Textbooks and Test Scores: Evidence from a Prospective Evaluation in Kenya[1]

Work-in-Progress

Paul Glewwe, Michael Kremer, and Sylvie Moulin[2]

December 1, 1998

Abstract

Although there is intense debate about the effect of increased expenditure on education in developed countries, there is widespread consensus that provision of textbooks can substantially increase test scores in developing countries. This paper evaluates a program through which a Dutch non-profit organization provided textbooks to 25 rural Kenyan primary schools that were chosen randomly from a group of 100 candidate schools. After one school year, average test scores did not differ substantially between program and comparison schools. However, for those students in the top quintile of the distribution of initial academic achievement, the program raised test scores by at least 0.2 standard deviations.

[1] This project was a joint undertaking of many people. The data were collected and analyzed by Charles Asoka, Najy Benhassine, Marcos Chamon, Edward Drozd, Dan Levy, Irene Majale, Sean May, Ted Miguel, Robert Namunyu, Caroline Nekesa, Stacy Nemeroff, Jaypee Sevilla, Michael Wambetsa, Polycarp Waswa, and Maureen Wechuli. Invaluable assistance was provided by the staff of Internationaal Christelijk Steunfonds: Chip Bury, Jos Huizinga, Paul Lipeyah, Japheth Mahagwa, Rebecca Mbaisi, and Susan Walji. We are grateful to Joshua Angrist, Angus Deaton, Paul Gertler, Claudia Goldin, Eric Hanushek, Lawrence Katz, Marlaine Lockheed, Richard Murnane, and Steve Pischke for advice and comments. Finally, we are grateful to Mr. Munala, the District Education Officer of Busia, to Mr. Buluma, the inspector of primary schools in Busia, and to the headmasters, teachers, and students of the participating schools. The costs of this evaluation were covered by the National Science Foundation and the World Bank research committee. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the World Bank, its Executive Directors, or the countries they represent.

[2] Glewwe and Moulin, World Bank and Kremer, MIT, NBER and HIID.

In 1996, a Dutch non-profit organization, Internationaal Christelijk Steunfonds (ICS), began a program to assist one hundred rural Kenyan primary schools over four years. In the first year, 25 of these schools, selected randomly, received textbooks. This paper reports on the results of an evaluation of the effect of the program.

There is intense debate about the effect of education expenditure on test scores in developed countries [Hanushek, 1995; Hedges, Laine, and Greenwald, 1994; Card and Krueger, 1992]. However, even writers who are skeptical about the effects of educational inputs in developed countries are optimistic that provision of textbooks in developing countries can substantially increase test scores. Lockheed and Hanushek [1988] summarize studies of textbooks in developing countries; the four studies that they examine in detail report that textbooks improved test scores by 0.34, 0.36, 0.30 and 0.06 standard deviations of individual test scores. Other reviews of the literature cast a wider net, including some studies of perhaps lower quality. Heyneman, Farrell, and Sepulveda-Stuardo [1978] find that textbooks had a positive effect on academic achievement in 15 of 18 studies. They argue that studies typically indicate that textbooks have a greater effect on students from poor backgrounds. Fuller [1986] reports significant effects of textbooks in 14 of 22 studies. A more recent study by Fuller and Clarke [1994] found significant effects in 19 of 26 studies. Based in part on these studies, an influential World Bank volume on education recommends that textbook provision be given high priority in developing countries [Lockheed and Verspoor, 1991].

Unfortunately, most of the existing literature is based on retrospective studies and therefore potentially suffers from omitted variable bias. Retrospective studies will overestimate the effect of textbooks if schools with more textbooks typically have other, less easily observed, advantages, such as parents who are more committed to education. On the other hand, retrospective studies will underestimate the effect of textbooks if textbooks are provided to schools which are seen as particularly needy. For example, in 1994 the World Bank and the Jomo Kenyatta Foundation provided textbooks to Kenyan schools identified as being particularly disadvantaged.

A study in Nicaragua [Jamison, et al 1981] attempted to overcome the problem of omitted variable bias through a prospective evaluation. In the Nicaraguan study, 48 first-grade classrooms received radio mathematics education, 20 classrooms received mathematics workbooks, and 20 classrooms served as controls. Teachers in the schools that received assistance attended a three hour training session to learn how to use the materials. After one year, pupils who received workbooks scored one-third of a standard deviation higher than the control group. This difference was significant at the 1% level (t-statistic of 2.74).[3] Pupils who received radio education scored more than one standard deviation higher than the control group, a difference that was also highly significant. Jamison, et. al. [1981] find no significant interactions between pre-test scores and the effect of textbooks, but they do find that provision of textbooks narrowed gaps between rural and urban students.

Heyneman, Jamison and Montenegro [1983] examined a World Bank project in the Philippines which provided one textbook for every two students. For a group of randomly selected schools they provided textbooks at a one-to-one ratio. They found little difference in test scores between these schools and schools that received textbooks at a one-to-two ratio. In many of their comparisons, the schools that received one textbook for every two students actually performed better than the schools that received one textbook for every student. Heyneman, Jamison, and Montenegro also compared test scores in schools receiving textbooks at a one-to-two ratio with test scores in the previous year. They found an average effect over two grades and three subjects of 0.40 standard deviations, with students from poor families benefitting the most. They conclude that textbook provision greatly improves test scores, but that there is no advantage of a one-to-two ratio over a one-to-one ratio.

In contrast to previous studies, this study finds no evidence that provision of textbooks in Kenyan primary schools led to a large positive impact on average test scores after two years. We report estimates of the effect of the program on average test scores based on: i) differences between textbook and comparison schools in post-test scores; ii) differences between textbook and comparison schools in differences between pre-test scores and post-test scores; and iii) differences between textbook and comparison schools in differences between normalized test scores in subject-grade combinations which did and did not receive textbooks. The estimated average impact across all subjects and grades is close to zero, according to all three estimates, and, depending on the estimator, it is sufficiently precise estimate to reject the hypothesis that the textbooks increased test scores by 0.22, 0.14, or 0.07 standard deviations. However, textbook provision did improve the scores of students with pre-test scores in the top quintile of the distribution by 0.22 standard deviations. The actual benefit to the top quintile of students by

[3] Jamison et. al. [1981] used the classroom as the unit of observation in calculating the standard errors.

initial academic achievement is likely to be closer to a third of a standard deviation, because the pre-test was noisy, so that the top quintile of students on the pre-test is likely to include many students who were not in the top quintile of academic achievement.

The results do not appear to be statistical artifacts. The treatment and comparison schools were fairly similar in geographic location, enrollment, and test scores prior to the introduction of the program. Problems of selection and attrition bias do not appear to drive the results, and attempts to correct for these problems do not change the results appreciably. Many of the tests used during the first year of the evaluation were difficult, which may mask program effects among students with low initial academic achievement. However, when attention is restricted to the easier tests the estimated program effect on average scores remains small and statistically insignificant, and a second year of data using less difficult tests yields similar results.

While average test scores did not increase, it is important to note that the introduction of textbooks changed various aspects of classroom behavior and interaction. Classroom observation data indicate that in the schools which received textbooks, children were more likely to read aloud from their textbooks, teachers spent less time writing exercises on the blackboard and using the blackboard as a visual device, and teacher absenteeism declined.

There is some evidence that provision of textbooks crowded out other fundraising by schools, with each dollar of assistance from ICS reducing funds raised from other sources by 39 cents. However, this estimated crowding out is not statistically significant and consists of a reduction in contributions to harambees, Kenyan fundraising events used to finance construction of new classrooms. It is doubtful that this led to much reduction in the flow of non-textbook educational services over the time period examined in this paper. Moreover, crowding out appears to have taken place only at small schools, and our results do not change if we focus only on large schools.

The remainder of the paper is organized as follows. Section I describes primary education in Kenya, the textbook program, and the data collected. Section II presents a pre-treatment comparison suggesting that the treatment and comparison groups were similar prior to the textbook program. Section III discusses the evidence on how textbooks were used, and how they have affected pedagogy. Section IV compares average test scores in treatment and comparison schools at the end of the first year of the textbook. Section V repeats the analysis, disaggregating by initial student performance and finds that the program improved test scores by 0.2 standard deviations for students who started out in the top quartile of the distribution of students on the pretest. Section VI examines potential sources of selection and attrition bias and argues that they are unlikely to account for our results. Section VII examines the effect of the program on fundraising through harambees. Section VIII examines the information content of the test score data. Section IX compares the results of the randomized evaluation with results based on applying other techniques to data from Busia and with results from earlier literature on textbooks. In particular, it argues that a difference-in-difference analysis of an earlier textbook distribution program would have produced misleading estimates of the effect of textbooks.

