

IMPACT OF NEW HORIZONS FOR PRIMARY SCHOOLS ON LITERACY AND NUMERACY IN JAMAICA 1999-2004¹

**Marlaine Lockheed, Ph.D., Abigail Harris, Ph.D.,
Paul Gammill, Karima Barrow
The Academy for Educational Development**

Citation

Lockheed, Marlaine; Harris, Abigail; Gammill, Paul; Barrow, Karima (2006). Impact of New Horizons for Primary Schools on Literacy and Numeracy in Jamaica 1999-2004, *Journal of Education for International Development*, 2:1. Retrieved from <http://www.equip123.net/JEID/articles/2/NewHorizons.pdf> on [insert month] [insert day], [insert year].

Abstract

In school year 1998-1999, the United States Agency for International Development fully rolled out the New Horizons for Primary Schools Program (NHP) for approximately 10% of the most poorly performing schools in Jamaica. The program was designed to improve the quality of teaching in these primary schools, to raise literacy and numeracy levels, to increase school attendance and to strengthen school management. The first cohort of students to attend NHP schools for all or most of Grades 1-6 completed Grade 6 in 2004. This report examines how NHP affected student learning achievement and discusses the data requirements for more rigorous analyses.

Section 1. Background and Key Research Questions

The New Horizons for Primary Schools program (NHP) was initiated in school year 1997-98 and fully rolled out in school year 1998-99. The objective of the program was to improve the language and mathematics performance of 72 of Jamaica's poorer performing schools through a school-based model of intervention. School improvement plans (SIPs) were to be developed for each school, with interventions selected from a menu of ten project interventions (see Box 1) in accordance with each school's need.

Box 1: New Horizons Interventions

1. Developing innovative mathematics and literacy programs
2. Providing in-service teacher training in reading and mathematics
3. Providing governance and leadership training for schools, communities, parents
4. Offering parent education and training
5. Facilitating selective nutrition and health programs
6. Providing reading and mathematics materials
7. Establishing computer use in school and training teachers in educational technology
8. Training resource teachers
9. Integrating databases using MIS
10. Improved school management through EMIS

¹ This paper was presented as a project evaluation to USAID.

Lead institutions were identified for each of the NHP interventions, with an institutional contractor taking responsibility for all but three of the interventions. The Professional Development Unit (PDU) of the Ministry of Education, Youth and Culture (MOEYC) was responsible for training of resource teachers and the MOEYC in partnership with the National Council on Education (NCE) was responsible for providing governance and leadership training for schools, communities and parents and for offering parent education and training. In the spring of 1999, a diagnostic survey of all NHP schools was undertaken to assess the schools' training and other needs (Juarez and Associates, June 1999).

Systematic evidence regarding the extent to which various interventions were implemented in the NHP schools is modest but shows improvement over time. Some evidence comes from evaluations of School Development Plans (SDPs), which are the "focal point of NHP's approach to school governance" and also are essential in the needs assessment process (Dye *et al.*, 2002, p. 11). One evaluation codified the quality of literacy and numeracy initiatives in SDPs for NHP schools as of November 1999, based on SDPs received from approximately three-quarters (56 of 72) of the NHP schools; of these, the quality of fewer than 20 percent was judged "satisfactory" and the quality of approximately 40 percent was judged "weak" (Juarez and Associates, December 1999). Approximately half the SDPs included a statement of actions that the school would take to reach their specific literacy or numeracy attainment target. The report notes that "very few schools appear to be in a stated position of readiness to deal with literacy and numeracy in their schools" (Juarez and Associates, December 1999). This number was apparently higher a few years later. An analysis of 56 SDPs in 2003 judged all but four of them to be "good" or "very good" (Summary Evaluation of School Improvement Plans (SIP) of NHP, Spring 2003). This later evaluation, however, noted that half (28 of 56) of the NHP schools for which SIPs were available lacked the desired three-year action plan for implementation of the program.

As anticipated, the NHP interventions were not implemented uniformly across all 72 schools; implementation varied across schools in accordance with local needs. For example, only 14 of the 72 schools received breakfast programs. The intensity of the interventions also varied, with training program duration lasting from a few hours to several days. Table 1 summarizes the main features of the implemented program.

Table 1. Features of NHP as Implemented by 2003

Intervention	Implementation
Developing innovative mathematics and literacy programs	<ul style="list-style-type: none"> ▪ 100s of site visits, deployment of 16 "NHP Associates" to work at classroom level
Providing in-service teacher training in reading and mathematics	<ul style="list-style-type: none"> ▪ Consolidated with #8
Providing governance and leadership training for schools, communities, parents	<ul style="list-style-type: none"> ▪ Procurement of governance and Leadership Coordinator and Officers ▪ Examination of SDPs ▪ Site visits in 60 NHP schools ▪ Finalize Manual on Governance and Leadership Training for School Boards and Principals

	<ul style="list-style-type: none"> ▪ NHP Principals' Workshops ▪ Other training
Offering parent education and training	<ul style="list-style-type: none"> ▪ National Parenting Conference (1999, 2002)
Facilitating selective nutrition and health programs	<ul style="list-style-type: none"> ▪ Subsidy of breakfast program in 14 schools ▪ Community mobilization to sustain program ▪ Teacher training on integrating health and nutrition in teaching of core subjects ▪ Nutrition Specialist and community development specialist.
Providing supplementary reading and mathematics materials	<ul style="list-style-type: none"> ▪ Supplementary materials and equipment distributed to schools
Establishing computer use in school and training teachers in educational technology	<ul style="list-style-type: none"> ▪ Five "technology-intensive" NHP schools established ▪ Two three-day and one overlapping six-day Educational Technology Workshops held for teachers in 5 NHP schools (2002) ▪ One-week consultancy on use of technology for student literacy ▪ Consultations with 72 school principals on incorporating technology into SIPs
Training resource teachers	<ul style="list-style-type: none"> ▪ Trained 180 Mathematics and Literacy Resource Teachers in workshops and in-school training activities
Integrating databases using MIS	<ul style="list-style-type: none"> ▪ Jamaica School Administration System software 5.0 was developed and used in NHP schools (and a staged rollout across Jamaica is planned). ▪ Support guides and training manuals for 200 non-NHP schools prepared
Linking Project Schools to EMIS Network	<ul style="list-style-type: none"> ▪ 25 large and medium schools received additional computers (2002) and 140 computers were networked

Source: O'Neil, October 2003

Over the period of implementation of the NHP a large number of formative and other evaluations have been carried out; nearly 100 have been catalogued by the Curriculum and Support Services Unit of the MOEYC (O'Neil 2003). However, none of these studies has addressed, in a comprehensive manner, a series of questions posed by USAID:

1. Have NHP schools made achievement gains over the years under review?
2. What factors in the project schools may have affected gains or lack of gains?
3. Is the use of mastery/near-mastery/non-mastery categories on the Grade Six Achievement Test (GSAT) masking real gains in student achievement in schools?

4. Is the GSAT the best measure of student performance for the project schools, considering that it is based on the content delivery system of the old curriculum?
5. How can valid measures of students' computational skills in numeracy be assessed for students who are unable to comprehend the language in which most numeracy items are couched in the GSAT examination?
6. How has "social promotion" to Grade 6 affected average performance results among students?
7. How effective were the indicators used for tracking the results of the NHP and what suggestions could be made for the future?
8. How effective are the methodologies used to collect data?

The present report addresses these questions. Section 2 examines the effects of NHP on student achievement, 1999-2004 and explores school factors that may have affected achievement changes over this time frame. Section 3 addresses issues related to student performance measures and social promotion and achievement. In Section 4 we consider a number of issues related to data collection and indicators. Section 5 presents our conclusions and recommendations.

In carrying out this evaluation, we utilized six school years of archival data, 1999-2004, from school censuses and the Grade Six Achievement Test (GSAT). We also reviewed key implementation and evaluation documents related to the NHP.

Section 2. Effects of NHP on Student Achievement

Two key evaluation questions dealing with achievement were posed by USAID:

- Have NHP schools made achievement gains over the years under review?
- What factors in the project schools may have affected gains or lack of gains?

In addressing these questions, we compare the performance of NHP schools both with that of all other government schools in Jamaica having primary sections and with that of a set of matched schools in which the NHP was not implemented. The only student performance data that were available for comparison across the two groups were Grade Six Achievement Test data. Two tests assessing performance in earlier grades – the Grade Three Diagnostic test and the Grade Four Literacy test – were administered in all schools in Jamaica, but results were not consolidated nationally and therefore could not be used in this analysis. Since the NHP program is intended to improve literacy and numeracy, we focus on four of the GSAT tests that measure these skills: language arts, mathematics, and writing (Communications Task I and Communications Task II).

Comparing NHP schools with other government schools allows us to see whether the apparent decline in GSAT scores in NHP schools is unique to program schools or is a phenomenon shared across schools in Jamaica. Comparing NHP schools with a matched set of non-NHP schools allows us to address the central question of the NHP's impact on student achievement as measured by GSAT.

How do NHP Schools Compare with all Government Schools in Jamaica Having Primary Sections?

