ELSEVIER

Journal of Development Economics 74 (2004) 251–268

JOURNAL OF Development ECONOMICS

www.elsevier.com/locate/econbase

Retrospective vs. prospective analyses of school inputs: the case of flip charts in Kenya

Paul Glewwe[a,*], Michael Kremer[b], Sylvie Moulin[c], Eric Zitzewitz[d]

[a] *Department of Applied Economics, University of Minnesota 1994 Buford Ave., St. Paul, MN 55108, USA*

[b] *Harvard University and NBER, USA*

[c] *The World Bank, USA*

[d] *Stanford University, USA*

Abstract

This paper compares retrospective and prospective analyses of the effect of flip charts on test scores in rural Kenyan schools. Retrospective estimates suggest that flip charts raise test scores by up to 20% of a standard deviation. Yet prospective estimators based on a randomized trial provide no evidence that flip charts increase test scores. One interpretation is that the retrospective results suffered from omitted variable bias. If the direction of this bias were similar in other retrospective analyses of educational inputs in developing countries, the effects of inputs may be more modest than retrospective studies suggest. A difference-in-differences retrospective estimator seems to reduce bias, but it requires additional assumptions and is feasible for only some educational inputs.

JEL classification: I21; O15

Keywords: Education; Flip charts; Developing countries; Kenya; Randomized trial; Program evaluation

1. Introduction

Most analyses of the effect of educational inputs are based on retrospective studies, which compare schools with different levels of inputs (Hanushek, 1995). One potential weakness of this approach is that observed inputs may be correlated with omitted variables that affect educational outcomes. This could potentially bias outcomes in either direction. For example, if parents who provide better home environments for children (a characteristic which is typically unobserved) tend to organize to obtain more observed school inputs

* Corresponding author. Tel.: +1-612-625-0225.

E-mail address: pglewwe@apec.umn.edu (P. Glewwe).

0304-3878/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.jdeveco.2003.12.010

for their children, estimates of the effect of those school inputs on test scores may be biased upwards. On the other hand, if compensatory programs provide schools in disadvantaged areas with additional observed inputs, and some aspects of their disadvantaged status are unobserved, then retrospective studies may underestimate the effect of these inputs. The direction and severity of these biases is ultimately an empirical question. This paper compares retrospective and prospective estimates of the effect of flip charts in Kenyan primary schools.[1] It finds that prospective estimates are much smaller than retrospective estimates, suggesting that retrospective estimates are subject to serious upward omitted variable bias, even when controlling for other observable inputs.

Straightforward OLS regressions using retrospective data on test scores in subjects in which flip charts can be used suggest that flip charts raise student test scores by 20% of a standard deviation. Controlling for other observed school inputs affects these estimates only slightly. Difference-in-differences retrospective estimates that compare the impact of flip charts on the relative performance of students in flip-chart and non-flip-chart subjects suggest a smaller effect of about 5% of a standard deviation, an effect that is still significant, though sometimes only at the 10% level.

These retrospective results contrast with those obtained from a prospective, randomized evaluation comparing 89 schools that were randomly chosen to receive flip charts with 89 schools that did not receive flip charts. After 2 years, test scores in subjects where flip charts can be used are virtually identical in the two types of schools (0.6% of a standard deviation lower in the schools that received flip charts, with a standard error of 4.8%). The analogous retrospective estimate of an increase of 20% of a standard deviation is decisively rejected. A difference-in-differences prospective estimator that compares the impact of flip charts on the relative performance of students in flip-chart subjects and in other subjects also yields an estimate that is effectively zero (0.8% of a standard deviation, with a standard error of 3.1%).

These results suggest that using retrospective data to compare test scores in subjects covered by flip charts between schools with and without charts seriously overestimates the charts' effectiveness. A difference-in-differences approach that compares relative performance across subjects reduces but does not eliminate this problem. Moreover, it is not clear that such a difference-in-differences approach has general applicability.

Given the scarcity of compensatory programs in developing countries, it seems reasonable to hypothesize that omitted variable bias will typically be positive in retrospective estimates of school inputs in developing countries. This suggests that the effect of large-scale programs to provide inputs may be even smaller than suggested by retrospective studies, which often find little or no effect of inputs (Hanushek, 1995).

The remainder of this paper is divided into six sections. The first describes the primary education system in Kenya, the flip charts, and the data collected. The second section sketches the analytic framework. The third section presents retrospective estimates of the flip charts' effect on test scores. The fourth section presents prospective estimates. A fifth section discusses potential biases from missing data, and a final section discusses potential explanations of the difference between the retrospective and prospective results and concludes the paper.

[1] We thus follow the approach LaLonde (1986) used in the context of U.S. job training programs.

2. Background

The vast majority of Kenyan children attend primary school, although in rural areas less than half reach grade 8, the final grade. Entrance into secondary school is highly competitive, based on students' performance on the Kenya Certificate of Primary Education (KCPE) exam, which is taken at the end of grade 8.