I. Background

Primary Education in Kenya

The overwhelming majority of children in Kenya attend at least some primary school. Of those who enroll in first grade, 80% complete fourth grade and 43% complete seventh grade [CITE]. At the end of eighth grade, Kenyan students take a national exam that determines which secondary schools they are eligible to attend. In many primary schools, only the best students are allowed to go on to grade eight, while the rest repeat grade seven or drop out.

In grades one through three, school is taught in a combination of English, Swahili, and the local language, which is Kiluhya in approximately 70% of the schools in our sample, and Ateso in the remainder. From fourth grade on, the language of instruction is English.

In Kenyan schools, the Ministry of Education sets the curriculum, administers national and district level exams, and hires, transfers, and pays teachers. Local school committees, which are composed primarily of parents, are responsible for raising funds for all other costs of running the school. Most of these funds are raised through various fees that parents are required to pay. In practice, there is often a complex bargaining process between headmasters and parents over how much of the official fee a particular family will actually pay. Physical facilities at these schools are minimal. Many students lack school-benches and therefore sit on the floor, and classrooms are often dilapidated, and in some cases non-existent.

Schools usually have textbooks for teachers to use, but few textbooks for children.. According to our data, in grades three, four and five, one out of every six students has textbooks in the most important subjects (English and mathematics), while in grades six and seven, about one out of four students has textbooks in these subjects. Very few students have textbooks in other subjects. Students who go on to grade eight have substantially more textbooks than students in the lower grades -- about 40 percent of students have textbooks in mathematics and English. About 80-90% of the textbooks that students have were purchased by their parents, rather than by the school.

School Selection

The project took place in Busia district, an agricultural region on the border of Uganda.[4] In 1995, Busia was 26th out of 50 districts in Kenya in average test scores on the national primary-school leaving exam. The 333 primary schools in the district serve approximately 63,000 boys and 60,000 girls in grades one through eight.

In late 1995, the Ministry of Education district office selected 100 schools in Busia for the School Assistance Program (SAP), a program was funded by Internationaal Christelijk Steunfonds (ICS), a Dutch NGO. These schools, which we will refer to as SAP schools, were chosen because they were considered to be particularly in need of assistance, yet had not been assisted by an earlier World Bank textbook assistance program, which was supposed to be targeted to the most needy schools. The median school average test score among SAP schools on the sixth and seventh grade 1995 district-wide exam was at the 40th percentile of the school average test scores in the district as a whole. On the eighth grade exam, the median SAP school

[4] In 1995 Busia was split into two districts, Busia and Teso. We refer to both as Busia, unless otherwise stated.

was at the 33rd percentile of schools in the district as a whole. Schools in our sample are smaller than average for the district, with an average size of about 200 students.

The one hundred SAP schools were randomly divided into four groups as follows. First, schools were grouped according to their geographic division within Busia. At that time, Busia had seven administrative divisions. (In late 1995, Busia district was split; five districts remained in Busia and two became the new Teso district). A single list of 100 schools was formed by starting with the (alphabetically ordered) list of schools in the first division, adding to it the (alphabetically ordered) list of schools in the second division, and so on. From this list, every fourth school, beginning with the first, was assigned to group one. Similarly, every fourth school beginning with the second, third, and fourth was assigned to groups two, three, and four, respectively. Group one schools received textbooks at the beginning of 1996. Group two schools received a grant, but no textbooks at the beginning of 1997, group three schools received a grant, but no textbooks at the beginning of 1998, and group four schools will receive financial assistance in 1999.

Textbook Provision

The 25 schools in group one received textbooks in January, 1996.[5] English textbooks were provided in grades 3 through 7. Math textbooks were provided in grades 3, 5 and 7. Since almost half of grade 8 students already had math and English textbooks, ICS provided science textbooks in grade 8. In January, 1997, math textbooks were provided to grades 4 and 6, and agriculture textbooks were given to grade 8. For each grade and subject in which textbooks were given, ICS also provided one copy of the accompanying teacher's guide. ICS provided the official government textbooks, published by the Kenya Institute of Education, which are structured around the official curriculum.

Textbooks were supplied at less than a one-to-one ratio, based on the results from the Heyneman, Jamison, and Montenegro [1984] study in the Philippines. In Kenya, it is standard practice for primary school students to share textbooks. Two or three students typically share a bench and desk, so it is easy for students to share texts in class. A 60 percent textbook per pupil ratio was provided in English and science, and a 50 percent textbook ratio was provided in math.

Data Collection

In Busia, district-wide exams covering all the subjects in the Kenyan curriculum are given to students in the upper grades by the Ministry of Education. In 1996, exams were given in grades 5 through 8, but in 1997 they were given only in grade 8.[6] The exams are achievement

[5] The Kenyan school year runs from January through November.

[6] This change in which grades were tested was caused by a national decree that was completely unrelated to the

tests designed to measure whether students have learned the official government curriculum. The district-wide exams are administered in October.[7] For the 100 SAP schools, the October 1996 district-wide exams in grades 5 through 8 were supplemented with tests in grades 3 and 4, and in October, 1997, the grade 8 exam was supplemented with tests in grades 3-7. Moreover, pre-tests in English, math, and science were administered for grades 3 through 8 in the 25 textbook schools (i.e. group 1 schools) and the 25 schools in group 4. The pre-test was administered in January 1996, before textbooks were distributed. These supplementary tests were prepared by the district education office using the same procedure that it uses to design the district-wide exams.

The 25 schools that received textbooks can be compared to several sets of comparison schools. After one year, the textbook schools can also be compared to all schools that did not receive assistance in year 1 (groups 2, 3, and 4). We will refer to these schools as the 75-school comparison group. For outcomes after 2 years, they can be compared to the 50 schools that did not receive assistance in year 1 or year 2 (groups 3 and 4). We will refer to these schools as the 50-school comparison group. Finally, when disaggregating results by the pre-test score for either year 1 or year 2, they can be compared to the 25 schools that were not assisted in the first two years and also took the January pretest (group 4). We will refer to these schools as the 25-school comparison group.

As discussed in Sections III and VII, we also collected information on school finances, and visited classrooms and interviewed students to collect information on pedagogical methods.

SAP program.

[7] There is one exception. In grade 8 the district-wide exam is given in July, and the national primary school leaving exam

II. Initial conditions in textbook and comparison schools

The 25 textbook schools and the two sets of comparison schools had fairly similar enrollment and test scores before the intervention. Table 1 examines levels and changes in enrollments. In August, 1995, about five months before the program started, mean enrollment in the 25 textbook schools (204.8 students) was about 6% higher than mean enrollment in the 75-school comparison group (193.7 students). However, this difference of 11.1 students was far from statistically significant (the standard error of the difference is 23.2), and the difference in mean enrollment between the 25 textbook schools and the 25-school comparison group was trivial (204.8 vs. 205.8) and completely insignificant. Table 1 also shows that the mean percentage decline in enrollment in the 25 textbook schools was about 1% from August, 1995, to March, 1996, which compares to a drop of about 4% for both the 75-school comparison group and the 25-school comparison group.[8] Again, these differences are not statistically significant, but it may be that the program caused enrollment in the textbook schools to be about 3% higher than it otherwise would have been.[9]

Initial test scores are examined in Table 2. To check whether there are significant differences between the textbook schools and the comparison schools, we estimate regressions

(KCPE) is given in November. The results presented here use the November tests.

[8] These mean percentages of growth in enrollment are not exactly equal to the percentage growth in mean enrollment, since changes in means generally differ from means of changes.

[9] Overall enrollment in Busia district fell by about 5% between 1995 and 1996, perhaps in part because school fees increased in 1996. Specifically, before 1996 activity fees of 100 Ksh (approximately $2) were collected only for grades 6-8, but in 1996 activity fees of 80Ksh were collected in all grades.

with random effects at the school level. Because the units in which test scores are measured are arbitrary, for each subject and grade combination we normalize all test scores by subtracting the mean test score in the comparison schools and then dividing by the associated standard deviation from those schools. Thus, a student who scored 0.1 was 0.1 standard deviations above the mean score in the comparison schools. For reference, it is worth noting that in a normal distribution an increase of 0.1 standard deviations would move a student from the 50th percentile to the 54th. (At parts of the distribution with lower density, the percentile change is smaller.)