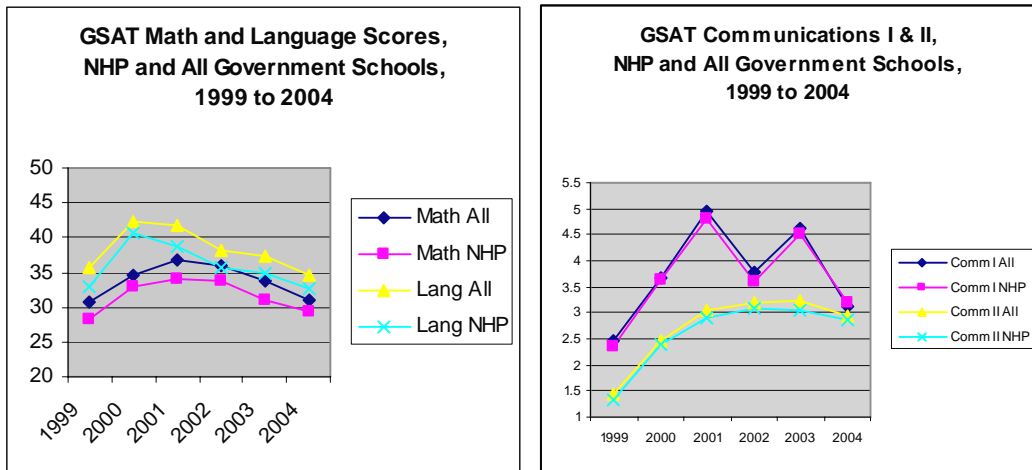
NHP was initiated in 1998 and rolled out in 1999. The program itself may have been fully operational in only about three-quarters of the schools as late as school year 2003. The effects of NHP on student performance on the Grade Six Achievement Test (GSAT) are not likely to be observed until after this time for two reasons. First, even if the program had been fully operational since 1999, the first cohort of primary students given the opportunity to attend NHP schools for all or most of grades 1-6 would have entered grade 1 in 1998 and completed Grade 6 only in 2004; they

would have taken the GSAT in that year. Second, incomplete implementation until after 2003 would have delayed observing NHP effects on GSAT even further; students entering grade 1 in 2003 would not take the GSAT until 2008. If student performance data from earlier grades were available, effects of NHP might be observed for student cohorts entering grade 1 in 2000 or 2001; these data, while collected locally, are not routinely aggregated centrally and were therefore not available for analysis.

GSAT performance of students in NHP schools parallels that of students in all government schools in Jamaica having primary sections. Raw GSAT scores for all schools increased from 1999 to 2001 and then dropped gradually through 2004 (figure 1). This is largely due to the impact on scores of a changing population of test-takers, following the elimination of the Common Entrance Examination (CEE 11+) after 1998. The CEE 11+, a selection test for secondary education, was administered to students within a given age range, without regard to the grade in which they were enrolled. As a consequence, a significant share of students took the CEE in Grade 5; if they passed the examination in Grade 5 they advanced to Grade 7 at the secondary level without taking Grade 6. Thus, the CEE 11+ skimmed off the higher performing student from Grade 6. Often students who did not pass the CEE 11+ the first time had the opportunity to retake the test the following year.

When the GSAT replaced the CEE 11+ in 1999, the higher performing students who would have been in Grade 6 in 1999 were already in Grade 7 and did not take the GSAT, depressing average scores. Scores on GSAT would naturally rise over the next two years, as Grade 6 included a higher share of higher performing students than in the past. We discuss this phenomenon in Section 3 in the context of changes in the age composition of those taking the GSAT. The mean scores of students in NHP schools showed the same pattern of rise and fall, although at a lower level in most years.

Figure 1. GSAT Mathematics, Language arts and Communications I & II scores in NHP and government schools with primary sections, 1999-2004



Moreover, since GSAT scores were not equated from 1999 to 2004 (sometimes intentionally), the 1999-2004 trend data do not accurately represent change in achievement over time but rather represent changes in the tests as well as changes in the test takers. Equating issues are discussed in Section 3. It is necessary to compare the performance of students in NHP schools with the performance of students in comparable non-NHP schools to accurately assess the impact of NHP on performance.

How do NHP Schools Compare with Matched non-NHP Schools?

We use propensity score matching to identify a set of government schools in Jamaica that were matched with NHP schools in 1999 across a wide range of characteristics, but that did not participate in the program. Propensity score matching is discussed in Annex A.

Of the 72 NHP schools, we were able to match 70 with equivalent schools not participating in the program. To the extent possible, using school census data available for all government schools in Jamaica, we matched schools on the basis of eight criteria that were initially used to place schools in the program: (a) performing at or below the national mean in language and mathematics in the National Assessment Programme, (b) performing at or below the national mean in language, mathematics, science and social studies in the National Assessment Programme, (c) geographic location, (d) evidence of Board, or principal and teachers taking action to address the under-achievement of students in the school, (e) active functioning School Board or SCOPE Committee, (f) recipient of grant for Jamaica School Investment fund or civil works in the IDB PIEP project, (g) potential for providing inspiration and leadership in the project, (h) participation in other initiatives complementary to the project. In 1998, 194 schools met these criteria (PIU, December 1998) and were eligible for selection; data related to some of these criteria were included in the 1998 School Census for all schools.

In addition, we identified four other characteristics of schools and teachers that we hypothesized were important determinants of student achievement and for which we could obtain data from the School Census: (a) teacher quality, (b) teacher experience, (c) poverty level of school community, and (d) size of school. These twelve characteristics and the data sources for each are summarized in table 2.

Because not all selection criteria were supported by data from the annual School Census, and because some of the selection criteria required expert judgment, we included “program eligibility” in the logit regressions for establishing the matched non-NHP comparison group.

The propensity score matching worked relatively well, with 97 percent of the NHP schools matched with non-NHP schools; approximately two-thirds of the matches were very close, with scores matched to the second decimal point or better. Detail on the propensity score matching approach can be found in Annex A.

The non-NHP schools were well-matched with the NHP schools at the outset of the program. Inspection of school, teacher and student characteristics of the NHP and matched non-NHP schools as of 1999 demonstrates the similarity of the two sets of schools. For none of the initial 1999 characteristics, including average student performance on the GSAT, are there statistically significant differences between the two groups (see Annex A for details). By 2004, however, differences are emerging on the GSAT. Since we theorize that the impact of NHP would not be observable in GSAT scores prior to 2004, we concentrate on this year.

Table 2. Criteria for School Selection into NHP Program

NHP Selection Criteria	Indicator from School Census or GSAT
Performing at or below the national mean in language arts and mathematics in the National Assessment Programme,	School mean GSAT Scores on language arts and mathematics, school year 1998-99
Performing at or below the national mean in language arts, mathematics, science and social studies in the National Assessment Programme,	School mean GSAT Scores on language arts , mathematics, science and social studies, school year 1998-99
Geographic location	School Census: Classification of school as rural, remote rural
Evidence of Board, or principal and teachers taking action to address the under-achievement of students in the school,	None
Active functioning School Board or SCOPE Committee,	School Census: presence of School Board or SCOPE Committee
Recipient of grant for Jamaica School Investment fund or civil works in the IDB PIEP project,	None
Potential for providing inspiration and leadership in the project,	None
Participating in other initiatives complementary to the project.	None
Other school factors	
Teacher quality in Grades 1-6	School Census: Percent teachers with CXC as highest level of school attainment
	School Census: Percent master teachers in school
Teacher experience in Grades 1-6	School Census: Average number of years experience as a teacher
	School Census: Average number of years experience in the school
	School Census: Percent teachers with less than two years experience
Poverty level of school community	School Census: School breakfast program
Size of school	School Census: Number of teachers, Grades 1-6
	School Census: Number of students, Grades 1-6
	School Census: School on Shift
PTA	School Census: presence of PTA

We compare NHP schools with non-NHP schools in three different ways. First, we examine the mean scores of schools in 2004, using matched pair t-tests, which are more sensitive to change than simple t-test comparisons of means. Second, we use a simple OLS regression of school means, where the dependent variable is the school mean GSAT score in 2004 and the NHP program is considered an independent variable, with the school mean in 1999 as a control. Finally, we classify the schools

according to their mean scores and compare the two groups (NHP and matched non-NHP) according to the share of each group in high, medium and low categories of achievement.

Matched Pair t-tests

In the first analysis, we found that NHP schools outperformed matched non-NHP schools on one of four GSAT 2004 tests.² We compared average 2004 GSAT scores in language arts, mathematics and writing (Communications Task I and Communications Task II) for 70 NHP schools with those of a matched set of 70 non-NHP schools. The approach used was matched pair t-tests (difference of differences). The figures in table 3 are raw scores, and the maximum score for the mathematics and language arts tests (80 points) is different from the maximum score for the writing tests (6 points). The GSAT-2004 results show a possible positive impact of the NHP program with respect to improvements in writing at the “basic” level (Communications Task I). The average score of NHP schools in 2004 is nearly 15 percent higher than that of the matched non-NHP schools and this difference is statistically significant ($t = 2.31, p < .03$). No significant difference was observed for the other tests, however (table 3).

Table 3. NHP program effects on school mean GSAT scores in 2004, matched pair t-tests

<i>2004 GSAT</i>	New Horizon Program School	Matched Non-New Horizon Program School	Matched pair t-test
Mathematics	29.3	29.2	n.s.
Language Arts	32.6	32.9	n.s.
Communications Task I	3.2	2.8	2.31, $p < .03$
Communications Task II	2.9	2.8	n.s.

Scatter plots showing the 1999 mean school achievement and the 2004 mean school achievement, for NHP and matched non-NHP schools can be found in Annex C. They show the slight positive effect of NHP on achievement in these initially lower performing schools.

OLS Regressions

OLS regressions confirmed the effect of NHP on Communications Task I scores. We tested for the impact of NHP on 2004 GSAT scores through OLS regressions, controlling for school mean 1999 scores on the same tests. The results of these regressions are shown in table 4, which show the unstandardized regression coefficients (standard error in parentheses) and confirm the previous analysis. Controlling for school average student performance on the GSAT Communications Task I in 1999, student performance in schools participating in the NHP program was higher in 2004 on the Communications Task I assessment. Program effects are not statistically significant for other tests.

² Because the objective of the NHP was to improve literacy and numeracy, this analysis focuses on tests measuring these skills. We did not, therefore, analyze GSAT Science and GSAT Social Studies tests. Results for all tests appear in Annex B.