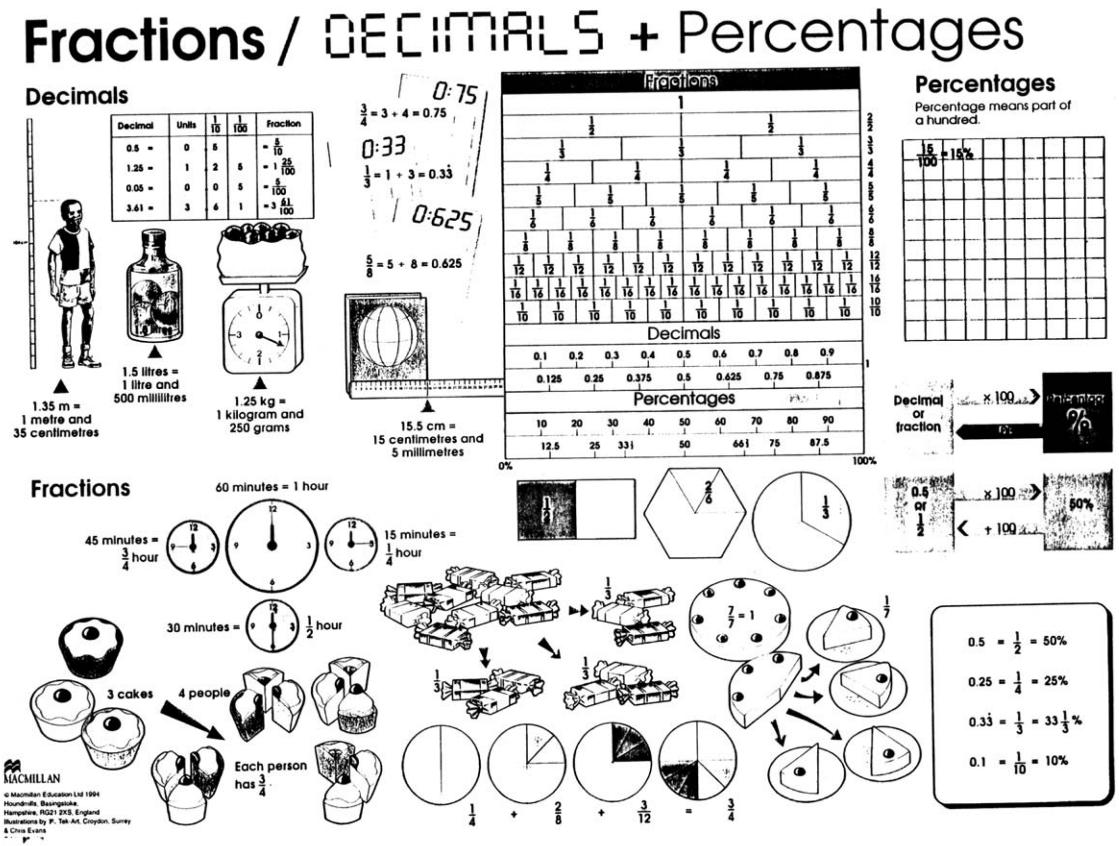
The schools in this study are located in Busia and Teso, two neighboring agricultural districts on the border with Uganda, both of which have below-average income for Kenya. Flip charts and other visual aids are rare in schools in these areas, and less than one-third of the schools had any flip charts before the study. Even textbooks are rare in these schools. In grade 8, which is selective, about 40% of students had textbooks in math and English, but 15% or less had textbooks in science and other subjects. In lower grades, textbooks are even rarer.

A Dutch NGO, Internationaal Christelijk Steunfonds, provided the flip charts distributed in the prospective study: two sets of science charts (one covering agriculture and the other covering general science), as well as a teacher's guide for science, one set of charts for health, one set of charts for mathematics, and a wall map of East Africa for geography. Each set of charts contains about 12 individual charts spiral bound together. Each individual chart covers different aspects of the topic (Fig. 1 shows a typical flip chart for mathematics). The charts are not kept in the classroom, but rather are brought in when they are relevant to the day's lesson, and can therefore be used in more than one classroom on any given day. The science charts are appropriate for grades 5–8, while the simplest math charts could, in principle, be used in grade 3. In practice, the grade 7 and 8 teachers have priority over the usage of the charts, and account for roughly 60–75% of total use, based on a survey in which teachers reported the number of times they had used the charts.

There are several reasons why visual aids such as flip charts might promote learning. Almost all students recall having seen pictures more often than having read words or sentences (Shepard, 1967). In addition, learning styles vary across students, so adding visual aids to traditional auditory presentations of material may reach a broader range of students.[2] Studies have found that supplementing textbooks with visual aids promotes learning in many different subjects, such as social studies (Davis, 1968), anatomy (Dwyer, 1970), ecology (Holliday, 1973), and reading (Samuels, 1970). Live presentations also benefit from supplementation with visual aids (see Dwyer, 1970; Holliday and Benson, 1991). For caveats and alternative views, however, see Dwyer (1970), Holliday and Benson (1991), Levin et al. (1976), and Lookatch (1995).

Flip charts may be particularly attractive in the rural Kenyan setting, where textbooks are too expensive for most students and many students have limited proficiency in English, the medium of instruction in Kenya and the language in which all Kenyan textbooks are written. Glewwe et al. (2004) find that textbooks improve scores only for students in the top two quintiles of the distribution of pre-test scores.

[2] For example Dunn et al. (1989) find that over 40% of students in the United States are visual learners, compared with under 10% auditory and about 20% tactual (touch) and 30% kinesthetic (activities). Wallace (1995) finds similar results for the Philippines.



