	53.1(13.6)	43.4(13.6)	47.3(11.4)	60.9(12.1)
Grade 5	46.2(9.7)	48.6(10.9)	55.1(14.6)	51.9(8.3)	48.6(10.3)	61.7(15.3)
Grade 7	50.6(10.6)	52.0(9.7)	53.7(11.1)	55.2(9.4)	52.9(9.3)	51.1(9.2)
Grade 9	48.6(10.5)	51.0(10.7)	50.6(10.1)	53.5(7.9)	55.0(10.2)	46.5(8.0)
All						
Grade 3	42.0(12.7)	48.4(14.4)	52.8(13.6)	44.2(13.5)	46.9(12.2)	61.5(11.8)
Grade 5	46.2(9.2)	48.0(11.1)	55.4(14.6)	50.8(8.4)	48.6(9.6)	60.6(15.8)
Grade 7	50.2(10.7)	51.5(9.3)	53.4(11.0)	54.7(9.7)	52.8(9.1)	51.8(8.9)
Grade 9	48.6(10.3)	51.0(10.3)	50.4(9.9)	52.0(8.3)	53.2(10.9)	48.6(8.2)

4.2 Overall project effects in Chinese by regression model analysis

The change patterns of students' mean ability score of girl and boy pooled as shown in Table 8 are described in Figure 1, suggesting a steady progress over time by Grades 3 and 5 students and greater progress made by Grades 7 and 9 students in project counties than those in non-project counties. The large leap at the EoP for Grades 3 and 5 students in non-project counties could be due to selection bias as discussed in the previous session of the report.

Figure 1: Trend of mean Chinese scores between project and non-project schools



To further quantify and test the trends observed in Figure 1, estimates in the differences of mean changes at the MTR and EoP between project and non-project counties based on multilevel meta-regression models are presented in Table 10. All estimates are statistically significant with $p<0.01$ by Wald test. A positive value indicates more progress made by students of project schools than by those of non-project schools. In the comparison, students of all grades from project counties demonstrated better improvement in their learning ability score at the MTR, and such improvement is continued to the EoP for students in Grades 7 and 9, but less so for Grade 5 students.

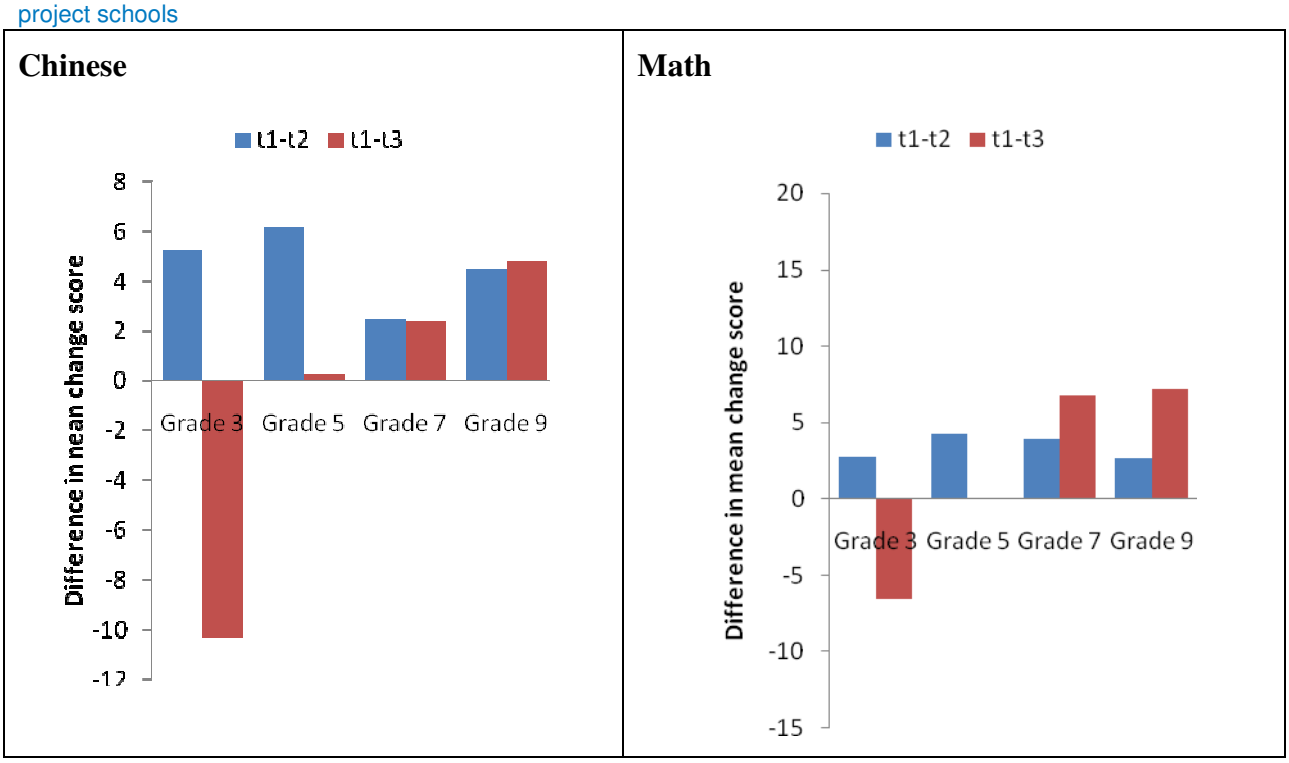
Table 10: Model estimated mean changes and differences in mean changes between project and non-project schools