Table 4. NHP program effects on school mean GSAT scores in 2004, OLS regressions

Dependent variable:	Unstandardized regression coefficient (standard error in parentheses)			
	2004 Math	2004 Language	2004 Comm. I	2004 Comm II
1999 School mean score	0.39** (0.11)	.27 (.96)	-.005 (.18)	.32* (.16)
School in NHP	-.006 (0.93)	-.47 (9.6)	.37* (.16)	.009 (.100)
Constant	18.4** (3.17)	24.24** (3.08)	2.93** (.45)	2.35** (.22)
R-square	.08	.06	.04	.04
Adj. R-square	.07	.04	.03	.02
Cases	140	140	140	140

** $p < .01$; * $p < .05$

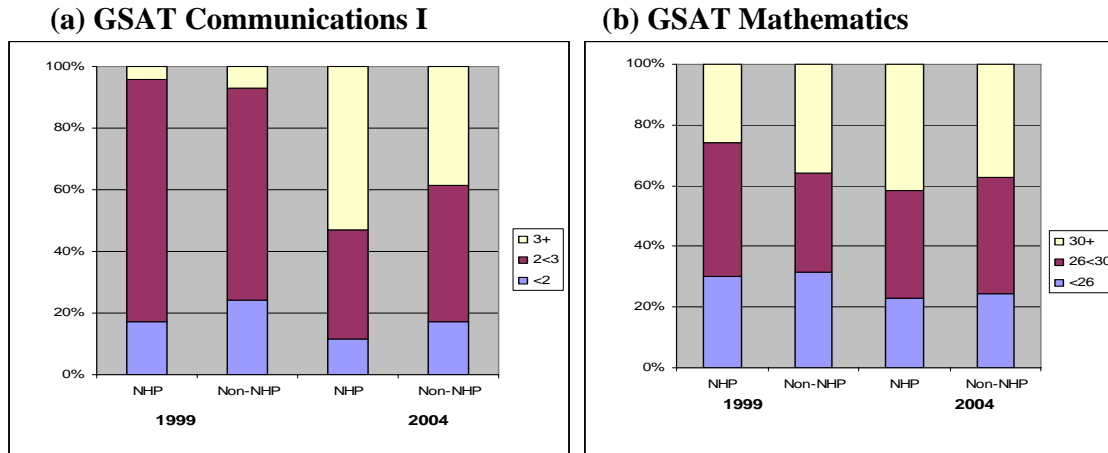
Non-parametric Tests of Differences

Our final test, which involved classifying schools according to their mean performance on GSAT tests, and comparing the NHP schools with the matched non-NHP schools in 1999 and 2004, revealed a small gain for GSAT Mathematics, in addition to Communications Task I.

For Communications Task I, the share of NHP schools with mean scores greater than three points (on a six-point scale) increased from four percent in 1999 to 53 percent in 2004; the share of matched and non-NHP schools with mean scores greater than three points increased from seven percent to 39 percent (figure 2a). Again, the tests may have not been fully equated, but the difference between NHP and matched non-NHP schools in gain was statistically significant ($p < .05$).

Surprisingly, for GSAT Mathematics, the share of NHP schools in the higher category also improved, from 26 percent of schools having means GSAT Mathematics scores of 30 points or more in 1999 to 41 percent of schools in 2004, compared with essentially no gain for matched non-NHP schools (36 percent of schools in 1999 with mean scores of 30 points or more to 37 percent of schools in 2004). This difference was also statistically significant (figure 2b).

Figure 2. Share of schools at three levels of performance on two tests, NHP and matched non-NHP schools, 1999 and 2004



By comparing NHP schools with a matched group of schools that did not participate in the program, and by examining performance improvements that occurred well below the “near mastery” level, it is possible to see a small achievement impact from the program. NHP appears to have boosted student writing skills and mathematics performance, albeit at lower levels of performance, as assessed by the GSAT.

School Factors that may Explain NHP Performance Advantage

We examined two sets of factors that may have explained the NHP performance advantage: (a) School Development Plans, and (b) inputs in non-NHP schools that may have been provided by other programs.

School Development Plans

School Development Plans (or School Improvement Plans) are a key aspect of the NHP program and may also be considered a proxy indicator for principal leadership. The objective of SDPs/SIPs is for the school to analyze its needs and set out an action plan to address these needs. We noted above that 20 percent of schools had actionable SDPs in 1999 and that this had increased to 72 percent of schools in 2003. We selected the sub-group of NHP schools that had “good” or “very good” SIPs in 2003 to see if the effects of NHP were more pronounced in these schools than in the remaining NHP schools. We found that the effects for NHP were similar in these schools to those for all schools (table 5). That is, the 50 NHP schools with good SDPs/SIPs achieved higher performance than their matched non-NHP schools, but the raw scores are no different from those reported above in table 3.

We conclude that SDPs/SIPs, and the principal leadership they imply, may account for the NHP advantage, but that other factors may also be responsible. We do not have information about whether the matched schools also had SDPs/SIPs.

Table 5. Impact of SIPs on school performance, GSAT 2004

<i>2004 GSAT</i>	New Horizon Program Schools (50 with 2003 SIP)	Matched Non-New Horizon Program Schools	Matched pair t-test
Mathematics	29.5	29.4	n.s.
Language Arts	32.6	32.9	n.s.
Comm. Task I	3.2	2.8	2.09, $p < .05$
Comm. Task II	2.9	2.8	n.s.

Other Factors Affecting Achievement

Isolating the effects of various inputs to the NHP program is difficult for two reasons. First, other programs may have provided similar inputs to other, poor-performing non-NHP schools, thus mitigating the unique effects of the NHP program; there is no systematic record of these inputs for non-NHP schools, however, to enable this hypothesis to be tested. Second, implementation of NHP was not systematically monitored, so even at the NHP school level basic information is not available on all schools.

We nevertheless attempted to estimate the achievement effects of a small set of characteristics of the schools, the teachers and the students for which data were available across all 791 government primary schools in Jamaica. In these analyses, we focus only on school mean achievement.³

We estimate school mean achievement separately for GSAT language arts and mathematics scores, as a function of school characteristics, teacher characteristics, and student achievement. In addition, we include an indicator variable for participation in the NHP program, to test for program effects on the full population of schools. Results are presented in Table 6.

³ We recommend that subsequent analyses be carried out at two-levels: students within schools.

Table 6. School, Teacher and Student Effects on School Mean GSAT scores in 2004, OLS Regressions

Dependent variable:	GSAT Language 2004			GSAT Math 2004		
	B	Std. Error	t	B	Std. Error	t
School characteristics 1999						
Total enrollment grades 1 to 6	0.003	0.002	1.164	0.002	0.002	1.026
Count of teachers grades 1 to 6	0.045	0.089	0.512	0.074	0.086	0.865
School Shift (1 = yes)	-3.377	0.899	-3.755***	-4.128	0.868	-4.759***
Rural location (1= rural)	-0.763	0.593	-1.286	-0.962	0.572	-1.681*
Remote Rural location (1= remote rural)	0.197	0.736	0.268	-0.042	0.710	-0.059
Breakfast (1 = yes)	0.110	0.606	0.181	0.708	0.585	1.211
Active PTA (1 = yes)	4.091	0.985	4.154***	3.730	0.950	3.928***
School has School Board or Scope (1 = yes)	-0.337	0.663	-0.508	-0.690	0.640	-1.079
NHP school (1= yes)	-1.056	0.723	-1.460	-0.643	0.698	-0.922
Teacher characteristics 1999						
Percent teachers with CXC (highest level of teacher training) only	-3.299	2.324	-1.419	-3.587	2.242	-1.600
Percent teachers with Certificate only	-3.265	2.045	-1.597	-4.077	1.972	-2.067**
Percent master teachers in school	0.164	0.321	0.511	0.046	0.309	0.149
Mean years experience at grades 1 to 6	0.043	0.012	3.686***	0.036	0.011	3.241***
Mean years experience in school (grade 1 to 6)	0.007	0.038	0.174	0.039	0.037	1.073
Student achievement 1999						
GSAT Mathematics 1999	0.058	0.070	0.823	0.241	0.067	3.564***
GSAT Lang 1999	0.276	0.061	4.539***	0.135	0.059	2.295**
(Constant)	20.673	2.882	7.174***	17.299	2.779	6.224***
R-square	0.244			0.273		
Adj. r-square	0.228			0.258		
Cases	791			791		

*** p < .01, ** p < .05, *p < .10

Factors associated with higher scores on 2004 GSAT Mathematics test suggest the importance of teachers and community, as well as historical trends in achievement. The school's prior achievement in mathematics and language arts as indicated by 1999 GSAT scores was positively associated with its subsequent performance in mathematics in 2004. The fact that language arts achievement is a strong predictor of mathematics achievement underscores the importance of disentangling the verbal and computational components of the GSAT mathematics test.

The single school characteristic that was positively associated with higher performance in mathematics was the presence of an active PTA, which boosted scores by nearly 4 points. Two school characteristics were associated with lower performance. Schools on multiple shift programs scored more than 4 points lower on GSAT mathematics in 2004 than schools without shifts, while rural (but not remote rural) schools scored about one point lower. Factors unrelated to 2004 GSAT performance were presence of a school board or SCOPE program, presence of a breakfast program, school size and remote rural location.

Teacher characteristics associated with higher performance included both teacher experience and teacher qualifications. Schools with more experienced teachers and those with a higher share of teachers with qualifications higher than either a CXC (highest level of teacher training) alone or a teacher certificate alone achieved higher GSAT mathematics scores in 2004, holding constant GSAT mathematics scores in 1999.

A similar pattern was found for language arts achievement, with three exceptions. The 1999 mean school GSAT mathematics scores was not predictive of the 2004 mean school language arts score; the effect of teacher qualifications, while in the same direction as for GSAT mathematics, was not statistically significant; and the effect of rural location, while also in the same direction, was not statistically significant.