Decimal	Units	$\frac{1}{10}$	$\frac{1}{100}$	Fraction
0.5 =	0	5		= $\frac{5}{10}$
1.25 =	1	2	5	= 1 $\frac{25}{100}$
0.05 =	0	0	5	= $\frac{5}{100}$
3.61 =	3	6	1	= 3 $\frac{61}{100}$

Fig. 1. Example of Mathematics Flip Chart (Macmillan Education Ltd. 1994)

2.1. Test score data

The data available for this study are the test scores of grade 8 students on the KCPE in November 1997 and November 1998, the scores of grade 8 students on practice exams given in July 1997 and 1998, and the scores of grade 6 and 7 students on practice exams given in October 1998.[3] Practice exams closely follow the format of the KCPE and are set at the district level by the Ministry of Education. Each exam covers seven subjects: English, Swahili, Math, Science/Agriculture, Geography/History/Civics/Religion (GHC-RE), Arts/Crafts/Music (ACM), and Home Science/Business Education (HS–BE). Of these seven subjects, the flip charts received were relevant to four: Math, Science/Agriculture, Home Science/Business Education (which includes health), and Geography/History/Civics/Religion (the wall map). Each subject exam consists mainly of multiple-choice questions with four possible answers, although the English and Swahili exams also require students to write a composition. The 1998 practice exams were administered only in Busia district (they are missing for the two divisions that split off during 1997 to form the new Teso district), but the 1997 practice and KCPE exams and the 1998 KCPE exam are available for both districts. In addition to these data on (post-intervention) performance in 1997 and 1998, there are also (pre-intervention) data on school-level performance, averaged across all subjects, from practice exams in 1996. There are no data for individual subjects or individual students in 1996.

All test scores are standardized using the individual-student mean and standard deviation for each grade–test–subject combination in the comparison schools. A score of 0.2, for example, represents someone who scored 0.2 standard deviations above the average in the 89 comparison schools. For reference, it may be useful to note that a movement from the 50th to the 54th percentile of the distribution corresponds to an improvement in test scores of 0.1 standard deviations (10% of a standard deviation).

3. Analytic framework

We assume that the production function for the academic achievement of a student, as measured by test scores, is

$$A = f(c, x),$$

where c and x denote flip charts and other inputs, respectively. Other inputs include both student and school characteristics. This relationship is structural; it holds regardless of the actions of schools, parents and students. The impact of flip charts on academic achievement that is embedded in this relationship is a structural parameter.

Exogenous provision of flip charts may cause schools and households to adjust their provision of other inputs, which implies that many of those other inputs can be treated as functions of the number of flip charts in the school. Substituting out all such endogenous inputs from the production function leads to a reduced form relationship in which the

[3] Unlike 1996 and 1998, in 1997 this practice exam was given only to grade 8 students.

impact of flip charts incorporates both direct (structural) and indirect (behavioral) effects; estimation of this relationship yields a reduced form parameter. The structural parameter that measures the impact of flip charts is a partial derivative of the production function, while the reduced form parameter is a total derivative that reflects both the partial derivative and agents' optimizing responses.

Policymakers may be interested in both the partial and the total derivatives. The total derivative is useful because it measures what will happen to the output variable (A) if the input is exogenously provided. On the other hand, partial derivatives may better capture overall welfare effects. Intuitively, if parents respond to the provision of flip charts by reducing their textbook purchases they will be able to raise their welfare by purchasing more of some consumer good; the reduced form impact (total derivative) of flip charts on student test scores ignores this welfare benefit. More formally, suppose that schools choose c and x to maximize

$$V = p_{\mathrm{A}}A - p_{\mathrm{c}}(c - c_{\mathrm{e}}) - p_{\mathrm{x}}(x - x_{\mathrm{e}}),$$

where the p's denote prices, c_{e} and x_{e} represent the school's initial endowments (which we assume are difficult to sell), and p_{A} measures the value that parents and schools ascribe to A. (We include c_{e} and x_{e} since some schools may have received flip charts or other inputs from a donor, rather than having purchased them.) The partial derivative of academic achievement with respect to the school's endowment of flip charts, $\partial A/\partial c_{\mathrm{e}}$, is a valid measure of the welfare effect of a change in c_{e}, namely $\partial V/\partial c_{\mathrm{e}}$, because by the envelope theorem this welfare effect equals, by a first-order approximation, p_{A} multiplied by $\partial A/\partial c_{\mathrm{e}}$. Alternatively, the welfare impact can be written as the total derivative times the value of the output minus the change in other inputs times the value of those inputs.

Prospective studies measure the total derivative of the output variable with respect to a change in the experimentally modified input. This is the case because they examine the output variable after some time has passed, time that allows the various actors to adjust their provision of other inputs. However, since some inputs can be adjusted more quickly than others, the total derivative may differ in the short run and in the long run.

What retrospective studies estimate depends on the source of variation in inputs. For example, if all the variation in unobserved inputs is determined by a random process that occurs after decision-makers have chosen the amounts for the observed inputs, the two sets of inputs are orthogonal to each other and retrospective studies give unbiased estimates of the partial derivative of the production function (the structural impact of the input). On the other hand, it is more plausible that variation in the input of interest is influenced by variation in the unobserved inputs, or that unobserved inputs vary in response to the (observed) input of interest, in which case retrospective estimates will be biased.