The results in Table 2 show no systematic differences in scores on the January 1996 pre-test between the 25 schools that received textbooks the following month (February, 1996) and the 25 schools in group 4 that took the same January tests. More specifically, Table 2 begins by presenting separate estimates of the difference in test scores between the textbook schools and the 25-school comparison group for each grade in each of the three subjects for which textbooks were given: English, mathematics, and science. Those subject-grade combinations in which textbooks were given are marked with a T in the "Grade" column. In most cases the difference between the textbook schools and the comparison schools is less than one tenth of a standard deviation of the normalized scores. None of the differences are statistically significant, even at the 10% level.

The rows labeled "All Grades" for English, math and science combine the data from all grades for each subject, which amounts to stacking the regressions that were estimated separately for each grade.[10] The "All Subjects Combined" rows shown at the bottom of the table report regressions that stack across different subjects in the same grade (the rows labeled 3-8) and a final regression that stacks across both grades and subjects. Finally, the last row of Table 2 shows an estimate that stacks across grades and tests. The difference between the textbook schools and the 25-school comparison group is only 0.04 standard deviations, and it is completely insignificant.

In summary, textbook and comparison schools seem comparable in terms of their initial enrollment and pre-test scores. Note also that because the randomization was stratified by geographic area, the ethnic composition of the schools is similar; 32% of the schools in each group were in Teso areas, with the remainder in Luhya areas.

III. Use of Textbooks

Much of the remainder of this paper will examine the impact of textbooks on student performance as measured by achievement tests. Before examining student test scores it is useful to examine the extent to which textbooks were used in the classroom, and what impact they had on classroom activities. There are two sources of information regarding whether the textbooks provided to the 25 Group 1 schools were used and whether teaching practices changed in response to receipt of textbooks. The first is classroom observation data. In 1997, each of the 100 SAP schools was visited by trained observers who recorded activities for about 15 minutes in classrooms for which textbooks were given. The second source of information is a pupil

[10] These stacked regressions allow for random class and subject effects, as well as random school effects. They also include dummy variables for each grade and subject combination. This is explained in detail in Section III.

questionnaire, which ICS field staff filled out by interviewing 4-5 randomly selected students per class.

The classroom observation data consist of minute by minute notes taken by observers. For each minute, the following information was recorded: i) the activity of the teacher (22 possible codes); ii) use of "visual aids" (blackboard, a chart or worksheets distributed to students) by the teacher; iii) use of a textbook by the teacher (yes or no); iv) who the teacher interacted with (no one, individual student, group, whole class); v) language used by the teacher; vi) pupil activity; vii) use of textbooks by students (yes or no); viii) number of pupils engaged in the pupil activity (none, one, several, all); and ix) language used by pupil(s). For evaluating whether textbooks are used, and their impact on teaching practices, the most useful information is that on teacher activity, use of visual aids and of textbooks by the teacher, pupil activity and use of textbooks by students.

From each of these "samples" of 15 minutes of time, the percentage of time spent by teachers and students on particular activities of interest can be calculated. For the 25 textbook schools and the 50 school comparison group, data are available from 551 class periods.[11] To test whether the percentage of time spent on a particular activity was significantly different between the textbook and comparison schools, a simple random effects model was estimated by regressing the percentage of time spent on a constant term and a dummy variable for textbook schools.

[11] Observers visited as many class periods as they could in one day, out of 12 possible per school (one for every grade/subject combination that had received textbooks by 1997). Only two observers visited each school, which made it difficult to complete the form for all 12 "target" class periods. For example, if four "target" subject/grade combinations met at the same time period, data could only be obtained from two (each period lasts 25 minutes).

School random effects were used. Because the exact number of minutes during which observations were recorded varied across class periods, weighted least squares regressions were run with the minutes of observation used as weights. Admittedly, these weights are not quite right because the minute by minute activities are correlated.

Over 100 regressions were estimated for a variety of student and teacher activities. Due to the large number of regressions, and the fact that most had statistically insignificant results, Table 3 shows only the statistically significant results. There are two statistically significant results regarding student activities: students in textbook schools are slightly more likely to be distributing textbooks or notebooks in class (an increase from 0.3% to 0.6% of students' time) and more likely to be reading aloud from a textbook in class (increase from 3.4% to 6.2%). Several statistically significant changes occurred in teachers' time: teachers in textbook schools are less likely to write notes or exercises on blackboards (decrease from 6.1% to 3.8%), slightly more likely to spend time assigning homework (increase from 0.0% to 0.2%), much less likely to be absent from school (decrease from 28.4% to 16.6%), more likely to teach without any visual aids (increase from 62.2% to 71.1%) and less likely to use blackboards as visual aids (decrease from 35.9% to 28.2%). Finally, observations of whether certain events took place at all during the 15 minute observation period showed more use of textbooks (increased probability of 20%) and a higher likelihood of the teacher being present (13% increase in probability). Overall, these results show less use of blackboards, more use of textbooks and less absences of teachers.

The results from the pupil interview data are similar. In particular, pupils report a 16% drop in the probability that they worked on exercises written on the blackboard, a 19% increase in the probability that they had access to a textbook, a 6% decline in the probability that they shared a textbook (conditional on having access to one), a **XX**% increase in the number of days they can take the textbook home during the past week and a **XX**% percent decrease in the length of time that the teacher was in the classroom.

IV. Program Effect on Average Test Scores

This section presents a variety of estimates indicating that provision of textbooks had little effect on average test scores. Sub-section IV.A first reviews the random effects estimation procedure that we employ, and then presents a levels estimator based on comparing post-test scores across treatment and comparison schools. Sub-section IV.B compares differences between pre- and post-test scores between treatment and comparison schools. Sub-section IV.C compares the difference between test scores in subject-grade combinations in which textbooks were and were not given in textbook schools to the same difference in comparison schools. All three estimators suggest little effect of textbooks on average test scores. The levels, difference, and subject-based estimators are precise enough to allow us to reject the hypothesis that textbooks raised school average test scores by 0.22, 0.14, and 0.07 standard deviations, respectively. This contrasts sharply with, for example, the retrospective studies reviewed by Lockheed and

Hanushek [1988], which typically found that textbooks increased test scores by about one third of a standard deviation.

IV.A. Levels Estimates

Because test scores are likely to be correlated among students in the same class and school due to unobservable characteristics of teachers and headmasters, we use an error components econometric model with school, grade, and subject random effects. Failing to account for these correlations would lead to underestimates of standard errors of the coefficients. More specifically, consider a regression of the impact of textbooks on children in grade j in subject k:

$$t_{ijks} = \alpha_{jk} + \beta_{jk}p_s + u_{jks} + e_{ijks} \qquad j = 3, 4,\ldots 8, \quad k = \text{English, Math, Science.} \qquad (1)$$

The test score of student i in grade j in subject k in school s is t_{ijks}. The constant term α_{jk} reflects the average score in schools that did not receive textbooks.[12] The variable p_s is a dummy variable that equals 1 if school s is a textbook school (i.e. a school that received textbooks) and 0 if it is not. Thus, for the subject-grade combinations in which textbooks were provided, the coefficient β_{jk} reflects the impact of textbooks on test scores in subject k in grade j. Equation (1) is

[12] This constant term may not equal to zero even though the average test score in the comparison schools is normalized to zero (as explained in Section II) because the regression is estimated with a generalized least squares specification that incorporates random effects, so a_{jk} represents a *weighted* average of test scores in comparison schools, where the weight on each school increases less than linearly with the number of pupils in the school.

estimated separately for each subject and grade combination.[13] The error term contains two components, a school specific effect (for grade j and subject k), u_{jks}, and a child specific effect, e_{ijks}, which implies a random effects estimation procedure. Since the schools that received textbooks were randomly selected, neither component of the residual term should be correlated with p_s.