Grade	Period	Chinese			Math		
		Proj	Non-proj	Diff (SE)	Proj	Non-proj	Diff (SE)
3	t1-t2	2.6	-2.6	5.2(0.07)‡	6.4	3.6	2.8(0.07)‡
	t1-t3	5.1	15.4	-10.3(0.08)‡	11.0	17.5	-6.5(0.07)‡
5	t1-t2	4.3	-1.9	6.2(0.08)‡	-1.6	-5.9	4.3(0.08)‡
	t1-t3	3.4	3.1	0.29(0.08)‡	-8.2	-8.3	0.08(0.07)
7	t1-t2	2.3	0.2	2.5(0.08)‡	-2.4	-6.3	3.9(0.08)‡
	t1-t3	0.6	-1.8	2.4(0.11)‡	-14.2	-21.0	6.8(0.11)‡
9	t1-t2	4.5	0.0	4.5(0.07)‡	-1.2	-3.9	2.7(0.08)‡
	t1-t3	4.8	0.0	4.8(0.11)‡	-15.7	-22.9	7.2(0.12)‡

‡ P<0.001; t1 for Baseline, t2 for MTR and t3 for EoP.

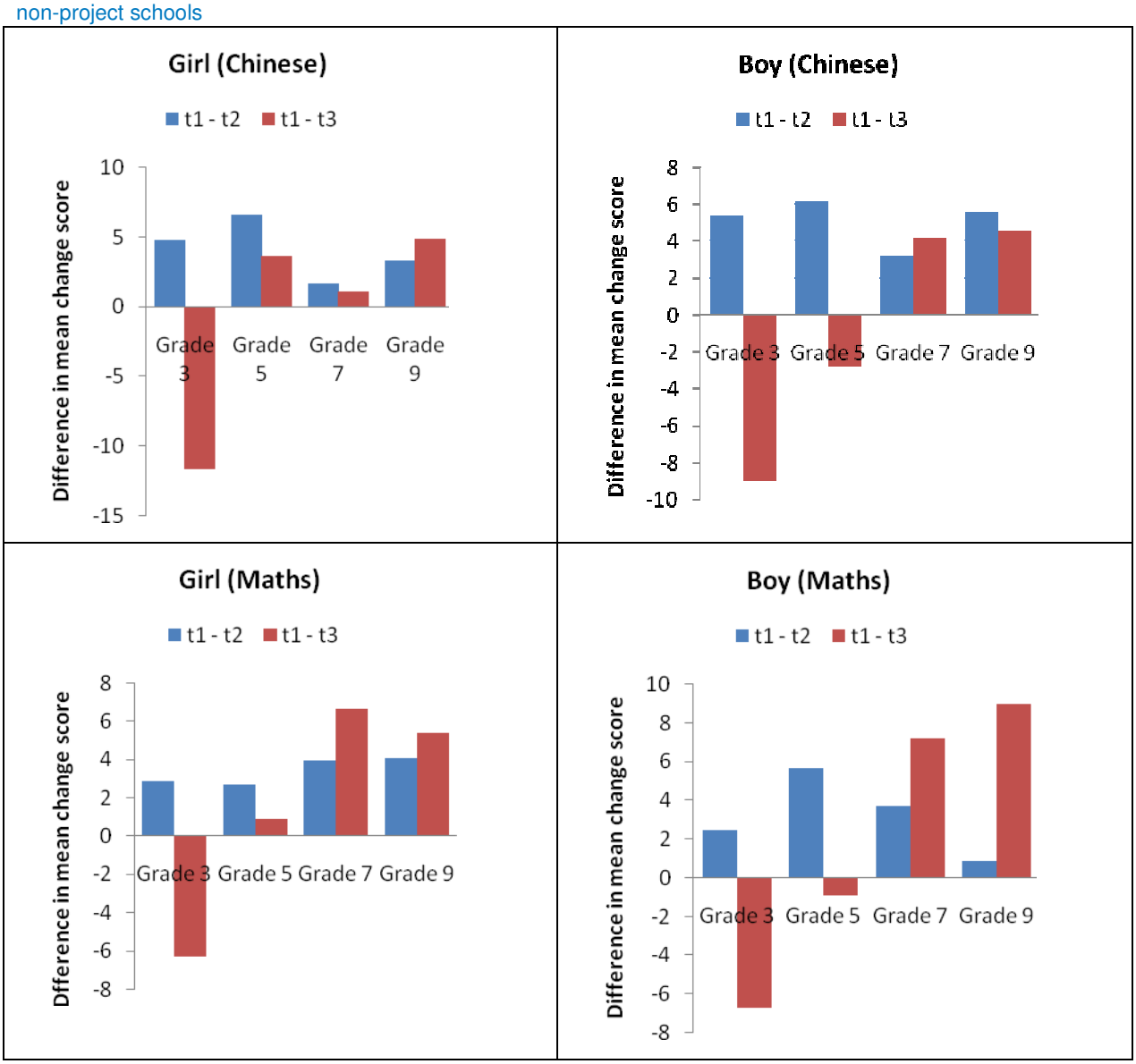
The model estimated project effects can be observed clearly in Figure 2 which shows positive effects of the project on students in Grades 5, 7 and 9 at the mid-term but much great positive impact at the end of the project. However, an opposite finding is observed for Grade 3 students who demonstrated a significant project effect at the mid-term but a large negative effect at the end of the project. This is most likely due to selection bias in the sample of non-project schools at the end of project survey. More detailed description of the sample issue can be found in the early part of this report.