One implication of this finding is that policies that support local PTAs and that bring better qualified teachers to poorly performing schools could boost achievement in such schools.⁴

Conclusion on Test Results

Compared with schools not in the NHP program, schools participating in the NHP program showed higher performance at the lower ends of two tests of achievement measured at Grade 6: writing (GSAT Communications Task I) and mathematics. Average Communications Task I scores were higher in NHP schools than in matched non-NHP schools in 2004, and in 2004 the share of NHP schools with average GSAT mathematics scores of 30 or more points (out of a possible 80) was greater than the share of non-NHP schools with average GSAT mathematics scores of 30 or more points. In both cases, the range of improvement occurred below the levels designated “near mastery.” Factors associated with higher literacy and numeracy in 2004 included the presence of a good quality School Improvement Plan (School Development Plan) in 2003, an active PTA, and higher performance in 1999. Schools with more qualified and more experienced teachers in grades 1-6 scored higher than those with less experienced and less qualified teachers. Schools on multiple shifts achieved less than those on single shift, while rural schools also scored less well than urban schools. Factors unrelated to 2004 GSAT performance were presence of a school board or SCOPE program, presence of a breakfast program, school size and remote rural location.

Section 3. Student Performance Measures

Three evaluation questions dealing with performance measures were posed by USAID:

1. Is the use of mastery/near-mastery/non-mastery categories on the GSAT masking real gains in student achievement in the schools?
2. Is the GSAT the best measure of student performance for the project schools, considering that it is based on the content delivery system of the old curriculum?

⁴ We recognize that deriving policy implications from correlational analyses is dangerous, as issues of causal attribution remain. However, as these findings tend to support important dimensions of the NHP and other program to improve primary education in Jamaica, it is worthwhile mentioning them.

3. How can valid measures of students' computational skills in numeracy be assessed for students who are unable to comprehend the language in which most numeracy items are couched in the GSAT examination?

Background on the GSAT and the NAP

The Grade Six Achievement Test (GSAT) was designed originally as part of the National Assessment Programme as a low-stakes test to be used primarily for national monitoring and evaluation. The content and skills assessed by the exam reflected a national curriculum from 1980 and the exam was designed for all children in Grade 6 regardless of their age or ability. Its stated intent was to measure “achievement of skills for continuing learning in Grade 7” (Russell, 1996, p. 94). The test was first administered in 1988. Subsequently, while some years the test was administered nationally, in other years only samples of Grade 6 students participated. Some adjustments were made in the test specifications to reflect changes in the national curriculum (primarily changes in the topics for science and social studies), but fundamentally the exam has maintained its original structure.

Until 1999, the GSAT existed alongside the Common Entrance Examination 11+ (CEE). The CEE was used in Jamaica from 1958-1998 to select children for admission into secondary high schools. It tested English, Mathematics and Mental Abilities and was not aligned with the national curriculum. Any child between the ages of 11 and 13 from Grades 4, 5 or 6 could sit the CEE and any child who “passed” the CEE (i.e., was awarded a place in a secondary program) was expected to enter Grade 7 the following year. Many children entered directly from Grade 5. Thus, for example, in 1994, 13,459 places were awarded to high schools but only an estimated 40 percent of these places were awarded to children from Grade 6 (Russell, 1996).

In 1999, the GSAT replaced the CEE as the mechanism for secondary school selection/placement, and students were allowed to sit the exam only once, in Grade 6. Overnight, the GSAT changed from a very low stakes test to a very high profile, high stakes test. At first, the number of students registering for the GSAT (41,932 in 1999) was lower than the number in the age cohort (estimated as 48,000-50,000 in 1999), a consequence of prior practices with the CEE, whereby the higher performing students of 1997 and 1998 would have proceeded directly to Grade 7 from Grade 4 or Grade 5, thus skipping Grade 6. In subsequent years, this “creaming” no longer occurred, and GSAT registrations both increased and stabilized: 46,746 in 2000; 47,889 in 2001; 50,547 in 2002; 49,281 in 2003; and 48,799 in 2004. A possible explanation for the slight decline in registrations in 2004 is the introduction of the Grade 4 Literacy Test in 2001, which may have slowed student progress to Grade 6 in 2004. An inspection of the ages of students taking the GSAT over these years shows that the share of those aged 12-13 increased slightly while the share of those aged 11-12 dropped slightly (table 7).

Table 7. Ages of students reporting scores from the GSAT, 2000-2004

<i>Age Group</i>	GSAT Test Year				
	2000	2001	2002	2003	2004
<11	1%	1%	0%	0%	0%
11<12	53%	53%	49%	51%	51%
12<13	45%	46%	49%	48%	48%
13+	1%	1%	1%	1%	1%

When the GSAT replaced the CEE, an attempt was made to expand the pool of secondary school places and effectively “place” the majority of Grade 6 students. Schools were instructed to include all Grade 6 students not just those who would likely “pass” and thereby keep up the school average. Regulations about grade repetition were debated. Should children be promoted with their age cohort regardless of their skill level (social promotion)? Should they be allowed to repeat Grade 5 to delay taking the GSAT or be allowed to repeat Grade 6 in order to repeat the GSAT? Regulations restricting test registration were strengthened and safeguards put in place to enforce the rule that students could only take the test once.

One challenge encountered by the MOEYC Student Assessment Unit (SAU) once the GSAT became a high stakes test was that there were insufficient items in the difficult or high end of the GSAT scale. Students were scoring 100 percent, making it impossible to distinguish among them. This was not surprising since estimations of item difficulty used in developing the test were based on data from pre-testing of items before the test carried meaning for the students. In 2002 and thereafter, the SAU modified the test specifications and added more difficult items to the test in order to be able to more accurately differentiate amongst high scoring candidates.

Mastery Levels and Measurement of Performance

The use of GSAT mastery levels for NHP program evaluation is overly ambitious. A common practice in program evaluation is to establish a criterion level of performance and to measure success against this criterion. In the absence of a meaningful comparison group, this method provides a stable means of measuring change. An example illustrates the potential utility of this approach. Consider the hypothetical situation in which prior to intervention 25 percent of the students completing Grade 1 can write their names without help and after the intervention 50 percent of students completing Grade 1 can write their names without help. In very concrete terms this means that there has been a 100 percent improvement in this skill. While it is not possible to attribute causality exclusively to the intervention, the change is objectively measurable.

Similarly it is sometimes possible to identify a specific curriculum objective (e.g., simple 1-digit addition) and measure student performance using a sample of items representative of the domain (e.g., $1 + 1$, $2 + 7$, $8 + 3$). Once again, the interpretation is fairly transparent: a student who gets less than 50 percent correct has not mastered the skill, a student with 50-75 or 85 percent correct shows partial understanding and a student who gets a high proportion of problems correct (usually in the range of more than 85 percent correct), has mastered the skill or concept. Mastery levels that rely on a score that is the aggregate of performance on several skills or curricular domains are not as easily interpretable. For example, achieving a score in the non-mastery range could mean the student mastered some skills but not others or it could mean the student has partial mastery of some skills and no master of others covered by the test. Generally, it is not possible to link the aggregate score to mastery of a particular skill.

Another issue is the use of “mastery” in the context of “high stakes” testing. What does it mean to get a score of 50 percent correct when the test is high stakes and the precision is most needed at the top? If the test is designed to concentrate on the high end of the scale (difficult items), sensitivity at other parts of the scale (for example, around 50 percent) is not as critical and consequently there may be fewer items and less sensitivity to finer distinctions in the middle and lower ranges of the scale.

In the GSAT, the high profile distinctions are at the top of the scale (that is, above 75 percent correct). Scores at this end of the range are used to identify the children who will be admitted into the prestigious high schools. A small fraction of Grade 6 students achieve at this level; most children’s

performance falls below the 50 percent correct mark (for example, in 2004, 68 percent of the test takers scored below 50 percent correct in GSAT mathematics and 60 percent scored below 50 percent correct in language arts), and the students in NHP schools on average performed substantially less well. Thus, a mastery level that defines 0-50 percent correct as “non-mastery” will lump all of these children together. The likelihood of moving out of this range, particularly if you are a student in a low performing school who started with lower skills than children in middle and high performing schools, is very slim and far too ambitious as the primary indicator of a reform’s success.

The GSAT as a Measure of Student Performance for Project Schools

As noted above, the GSAT is a curriculum-based test, which was initially developed in the 1980s. It has been continuously revised, in part to reflect changes in the curriculum and in part to reflect its increasing emphasis on selection/placement. Its utility for assessing program impact is limited by (a) the share of test questions that are comparatively more difficult, and therefore not sensitive to changes in more “basic” skills, (b) its inability to assess growth in the early grades, and (c) the weakness of its horizontal equating, rendering changes in scores difficult to interpret.

Given the pace of implementation of the NHP program, effects are most likely to be observed in the early grades, and tests designed to capture changes in beginning literacy and numeracy may be more appropriate as a measure of program impact. For example, the Grade Three Diagnostics Test was used in the formative evaluations of NHP and achievement data were systematically collected for students in project schools. Unfortunately, performance data for this test are not uniformly available for non-NHP schools. Although the MMOEYC requests that all schools submit the data to the Student Assessment Unit, many schools do not follow through. Hence it was not possible to compare third grade performance of students in NHP schools with performance of students in their matched comparison schools or of students nationally. Similarly, it is possible that the Grade Four Literacy Test is designed to measure skills addressed by an intervention targeting lower performing schools. However, this test was first administered mid-way through the NHP implementation and the psychometric properties of the test are not known.