The effect of textbooks may differ across subjects and grades. Stacking across grades and subjects allows us to measure the weighted average impact of textbooks on test scores across all grades and subjects in which texts were provided. Grades with more students will be given more weight in the stacked regression, although the weight will increase less than linearly with the number of students due to the random effects error structure. Conceptually, there is little difference between aggregating potentially disparate effects across grades and subjects and aggregating potentially disparate effects across students, as is routinely done. Just as textbook effects may differ across grades or subjects, they may differ across students. Empirically, as will become clear below, we find little evidence that the effect of textbooks differs between subjects, but fairly strong evidence that the effect of textbooks differs between weak and strong students.

Consider now a regression that combines several grades to measure the impact of textbooks for a given subject k:

$$t_{ijks} = \alpha_{3k}D_{3i} + \alpha_{4k}D_{4i} + \ldots + \alpha_{8k}D_{8i} + \beta_k p_s + u_{ks} + v_{jks} + e_{ijks} \qquad k = \text{English, Math, Science.} \qquad (2)$$

[13] Thus, with 6 grades (3-8) and 3 subjects (English, math and science) we run 18 separate regressions.

If the impact of textbooks varies across grades, β_k will measure the (weighted) average impact of textbooks across all grades. The dummy variables D_{ji} indicate whether child i is in grade j. In effect, one stacks the data for all grades (in this case grades 3-8). Note that because there is now more than one grade per school, the school-specific error term in equation (1) has been decomposed into a school-specific term u_{ks}, and a grade-specific term conditional on being in that school, v_{jks}.[14]

We also estimate regressions that combine more than one subject in the same grade, to measure the average impact of textbooks in a given grade j:

$$t_{ijks} = \alpha_{jE}D_{iE} + \alpha_{jM}D_{iM} + \alpha_{jS}D_{iS} + \beta_j p_s + u_{js} + v_{jks} + e_{ijks} \qquad j = 3, 4, \ldots 8. \qquad (3)$$

where the E, M and S subscripts represent the three subjects (English, math and science). In this regression β_j will measure the average impact of textbooks across subjects for a given grade. Similar to equation (2), D_{iE}, D_{iM} and D_{iS} are dummy variables indicating which test, English, Math, or Science, pertains to a given observation for child i.

Finally, we also estimate regressions that combine all grades and subjects, which measure the weighted average impact of textbooks across all grades and subjects:

[14] Random effects statistical models along these lines are known as multi-level models or hierarchical linear models

$$t_{ijks} = \alpha_{3E}D_{3Ei} + \alpha_{3M}D_{3Mi} + \alpha_{3S}D_{3Si} + \ldots + \alpha_{8E}D_{8Ei} + \alpha_{8M}D_{8Mi} + \alpha_{8S}D_{8Si}$$

$$+ \beta p_s + u_s + w_{js} + v_{jks} + e_{ijks}. \quad (4)$$

We estimate these equations using Generalized Least Squares (GLS) without imposing a specific distribution (e.g. the normal distribution) on the error terms.[15]

The regressions we report below include controls for sex and geographic division within Busia. Given the prospective design of the program, regressions without such controls are consistent, but adding controls to the regression increases the precision of the estimates. As a check, we ran regressions without the controls for region and sex; they yield similar results. This is not surprising because we stratified the sample so that geographic division is (up to integer constraints) orthogonal to treatment, and the sex ratio is nearly identical between treatment and comparison schools.

Table 4 presents random effects regressions of post-test scores on a dummy variable for textbook schools, and dummy variables for region and the sex of the student. The sample consists of all students tested in October, 1996, who were enrolled in January, 1996, in either the 25 schools that received textbooks (which occurred in February, 1996) or the comparison group of

in the education literature. See Goldstein [1987] and Bryk and Raudenbusch [1992].

[15] This structure is appropriate if students have the same teacher for all subjects, so that w_{js} is a teacher-specific effect and v_{jks} is a subject-specific effect conditional on having that teacher. In the upper grades, each teacher specializes in a given subject. In this case the error term should be specified as $u_s + w_{ks} + v_{jks} + e_{ijks}$, so that w_{ks} is a teacher specific effect for the teacher(s) that teach subject k to all grades, and v_{jks} is the grade specific impact of that teacher. In practice, these two different error structures for equation (4) yield similar results. Adding a random effect at the individual level when stacking across subjects, i.e. for equations (2) had almost no effect on the point estimates, though it occasionally reduced estimated standard errors slightly.

75 schools that did not receive textbooks.[16] The first 6 rows of Table 4 show estimates of β from equation (1) for the English test in grades 3-8. Recall that grades 3-7 received English textbooks (in Table 4 those grades that received textbooks are marked with a T). None of these five grades shows a significantly higher score in textbook schools relative to comparison schools. In fact, in two cases (grades 3 and 6) the point estimate is negative, though insignificant. When all five grades that received textbooks are combined into a single regression (i.e. equation (2)), the average impact of textbooks is very close to zero (0.01 standard deviations) and completely insignificant. This is shown in the row labeled "Grades w/ Texts."

The results for mathematics and science in Table 4 also provide little evidence for any large impact of textbooks on test scores. In each of the three grades that received mathematics textbooks (3, 5 and 7), the point estimate of the treatment effect is less than 0.1 standard deviations and is statistically insignificant. When these three grades are combined into a single regression, the estimated average impact is only 0.07 standard deviations, which again is statistically insignificant. Finally, for the single grade that received science textbooks, grade 8, the estimated impact is positive (0.08 standard deviations), but again not statistically significant.

The "All Subject-Grade Combinations with Textbooks" panel at the bottom of Table 4 shows the estimated textbook effect by grade, stacking across subjects in those grades in which

[16] A few students who transferred between these two groups of schools were also dropped. Note also that in some cases we do not have data from all 100 schools. Specifically: 1. For grade 5, two schools are missing data, one because the data were lost by the district office of the Ministry of Education after the test was administered, and the other because the test was never administered because the school did not pay the district office for the cost of the tests; 2. For grades 6 and 7, one school only goes up to grade 5 and in another the test was never administered; and

textbooks were given (i.e. equation (2)). None of the estimated parameters are large or statistically significant. Regressions stacking all subject-grade combinations that received textbooks (equation (4)) yield an estimated program effect of only 0.038 standard deviations, which is not statistically significant. The standard error of 0.088 standard deviations implies that one can reject (at the 5% significance level) the hypothesis that the true (average) effect was 0.22 standard deviations or higher.

IV.B. Differences-in-Difference Estimator

Comparing the difference between post- and pre-test scores across the textbook and comparison schools also provides no evidence for a large impact of textbooks on test scores. Note that unlike the levels estimator, the difference estimator is valid only under the assumption that a student who starts half a standard deviation ahead of another student would remain half a standard deviation ahead if both students were treated identically. It is possible that initial differences in test scores would narrow or widen over time given identical treatment. Imposing the assumption buys precision, however, by controlling for those discrepancies in initial conditions between the two groups that did not average out through randomization.[17]

Table 5 shows difference estimates. The comparison group is now only 25 schools, those that participated in the January, 1996, pre-test. As before, we restrict attention to the students

3. For grade 8, three schools do not go up to grade 8 and in one the test results were lost by the district office.

[17] Theoretically, levels estimates could be more precise because differencing implies that the error term will contain not only the noise in the post-treatment measurement but also the noise in the pre-treatment measurement, but this effect does not dominate in practice.

who were enrolled in January, 1996, in order to avoid bias caused by non-random enrollment of new students into program schools. All of the estimates for individual grade-subject combinations that received textbooks are statistically insignificant, as are regressions that stack across grades or subjects. The estimate of equation (4), which stacks across all grades and subjects and is shown in the bottom row of Table 5, yields an estimated treatment effect of 0.024 standard deviations, with a tight standard error of 0.059 standard deviations, implying that we can reject (at the 5% significance level) the hypothesis that the impact of textbooks on average test scores is 0.14 standard deviations or higher.

IV.C. Comparing Subject-Grade Combinations that Received and Did Not Receive Textbooks

A third estimator is based on regressing test scores on dummy variables for whether students were in textbook schools, and whether they were in subject-grade combinations that received textbooks. To understand the intuition, first think of calculating the difference in normalized test scores between treatment and comparison schools in subject-grade combinations in which textbooks were provided. Then calculate the difference in test scores between treatment and comparison schools in subject-grade combinations in which textbooks were not provided. Now take the difference between these differences. Random differences in school quality, for example due to differences in headmaster quality, could create some differences in test scores between treatment and comparison schools in subjects in which textbooks were not provided, but if the difference between treatment and comparison groups in those subjects was significant, it would suggest that textbooks had a spillover effect on other subjects. An additional benefit of exploiting variation within schools is that problems of sample selectivity are reduced. When comparing subjects that vary in whether or not a textbook was received within a single grade, one effectively compares test scores in different subjects for the same student, so differences in student composition across textbook and comparison schools (due to differential attrition) will not affect the estimator (although the estimator would be biased if students who are particularly good in a specific subject are prevented from dropping out by provision of a textbook in that subject). As explained below, however, this estimator will only be unbiased under a somewhat narrower set of assumptions.