Figure 2: Project effects: changes of mean ability scores over time in project schools below or above those of non-project schools



The gender-separated change patterns are shown in Figure 3. For the period of Baseline to mid-term, both boys and girls in all grades of project schools demonstrated greater improvement in their test scores than did non-project students. However, in the period from Baseline to EoP, girls in Grade 5 continued to show greater progress, but for boys the trend was in opposite direction. The reason is yet to be investigated.

Figure.3: Gender differentiated changes of mean ability scores over time in project schools below or above those of non-project schools

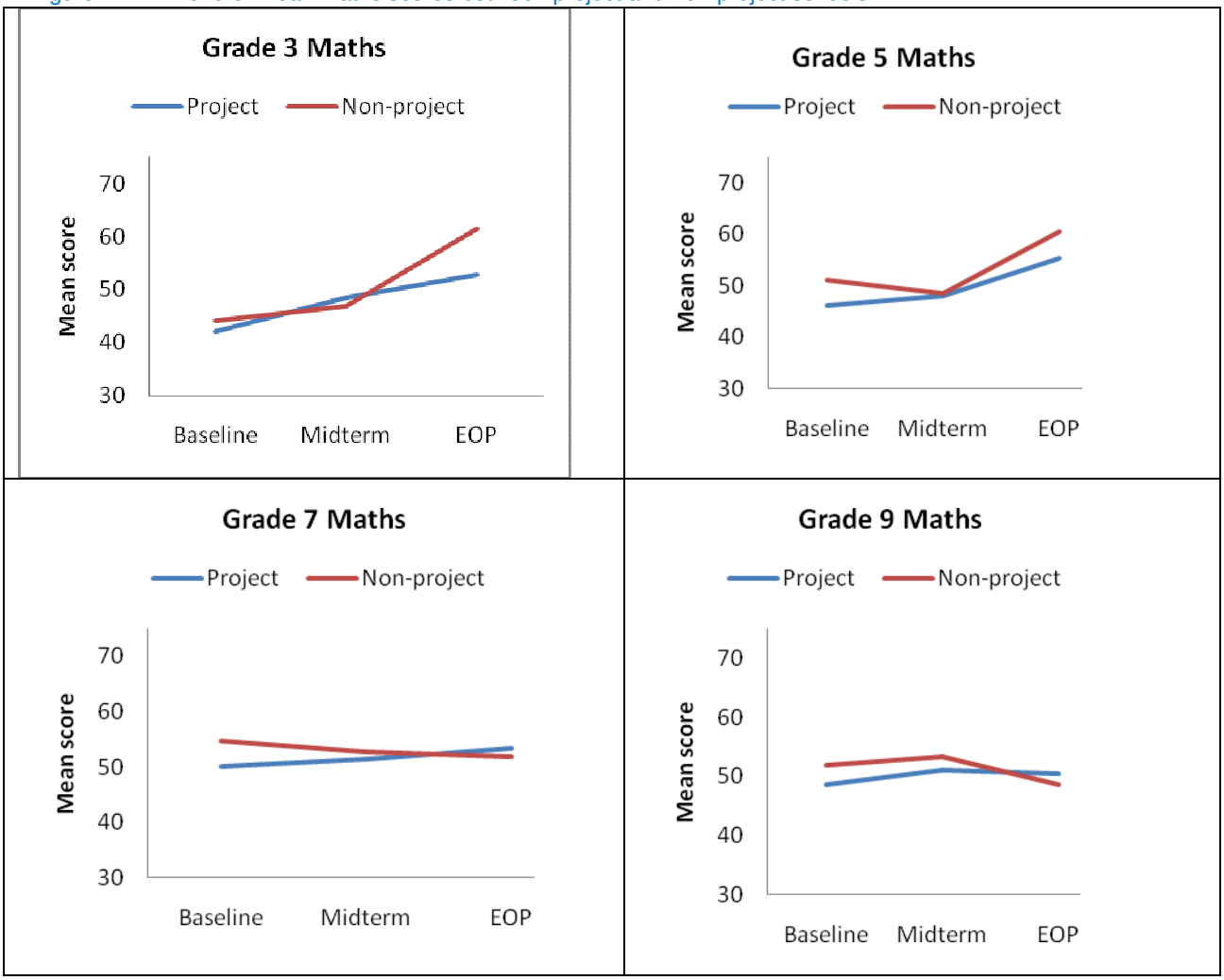


4.3 Overall project effects in Maths by regression model analysis

The change patterns in Maths in Figure 4 suggest increased performance over time among Grades 3 and 5 students in project schools and a decreased pattern among students in Grades 7 and 9. The cross-over in the performance curves between project and non-project students of Grades 7 and 9 clearly suggests greater progress made by the former group than the latter. Model-estimated mean differences in changes

of ability scores between project and non-project groups are presented in the last column in Table 10, with statistical significance at p<0.001 for all except for Grade 5, period Baseline to EoP, by Wald test. Although, negative changes in mean ability scores over time were observed for most grades of the two groups, the project group still showed greater progress in relative terms than the non-project group. More specifically, Grade 9 students of the project group demonstrated faster improvement at both mid-term and end of the project stages. Students of Grades 5 and 7 in project schools demonstrated faster improvement than did their counterparts, but only at the end of project.

Figure 4: Trend of mean Maths scores between project and non-project schools



Gender differentiated change patterns in Maths are presented in Figure 3. The only gender difference was among Grade 5 students for the period from Baseline to the EoP, similar to that observed in Chinese test score.