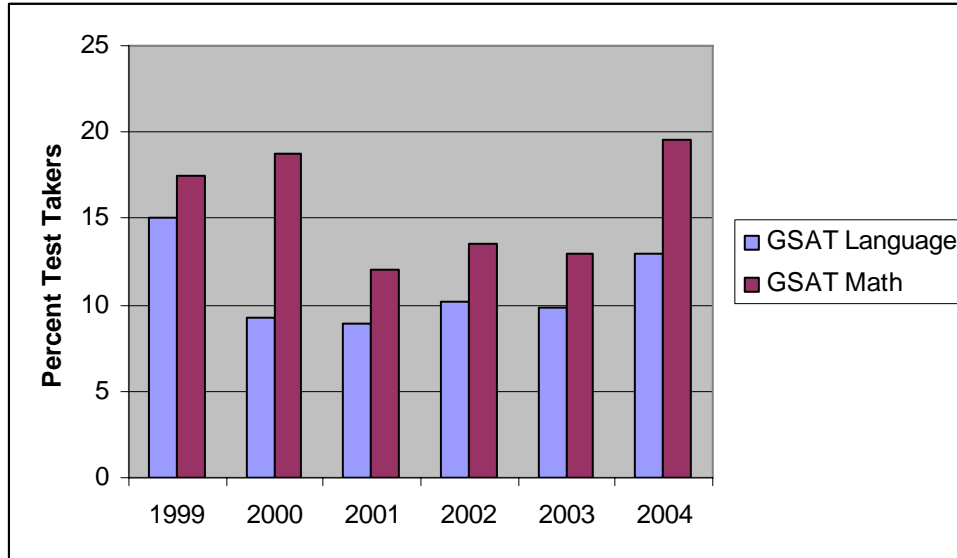
As might be expected, annual variations in the composition of GSAT test takers (e.g., absence of those who had skipped Grade 6, repeaters, etc.) yielded overall differences in average performance on the test from one year to the next. As was discussed earlier, during the implementation of the New Horizons Program, average performance increased initially and then declined. These trends, rather than indicating actual changes in student learning, largely reflected changes in the student population taking the exam.

Annual variations also present challenges for those who wish to interpret annual mean score changes. The MOEYC Student Assessment Unit (SAU) is cognizant of the importance of test equating and, when it does not defeat the tests’ purpose and resources are available, it employs techniques to horizontally equate its high volume tests. The SAU has clearly defined test specifications, pretest items to calibrate item difficulty, and target difficulty levels for items and tests. When it chose to add more difficult items to the test in 2002, it appears that moderately difficult items were replaced with very difficult items. As a consequence, the distribution of easier items was similar across tests. This was an astute decision and one not usually made by less sophisticated test developers. Nonetheless, the SAU relies primarily on classical test theory for their equating and they would benefit from more training and support in item response theory and its applications.

At the same time, we found substantial variations in the share of students reporting scores who scored at or below the “chance level,” or what they would have achieved by simply guessing (Figure

3). In 1999, more than 15 percent of GSAT test takers scored at or below “chance” levels. This share dropped sharply in 2000 for GSAT language arts and in 2001 for GSAT mathematics and remained at a lower level until 2004 when the share increased again. The variation reflects both changes in the test difficulty and changes in the test taker population caused by the phasing out of the CEE.

Figure 3 Percent of GSAT test takers scoring at or below chance in mathematics and language arts, 1999-2004



Literacy Effects on Mathematics Performance

Literacy, as assessed through the GSAT language arts test, is a powerful predictor of performance on the GSAT mathematics test. Assessment of numeracy, independent of literacy, is possible for non-verbal computation skills. However, such an assessment may not represent the full range of skills taught through the curriculum or intended to be measured. Word problems typically form a large share of mathematics assessments, and it is expected that the GSAT is no different in this respect. This is especially true if the GSAT has been adjusted to reflect the newer curriculum objectives in mathematics that emphasize problem solving and mathematics applications rather than simply mathematics computation. We were not able, however, to directly examine the GSAT and assess the language demands of the items and its use of word problems. Further, although we had subscores for each of the multiple choice subject tests (mathematics, language arts, science and social studies), we did not have information on which items formed the subscales within each subject test. This made it impossible to investigate possible mastery of subskills within a subject area.

Initially, we anticipated receiving from the MOEYC copies of the GSAT item questions and response alternatives so that these documents could be content analyzed to address the question of how literacy affects mathematics performance. Because the GSAT questions were not released to the evaluation team, we were not able to carry out this analysis. The Student Assessment Unit had begun such an analysis and expressed interest in pursuing funding to continue their efforts. However, we did explore the question in two ways. First we examined individual performance on GSAT mathematics as a function of GSAT language arts and prior school GSAT achievement for all GSAT test takers. Second, we explored the question through an analysis of the determinants of 2004 GSAT scores at the school level, as described in Section 2 for NHP, and matched non-NHP schools. In this section we discuss these findings, in lieu of the analysis we originally intended to carry out.

At the individual level across all GSAT test takers, the correlation between GSAT language arts and GSAT mathematics is very high: .84 for 2004. Controlling for school quality, as indicated by school average GSAT scores in 1999, does little to alter this relationship (table 8). For every point on the GSAT language arts test, the student's score on the GSAT mathematics test increases by .84 points. The school average GSAT language arts score in 1999 is, surprisingly, negatively related to individual achievement in 2004. Although the coefficient is significant, it is relatively small, suggesting a regression to the mean at the higher end of the continuum. The school average mathematics score in 1999, however, is positively related to individual achievement in 2004, and the coefficient is similar to the previously reported.

Table 8. Individual GSAT mathematics score in 2004 as a function of individual GSAT language arts score in 2004 and school average GSAT mathematics and language arts scores in 1999

Dependent Variable	GSAT Mathematics 2004		
	Coefficient (standard error)	t	Sig.
Individual GSAT language score 2004	.8414 (.0026)	322.78	.000
School average GSAT language score 1999	-.0855 (.0142)	-6.016	.000
School average GSAT math score 1999	.2546 (.0173)	14.57	.000
Constant	-2.6020 (0.2287)	-9.0126	.000
R-square	0.7227		
Adjusted R-square	0.7227		
Cases	43454		

At the school level for NHP schools and matched non-NHP schools, the mean school average 1999 GSAT language arts score is also included as a predictor of the school average GSAT mathematics score in 2004 (see Table 4). The language arts score in 1999 was a strong predictor of mathematics score in 2004, even when the mathematics score in 1999 was statistically controlled. For every point on the 1999 language arts test, the school average mathematics test score improved by .13 points in 2004.⁵ The effect was half the size of the effect of 1999 mathematics score and nearly as statistically significant: For every point on the 1999 language arts test, the school average mathematics test score improved by .14 points in 2004 compared with an improvement of .24 for every point on the 1999 mathematics test.

We simulate what the effect would be of boosting literacy to the “near mastery” level on mathematics performance (Table 8). Improving the school average language arts scores of the students in NHP schools in 1999 to “near mastery” would have raised the school average mathematics scores by one full point on the mathematics test in 2004. This would not, however, have

⁵ Caution must be taken in interpreting these coefficients, as the student taking the tests come from different cohorts and – for reasons discussed in this section – may not be similar at the two points in time. Moreover, the tests themselves are not fully equated.

been sufficient to raise the mathematics scores to “near mastery” and the effect is not nearly as great as raising the school mean mathematics scores to “near mastery.”

Table 9. Estimated effect on 2004 GSAT scores of improving 1999 GSAT from NHP school means to “near mastery” levels (simulation)

Simulation	Actual NHP School Average 2004 GSAT Math	Estimated NHP School Average 2004 GSAT Math
1999 GSAT Math increased to “near mastery” from actual	28	31
1999 GSAT Language arts increased to “near mastery” from actual	33	34

Source: Table 4

Conclusion on Effective Measures of Performance

We conclude that the GSAT was not an effective measure of the impact of the NHP on student learning in mathematics and language arts for two main reasons: (a) imperfect equating and (b) insensitivity to modest achievement gains by lower performing students.

With respect to imperfect equating, annual variations in the composition of GSAT test takers (e.g., absence of those who had skipped Grade 6, repeaters, etc.) and changes in the difficulty of test items and specifications yielded overall differences in average performance on the test from one year to the next. As was discussed earlier, during the implementation of the New Horizons Program average performance increased initially and then declined. These trends, rather than indicating actual changes in student learning, largely reflected changes in the student population taking the exam as well as variations in the item difficulties represented in the test. Although the Student Assessment Unit of the MOEYC works to equate its high volume tests and render comparable results from one year to the next, sometimes the need to adapt the test takes precedence.

With respect to insensitivity to modest achievement gains by low performing students, the test is sensitive at the higher performance level. In order to serve its high stakes purpose the test must cover a broad range of skills and largely focus on the skills needed for secondary school success. The GSAT was never intended to cover beginning literacy and numeracy skills. Consequently, it is not surprising that many students are scoring at a level suggesting that they guessed at the majority of the items and that the test is not adequately sensitive to changes in achievement of lower performing students.

Section 4. Data Considerations

Two questions dealing with data were posed by USAID:

1. How effective were the indicators used for tracking the results of the NHP and what suggestions could be made for the future?
2. How effective are the methodologies used to collect data?

Indicators for Tracking Results

The results of the NHP were intended to be improved literacy and numeracy in schools participating in the program, which provided schools a menu of ten interventions from which they could choose to help achieve results. Data on how schools picked from the list of ten potential interventions and how

effective the school was in the implementation of these interventions were not available and may not have been collected. The only Jamaica-wide gauge available to measure mathematics and language arts performance was the GSAT, whose limitations for measuring growth in the lower performing schools have been discussed in Section 3.

In addition, while anonymous GSAT data were available at the individual student level, other useful demographic information (other than birthdate) was not available.⁶ Examples of such data for individual students include data on gender, family socio-economic status, past academic performance, mobility among various schools, school attendance, grade repetition, and special needs status. High quality, individual student records can be extremely valuable in controlling for exogenous factors affecting student performance and in explaining the effectiveness of educational programs. More importantly, availability of past performance data for individual students allows for measurement of individual student growth in achievement over time, which is essential for assessing program impact.

Data that permit statistical controls to adjust for exogenous variables and endogeneity greatly increase the rigor of the analytic methods that can be used to assess program effects. The implementation of the Student Information System piloted by the NHP schools and currently being implemented in some additional Jamaica schools could help enable future evaluations with access to more robust student-level data.

Data Collection Methodologies

The team encountered four important challenges in working with data to conduct this evaluation: (a) lack of data integration, (b) lack of codebooks adequately describing the School Census in various years, (c) complexities of the teacher files, and (d) inconsistent documentation regarding the NHP schools and the qualified pool of schools from which the program participants were selected. These have implications for future data collection methodologies.