4. Retrospective analysis

The retrospective analysis uses data for 100 schools involved in a separate study that provided textbooks and grants to randomly selected schools (described in Glewwe et al., 2004). Data for these schools on flip charts and other school inputs were collected in early

1998 and the effect of the inputs was estimated using the 1998 practice and KCPE exam data for grades 6–8. Data were available only on the total number of flip charts, not their subject, and wall maps were not included. Since wall maps were not included for the purposes of the retrospective analysis, the flip charts could potentially be relevant to three subjects: Math, Science/Agriculture, and Home Science/Business Education. Data on the availability of flip charts were available for 83 schools; when controlling for other inputs, the sample drops to 79 schools.

Because the data for these schools provide information only on the total number of science, math, and health science-business education (HS–BE) charts in the school, not the number in each subject, the paper estimates the average effect of charts across all three subjects. Since the program evaluated in the prospective study distributed four flip charts (2 Science, 1 Math, 1 HS–BE), the number of charts variable is divided by four to generate coefficients that are comparable with the retrospective analysis.

Table 1 presents results from regressions of test scores on flip charts and other school inputs. In all regressions, data from multiple subjects and four grade–test combinations (practice exam for grades 6–8 and KCPE for grade 8) are combined into a single regression. Columns 1–4 estimate the effect of flip charts in the three flip-chart subjects. Columns 5 and 6 present results from a difference-in-differences specification that compares the impact of flip charts on the relative performance of students in the three flip chart and the four non-flip-chart subjects. All regressions include subject and grade fixed effects, and controls for whether the school was in a group that received textbooks or grants through another program, (the omitted category is the comparison group for that program). Thus, the coefficient on books per pupil reflects only variation in textbooks due to other factors, primarily the number of books prior to the program. Regressions also include school random effects to allow for within school correlation in test scores for example, due to differences in headmaster quality.

Column 1 presents an estimate of the impact of flip charts without controlling for other educational inputs. This indicates that provision of flip charts will increase student achievement in flip chart subjects by 19% of a standard deviation, a result that is significant at the 1% level. Of course, if other inputs are correlated with flip charts, this estimate of the partial derivative (structural impact) of flip charts could be biased. To (partially) correct for this, the estimates in columns 2 and 3 control for a variety of other educational inputs. They suggest that, holding all other inputs constant, adding flip charts raises test scores by about 20% of a standard deviation in flip-chart subjects, an estimate that is also significant at the 1% level. This result is virtually identical to the estimates in column 1, which suggests little correlation between flip charts and these other inputs (blackboards, textbooks, non-leaking roofs, desks, teacher training and class size).

The estimators in columns 4–6, all of which control for other educational inputs implicitly compare the relative performance of flip charts in flip-chart and non-flip-chart subjects. The effect of flip charts is estimated by comparing, across schools with and without flip charts, the difference between scores in subjects where flip charts are used and scores in other subjects. The validity of this approach is open to question. Some question whether it is possible to add and subtract test scores in different subjects, given their ordinal, rather than cardinal nature (Krueger and Whitmore, 2000). Aside from this issue,

Table 1
Retrospective estimates of effect of four flip charts in grades 6–8

Dependent variable: normalized 1998 test scores

Specification		(1)	(2)	(3)	(4)	(5)	(6)
	Mean(S.D.)	Level estimates				Diffs-in-diffs	
Random effects							
School		Yes	Yes	Yes	Yes	Yes	Yes
School × subject	No	No	No	No	No	No	Yes
Schools		83	79	79	79	79	79
Pupils		5152	4998	4998	4998	4998	4998
Grades included		6–8	6–8	6–8	6–8	6–8	6–8
Subjects included		Sc, Mat, HS	Sc, Mat, HS	Sc, Mat, HS	Sc, Mat, HS	All	All
Flip chart variable							
Number of charts in school	1.1 (2.4)	0.192***	0.194***	0.205***	0.076*	0.154***	0.157***
(divided by four)		(0.080)	(0.065)	(0.064)	(0.041)	(0.057)	(0.056)
Charts × flip-chart						0.049**	0.040*
subject (Science/						(0.021)	(0.024)
Agr., Math, HS–BE)							
Other variables							
Indoor classroom	0.97 (0.17)		0.454**	0.399**	0.031	0.506***	0.503***
			(0.114)	(0.147)	(0.151)	(0.123)	(0.123)
Roof does not leak	0.98 (0.13)		0.161	0.063	− 0.029	0.205	0.203
			(0.375)	(0.291)	(0.105)	(0.479)	(0.480)
Blackboard (1 = good	0.92 (0.18)		0.298	0.386*	0.038	0.294	0.293
cond., 0.5 = bad cond.,			(0.188)	(0.228)	(0.065)	(0.180)	(0.180)
0 = none)							
Textbooks per pupil	0.21 (0.24)		0.096*	0.119*	0.133***	0.063	0.089
			(0.051)	(0.069)	(0.028)	(0.047)	(0.065)
Desks per pupil	0.39 (0.16)		− 0.018	0.098	− 0.254***	0.246	0.247
			(0.339)	(0.418)	(0.098)	(0.327)	(0.327)
Teacher training level	2.1 (0.8)		− 0.039	− 0.051	0.033*	− 0.023	− 0.023
(0–6, 6 = high)			(0.033)	(0.045)	(0.018)	(0.033)	(0.032)
Class size	33 (16)			− 0.001			
				(0.004)			
Pupil age	14.3 (2.0)			− 0.069***			
				(0.009)			
Pupil's average score					0.770***		
on non-flip-chart subjects					(0.009)		