Consider a regression that combines the three different subjects (English, math and science) for a given grade:

$$t_{ijks} = \alpha_{jE}D_{iE} + \alpha_{jM}D_{iM} + \alpha_{jS}D_{iS} + \beta_j p_s + \gamma_j T_{jks} + u_{js} + v_{jks} + e_{ijks} \qquad j = 3, 4, \ldots 8. \qquad (5)$$

Here T_{jks} indicates that students in grade j in school s received a textbook in subject k. Theoretically, textbooks could create both a direct within-subject effect on test scores by conveying information in the subject for which the textbook was provided, and also a general program effect in all subjects perhaps by making students more interested in school (i.e. a morale effect), or providing them with practice in reading. In equation (5), β_j is an estimate of general program effects, while γ_j is an estimate of the direct within-subject effect of textbooks. The total effect of distribution of textbooks in subject j on test scores in subject j is $\beta_j + \gamma_j$.

A variant of equation (5) is a regression that includes several grades within a given subject:

$$t_{ijks} = \alpha_{3k}D_{3i} + \alpha_{4k}D_{4i} + \ldots + \alpha_{8k}D_{8i} + \beta_k p_s + \gamma_k T_{jks} + u_{ks} + v_{jks} + e_{ijks} \quad k = \text{English, Math, Science.} \qquad (6)$$

Here T_{jks} indicates that textbooks in subject k were distributed in grade j of school s. The direct within-subject effect of receiving textbooks in a particular subject-grade combination is picked up in the coefficient γ_k, while any general program effect of being in a program school is picked up in the coefficient β_k. A general program effect could exist if grades that did not receive textbooks in a particular subject showed improvements in that subject due to receiving textbooks in other subjects.

Finally, we estimate a single regression by stacking across both subjects and grades:

$$\begin{aligned} t_{ijks} = {} & \alpha_{3E}D_{3Ei} + \alpha_{3M}D_{3Mi} + \alpha_{3S}D_{3Si} + \ldots + \alpha_{8E}D_{8Ei} + \alpha_{8M}D_{8Mi} + \alpha_{8S}D_{8Si} \\ & + \beta p_s + \gamma T_{jks} + u_s + w_{js} + v_{jks} + e_{ijks}. \qquad (7) \end{aligned}$$

Note that the estimates in this subsection are valid only under stronger functional form assumptions than those required in the estimates in the previous sub-section. We assume that all textbooks affect text scores in the same way, both in terms of direct within subject and general program effects. If, for example, it is more difficult to change English scores than math scores, then it is conceivable that providing an English textbook will lead to a greater effect on math scores in the same grade than on English scores. In this case, a negative direct within-subject effect of textbooks will be estimated in English. To put it another way, it may not be legitimate to add and subtract test scores in different grade-subject combinations. Violations of the functional form assumptions may help explain why the point estimates below differ considerably across subject-grade combinations, and even specifications, although the calculated standard errors are small.

The first three columns of Table 6 show, for each grade, estimates of equation (5) where the dependent variable is the post-test score. As in previous tables, dummy variables for sex and region were added to obtain more precise estimates. Note first that the coefficients corresponding to being in a textbook school are generally small and insignificant, providing little support for the existence of such general program effects. Although general program effects are not estimated precisely, because they are estimated based only on the variation between schools, and not variation within schools, large spillover effects across grades are not particularly plausible, since textbooks were used only in the grades where they were given. Interpreting this insignificant coefficient as resulting from random variation in school quality seems more reasonable.

The overall effect of receiving a textbook, when all grades and subjects are combined into a single regression (equation (7)) is slightly negative, at –0.01 standard deviations. Because the standard of this regression is fairly precisely estimated, we can even reject (at the 5% level) the hypothesis that textbooks raise the average text score by as much as 0.07 standard deviations.

In summary, the results in Table 6 support the hypothesis that textbooks did not have substantial effects on average test scores. In some cases, one could imagine complicated spillover effects that might yield these results, but in many cases these would be implausible. For example, is difficult to imagine a story about general program effects that could explain why program schools did not have particularly high scores in mathematics in the grades in which they received mathematics textbooks.

V. Program Effects Disaggregated by Initial Test Score

Although Tables 4, 5, and 6 suggest that the program did not have a large positive effect on average test scores, there is evidence that the program had a substantial positive impact on test scores of students who started out with high levels of academic achievement. Table 7 reports regressions of post-test scores on a dummy variable for textbook schools, the student's average pre-test score across all subjects in which the student took the pretest, and the interaction between these two variables.[18] The coefficients on the interaction term are usually positive in the different subject-grade combinations, and in several cases significantly so. If one aggregates across all subject-grade combinations, the interaction term is highly significant.[19]

[18] As in Section IV, all regressions also included region and sex dummy variables and employed a random effects error term structure.

[19] However, in one subject-grade combination that did not receive textbooks (math grade 4) a highly significant positive interaction coefficient was also found, which suggests that some other factor may be involved. To check

Given the normalization of the pre-test scores, the coefficient on the program dummy can be interpreted as the estimated effect of the program for someone with an average score on the pre-test. The program effect for someone who scored one standard deviation above average on the pretest is equal to the average program effect plus the coefficient on the interaction term, while the program effect for someone who scored one standard deviation below average on the pretest is the average program effect minus the coefficient on the interaction term.

Test scores can be interpreted as a noisy signal of underlying academic achievement, and the regressions as indicators of the effect of initial academic achievement and the interaction between initial achievement and textbooks. Since the pre-test is a noisy measure of achievement, the coefficients on the pre-test and on the interaction between the pre-test and treatment will underestimate the direct effect of initial achievement and the interaction effect between initial achievement and the presence of the program. If one assumes that the true coefficient on initial academic achievement is one, as in standard difference-in-difference specifications, then this attenuation bias is major, since it causes the coefficient on the pre-test score to be only 0.408. If a similar correction factor were applied to the estimated interaction effect it would imply that the true interaction effect is 16% of a standard deviation and hence that the difference in the impact of textbooks for a student 1 standard deviation below and 1 standard deviation above the mean academic achievement is 0.32 standard deviations.

this possibility, a regression that stacks grades that did not receive textbooks (English grade 8, mathematics grades 4, 6 and 8, and science grades 3-7) was estimated; the interaction effect was positive but statistically insignificant, which suggests that the significantly positive interaction effect for the subject-grade combinations that did receive textbooks was indeed due to those textbooks.

One way to correct for the attenuation bias caused by measurement error in pre-test scores is to instrument for them. Table 8 repeats the regressions shown in Table 7 with one difference: the pre-test score for the particular subject (as opposed to the average over three subjects) is used, while the other two pre-test scores are used as instrumental variables. Similarly, the interaction between the textbook school dummy variable and the pre-test score is instrumented with interactions between that dummy variable and the other two pre-test scores. Since skills in English may help students learn math and science, and math skills maybe useful in science, it is possible that the exclusion restriction is violated for these subjects. However, it is unlikely that math and science skills help in English. If one focuses only on English, the estimated interaction effect is 0.2 standard deviations, suggesting that a student with initial academic ability one standard deviation above the mean benefits from textbooks by 0.4 standard deviations more than a student who starts one standard deviation below the mean. If one looks at all subjects and grades together, the estimated interaction effect is 0.17 standard deviations.

The specification used in Tables 7 and 8 constrains the interaction between pre-test scores and the treatment variable to be linear. It also constrains the random school effect to be the same for students of every ability level; that is, it does not allow for the possibility that some schools may be particularly good for high-ability students but bad for low-ability students. These assumptions may be too restrictive, so in Table 9 we disaggregated the sample into quintiles, as determined by average pre-test scores, and re-estimated the regressions in Table 4 for each quintile.[20] This allows for separate treatment effects for students in each quintile. As usual, we allow for school random effects. Moreover, because separate regressions are estimated for each quintile, the random effect for any given school can vary by quintile, so that some schools can be particularly good for strong students and others can be particularly good for weak students.