One challenge was that the data were not housed in one location nor were the various data sets integrated. Some data were located with the MOEYC, while other data were housed with a Jamaican data processing organization. The lack of data integration required the evaluation team to conduct considerable work identifying common linking variables and matching different types of data files to conduct the detailed statistical evaluation required.

This integration and matching effort was complicated because the Ministry was unable to provide a detailed codebook for any of the School Census files, although copies of the survey document were provided for the 2003 School Census. We found that the School Census Survey differed each year, 1999 to 2003, in the actual number of variables, type of variable, and the location of these variables in the data record. We were advised that the data file generally followed the survey but that some of the data elements were different. We were provided with screen shots of some of these differences. Based upon the 2003 School Census Survey instrument, the screen shots, and some investigative work, such as running frequencies and cross-tabulations on variables whose definitions were uncertain, we were able to clarify most of the over 170 school level variables included in the annual School Census file. One notable exception was the use of computers in the schools, including teacher training for computer use, which we were unable to identify in any of the School Census files (although the questions appeared on the survey). Increased use of computers for administrative and

⁶ Individual student identity was ensured by removing all individual identifying information for all student level records

educational purposes is one of the key interventions of the NHP program and, because Jamaica-wide information was not available on this, we were unable to include this variable in our analysis. The Ministry would benefit from the development of a code book for each file they maintain. Such a code book should explain what each grouping of data represents and the metadata about that data.

Teacher files were complex and some of the variables describing teachers were inconsistent across years. The Annual School Census survey gathers information about teachers and staff in the school. This file contains approximately 50 variables and is different in that it is a file at the individual teacher level. Thus, teachers are nested within schools, which adds to the complexity of managing the data files. In addition, since definitions of teacher variables varied over years, the absence of a codebook hampered our work.

We found inconsistencies in the identification of NHP program schools. Two documents described the NHP schools and the original 194 schools from which these 72 NHP schools would be selected. We found small inconsistencies in what should have been straight forward matches and generation of matching code. In some cases, schools identified as participating in the NHP were not actual participants, and in other cases, school IDs were inconsistently used. In measuring program effects with small samples, it is critical to be as precise as possible. The discrepancies were found may reflect a need for better documentation or more available documentation on actual program implementation or the documentation of modifications to original plans.

Conclusion on the Effectiveness of Data Collection Methodologies

Effectiveness of data collection methodologies could be improved by enhanced documentation of codebooks, full documentation of data integration, and a system-wide EMIS similar to the Jamaica School Administrative System that was implemented in the NHP schools. We understand that the MOEYC is aware of these issues and is moving forward with wider application of the EMIS, beginning with 200 additional schools.

Section 5. General Conclusion and Recommendations

With the support of USAID, the New Horizons for Primary Schools program was initiated in 1998 and rolled out in 1999 to improve the quality of teaching in order to raise literacy and numeracy at the primary level, improve school attendance and strengthen school management. The first cohort of students who attended NHP schools for all or most of Grades 1-6 completed Grade 6 in 2004. This report examines the effect of the NHP on the learning achievement of those students and addresses eight questions posed by USAID Jamaica.

1. Have NHP schools made achievement gains over the years under review?
2. What factors in the project schools may have affected gains or lack of gains?
3. Is the use of mastery/near-mastery/non-mastery categories on the GSAT masking real gains in student achievement in schools?
4. Is the GSAT the best measure of student performance for the project schools, considering that it is based on the content delivery system of the old curriculum?
5. How can valid measures of students' computational skills in numeracy be assessed for students who are unable to comprehend the language in which most numeracy items are couched in the GSAT examination?
6. How has "social promotion" to Grade 6 affected average performance results among students?
7. How effective were the indicators used for tracking the results of the NHP and what suggestions could be made for the future?
8. How effective are the methodologies used to collect data?

Gains in Achievement

Compared with schools not in the NHP program, schools in NHP showed higher performance at the lower ends of two tests of achievement measured at Grade 6: writing (GSAT Communications Task I) and mathematics. Average Communications Task I scores were higher in NHP schools than in matched non-NHP schools in 2004, and in 2004 the share of NHP schools with average GSAT mathematics scores of 30 or more points (out of a possible 80) was greater than the share of non-NHP schools with average GSAT mathematics scores of 30 or more points. In both cases, the range of improvement occurred below the levels designated “near mastery.”

Factors Affecting Gains in Achievement

Factors associated with higher literacy and numeracy in 2004 included the presence of a good quality School Improvement Plan (School Development Plan) in 2003, an active PTA, and higher performance in 1999. Students in schools with more qualified and more experienced teachers in Grades 1-6 scored higher than those with less experienced and less qualified teachers. Students in schools on multiple shifts achieved less than those on single shift, while students in rural schools also scored less well than those in urban schools. Factors unrelated to 2004 GSAT performance were presence of a school board or SCOPE program, presence of a breakfast program, school size and remote rural location.

Use of Mastery/Near-mastery/Non-mastery Categories

The use of mastery/near-mastery/non-mastery categories on the GSAT hides most observable gains. Nearly two-thirds of students taking the GSAT score below the near mastery cut-off score (50 percent correct) on the key achievement tests used to evaluate NHP: mathematics and language arts. Since the NHP schools were selected from among those in which the average school achievement was below the national average, the average student performance in these schools is well below near-mastery. A test that included more comparatively easy items would be more sensitive to change than a harder test, particularly when results are aggregated into such broad “mastery” categories. We found that the reliability and discrimination of hypothetical subscales based on “easier” items was acceptable. However, analyses of reliability, discrimination and differential item functioning based on actual GSAT subscales, which would provide more information regarding sensitivity of the GSAT at the lower levels, were constrained by lack of access to actual test questions and response options and to information on subscale composition.

The GSAT Curriculum

The GSAT was designed originally to reflect the 1980 national curriculum and to assess skills thought to be necessary for secondary school success. Unlike its predecessor the CEE, it is curriculum based and covers the major elements of the upper primary curriculum: Mathematics, Language Arts, Science, Social Studies and Communication (writing). With curriculum reform in the last decade, the test has been adjusted to reflect changes in emphasis. Primarily the differences are in the topics covered in Science and Social Studies. Fundamentally the basic structure of the test remains unchanged.

Assessing Numeracy net of Literacy

Using the GSAT to assess student gains in mathematics, independent of students’ language skills, is possible, in two ways. First, scores on mathematics could be statistically controlled for language arts performance, in multivariate analyses. We adopted this approach at both the individual student and school level, and found strong correlations between achievement on GSAT language arts and GSAT mathematics at both levels. This suggests that the GSAT mathematics test has a strong verbal component, but may also indicate underlying skills common to performance on both tests.

Second, GSAT mathematics items could be analyzed for verbal content, and those items lacking high verbal content could be selected for analysis. We were unable to do this for the 2004 GSAT, as the MOEYC would not grant us permission to review the item questions (stems and response options). Apparently, such an analysis had been initiated in 2003; however, the investigation was not completed because of other priorities within the Student Assessment Unit and limited resources to conduct the workshops needed for item classification. This analysis could be undertaken by the Student Assessment Unit of the MOEYC, serving a dual purpose: investigating the role of language in mathematics performance and building capacity within the MOEYC to conduct and utilize data from this kind of study.

Social Promotion

The data provide little evidence of “social promotion” in primary schools in Jamaica. If “social promotion” had been in place in the early 2000s, we would have expected to see a slight decline in the share of older students in Grade 6, as they would not have been held back in earlier grades. Instead, the share of students reaching Grade 6 at a slightly older age increased from 2000 to 2004. There are two possible explanations. First, introduction of the Grade 4 literacy test in 2000 may have resulted in students repeating that grade.

Second, the elimination of the CEE 11+ in 1999 also eliminated the possibility of students advancing to Grade 7 prior to completing Grade 6. In 2000 and 2001, years in which the GSAT test takers were slightly older, some students who were successful on the CEE 11+ in 1998 or 1999 had already advanced to Grade 7, and were therefore not included in the GSAT populations.

NHP Indicators

The indicators used in tracking the NHP were ineffective in two ways. First, the impact of the program should have been monitored in earlier grades and through tests that were not “high stakes.” Collection of Grade 3 Diagnostic test results from all schools in Jamaica would have enabled a more robust analysis of the effects of the program. Second, indicators of NHP implementation (and the implementation of similar interventions in non-NHP schools) were not collected.

Data Methodologies

While the data systems that were developed for NHP have many positive features, rigorous evaluations would require comparable information to be available across a set of comparison schools, if not for all Jamaica. In addition, use of the data would be facilitated by the preparation and dissemination of comprehensive codebooks for all data sets, on an annual basis. Finally, the MOEYC should be encouraged to establish unique codes for all schools, and to discontinue the practice of “recycling” school codes, which leads to confusion in the use of school level data.

Recommendations

On the basis of the analyses in this report, we recommend the following for improving future evaluation designs and processes:

- Comparison schools are essential and should be identified at the outset and monitored simultaneously with NHP program schools.
- Data collection should include indicators that assess all main program objectives, including: achievement, attendance and school management.
- Indicators should be collected for all project schools and matched comparison schools.
- If subsets or samples of project schools are included for special evaluations, the subsets or samples should remain the same over time, to monitor trends.
- Data need to be reliably collected, aggregated and reported centrally.

- Multivariate and hierarchical linear modeling techniques should be used for analysis purposes.
- We also recommend the following for evaluation indicators.
- Monitoring of NHP implementation at the school level is essential.
- Third grade achievement test data should be collected nationally (or, minimally for all NHP and matched comparison schools) and reported centrally to allow for earlier assessment of impact.
- Performance scores should be adjusted to correct for annual imprecision in test equating.
- Comparisons should use actual scores rather than collapsing scores to “mastery levels.”