Regressions contain one observation per pupil for each subject. Columns 1–4 include flip-chart subjects only; columns 5 and 6 include all seven subjects. All regressions contain school random effects, subject and grade fixed effects, and controls for the assistance received through the textbook and grant programs. Column 6 includes school × subject random effects. Since the data for these schools only provide information on the total number of science, math, and home science/business education (HS–BE) charts in the school, not the number in each subject, we estimate the average effect charts across all three subjects. Since the program evaluated in the prospective study distributed four flip charts (2 Science, 1 Math, 1 HS–BE), the number of charts variable was divided by four to generate coefficients that are comparable with the retrospective analysis. Standard errors are heteroskedasticity robust. Statistical significance at the 10%, 5%, and 1% level is indicated by 1, 2, and 3 asterisks, respectively.

these estimators will be valid only if flip charts have no effect on test scores in non-flip-chart subjects, and if other factors correlated with flip charts that could influence scores do so equally across all subjects. Each of these assumptions is open to question. Flip charts could potentially either raise or lower test scores in other subjects. They could raise test scores by improving pupils' general interest in school, and thus attendance, or they could lower scores by diverting pupils' or teachers' attention from non-flip-chart subjects.[4] Moreover, since different tests were given in different subjects, an omitted variable correlated with flip charts, such as headmaster characteristics, could potentially differentially affect test scores in different subjects.

Column 4 directly controls for the performance of students in non-flip-chart subjects; this reduces the estimate to 7.6%, which is significant at the 10% level. The difference-in-differences estimates in columns 5 suggest that providing four flip charts would raise test scores by 4.9% of a standard deviation in the three flip-chart subjects. This is significant at the 5% level. Given that these regressions compare results across subjects, and that the performance of students in a particular school in a particular subject may be correlated due to teacher ability, column 6 allows for random effects at the level of interaction between schools and subjects. This reduces the point estimate to 4%, and reduces the significance level to 10%. The two difference-in-differences regressions also suggest that flip charts raise test scores by 15–16% in non-flip-chart subjects, suggesting either that flip charts have a positive effect in non-flip chart subjects or that the direct estimators are inflated by an omitted variable bias problem that controlling for other observable school inputs does not alleviate.[5]

These retrospective results imply that flip charts are cost effective compared to textbooks. The per-pupil cost of providing four charts is only 10% of the cost of providing a textbook for every pupil in each of the three subjects,[6] but the retrospective estimates suggest that the flip-chart effect is twice as large as the effect of providing textbooks for each pupil in three subjects (from column 2, comparing 0.194—the effect of four charts—with 0.096). Since flip charts are much less expensive, their cost-effectiveness is much higher than that of textbooks; indeed, they are about 20 times more cost-effective than textbooks, in terms of dollars per average test score gain. Even though the difference-in-differences estimate is much smaller than the direct estimate, it still suggests that flip charts are 4–5 times as cost-effective in raising average test scores as textbooks. As discussed below, a prospective analysis does not support this conclusion.

[4] Note, however, that in the upper grades, the school day is divided into separate periods with different teachers for different subjects.

[5] A final caveat to the retrospective analysis is that the significance of some of the results is caused by the inclusion of one school with well above average test scores and 15 charts (compared with an average of 1.1 per school). Although we have no reason to doubt this data, treating the school as having only five charts reduces the estimates in column 2 to 18% with a standard error of 16%. The differences-in-differences estimate in column 4 remains significant and increases slightly in magnitude, however.

[6] Flip charts cost about US$20 each, so four would cost US$80. Textbooks in Kenya cost approximately US$3.33; it would therefore cost about US$800 to provide one textbook per pupil in each of three subjects to the 80 students in grades 6–8 at the average-sized school in the sample. These cost figures are from 1997 and are converted to US$ at the then current exchange rate of 60Ksh/$.

5. Prospective analysis

Internationaal Christelijk Steunfonds (ICS), a Dutch non-governmental organization, distributed flip charts to selected schools in Busia and Teso in early 1997. One hundred and seventy-eight schools were potentially eligible. Schools that ICS had previously assisted through the textbook and grant program or through other programs were ineligible, as were a smaller number of relatively well-off schools. Since ICS began assisting schools that were relatively poor, those participating in the textbook/grant study tended to be slightly worse performers than those in the flip-chart study, and since the best off schools were excluded, those in the flip-chart study had roughly similar mean scores as the district as a whole.[7] Table 2 shows that the average pre-intervention characteristics of these 178 schools were nonetheless fairly close to those of the district as a whole. The assignment of the 178 schools into flip chart and comparison groups was done as follows; the schools were sorted alphabetically, first by geographic district, then by geographic division, and then by school name. Then every other school on that list was placed in the flip-chart group. The two types of schools will henceforth be referred to as flip-chart schools and comparison schools, respectively. The flip-chart program was announced in January 1997, after the start of the 1997 school year (in Kenya the school year runs from January to November), and the charts were distributed in early February 1997. Each school received two sets of science charts (including a teacher's guide), one set of charts in math, one set in health, and a wall map.