When all subject-grade combinations with textbooks are considered together, the estimated effect of textbooks on test scores is -0.04 standard deviation for students for students whose pre-test scores were in the lowest quintile, -0.05 standard deviations for students whose initial test scores were in the second quintile, 0.07 standard deviations for students whose test scores were in the third quintile, 0.12 standard deviations for students whose test scores were in the fourth quintile, and 0.22 standard deviations for students whose pre-test scores were in the top quintile. Only the last of these is statistically significant. Note that since the pre-tests were noisy, some of the students in the top quintile of pre-test scores were not necessarily in the top quintile of initial achievement, and if the effect of textbook provision depends on initial achievement, the true effect of textbooks on students in the top quintile of achievement will be greater than 0.22 standard deviations. For example, if 1/3 of the students in the top quintile of pre-test scores came from lower quintiles of the achievement distribution, and if textbooks had no effect on test scores in these lower quintiles, then the true textbook effect for students in the upper quintile of the achievement distribution would be approximately 0.33, about the same as the typical effect found in Lockheed and Hanushek's (1988) review of retrospective studies.

[20] Note that we are using the 25-school comparison group, not the 75-school comparison group, because pre-test scores exist only for the former comparison group.

Note that the relationship between the effect of textbooks and the quintile of the pre-test score does not hold across all grades. In particular, textbooks seem to raise test scores in both lower and upper quintiles among grade 8 students. This may be a feature of science, or the science textbooks, but it may also be because grade 8 students are a selected group, almost all of whom have fairly high academic achievement.

Textbooks may have benefitted only the stronger students either because only the strong students were able or willing to read textbooks, or because the particular textbooks provided were too difficult for many of the students. As explained above, the Ministry of Education directed ICS to provide schools with the official textbooks published by the Kenya Institute of Education. Since the textbooks are written to cover the official curriculum and expected progression through grade levels, and since textbooks may be designed with their primary markets, teachers and rich students in cities, in mind, it seems plausible that textbooks would be too difficult for students in Busia's poorer rural schools.

It is worth noting that if textbooks are more useful for better students, and if parents respond to this by purchasing textbooks for children who perform well in school, then retrospective estimates of the effect of textbooks are likely to be biased not only because of heterogeneity among parents in commitments to education, but also because of variation in ability among students. Thus, for example, controlling for a family effect by differencing across siblings will not eliminate omitted variable bias.

To summarize Sections IV and V, Tables 4, 5, and 6 show no significant impact of textbooks on average student achievement, while Tables 7, 8 and 9 show a significant impact on students that were already performing well. The next sections check for several potential problems with the analysis. Section VI argues that neither selection nor attrition bias is likely to account for the results. Section VII checks whether provision of textbooks from ICS crowded out assistance from other sources. Section VIII argues that although the difficulty of the tests may have made it hard to measure improvement in academic performance among weak students, this is not likely to fully account for the pattern of treatment effect by initial test scores.

VI. Selection and Attrition Bias

This section argues that the results of the previous two sections are unlikely to be driven by selection or attrition bias. There were some changes in the composition of the student body due to the program, but they were not large, and attempts to correct for these changes do not significantly affect estimates of the impact of textbooks on test scores.

Selection

The results presented in Sections IV and V excluded children not enrolled in the 100 SAP schools in January 1996. Since textbooks were distributed in February 1996, limiting the sample to children who were enrolled in January of that year would seem to prevent selection bias due to new students entering the school after textbooks were provided. Yet selection bias could still be a problem because the program was announced in December 1995, so some children may have left or transferred into the textbook schools at the beginning of January 1996 based on this information.

Another more subtle selection bias problem is that there is some evidence that textbook schools were more likely to promote students from one grade to another between 1995 and 1996, perhaps because headmasters thought that the availability of textbooks would help weaker students keep up with their peers.

Tables 10 examines the extent of selection bias. The top half of Table 10 shows the 1995 status of all 1996 students. The overall promotion rate, including transfer students, was 71.0% in textbook schools and 67.2% in comparison schools, a difference of about 4 percentage points. Conversely, the repetition rate is about 4 percentage points lower in the textbooks schools (21.2%, compared to 25.6% in the comparison schools).[21] This suggests a potential selection problem due to differential promotion. These data also show that the percentage of 1996 students who had transferred into the textbook schools (15.6%) is about three percentage points higher than the same figure for the comparison schools (12.6%). The differences in transfer rates are significant at the 10% level, while differences in promotion and repetition rates across textbook and comparison schools are significant at the 1% level when all grades are combined. They are particularly common in the lower grades (3, 4 and 5). Headmasters in textbook schools may have been more likely to promote children from grade 2 (in 1995) to grade 3 (in 1996) because

[21] The repetition and promotion rates do not add up to 100% because in both schools about 7-8% of the pupils are transfer students for whom there are no data on repetition or promotion.

students in grade 2 in 1996 received no textbooks, while students in grade 3 received two textbooks (English and Math), so the school as a whole would appear to receive more textbooks when more students are in grade 3 by January 1996. Even though headmasters were specifically told, when the program was announced in December 1995, that the numbers of textbooks distributed in January 1996 would be based on 1995 (as opposed to 1996) enrollments, some headmasters may have overlooked this point or may not have believed it. This boost in grade 3 enrollments may have had "spillover" effects that created pressure to promote more children from grade 3 to 4, and even from grade 4 to grade 5. Another reason why headmasters in textbook schools may have been more likely to promote marginal students is because they thought that textbooks would help them keep up with the curriculum.

Another potential source of selection bias would be asymmetries in the dropout and transfer out rates between 1995 and 1996 between treatment and comparison schools. Of the 100 schools in the sample, only 20 (10 textbook schools and 10 comparison schools) have complete data on enrollment in 1995.[22] The bottom half of Table 10 shows, separately for 10 treatment and 10 comparison schools, the status of students in 1996 conditional on enrollment in 1995. The textbook schools did not retain a greater percentage of students (fewer dropouts and transfers out) than the comparison schools. Combining both types of dropouts (those who dropped out before January 1996 and those who dropped out later), 12.3% of students in the textbook schools

[22] In particular, information on students who dropped out in 1995 was collected from only 20 schools in 1996. While the same information was collected from the other 80 schools in 1997, it appears to be much less accurate for those schools. Thus while the data on where 1996 students were in 1995 is reliable for all 100 schools, the data on where 1995 students were in 1996 is reliable for only 20 schools.

dropped out, compared to 11.6% of the students in the comparison schools. Transfers out are slightly higher in the textbook schools (3.3%) than in the comparison schools (1.8%). Although this difference is statistically significant at the 5% level, it seems unlikely that students were actually led to drop-out by the prospect of receiving textbooks.

Several checks of the data suggest that these differences in transfer and promotion rates between textbook and comparison schools do not account for the pattern of results. First, differences in student composition in 1996 due to transfers or differential promotion before January 1996 should appear as differences in pre-test scores, but no significant differences were observed in Table 2. Second, the estimator, based on comparing relative performance of textbook and comparison schools in subject-grade combinations which did not receive textbooks would not be biased by general differences in ability. Finally, re-estimation of Tables 4, 5 and 9 after restricting the sample to children who were in their current schools in 1995 does not give significantly different results. For brevity, the complete results are not presented here, but the results when all grades and subjects are stacked show how little the results change. For Table 4, restricting the sample in this way changes the overall impact from 0.038 to 0.040 standard deviations. For Table 5, the change is from 0.024 to 0.015 deviations. For Table 9, recall that for the top quintile there was an impact of 0.220 standard deviations, which was significant at the 5% level; the estimate when the sample is restricted to students who were present in 1995 is 0.195, which is significant at the 10% level.[23]

[23] CHECK IF INTERACTION EFFECT IS STILL SIGNIFICANT?