4.4 Project effects on performance of disadvantaged students

Table 11 presents estimated performance of students by their background from 3-level models analysis with county at level 3, school at level 2 and students at level 1 (Model 2) by grade and by subject separately. Out of 27 project counties in the four provinces, 24 counties were included in the analysis.

A positive value suggests a higher ability score of students at the Baseline, hence better performance than that of the reference group. We can see in the table that all associations are as what one would expected for in general, for example, girls tended do better in Chinese than boys and less well in Maths, older students did worse than younger ones in the same grade; worse performance was associated with minority students, those with longer distance from home to school, from poor SES background, more siblings, not mandarin speaker and so on. However, a rather strong positive effect has been found in high expectation of education by students of all grades.

Table 11: Student background in association with test scores (adjustment for each other in Model 2)

	Chinese				Math			
Grade	3	5	7	9	3	5	7	9
No. schools	365	339	87	85	365	339	87	85
No. students	28802	23930	7280	6739	29162	23701	7165	6676
Girl	.847‡	.937‡	.044	0.602*	-.276	-.288	-.831†	-.358
Minority	-.485	-.197	-.931*	-.111	-.641	-.468	-1.38‡	.514
Disabled	-.702	-1.27*	-.547	-0.804	-.239	-.792	-.416	-.326
Boarder	-.606†	-.603‡	-.134	.447	-.454	-.637‡	.783†	.192
Student SES	-.153‡	-.124‡	-.101*	-.277‡	-.212‡	-.107‡	-.160†	-.136*
Long distance to school	-.506‡	-.420‡	-.223†*	-.271‡	-.485‡	-.249‡	-.029	-.258†
Many sibling	-.208‡	-.073	-.163*	-.059	-.135†	-.059	-.169*	.132
Not mandarin speaker	-.911*	-1.228‡	-.027	-.222	-.677	-1.20‡	-1.07†	-.960*
Expect	.166‡	.436‡	.634‡	.736‡	.166‡	.352‡	.530‡	.749‡
Age	-.155*	-.433‡	-.527‡	-.269*	.088	-.319‡	-.553‡	-.089

* P≤0.05; † P≤0.01; ‡P≤0.001.

Results in Table 12 are estimates of independent overall effects on students test scores of two school level variables, SDP schools and percentage of school teachers trained by the SBEP project. The observation is that: (i) the effect of project interventions by teaching training measure was not consistent, being positive impact on Grade 5 students and negative on students in other grades and (ii) students of Grades 3, 5 in SDP schools did not improve performance over time, but those of Grades 7 and 9 showed some degree of improvement over time.

Table 12: School factors in association with test scores of students

	Chinese				Math			
	g3	g5	g7	g9	g3	g5	g7	g9
SDP school	-1.67	0.732	-2.83	-2.73	-3.62‡	-.419	-1.13	-.806
%teacher training	-.001	-.012‡	.038‡	-.008*	-.015‡	-.006*	.026‡	-.034‡
(SDP)×(MTR)	-1.00*	-2.12‡	2.81†	-.221	.105	-.991*	-.180	-3.06†
(SDP)×(EoP)	-2.80‡	-3.15‡	2.91†	9.33‡	-2.33‡	-2.41‡	-.200	3.55‡

* P≤0.05; † P≤0.01; ‡P≤0.001.

Further analyses for differentiated project effect via SDP measure by four major disadvantaged groups, girl, minority, disabled and student SES scale, produced the results shown in Table 13. The estimates were adjusted for other variables including boarder, long distance to school, many siblings, not speaking mandarin at home, age and rural school type. They are terms interacted with both project period and SDP indicator. A positive value in Table 13 indicates faster changes or more improved mean scores of students

from SDP schools compared to that of non-SDP schools, hence possible specific effects of intervention on those groups of students brought in by SDP activities.

Table 13: Model estimated average change of standardised test score over time between SDP and non-SDP schools

Disadvantaged group	Baseline to mid term				Baseline to EOP			
	g3	g5	g7	g9	g3	g5	g7	g9
Girl (Chinese)	.403	.775*	1.17*	1.25*	.543	2.08‡	.965	-.094
(Math)	-.385	-.603	.008	-.938	.020	1.38†	.574	-1.12
Minority (Chinese)	-1.04†	.309	-3.74‡	-3.86‡	-1.45†	-1.74‡	-4.71‡	-10.5‡
(Math)	-1.84‡	.247	-2.77‡	.572	-1.10*	-1.26*	-6.41‡	5.76‡
Disabled (Chinese)	-3.73†	-1.451	-4.46*	-6.08†	-.030	-1.33	1.409	1.31
(Math)	-4.40†	-3.31†	-2.54	-.867	-2.04	-3.07	-.398	2.87
SES (Chinese)	.232‡	.250‡	.057	.350†	.126	.400‡	.144	.105
(Math)	.215†	.175†	.094	.130	.127	.592‡	.122	.749‡

* P≤0.05; † P≤0.01; ‡P≤0.001

Several observations can be made based on statistics in Table 13:

i. Girls of all grades in SDP schools demonstrated an improved performance in Chinese over those in non-SDP schools at both mid-term and end of the project stage, which was particularly true for Grades 5, 7 and 9 showing a significant improvement at the mid-term review, and further significant improvement at the end of the project time was shown among girls of Grade 5. The improvement of girls in Grades 3 and 7 at the EoP was marginally significant but with consistence. Improvement of girls in SDP schools in Maths was not observed at the mid-term review but evidenced at the EoP, among Grade 5 students especially.

ii. Minority students in SDP schools demonstrated worse progress overall for both subjects except for students in Grade 9 in Maths, who made significant progress.

iii. Disabled students of all grades in SDP schools made slower progress in Chinese than did those in non-SDP schools at the mid-term review, and made some catch up to the end of the project without reaching statistical significance. Some weak evidence of SDP impacts on Maths tests of this group of students was observed among Grade 9 students but without statistical significance.

iv. Students with low socio-economic background (SES) of all grades in the SDP schools showed a consistent pattern of greater progress in both subjects over the project life than did those in the non-SDP schools.