Table 3 contains the average raw scores out of 100 for each test and grade for the flip chart and comparison schools in the prospective study. The most straightforward method of evaluating the effect of the randomized distribution of flip charts is to compare the post-intervention (1997 and 1998) scores in the 89 flip-chart schools with the scores in the 89 comparison schools. The last four columns of Table 3 provide average test scores across all seven subjects for the flip chart and comparison schools for each test and grade combination. The flip-chart schools scored equal to or slightly below the comparison schools on all of the grade 8 exams and slightly above the comparison schools on the grade 6 and grade 7 exams. In all cases, the differences are much less than 10% of a standard deviation of the distribution for the comparison group.

Table 4 presents random effects regression estimates of the difference in test scores between flip chart and comparison schools for each subject. Recall that prospective regressions estimate the reduced form parameter, as explained in Section 2. Data from all tests are pooled to construct these estimators. School random effects are included to allow for correlation in the error term among students within a school, and these school random effects are allowed to vary by year. Results are presented with and without controls for pre-intervention (1996) school-average test scores. Controlling for pre-intervention test scores reduces the size of the school random effect, improving the efficiency of estimation. For science-agriculture, the subject for which two sets of flip charts and a teacher guide were

[7] If flip charts were more helpful for weaker students, part of the explanation for the larger estimated effects in the retrospective results could be that these schools had, on average, lower 1996 test scores. Interaction regressions in both the retrospective and the prospective sample found no evidence that the effect of flip charts was greater for schools with low initial test scores. The point estimates for the interaction coefficients suggested that the effect of flip charts would be between − 0.4% and 0.2% higher in the retrospective sample; these point estimates were statistically insignificant.

Table 2
Enrollment and prior year performance of Busia and Teso schools

	Entire district		Mean by study			Prospective study			
	Mean	S.D.	Prospective	Retrospective	Neither	Flip-chart	Comparison	Difference	*t*-stat
Number of schools	337		178	100	59	89	89		
Enrollment									
1997 Grade 8 (Feb.)	22.3	11.9	23.3	18.9	25.6	24.3	22.4	+ 1.9	1.06
1996 Grade 8 (March)	20.7	12.4	21.4	17.4	24.6	22.3	20.5	+ 1.8	0.96
1995 Grade 8 (July)	21.0	11.5	21.7	18.2	23.8	22.7	20.8	+ 1.9	0.56
1997 Grade 7 (Feb.)	36.2	20.2	38.7	28.7	42.3	39.8	37.7	+ 2.1	0.69
1996 Grade 7 (March)	36.7	21.7	39.3	29.4	41.7	39.9	38.6	+ 1.3	0.40
1995 Grade 7 (July)	36.4	22.4	38.5	30.1	41.9	39.2	37.7	+ 1.5	0.87
Pre-intervention school-average test scores									
1996 Practice (March)	308.4	34.5	314.1	295.8	325.2	314.1	314.0	+ 0.1	0.02
1996 Practice (July)	304.3	37.8	312.1	288.9	325.3	312.1	312.0	+ 0.1	0.02

Test scores are the sum of the raw scores which range from 0 to 100 on seven subject exams. *t*-test statistic is the same as the *t*-statistic from a regression of school enrollment/test scores on a constant term and a treatment-comparison dummy variable.

given, test scores for the flip chart and comparison schools were almost identical; the same is true for Geography/History/Civics/Religion. For Math and Home Science/Business Education, scores in flip-chart schools were 2–3% of a standard deviation below those of comparison schools. None of these differences is statistically significant. Even if the analysis is limited to the subject–grade combinations in which charts appear most promising, namely Math and Science in grades 6 and 7, a procedure that is obviously open to criticisms of data mining, none of the *t*-statistics obtained is greater than one. In summary, there is little evidence from the reduced form estimates in Table 4 that flip charts had a positive impact on test scores.

Tables 5 and 6 present estimates that pool across subjects. In Table 5, the estimate of the difference between flip chart and comparison schools is allowed to vary for the four flip-chart subjects and the three non-flip-chart subjects. This estimation includes random effects for school, school–subject combinations, and pupils.[8] Controlling for pre-inter-

[8] Due to computational constraints, pupil random effects could not be included in the regressions which include all subject–grade–test combinations. Despite the large size of the pupil random effects, the results for the single-test, multi-subject regressions change very little when pupil random effects are omitted. Flip chart effects change by no more than 0.5% of a standard deviation, and standard errors increase by 0.1% of a standard deviation for 8th grade and decrease by 0.04% for 6th and 7th grade, when pupil random effects are omitted.

Table 3
Sample size and summary statistics for the prospective analysis

Test	Grade	Students tested						Average test score			
		Received charts			Did not receive charts			(Percent correct on 4-choice test)			
		Both distr.	Busia	Teso	Both distr.	Busia	Teso	Charts	No charts	Difference	S.D.
Jul-97	8	1848	1263	585	1861	1357	504	45.5	46.0	−0.5	12.5
Nov-97	8	1790	1262	528	1843	1420	423	48.7	49.6	−0.9	13.3
Jul-98	8	1211	1211	0	1343	1343	0	42.7	42.9	−0.3	11.2
Nov-98	8	1737	1206	531	1891	1370	521	49.5	49.5	0.0	13.0
Oct-98	7	1734	1734	0	1798	1798	0	37.6	37.5	+0.1	11.3
Oct-98	6	1664	1664	0	1726	1726	0	37.3	36.9	+0.4	11.4

vention school-average scores and combining all tests, test scores in schools that received flip charts are estimated to be 0.6% and 1.4% of a standard deviation lower in flip-chart and non-flip-chart subjects, respectively, with standard errors of roughly 5% of a standard deviation. Again controlling for pre-intervention school-average scores, scores are 3–7% of a standard deviation lower in flip-chart schools in both groups of subjects in 1997, while they are 2–6% of a standard deviation higher on the 1998 practice exam. None of these