Recommendations for next steps

In order to gain a better understanding of the impact of NHP on student achievement and to the assess adequacy of alternative measures for monitoring future reform efforts, we recommend a series of further data collection and analysis activities:

- Collect from all NHP and a matched set of non-NHP schools the results from the 2004 (and possibly 2005) Grade 3 Diagnostic and Grade 4 Literacy tests.
- Evaluate the adequacy of the Grade 3 Diagnostic and Grade 4 Literacy tests as indicators for monitoring and evaluating program impact.⁷
- Survey all NHP and a matched set of non-NHP schools to identify program inputs that may boost literacy and numeracy.
- Analyze the resulting data using multivariate, including HLM, statistical techniques.

⁷ This assessment would include looking specifically at (a) procedures for test administration, (b) test development and item security, (c) targeted skills, (d) data management, (e) equating and (f) psychometric analyses of recent administrations based on convenience samples. It would also review MOEYC plans for continuing use of these two tests.

References

- Dye, Richard, Joan Jennings, Clement Lambert, Barbara Hunt and Gerald Wein (July 2002). *Evaluation and Recommendations for Strengthening and Extending the New Horizons for Primary Schools Project in Jamaica*. Washington, DC: Aguirre International.
- Enge, Kjell and Heather Simpson (October 2003). *New Horizons for Primary Schools/Jamaica: Formative Evaluation 2003*. Los Angeles, CA: Juarez and Associates, Inc.
- Juarez and Associates, Inc. (June 1999). *New Horizons for Primary Schools: Diagnostic Report on Project Schools*. Los Angeles, Ca: Juarez and Associates, Inc.
- Juarez and Associates, Inc. (December 1999). *New Horizons for Primary Schools: Report on the Status of Numeracy and Literacy Programs in Individual Schools*. Los Angeles, CA: Juarez and Associates, Inc.
- Ministry of Education, Youth and Culture, Planning and Development Division, Statistics Unit (October 2003). *Annual School Census Questionnaire 2003/2004: Public Schools*. Kingston, Jamaica: GOJ MOEYC
- New Horizons Activity Project Implementation Unit (December 1998). *Final Report: Project Schools Selection, New Horizons Activity*.
- No Author (Spring 2003). *Summary Evaluation of School Improvement Plans (SIP) of NHP Schools*.
- O'Neil, Ernest (October 2003). *New Horizons for Primary Schools: Special Report to Private Sector Organizations of Jamaica*. Kingston, Jamaica: New Horizons for Primary Schools
- Russell, Fitz-Albert R. (1996). *Issues of Validity Related to the Jamaican Grade Six Achievement Test*. Unpublished doctoral dissertation, Florida State University, Tallahassee.

Annex A: Propensity Score Matching

Previous evaluations of the New Horizons for Primary Schools program (NHP) have collected data on NHP schools only or have used small samples of NHP and comparison schools. In this study we use propensity-score matching techniques to create a set of schools that match, on a one-to-one basis, the schools in the NHP program. We then compare the NHP schools with these matched non-NHP schools to test for effects.

Propensity score matching

Propensity score matching is utilized to compensate for the absence of a pure experimental design, whereby treatment and control groups are established *a priori*. It helps correct for any “self-selection” that may have occurred in the identification of the NHP schools, as well as creates a comparable non-treatment group against which the NHP may reasonably be compared.

In order to establish predicted probabilities for schools to participate in the NHP program, we employed a logit regression with a bivariate variable indicating the school’s participation in the program as the dependent variable. We examined three sets of variables related to participation in the NHP program: (a) characteristics of the school in 1999, (b) characteristics of the teachers in the school in 1999, and (c) achievement of the students in the school in 1999. We used 1999 school year data, instead of actual pre-program data, as they were the first available for all government schools with primary (grade 1-6) sections. The most important predictor of being in the program was having been designated in the list of 194 schools that were “qualified” for the program as per *Final Report: Project Schools Selection, New Horizons Activity* (Project Implementation Unit, December 1998), followed by the size of the school, an indicator of the poverty level of the community (breakfast program) and presence of a master teacher. The results of the logit regression are presented in table A.1

Table A1. Logit regression predicting participation in NHP program (N = 791)*

	B	S.E.	Significance
Remote_Rural_Location	0.137	0.552	0.804
Rural_Location	-0.004	0.427	0.993
Boarandscope	-0.788	0.539	0.144
PTAexist	0.430	1.067	0.687
breakfastYES	0.780	0.477	0.102
Shift	0.694	0.695	0.318
teachergrade	0.166	0.074	0.026
TotalEnrollment1to6	-0.003	0.002	0.086
MathScr_mean1999	-0.064	0.068	0.348
Orginal195(1)	-6.113	1.066	0.000
LangScr_mean1999	-0.008	0.062	0.897
Tsk1_mean1999	0.306	0.651	0.639
Tsk2_mean1999	0.044	0.797	0.956
YrsOfServicegrade1to6	0.002	0.041	0.962
YrsInSchoolgrade1to6	0.002	0.025	0.923
CXC (highest level of teacher training) Percent	2.960	2.119	0.162
Certificate_percent	1.343	1.909	0.482
Master teacher percent	0.077	0.207	0.710
Constant	-1.570	2.761	0.570

*Through a series of regressions these indicators proved to be the most predictive

1. Remote_Rural_Location: Schools were categorized into three geographic locations Urban, Rural and Remote Rural. Remote Rural and Rural were identified as predictive
2. Rural_Location - see above
3. Boarandscope : a combination variable for a school that has a school board and or a SCOPE program present
4. PTAexist: variable for the existence of a functioning PTA at the school.
5. breakfastYES: Variable for a functioning breakfast program
6. Shift: Variable for schools on a shift program
7. teachergrade: Variable for teachers at each grade level
8. TotalEnrollment1to6: variable for total enrollment in grade 1 to 6. This was to try to control for schools that had grades that were different from grade 1-6, such as schools with grades 1-8
9. MathScr_mean1999: variable for the Mean Grade Six Achievement Test (GSAT) in 1999
10. Original195(1) : variable for the pool of schools that the 72 NHP schools were supposed to be selected from
11. LangScr_mean1999: variable for mean GSAT language score by school
12. Tsk1_mean1999: variable for the mean GSAT task 1 (a performance based language part of the GSAT)
13. Tsk2_mean1999: variable for the mean GSAT task 2 (a performance based language part of the GSAT)
14. YrsOfServicegrade1to6: variable for years of teaching experience of teachers in grade 1-6.
15. YrsInSchoolgrade1to6 : variable for years of teaching experience of teachers in grade 1-6 in this school.
16. CXC_Percent: variable for the percent of teachers who's highest degree was a high school degree
17. Certificate_percent:: variable for teachers who had a teaching certificate.
18. Master_teacher_percent: variable for schools that had at least one teacher that was classified as a master teacher
19. Constant

Propensities

In order to try to control for differences we used propensity score matching (PSM) to identify 72 "comparison schools" that were similar to the 72 NHP schools being studied at the program start year of 1999. Supposedly, the NHP schools would then be the treatment group and the PSM schools could act as a comparison group. Such efforts are not perfect. Table A2 shows how close the matches were: very close in some cases and not so close in others. This is an effort to be transparent and accurate. We only have 70 schools here as two schools, which came into the program a little later and did not have key data from the 1999 GSAT assessment, did not generate a match using PSM.

The propensities to participate in the programs predicted by the first stage logit regression were then used to match non-participating schools to the participating schools. We were able to match 70 of the 72 NHP schools (97 percent) to a non-participating school using "nearest neighbor" matching. The difference in propensity score was overall very small, with two-third of the matches the same up to two decimal points or better. The average difference across all pairs was 0.08 (Table A2).

Table A2. Propensity Scores, 70 pairs of schools, Jamaica 2004

Matched Pair	Non-NHP	NHP	Delta		Matched Pair	Non-NHP	NHP	Delta
1	0.5042513	0.5119474	0.0076961		36	0.2611024	0.2613415	0.000239
2	0.4172761	0.570651	0.153375		37	0.2678804	0.2698231	0.0019427
3	0.3520363	0.3514573	0.0005791		38	0.5679271	0.524	0.0439271
4	0.6025884	0.5326875	0.0699009		39	0.2720227	0.2711854	0.0008373
5	0.241238	0.2410233	0.0002147		40	0.2561346	0.25502	0.0011146
6	0.6510356	0.5416208	0.1094148		41	0.122819	0.1150471	0.0077719
7	0.6586419	0.5699107	0.0887312		42	0.3647079	0.8731876	0.5084797
8	0.4967678	0.4928689	0.0038989		43	0.2909035	0.291495	0.0005915
9	0.462168	0.4529585	0.0092095		44	0.340772	0.3409059	0.0001339
10	0.3323626	0.3315809	0.0007817		45	0.3785678	0.7521022	0.3735345
11	0.349836	0.3476235	0.0022125		46	0.3685919	0.8467132	0.4781213
12	0.3085159	0.3057116	0.0028042		47	0.3715102	0.7930032	0.421493
13	0.3566964	0.3569009	0.0002045		48	0.2732814	0.2771164	0.0038351
14	0.4503066	0.448539	0.0017677		49	0.3466962	0.344105	0.0025912
15	0.6066679	0.5348319	0.071836		50	0.2223232	0.223047	0.0007238
16	0.3210153	0.3168356	0.0041797		51	0.0029482	0.0029779	0.0000297
17	0.3215854	0.3187062	0.0028792		52	0.4369853	0.5381885	0.1012032
18	0.503142	0.5214987	0.0183567		53	0.4203101	0.4217206	0.0014105
19	0.0961448	0.0959445	0.0002003		54	0.3940174	0.6963069	0.3022895
20	0.5132916	0.5115588	0.0017327		55	0.379759	0.7477385	0.3679795
21	0.4708211	0.4802563	0.0094352		56	0.3898405	0.707387	0.3175465
22	0.4010899	0.6080119	0.2069221		57	0.2299804	0.229548	0.0004324
23	0.2735019	0.2767505	0.0032487		58	0.4650572	0.5284258	0.0633686
24	0.3281903	0.3274532	0.0007371		59	0.3771087	0.3771166	7.82E-06
25	0.3980607	0.6616915	0.2636308		60	0.3353627	0.3342119	0.0011508
26	0.2927829	0.2932756	0.0004927		61	0.1355779	0.1312219	0.0043561
27	0.4105894	0.4061338	0.0044556		62	0.3811791	0.3846298	0.0034507
28	0.4300269	0.5389318	0.108905		63	0.3000524	0.2986467	0.0014056
29	0.3639371	0.8834285	0.5194915		64	0.1548092	0.1548988	8.955E-05
30	0.4000816	0.6498142	0.2497326		65	0.3132456	0.3072834	0.0059622
31	0.5007799	0.4966614	0.0041186		66	0.4005574	0.62211	0.2215526
32	0.4626283	0.5330533	0.070425		67	0.4448428	0.4460456	0.0012028
33	0.4129256	0.4145678	0.0016423		68	0.4050324	0.4045277	0.0005047
34	0.1476972	0.149547	0.0018499		69	0.416801	0.5961867	0.1793857
35	0.3856446	0.3845955	0.0010491		70	0.9487294	0.6845186	0.2642108