4.5 Project impacts on sub-domain test scores of students in project schools

Since the analysis of sub-domain scores between the Baseline and MTR data had been done already and included in MTR report, this analysis was set to assess only relative changes in sub-domain scores of students from the MTR to the EoP, and compare such changes between students of SDP and non-SDP schools. The period assessed for changes is different from the analysis of total test scores where two changes are assessed in two periods of different length: Baseline to MTR for a 2 year period and Baseline to EoP for a 4 year period. The 2-year period from MTR to EoP is embedded in the 4-year period of the

analysis of the total test score. For this reason, we would not expect findings of this analysis to be fully consistent with those from analyses of the total score. Instead some different findings might emerge.

Results shown in Table 14 are from fitting the Model 3. Again the authors looked for estimates with positive sign of improved performance of particular groups of students due to the exposure to the SDP interventions, and differences in such estimates among different knowledge domains. Several observations are made from the results.

i. More positive estimates are found among girls of all grades for all sub-domains in both Chinese and Maths, with some reaching significant levels. No marked differences were found between sub-domains. They all suggest that SDP interventions seemed to be having positive effects on girls, who showed greater improvement during the last 2 years of the project period in all domains of school performance.

ii. More negative estimates are found among minority students of all grades for all sub-domains in both subjects. Half of the negative estimates reached significant level. The findings suggest that SDP interventions had not shown a positive impact on the learning ability of minority students.

iii. Positive effects are found among disabled students of Grades 3, 7 and 9, with more reaching significant level for Chinese sub-domains, but less so for Maths.

iv. Positive effects are dominant in the association of SDP interventions with students' SES status for all grades and all sub-domains. In general, poorer students in SDP schools made more progress overall during the last two years of the SBEP life than those in non-SDP schools.

v. No evidence of SDP interventions on specific subject domains of students' learning was found.

Table 14: Model estimates of project impacts on disadvantaged students by sub-domain test scores from multi-variate multi-level analysis (Model 3)

Grade No. Schools No. Students	(Disadvantage variable)× ×(EoP)×(SDP)	Chinese			Math		
		Y1	Y2	Y3	Y1	Y2	Y3
3 364/355 20163/19373	Girl	-.369	.352		.539	-.022	.374
	Minority	-.755	.728		.247	-.239	1.57†
	Disabled	3.88	2.25		.492	2.06	3.51*
	SES	.071	.070		-.066	.219	.110
5 336/336 15694/15418	Girl	1.65†	1.20*		1.27*	1.86†	1.02
	Minority	-2.16†	-4.12‡		-1.72*	-.837	-.178
	Disabled	-1.27	-1.89		-.144	-.769	-1.67
	SES	.712‡	.298		.585‡	.778‡	.330*
7 87/87 5759/5629	Girl	.087	-.046	1.00	1.95*	.477	.565
	Minority	-3.23‡	-1.93*	1.50	-6.68‡	-1.37	-5.71‡
	Disabled	5.85‡	5.06†	2.26	1.73	4.86	2.66
	SES	.075	.157	.176	.269	.107	-.152
9 86/86 4749/4673	Girl	.093	-1.72*	-1.37	-.672	2.67†	1.82
	Minority	1.08	-6.34‡	-7.22‡	-2.13*	-3.18†	-8.73†
	Disabled	4.10	6.45‡	5.18	-2.85	1.26	6.089*
	SES	.267	.027	-.109	.155	.202	-.552*

For Chinese: Y1 (Recognition of characters), Y2 (Reading), Y3 (Ancient Chinese)

For Math: Y1 (Algebra), Y2 (Space), Y3 (Practice)

* P≤0.05; † P≤0.01; ‡P≤0.001.

Furthermore, Table 15 presents the overall association of percentage of school teacher training and SDP with sub-domain test scores of students. These effects are adjusted for all major students' background factors as well as random effects in schools.

Table 15: School variables in association with performance of sub-domain test scores

Grade No. Schools No. Students	School variable	Chinese			Math		
		Y1	Y2	Y3	Y1	Y2	Y3
3 364/355 20163/19373	SDP	-1.27	-3.73†		-2.43	-1.89	-.958
	SDP×EoP	-2.45†	-2.89‡		-3.17‡	-4.03‡	-4.14‡
	%Teacher training	-.023‡	-.020‡		-.034‡	-.027‡	-.025‡
5 336/336 15694/15418	SDP	1.86	1.39		-.130	1.74	2.08
	SDP×EoP	-4.19‡	-1.12		-2.61†	-3.53‡	-1.14
	%Teacher training	-.031‡	-.025†		-.010‡	-.011‡	-.006
7 87/87 5759/5629	SDP	-4.26	-.579	1.73	-.198	-.929	-2.41
	SDP×EoP	-.489	1.44	-3.64†	1.05	-1.22	.533
	%Teacher training	.058‡	.062‡	.055‡	.071‡	.052‡	.043‡
9 86/86 4749/4673	SDP	2.35	-1.08	-4.59	-2.85	.374	-5.67
	SDP×EoP	-.979	6.46‡	9.27‡	1.91	-.757	6.48‡
	%Teacher training	.028‡	.028‡	.030‡	.005	.027‡	-.001

For Chinese: Y1 (Recognition of characters), Y2 (Reading), Y3 (Ancient Chinese)

For Math: Y1 (Algebra), Y2 (Space), Y3 (Practice)

* P≤0.05; † P≤0.01; ‡P≤0.001.