Table 4
Prospective estimates of effect of flip charts—single subject multi-test regressions

Dependent variable: normalized test score

Subject	Past perf.	Flip-chart school		Obs.
	Controls	Coeff.	S.E.	
Flip-chart subjects				
Science/Agriculture	No	0.0005	0.0752	20,446
	Yes	−0.0007	0.0591	
Math	No	−0.0201	0.0600	20,441
	Yes	−0.0212	0.0486	
Health Science/Business Ed. (HS–BE)	No	−0.0295	0.0728	20,434
	Yes	−0.0276	0.0559	
Geography/History/Civics/Religious Ed. (GHC)	No	0.0018	0.0714	20,450
	Yes	−0.0012	0.0553	
Non-flip-chart subjects				
English	No	0.0038	0.0737	20,433
	Yes	−0.0100	0.0576	
KiSwahili	No	0.0110	0.0790	20,448
	Yes	0.0146	0.0737	
Arts/Crafts/Music (ACM)	No	−0.0679	0.0758	20,417
	Yes	−0.0723	0.0589	
Memo				
Math and Science; grades 6 and 7 in 1998 only	No	0.0508	0.0828	13,836
	Yes	0.0534	0.0655	

Regressions include school and school × year random effects and test fixed effects. Past performance controls are controls for the school-average performance on the July 1996 practice exam.

Table 5
Prospective estimates of effect of flip charts—single test multi-subject regressions

Dependent variable: normalized test score

Test	Grade	Past perf.	4 Flip chart subjects		3 Non-flip chart subjects		Obs.
		Controls	Coeff.	S.E.	Coeff.	S.E.	
All tests	6–8	No	– 0.0117	0.0638	– 0.0149	0.0649	143,069
		Yes	– 0.0063	0.0484	– 0.0144	0.0498	141,698
Jul-97	8	No	– 0.0138	0.0716	– 0.0388	0.0751	25,939
		Yes	– 0.0347	0.0605	– 0.0627	0.0644	25,827
Nov-97	8	No	– 0.0474	0.0744	– 0.0516	0.0758	25,418
		Yes	– 0.0656	0.0601	– 0.0700	0.0617	25,418
Jul-98	8	No	0.0135	0.0848	0.0102	0.0866	17,882
		Yes	0.0325	0.0718	0.0291	0.0739	17,791
Nov-98	8	No	– 0.0018	0.0708	– 0.0134	0.0722	25,396
		Yes	0.0145	0.0575	0.0043	0.0591	25,060
Oct-98	7	No	– 0.0061	0.0910	– 0.0029	0.0925	24,708
		Yes	0.0327	0.0669	0.0268	0.0690	24,288
Oct-98	6	No	0.0708	0.1005	0.0612	0.1019	23,726
		Yes	0.0579	0.0799	0.0485	0.0817	23,314

Regressions include school, school × subject, and pupil random effects subject and test fixed effects. Pupil random effects cannot be included when results are estimated for all tests due to computational constraints. For the single-test results, excluding pupil effects changes point estimates by no more than 0.0045 and standard errors by no more than 0.001.

differences is close to being statistically significant (none has a *t*-statistic over 0.75), nor is the slight improvement from 1997 to 1998 statistically significant in regressions that estimate separate flip-chart effects for each year for 8th graders (not shown).

The results in Table 5 suggest that the overall performance of a school can vary from year to year across subjects. This variation in the cross-subject school effect adds noise to the estimated difference between test scores in flip-chart and non-flip-chart schools. An alternative approach is to assume that flip charts do not affect performance in non-flip-chart subjects and estimate the effect of flip charts by comparing the relative performance of flip-chart schools in flip chart and non-flip-chart subjects with the analogous relative performance in the comparison schools. Under the assumption that flip charts do not affect non-flip-chart subjects, the efficiency of the estimation can be increased by using the non-flip-chart subjects to better control for school effects. Table 6 presents estimates of the difference in the flip chart-comparison school performance differential between flip-chart and non-flip-chart subjects. Across all subjects and test–grade combinations and controlling for past performance, the effect of flip charts is estimated to be 0.8% of a standard deviation. The standard error of the difference-in-differences estimator is lower (3.1% of a standard deviation), but the estimated effect of flip charts is still far from significant.