New Horizon schools compared with matched non-New Horizon schools

The two groups of schools – NHP schools and matched non-NHP schools – were very similar in 1999, in most observed respects. However, on average NHP schools are 20 percent larger than matched non-NHP schools and are twice as likely to be on a multiple shift; 10 percent more NHP schools are urban as compared with matched non-NHP schools. Five percent more NHP schools have breakfast programs. This table shows the gross agreement between the NHP schools and the PSM comparison schools to try to be transparent and accurate. Most of the agreements were quite close.

However, we were only able to run this information against 70 schools because of the missing 1999 GSAT scores for two schools. In our follow on study, we estimated a score for these schools using scores from later GSAT assessments.

Table A3. NHP schools compared with matched non-NHP schools, various characteristics.

	NHP (N= 70)	Non-NHP (N = 70)
School Characteristics 1999		
Size: Enrollment in Grades 1-6	387	321
Size: Number of teachers in Grades 1-6	13.00	10.57
School Shift (percent)	13	7
Rural location (percent)	43	51
Remote rural location (percent)	20	21
Breakfast Program (percent)	19	14
Active PTA (percent)	97	94
Board and or Scope (percent)	87	91
On list of initially qualified schools (percent)	99	99
Teacher Characteristics 1999		
Qualifications: CXC highest (percent)	28	27
Qualifications: Certificate highest (percent)	67	68
Qualifications: Master teacher in school (percent)	39	39
Experience: Mean years of service at grade 1 to 6	15.13	15.27
Experience: Mean years of service at grade 1 to 6 in school	10.45	10.73
Experience: Percent of teachers in school with two or less years experience	22	25
Master Teacher in school (percent)	87	87
Student Achievement (GSAT school means) 1999		
Mathematics 1999	28.36	28.20
Science 1999	22.95	22.71
Social Studies 1999	33.99	33.57
Language Arts 1999	33.04	32.52
Communications Task 1 1999	2.37	2.34
Communications Task 2 1999	1.34	1.31
Student Achievement (GSAT school means) 2004		
Mathematics 2004	29.33	29.19
Science 2004	23.83	23.53
Social Studies 2004	34.51	34.28
Language Arts 2004	32.60	32.94
Communications Task 1 2004	2.87	2.74
Communications Task 2 2004	3.18	2.79

Annex B: Matched-pair t-tests

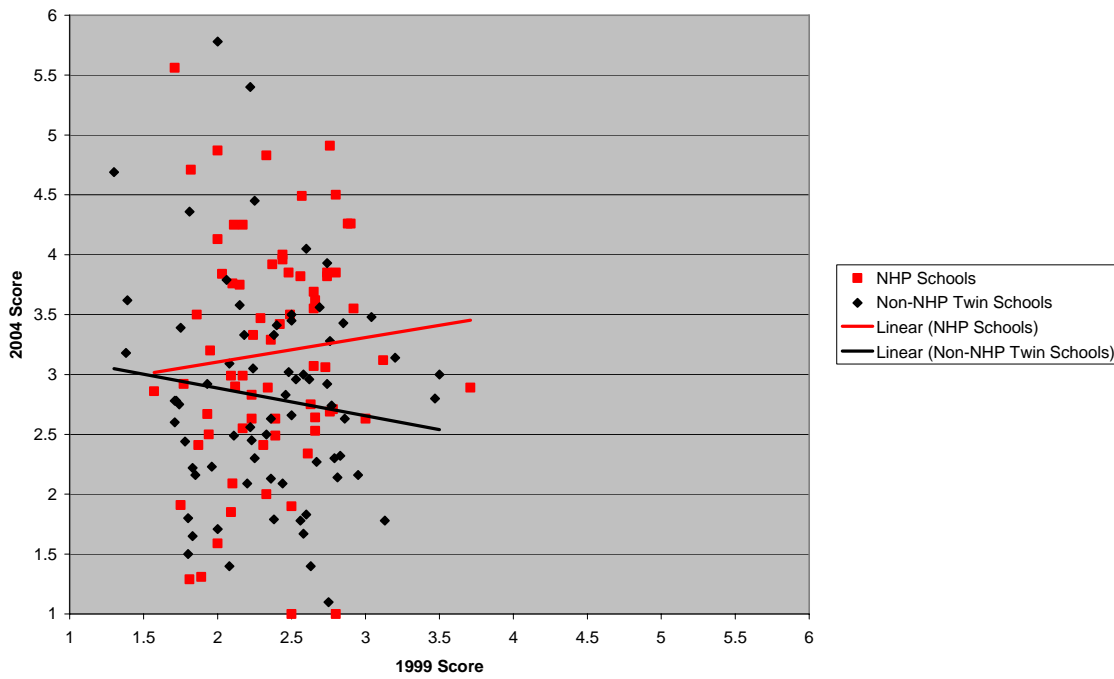
	Paired Differences					t	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference			
1999							
Mathematics	0.72948	5.93770	0.70969	-0.68632	2.14527	1.028	0.308
Science	0.26177	4.50051	0.53791	-0.81134	1.33487	0.487	0.628
Social Studies	0.58154	7.12427	0.85151	-1.11718	2.28027	0.683	0.497
Language arts	0.24470	6.43905	0.76961	-1.29063	1.78004	0.318	0.751
Communications Task I	0.02916	0.55211	0.06599	-0.10249	0.16080	0.442	0.660
Communications Task II	-0.00061	0.45083	0.05388	-0.10810	0.10689	-0.011	0.991
2000							
Mathematics	-0.55459	8.59018	1.02672	-2.60285	1.49366	-0.540	0.591
Science	-0.18564	6.84579	0.81823	-1.81796	1.44668	-0.227	0.821
Social Studies	-0.37767	9.97244	1.19194	-2.75551	2.00018	-0.317	0.752
Language Arts	-0.56014	8.63086	1.03158	-2.61810	1.49781	-0.543	0.589
Communications Task I	-0.07413	1.56290	0.18680	-0.44679	0.29853	-0.397	0.693
Communications Task II	-0.02159	0.67880	0.08113	-0.18345	0.14026	-0.266	0.791
2001							
Mathematics	0.94832	7.57391	0.90526	-0.857	2.75425	1.048	0.298

				62			
Science	0.91268	5.98222	0.71501	0.513 73	2.33909	1.27 6	0.20 6
Social Studies	0.64815	9.34531	1.11698	1.580 16	2.87646	0.58 0	0.56 4
Language Arts	1.87063	7.96543	0.95205	0.028 66	3.76992	1.96 5	0.05 3
Communications Task I	0.02722	0.91213	0.10902	0.190 27	0.24471	0.25 0	0.80 4
Communications Task II	0.07163	0.70583	0.08436	0.096 67	0.23993	0.84 9	0.39 9
2002							
Mathematics	0.71774	8.82749	1.05509	1.387 11	2.82258	0.68 0	0.49 9
Science	0.83691	5.66040	0.67655	0.512 77	2.18659	1.23 7	0.22 0
Social Studies	-0.23295	8.48650	1.01433	2.256 49	1.79058	0.23 0	0.81 9
Language Arts	0.60640	7.49033	0.89527	1.179 61	2.39240	0.67 7	0.50 0
Communications Task I	0.12098	1.08440	0.12961	0.137 59	0.37955	0.93 3	0.35 4
Communications Task II	-0.02482	0.67560	0.08075	0.185 91	0.13627	0.30 7	0.75 9
2003							
Mathematics	1.40608	7.77809	0.92966	0.448 54	3.26070	1.51 2	0.13 5
Science	1.06217	5.52374	0.66021	0.254 92	2.37926	1.60 9	0.11 2
Social Studies	0.79051	9.26840	1.10779	1.419 46	3.00048	0.71 4	0.47 8
Language Arts	0.65879	7.62605	0.91149	1.159	2.47716	0.72 3	0.47 2

				57			
Communications Task I	-0.01447	0.77600	0.09275	0.19950	0.17056	0.156	0.876
Communications Task II	-0.01729	0.68912	0.08237	0.18160	0.14703	0.210	0.834
2004							
Mathematics	-0.14194	8.21766	0.98220	2.10138	1.81749	0.145	0.886
Science	-0.29843	5.75582	0.68795	1.67086	1.07399	0.434	0.666
Social Studies	-0.22290	8.50060	1.01602	2.24980	1.80400	0.219	0.827
Language Arts	0.33563	8.49021	1.01477	1.68879	2.36005	0.331	0.742
Communications Task I	-0.39151	1.25750	0.15030	0.69135	0.09167	2.605	0.011
Communications Task II	-0.13329	0.88740	0.10606	0.34488	0.07830	1.257	0.213

Annex C: Scatter Plots, NHP and matched non-NHP schools, 1999 and 2004

Task 1 Mean Scores



Task 2 Mean Scores