The main findings of this analysis are that:

i. There is no clear evidence or patterns suggesting different effects on different knowledge domains of student learning, or more negative associations between SDP interventions and test scores at the MTR except for Grade 5.

ii. Students of Grades 3 and 5 in SDP schools did worse at the EoP, while students of Grades 7 and 9 in SDP schools did better at the EoP, compared to their counterparts in non-SDP schools.

iii. The percentages with teacher training were negatively associated with test scores of all sub-domains for Grades 3 and 5 students, but there was positive association for Grades 7 and 9 students. The pattern was consistent and highly significant.

5. Summary and discussion

5.1 Overall performance

Several important findings can be derived from the analysis of overall performance based on mean test scores.

i. Sharper and steadier increase of learning ability of Grades 3 and 5 students than that of Grades 7 and 9 students in both subjects over time, with a slight down turn in Maths score of Grades 7 and 9 students.

ii. Students of non project counties had higher learning ability in absolute measure than students of project counties overall on both Chinese and Maths, especially at both Baseline and MTR tests.

iii. Students of project counties demonstrated significant improvement in both Chinese and Maths by the MTR testing over that of students from non-project counties in terms of relative changes of their learning score from the Baseline. Such improved performance remained significant at the EoP for students in Grades 5, 7 and 9, and the same for girls and boys.

iv. Following a significant improvement on both Chinese and Maths learning at the MTR, Grade 3 students in project counties demonstrated a significant worse performance than did their counterparts in non project counties at EoP.

In this study, the team used the test score derived from item response theory model, equated for the 3 waves of testing and standardised for all grades. The credibility of the procedure was around 80% by an assessment at the MTR. It measures learning ability of students and is comparable across times and grades. The increased ability score over time for all implies improved quality in the education system of the four provinces in the Southwest China in general over the SBEP project period. Students in primary schools seemed to benefit more from the improved system than did those in middle schools. This is supported in the SBEP "End-of-project review; Quantitative Survey Report" (the Quantitative Report) where a proxy measure of quality (availability of equipment and materials) was found to have improved over the life of the project – especially in primary schools. A different measure of quality (teacher talk time in class) was found to have improved (reduced) strongly in both primary and junior middle schools equally (para 113)

This could also imply that middle schools in those counties were better resourced than primary schools in reality when the SBEP started, and hence there was less margin for improvement. The decreased pattern of learning ability of all students in middle schools in Maths could be due to difference in the sampled schools over time or equating procedures. Since this study is only interested in relative changes of the test score over time between student groups, whatever patterns shown in the absolute test score will make no difference to further analysis in this study. This will also apply to the fact stated in the finding (ii) above.

The comparability of the non-project group to the project group is key to disentangling potential project intervention effects. Lower test scores of project group reflected the fact that SBEP project was focused on the most deprived schools in the poorest counties. Schools entering the SBEP project had least resources and most inequality problems in the system. It is inevitable to see the 'control' group showing higher results on student performance than does the other, even it was matched by socio-economic measure and some student characters at school level with some details presented in Table 2 of the report. The SAS study was designed to examine changes over time and compare differences in changes between the

'experiment' and 'control' groups. High absolute score in one group will not matter. Potential confounding in the comparison in changes could be student background factors such as age, gender and ethnicity, as older students may make slower progress than younger ones, and girls might make faster progress in Chinese than might boys. Statistics in Table 2 showed no difference in the distribution of age and gender between the project and non-project groups, but considerable difference in proportion of minority students between the two. Further examination for possible impact of such difference on the outcome, we found from the Baseline data that for the project group, minority students had significantly lower test scores on both subjects than did Han students (47.7 vs 45.2 in Chinese and 46.1 vs 44.2 in Math). For the non-project group, no difference was found between Han and minority students in Chinese (49.2 vs 49.2) but a higher score in Maths in minority students than in Han (48.1 vs 52.2). Assuming the same patterns among the 3 phases of samples, we can see that minority students in the non-project group will not affect the comparison in the change score, or could affect it in favour of the non-project group.

Another finding was that in non-project group, the dramatic rising of test scores of Grade 3 students in this group at the EoP test was most likely due to selection bias from a small and unrepresentative sample size, which resulted in much greater progress for them than for those of the project group. Possible selection bias was also speculated for the non-project sample of Grade 5. At this point it would be reasonable to say that the analysis was not conclusive on the overall impact of the project for Grades 3 and 5 students in a 4-year period.

Although for the more able students in the non-project group, there could be possible selection bias at the EoP in favour of this group and more minority students who could do better than Han students in this group, the weighted regression analysis of mean scores still demonstrated significant improvement of students in the project group above the other group either for all students in the 1st two year period or for Grades 7 and 9 students in the whole 4-year period of the project.