For a given grade in a textbook school, the tendency to promote borderline students (i.e. students who would otherwise not be promoted) has two distinct effects: it adds lower performing students to that grade and it takes away students who would have been in that grade had they repeated. The first effect clearly biases downward the estimated impact of textbooks. The direction of the bias of the second effect is theoretically unclear, but since data from the 75 comparison schools (not shown here) shows that repeaters usually perform slightly better than non-repeaters, and since the receipt of textbooks would presumably remove the better performing of these repeaters, the second effect is also likely to bias estimates of program effect downwards. To adjust for the first effect, all repeaters were dropped from both the textbook and the comparison schools. To obtain an upper bound on the true effect of the program, in the presence of the second effect, we can make the extreme assumption that the “extra” students who were promoted in the textbook schools are the most marginal students, that is the ones who scored the lowest on the October 1996 tests, and drop these students from the analysis. Since the tests are a quite noisy measure of underlying achievement, dropping the worst students in each grade in the textbook schools would probably lead to substantial overestimation of the impact of textbooks, so that results based on such a sample would give a very safe upper bound to the true effects.

Table 11 repeats the regressions shown in Table 5, except the sample is altered as described above to obtain an upper bound. As expected, the point estimates are somewhat larger than those in Table 5, yet they remain statistically insignificant. The aggregate estimate at the bottom of Table 13 shows a positive impact of 0.10 standard deviations, with a standard error of

0.09. Thus even these estimates, which are upward biased, rule out an impact of 0.3 standard deviations, which is typical of many retrospective studies.[24]

VII. Crowding Out of Other Assistance

There is some evidence that receipt of assistance from ICS reduced local fundraising by textbook schools, but it is not statistically significant. Data on school finances in 1996 is available for all 100 SAP schools from a school questionnaire and a school committee questionnaire, both of which were administered in 1997. The 1997 school questionnaire asks for the amount of money raised by the school in 1996 through local fundraising events and about in-kind assistance received from NGO's and other donor groups. The 1997 school committee questionnaire asks only about money raised through harambees.[25]

Most of the large-scale fundraising by schools is through harambees which are large-scale events that typically raise money for construction of new schools or classrooms, with politicians and businessmen as prominent donors. Most schools report no harambees occurred in 1996. Of the 75 comparison schools, only 14 reported harambees in the school committee questionnaire, and about 95% of the funds raised were used for construction.

Table 13 shows the amount of assistance received from ICS and from other sources for the textbook and comparison schools. The average value of the textbooks given by ICS was

[24] DO THIS WITH PRE-TEST SCORES INSTEAD?

[25] The data analyzed in this subsection come only from the school questionnaire, which collects a wider range of information, with one exception: if the amount reported for harambees in the school committee questionnaire exceeds the amount reported for harmabees in the school questionnaire, the harambee amount from the former replaced the amount from the latter.

$485. The average amount of non-ICS aid received by comparison schools was $456, while the amount received by the textbook schools was $267. The difference of $189 suggests that receipt of textbooks from ICS reduced assistance from other sources. More specifically, it appears that about 39% of value of the assistance provided by ICS was crowded out by a reduction in assistance from other sources.

However, although the differences in non-ICS assistance reported in Table 3 are large, they are not statistically significant, even at the 10% level, in any of a wide variety of specifications. The high standard errors are due to the very large variation in the amount of assistance received from other sources. Most schools receive nothing, while a few schools received large amounts of assistance.

Even if assistance from ICS reduced the amount of money raised harambees, there may be little direct impact on educational outcomes in the short run, since it takes some time to construct the new facilities, and the flow of services from these facilities take place over a long period. Thus, for example, if provision of a dollar of textbooks led to a forty cent drop in funds raised for new classrooms, and if half of these classrooms were built during the year that the textbooks were provided, and if the flow of services from classroom construction takes place over a period that is five times as long as that of textbooks, the provision of one dollar more in textbook services immediately leads to approximately a four cent decline in classroom services in the first year. XX

Splitting the samples into small and large schools (based on median enrollment in the comparison schools) indicates that the difference in non-ICS assistance is due solely to differences among smaller schools. This makes sense, because it is the smaller, newer schools that are most likely to be incomplete and to still need additional classrooms.

We therefore re-estimated Table 5 for the larger schools only, for which there is no evidence of crowding out. For brevity, the only result reported here is for the regression in Table 5 that aggregates across all grades and subjects that received textbooks. The point estimate is 0.061, with a standard error of 0.130. This is both small and statistically insignificant, and not much different from the estimate of 0.038 presented in Table 5. Thus we find no evidence that crowding out explains the low impact of textbooks on average scores that we found in Sections IV and V.

VIII. How Much Information is Conveyed in Test Scores?

Noise in the tests may have made it more difficult to identify a treatment effect for students with low initial achievement, but it is unlikely to fully account for the pattern of treatment effects across grades. Although the tests are a fairly noisy signal of achievement, especially for low achieving students, the tests nonetheless convey some information. Moreover, preliminary results from the second year of testing with tests that were less difficult suggest similar effects.

Table 15 displays a variety of information about the tests given in 1996. The last column shows correlations between the pre-tests and post-tests, which range from 0.13 to 0.61. If either the pre-tests or the post-tests were completely noise, the correlation between pre-test scores and post-score would be zero. Table 8 also showed that pre-tests are a highly significant predictor of performance on post-tests.

The data suggest that it is possible to identify differences among schools in average test scores. Since treatment was randomized at the school level, our estimates are identified based on variation among schools in average test scores. Most of the variation in test scores at the individual level that is due to guessing will be eliminated at the school level. In a school with 200 students, for example, the variance in average test scores induced by guessing will be 0.005 times the variance in individual test scores induced by guessing. A glance at the between-school variance figures in Table 15 shows substantial variation in average test scores among schools, orders of magnitude above what one would see from pure guessing. The fact that average test scores differ substantially among schools implies that school level differences in the learning environment could be identified and measured.

While it is clear that the tests are not pure noise, their level of academic difficulty raises the possibility that we can detect improvements in performance only for the subset of students whose initial performance was above a certain level, which is one explanation for the results in Table 10. Perhaps all students benefited from the textbooks, but this benefit was only observable for the best students.

For each test, Table 15 shows the mean score on each exam and the expected score if all students guessed on all multiple choice questions and could not answer any other question.[26]

[26] Examination of the answer sheets showed that virtually all students answered all the multiple choice questions, implying that children who did not know the correct answer to a question guessed instead of leaving it blank. Kenyan examination

Scores were very low on many tests. Many of the weaker students may have been guessing on almost all the questions. (For all tests, Kolmogorov-Smirnov tests reject the extreme hypothesis that all students were guessing.) For each test that was multiple choice, Table 15 also shows the percentage of students whose scores were above a 95% confidence interval around the mean of a binomial distribution that would result if they guessed on all items. (2.5% of students would have scored above this level if all students guessed on all questions.) Of course, this does not imply that all other students were guessing. Many students who were not guessing on all questions probably fell in the 95% confidence interval of the guessing distribution.

If one confines attention to the tests with the least guessing, the point estimate of the treatment effect on average test scores is slightly higher, but still typically less than 0.1 standard deviations, and statistically insignificant. Specifically, among the English post-tests in grades that received textbooks, the highest scores are in grade 4. For the math post-tests, among the three grades that received textbooks, the highest scores are in grade 7. The estimated treatment effects on average scores in these tests shown in Table 5 are just 0.11 and 0.09 standard deviations, respectively. While scores were also high on the grade 8 science post-test, the estimated treatment effect was still only 0.075 standard deviations. Finally, reexamination of Table 10 shows that in grade 4 English and in grade 7 math, the treatment effect is essentially zero in the

graders commonly give scores of zero for poor essays.

first three quintiles, and highest in the fifth quintile. In grade 8 science the treatment effect shows no pattern across quartiles. However, recall that grade 8 students are already a select group.

A final check can be done by considering exams administered after two years of treatment in October 1997, which we designed, and on which average scores were approximately fifty. The results are summarized in Table 16. Of the 11 grade-subject combinations that received textbooks in either 1996 or 1997, none shows an impact that is statistically significant at the 5% level, although there are two cases where the impact is significant at the 10% level (mathematics grade 7 and science grade 8). To derive a more powerful aggregated test, the six grade subject combinations in which students had been exposed to textbooks for two years (grades 4-7 in English and grades 4 and 6 in math) were combined. The aggregate impact is only 0.086 standard deviations, with a standard error of 0.161.

Future work will look more closely at the 1997 test score data, adjusting for possible sample selection and other problems, but we do not expect the findings presented in the current draft of this paper to be substantially altered.