5.2 Project effects by intervention intensity

Hampered by the fact that data in non-project counties were only collected on test scores and a few variables (gender, age and ethnicity) with a lot of data on those variables missing due to lack of resources and manpower in tracking the non-project samples, for ascertaining project effects the team looked for variables that could differentiate project units in intervention intensity based on much more comprehensive and complete data collected from the project samples. The most intensively used measure for intensity of project interventions within project counties was "SDP", an indicator to differentiate students and schools from SDP and non-SDP schools in the project counties. Our study found that apart from SDP-related activities which took place only in SDP schools, SDP schools also differed from non-SDP schools in terms of the intensity of project interventions in some other aspects such as the percentage of students receiving SBEP aids and the percentage of teachers who have received training in the past two years.

The analysis aimed to test the assumption that if the project interventions via SDP activities were effectively focused on the most needy, deprived and poorest schools and students, one should be able to observe faster improvement in performance among those students who had real exposure to SDP interventions over time than in their counterpart who did not have exposure to SDP activities. This assumption could be tested using our Model (2) which takes into account random effects among counties and among schools and then adjusting for many confounding variables of student level that could obscure some possible SDP effects on special groups of students with disadvantages such as girls, minority, disabled and the SES scale as a composite measure of family background of students. These groups were chosen for investigation in order to reflect the fact that many project interventions were aimed at helping girls, minority students, the disabled and those from the poorest families.

The analysis found was encouraging. Students in SDP schools were generally low performers at the Baseline, remained low at the MTR but showed significant improvement at the EoP, for Grade 9 students in particular. Girls of all grades in SDP schools demonstrated much improved performance in Chinese over girls in non-SDP schools at both mid-term and end of the project, and small improvement in Maths at the end of project. Students with low socio-economic background at all grades in the SDP schools showed a consistent pattern of greater progress in both subjects over the project life than did those in the non-SDP schools, with half showing statistical significance.

These findings could be evidence of project effects - by having provided boarding subsidies for 220,000 disadvantaged students, particularly girls and students from the poorest townships/families (Output 1) and by training in child protection issues and pastoral care in boarding schools with particular emphasis on girls' safety and security in the best interest of girls and most disadvantaged students (Output 3 and 5).

In fact, there were significant increases in the proportion of girls at junior secondary level over the life of the project. The proportion of girls increased from 43.4% to 46.2% from baseline to the end of the project (Qualitative Research of Project Completion Review – National Summary Report para 39 ; the "Qualitative Report") and retention was also improved. The Qualitative Report found that living conditions for girl boarders at junior middle schools (all of whom did SDP) had "greatly improved" along with satisfactions ratings, especially for girls (p16). The report also found that SDP in some schools had led to better consideration of gender issues particularly for female teachers and their living and teaching conditions as well as for community engagement (p 42-43). Further, the project's emphasis on training and promoting female head teachers was felt to set good examples and positive role models for girls (p 55).

The evidence for SDP effects on minority and disabled students was patchy or inconsistent or weak. One possible explanation for unobserved impact on minority students could be that at least half of girls came from ethnic backgrounds, and the impact on girls would have included minority girls, i.e. a possible overlap in effects between the two. Further exploring for evidence of this speculation in the data could be helpful. A possible way forward could be deriving new variables based on needs of students for the project support. This variable would make exclusive categories out of the current identifiers of disadvantaged groups such as girl, minority, disabled and many others available in the dataset. The sample size for disabled students could be too small to detect moderate or small project effects from the mixed panel design. Future study of this group of students should consider full cohort design with well-matched parallel control for project effects. A longer time of follow-up for such a cohort might be desirable.

5.3 The SDP effects on sub-domain scores

Drawing on experiences from the UK-China Gansu Basic Education Project (GBEP, 1999-2006), the authors intended to link possible project effects on different types of knowledge or learning skills via teacher training for classroom teaching and curriculum development. An assumption was that the project might have brought in a change of teaching methods among teachers, a change of learning methods among students, which in turn might have impact on certain areas of knowledge such as practice in Maths or Reading in Chinese more than in other areas. With sub-domain analysis, we could test such a hypothesis. Since SDP had been expanded gradually to more schools by MTR stage and there was a much higher intensity of other project interventions (such as teacher training) also around that time, this analysis was focused on examining changes in disadvantaged groups in SDP schools over the period from MTR to EoP in comparison to those in the non-SDP schools in the same period.

The main findings were consistent with what was found in the total score analysis, i.e. most positive effects of the SDP interventions were on girls and poor students of all grades measured by the SES scale. The

reasons of these findings were discussed in the previous section of this report. A different pattern emerged of project effects on disabled groups – this was that a positive impact was found among this group of Grades 3, 7 and 9 students regardless of sub-domains for both subjects. There could be two reasons for this finding: different period for changes and different group of disabled students. It is possible that disabled students were from different schools at the Baseline, MTR and the EoP. Previous analysis of the total test score was to measure changes from Baseline to MTR and Baseline to EoP, while at the sub-domain analysis was for changes between MTR and EoP. With different student groups at different times, the authors were comparing means between cross-sectional samples, and project effects could be found at one time for one group but not at another time for another group. These patchy findings may still suggest some project effects on this group of students, but requires further research with a different design to confirm.

A lack of evidence in differentiated project effects on sub-domain knowledge or skills in learning ability of students might be explained by the fact that teacher training courses in the SBEP project were focused more on general pedagogy, classroom management and specific teaching techniques than on curriculum and teaching materials.

5.4 Lessons learnt

5.4.1 Project design and implementation

For large-scale field evaluation studies like the current one, it is extremely difficult to follow standard rules of experimental design without sufficient budget and manpower, in particular for follow-up studies. On balance, between the best possible designs and resources available for the study, the team chose school panel design instead of mixed student cohort design over the project period to take test scores at 3 time points. The carefully selected matched non-project control for the same length of follow-up of the same test outcome should be a merit of the study design. It was intended that this group would differentiate project effects over time from natural change trends over time in students' learning ability, providing the comparability of the two groups over time in terms of school SES and major student background.

The non-project sample size was adequately calculated detecting moderate mean changes as project effects. At the Baseline, there were very detailed technical requirements on data collection from the non-project schools on student's testing scores, age, gender and ethnicity of student and school type. Data at the MTR and EoP tests had to be collected in the same school and on different students for the keeping of school panel over time.

However, such design was not implemented in full with only one third of project schools followed up 3 times and another 1/3 two times. Most non-project schools had only one or two entries. The study ended up having mixed panel data. In addition to this, for the Baseline, MTR and EoP follow-up, data collected from the non-project schools contained only the school name, student test score (total score) and student gender. No-one seemed responsible for monitoring data collection in the schools of non-project counties, although non-project counties were also located in the same SBEP project provinces. As a result of limited data, the analysis of project overall effects could not be conducted by school type or by ethnicity for more explicit interpretation.

5.4.2 Data collection procedures

There was evidence that sampling and data collection in non-project counties were not well implemented. There are two possible reasons for the biased sampling: the first is that the Output 4 expert team did not

give the PMOs clear instructions as to how to conduct sampling in non-project counties (as they did do for the sampling in project counties), giving only some general criteria for the sample selected; there were no measures or procedures to guarantee that the sampling obtained the "qualified" student group as designed. The second is that the implementation party may have not taken the sampling principles or criteria into full consideration. In some provinces, the sampling in non-project counties seemed not to be very seriously undertaken, even schools in the provincial capital were selected.

There were also problems with data collection. Researchers found that, apart from other problems, the quality of the data from non-project counties was not ideal. One piece of evidence of this problem is that for the handful of variables associated with the student tests, many were missing values.

5.4.3 Data management and research procedures (in relation to SBEP project administration)

It was understood from the start that sampling and surveying in non-project counties might have difficulties: non-project counties didn't receive project funding and were not likely to feel obliged to do the surveys for the project. Lessons could be learned in this regard: as a government project, the provincial PMOs have some authority in non-project counties but they may not take the activities outside the project counties as seriously as those within project counties. Closer monitoring and guidance of the activities outside project counties from the national level would have been helpful to guarantee the same quality as in the project counties.

It must also be said that data management also has room for further improvement. For example, the problems with data from non-project counties only actually emerged at the MTR stage. Because there were no sound mechanisms for the checking of data quality, or for exchanges and discussions between the overall designer, the data management team and the analysts, the problems with data from non-project counties persisted until the project ended.

One other management problem was that the expert who designed the SAS study was only brought on-site at the Baseline stage, and did not participate in any of the on-site training of researchers for the two follow-up data collection, database quality and data analyses. The lack of international expertise on-site may have contributed to data collection problems. This was due to budget shortages for international consultants and to overestimates of the capability of the provincial and national teams to undertake this kind of assignment.

6. References

Alexander, R. (2008) Education For All, The Quality Imperative and the Problem of Pedagogy. CREATE Research Monograph (20). IoE London; Bennell, P. (2004) Teacher Motivation and Incentives in Sub-Saharan Africa and Asia. Brighton; Omari, I., Baser, H. (- in draft 2012) "Human Resources for Quality Education in Tanzania".

Barber , M and Mourshed, M (2007) How the world's best performing school systems come out on top http://www.mckinsey.com/App_Media/Reports/SSO/Worlds_School_Systems_Final.pdf

Barrera-Osorio, F. et al., 2009, World Bank Decentralized Decision-Making in Schools, The Theory and Evidence on School-Based Management; MOEVT/VSO Zanzibar (2011) Leading learning: a Summary report on effective leadership and quality education in Zanzibar

Goldstein, H. (2010). Multilevel Statistical Models, 4th Edition, John Wiley & Sons, UK

Goldstein, H., Yang, M., Turnar, R., Omar, R. & Thompson, S. (2000). Meta analysis using multilevel models with an application to the study of class size effects. Journal of Royal Statistical Society, C. 49, Part 3, 399-412.

Khattri, N Ling, C Shreyasi J (2010) The Effects of School-based Managementin the Philippines (An Initial Assessment Using Administrative Data). The World Bank Independent Evaluation Group,East Asia Education Sector Unit & World Bank Institute

Liu, Yunshan etc, 2009: Selection of Elites: Views from Social status, Geographical variation, and Capital Gaining: Case Study on Farmers' Children Who Get Admit ted into Peking University (1978- 2005), TSINGHUA JOURNAL OF EDUCATION, Vol 30, No. 5 Oct. 2009

Li Chunling: Socio-Political Change and Inequality of Educational Opportunities

--Influences of Family Background and Institutional Factors on Acquisition of Education (1940-2001), Chinese Education and Society, M.E.Sharpe Jan –Feb 2012,

Ministry of Education: The Outline of China's National Plan for Medium and Long-Term Education Reform and Development (2010-2020), 2010
http://www.moe.gov.cn/publicfiles/business/htmlfiles/moe/moe_177/201008/93785.html

SBEP (2011) Qualitative Research of Project Completion Review – National Summary Report. (SBEP Report)

SBEP (2011) End-of-project review; Quantitative Survey Report (SBEP Report)

Yang, M., Goldstein, H., Browne, W., and Woodhouse, G. (2002), Multivariate multilevel analyses of examination results, Journal of Royal Statistical Society, A, 165, Part 1, 137-153.