IX. Comparing Results of Randomized and Non-Randomized Studies

It is possible to compare the results from this prospective, randomized evaluation of textbooks with results from a cross-section regression based on existing variation across schools in the number of textbooks and results from a difference-in-difference analysis of a World Bank textbook program.

We also have data on variation among comparison schools in the number of textbooks, and plan to create an OLS estimate of the effect of textbooks using this data. XX

In a World Bank-funded project, the Jomo Kenyatta Foundation provided textbooks to 95 of the 334 schools in Busia in March, 1994,[27] at a ratio of roughly one textbook for every two pupils in English and math and one book for every four students in Swahili and Science. The schools were supposedly selected as being particularly in need of the textbooks. However, any tendency for weaker schools to be chosen was not reflected in the test scores of these schools. The median test score of the 6th, 7th, and 8th grades on the 1993 district exam among schools receiving textbooks from the Jomo Kenyatta Foundation was at the 54th, 45th and 48th percentile of the non-textbook schools in the district as a whole.[28]

A difference-in-difference analysis of this data suggests that textbooks only have an effect after two years, and not after one year, as shown in Table 17. In 1994, the program seems to have had a negative effect on test scores in grades 6 and 8, although a positive effect in grade 7. The results are dramatically different in 1995. There is a large positive effect in grades 7 and 8, and a moderate, but statistically insignificant negative effect in grade 6. One interpretation of this

[27] An additional 13 schools were given textbooks in October 1995, but, in order to keep the analysis clear, these

is that textbooks have only a small effect after one year, but a large effect after two. A less favorable interpretation is that the schools which the JKF allocated textbooks had some general increase in political influence, which allowed them to obtain other inputs, and perhaps be assigned better teachers and headmasters, and this explains the increase in test scores. In this case, the "natural experiment" approach of examining the JKF program would be misleading. Preliminary data on the second year of the SAP program supports this second interpretation.

Several factors may help explain why the results of this study differs from those found in previous studies, particularly the Jamison et al [1988] study in Nicaragua. The students in Nicaragua received workbooks, in which they could write the answers to mathematics problems. Workbooks allow for more active learning, but are much more expensive than textbooks. One workbook is needed per student, whereas one textbook may serve for two or three students in rural Kenyan schools. Moreover, a workbook can only be used for one year, whereas the same textbook can be used for three or four years. Thus the annual cost of workbooks per student may be six to nine times that of textbooks. Another difference is that whereas the Nicaraguan students were first graders who were likely starting off on an equal footing, the students in this study are considerably older and thus more likely to differ in their initial level of achievement. The Nicaraguan teachers received three hours of training in the use of workbooks whereas the Kenyan

have been dropped from the analysis.

[28] CHECK

teachers received none. Of course, the particular textbooks given and the schools and parental backgrounds differ between the studies.

This analysis may also be useful in interpreting the results of the Busia schools pilot study [Kremer, et. al., 1997]. That study examined a program which provided textbooks and paid for the school uniforms that children in Kenya are required to purchase. The program led to a 40% increase in enrollment. The earlier study suggested that the joint impact of textbook provision and a 40% increase in enrollment on test scores was small, but based on the pilot study it was impossible to determine whether textbooks had a large positive impact on test scores which was offset by a large negative impact of increased enrollment, or whether the effects of both textbooks and class size were small. This study suggests that it is more likely that both effects were small.[29]

[29] There are, however, several reasons why the impact of textbooks could differ between the two programs. First, the schools in the pilot program were typically somewhat poorer than the schools in this study. Second, the number of textbooks given and the distribution across grades differed between the two programs. In the pilot program, schools were allowed to allocate their budgets as they saw fit. Third, in the pilot program, schools generally chose to buy textbooks printed by private publishers whereas in this study, the Ministry of Education insisted that official textbooks of the Kenya Institute of Education be used.

References

Bryk, Anthony, and Stephen Raudenbusch (1992), *Hierarchical Linear Models.* Sage Publications: Newbury Park, CA.

Card, David, and Alan B. Krueger. 1992. "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States." *Journal of Political Economy* 100(February):1-40

Birdsall, Nancy. 1983, Strategies for Analyzing Effects of User Charges in the Social Sectors, Discussion Paper No. 1983-1989, Country Policy Department, World Bank Reissued as PHN technical note 87-94 and now available from population and human resources department (Washington, DC)

Finn, Jeremy and Charles Achilles (1990), "Answers and Questions about Class-Size: a State-Wide Experiment," *American Educational Research Journal* (Fall), 557-577.

Fuller, Bruce (1986), *Raising School Quality in Developing Countries: What Investments Boost Learning?* The World Bank, Washington D.C.

Fuller, Bruce and Prema Clarke (1994), "Raising School Effects While Ignoring Culture? Local Conditions and the Influence of Classrooms, Tools, Rules and Pedagogy", *Review of Educational Research*, 64(1), 119-157.

Goldstein, Harvey (1987), *Multilevel Models in Educational and Social Research*. Oxford University Press.

Hanushek, Eric A. (1995), "Interpreting Recent Research on Schooling in Developing Countries," World Bank Research Observer, 10 (August), 227-246.

Heckman, James J. Anne Layne-Farrar, and Petra Todd. 1994. "Does Measured School Quality Really Matter? Understanding the Empirical and Economic Foundation of the Evidence." University of Chicago, Department of Economics, Chicago. Processed.

Heckman, James J. and Jeffrey A. Smith (1995). "Assessing the Case for Social Experiments," *Journal of Economic Perspectives*, 9(2), pp. 85-110.

Hedges, L. V., Richard Laine, and Rob Greenwald. 1994. "Does Money Matter? A Meta-Analysis of Studies of the Effects of Differential School Inputs on Student Outcomes." *Educational Researcher* 23(3):5-14.
Heyneman, Stephen P., Joseph P. Farrell and Manuel A. Sepulveda-Stuardo. 1978. *Textbooks and Achievement: What We Know.* World Bank Staff Working Paper No. 298.

Heyneman, Stephen P., Dean T. Jamison and Xenia Montenegro. 1984. "Textbooks in the Philippines: Evaluation of the Pedagogical Impact of Nationwide Investment." *Educational Evaluation and Policy Analysis* 6(2):139-150.

Jamison, Dean, Barbara Searle, Klaus Galda and Stephen Heyneman. 1981. "Improving Elementary Mathematics Education in Nicaragua: An Experimental Study of the Impact of Textbooks and Radio on Achievement". Journal of Educational Psychology 73(4): 556-67.

Kagitcibasi, Cigdem, Diane Sunar, and Sevda Bekman (1993), "Long-Term Effects of Early Intervention," Department of Education, Bogadzdi University, Istanbul, Turkey.

Kremer, Michael, Sylvie Moulin, Robert Namunyu, and David Myatt. 1997. "The Quantity-Quality Tradeoff in Education: Evidence from a Prospective Evaluation in Kenya", unpublished.

Levin, Henry, and Marlaine Lockheed. 1993. *Effective Schools in Developing Countries.* Falmer Press. Washington, DC.

Lockheed, Marlaine E., and Eric Hanushek. 1988. "Improving Educational Efficiency in Developing Countries: What Do We Know?" *Compare* 18(1):21-38.

Lockheed, Marlaine, and Adriaan Verspoor (1991), *Improving Primary Education in Developing Countries*, NY: Oxford University Press.

Newman, John, Laura Rawlings, and Paul Gertler (1994),"Using Randomized Control Designs in Evaluating Social Sector Programs in Developing Countries," *The World Bank Research Observer*, (July), 181-202.

Olsen, Randall, J. and George Farkas (1990), "The Effect of Economic Opportunity and Family Background on Adolescent Cohabitation and Childbearing among Low-Income Blacks," *Journal of Labor Economics* 8(3).

Selden, Thomas and Michael Wasylenko (1995), "Measuring the Distributional Effects of Public Education in Peru." In Dominique Van De Walle, and Kimberly Nead eds., *Public Spending and The Poor.* Baltimore: The Johns Hopkins University Press for the World Bank.

Thobani, Matteen, (1983) "Charging User Fees for Social Services: the Case of Education in Malawi," World Bank Staff Working Paper no.527, (Washington DC).
UNDP. 1990. *Human Development Report*. Oxford University Press. New York.

World Bank. 1989. "Sub-Saharan Africa: From Crisis to Sustainable Growth". The World Bank. Washington, DC.

World Bank. 1990. *World Development Report*. Oxford University Press. New York.