Across all the different estimators in the prospective study, the effect of flip charts appears to be essentially zero. There is no evidence that this is because flip charts were not used. We interviewed 82 grade 7 and 8 teachers in flip-chart subjects at 21 of the schools that received flip charts. Ninety-eight percent of the teachers were aware that their school

Table 6
Prospective estimates of effect of flip charts—differences-in-differences estimator

Dependent variable: normalized test score

Test	Grade	Past perf. controls	FC school and FC subject		Flip-chart school		Obs.
			Coeff.	S.E.	Coeff.	S.E.	
All tests	6–8	No	0.0031	0.0312	0.0117	0.0638	143,069
		Yes	0.0080	0.0308	0.0063	0.0484	141,698
Jul-97	8	No	0.0250	0.0594	– 0.0138	0.0716	25,939
		Yes	0.0280	0.0581	– 0.0347	0.0605	25,827
Nov-97	8	No	0.0042	0.0381	– 0.0474	0.0744	25,418
		Yes	0.0044	0.0376	– 0.0656	0.0601	25,418
Jul-98	8	No	0.0033	0.0464	– 0.0135	0.0848	17,882
		Yes	0.0034	0.0468	0.0325	0.0718	17,791
Nov-98	8	No	0.0116	0.0367	– 0.0018	0.0708	25,396
		Yes	0.0102	0.0368	0.0145	0.0575	25,060
Oct-98	7	No	– 0.0032	0.0444	– 0.0061	0.0910	24,708
		Yes	– 0.0059	0.0442	0.0327	0.0669	24,288
Oct-98	6	No	0.0095	0.0448	0.0708	0.1005	23,726
		Yes	0.0094	0.0453	0.0579	0.0799	23,314

Regressions include school, school × subject, and pupil random effects subject and test fixed effects. Pupil random effects cannot be included when results are estimated for all tests due to computational constraints. For the single-test results, excluding pupil effects changes point estimates by no more than 0.0045 and standard errors by no more than 0.001.

had been given flip charts, and 91% claimed to have used the flip charts. In no cases had the flip charts been lost or stolen. Ninety-two percent of teachers claimed they found the charts helpful, and they reported that the average chart had been used in each class on 10–20% of school days in the current year (1998). Given that the charts were shared between grades 6–8 at least, this represents reasonably high utilization of the charts. One caveat is that although teachers were surveyed in private and told that their answers would be kept confidential and would not affect future aid to their school, the teachers may have nonetheless felt an incentive to bias their usage estimates upward. Yet over 90% of the teachers gave specific answers to questions that required some experience using the charts (e.g., which charts did they find most and least helpful, and why), which suggests that the charts had in fact been used.

The incentives faced by schools in Busia may have led them to use the charts for students in upper grades, who would soon take the KCPE exam on which schools are judged. The flip chart use survey revealed that charts were used an average of 13 days per 75-day term in grade 8 compared to 7 days each in grades 6 and 7. One potential hypothesis for the low estimated effect of flip charts is that the charts would have been more useful in lower grades. Thirty percent of grade 7 and 8 teachers reported that the charts helped the worst students the most, while only 3% reported that they helped the best students most. The fact that the estimated effect of the flip charts was highest for grade 6 students is at least consistent with the charts being more appropriate for those students. However, neither the estimated effect for grade 6 nor the difference in estimated effect between grade 6 and the higher grades is statistically significant. We also used quantile regressions to test whether flip charts had a greater impact for lower-ability students and found that the coefficients from the quantile regressions did not differ with those from mean regressions by more than 1% of a standard deviation and remained insignificant.

Note that even the lowest retrospective estimate implies that the program should have raised scores by 4% of a standard deviation, while the levels retrospective estimator suggests that it should have raised scores by 20% of a standard deviation. The latter possibility is rejected by the prospective study, although the former is within a 95% confidence interval.

6. Missing data and potential biases

The results of both the prospective and retrospective evaluations could be biased if the probability of observing the test score of pupils of different ability were affected differentially by the flip charts. In particular, both estimates could be biased downward if flip charts induced more low-ability students to take the exams. There are no data to check this for the retrospective estimates, but in the prospective study, absenteeism rates for each exam were very similar in the treatment and control schools. Probit regressions (not shown here) reveal that the differences in absenteeism between the two types of schools are not significant, and the magnitude of the differences is small enough that even if it were the worst students that missed the tests, the effect on the average result would be small.

More specifically, absenteeism rates for flip chart and comparison schools in 1997 were 2.2% and 2.4%, respectively, for the practice exam and 1.0% and 1.2% for the KCPE exam. Absenteeism rates for flip chart and comparison schools for grade 8 in 1998 were 6.3% and 3.8% for the practice exam and 3.5% and 3.1% for the KCPE exam; absenteeism for the practice exam was 10.8 and 10.8 for grade 6 and 9.4 and 6.9 for grade 7. The largest differences were for the 1998 grade 7 and 8 practice exams, where absenteeism for flip-chart schools was 2.5 percentage points higher. If the marginal student was one standard deviation below the mean on each individual test, this difference in absenteeism would lead to an *overestimate* of the relative performance of the treatment schools by 2.5% of a standard deviation on these two tests; the differences for the other exams would be trivial. The assumption that non-takers would score one standard deviation below the mean is probably extreme given that in 1998, 8th graders who did not take the KCPE scored only 0.1 standard deviations on the practice exam below those who did. It is therefore unlikely that absenteeism is responsible for the results.

In addition, due to illegible or lost score recording sheets or non-administration of the exam, 6, 14, 15, 14, 1, and 2 schools were missing scores for the 1997 grade 8 practice and KCPE, 1998 grades 6–8 practice, and 1998 grade 8 KCPE, respectively. Roughly half of the school missing data were comparison schools (4, 8, 6, 8, 1, and 2, respectively). The missing schools were roughly average performers on the 1996 tests, so their omission should not systematically affect the results, and any effect should be mitigated by controlling for 1996 test performance. For example, the difference between the average raw 1996 score for the schools with data and for all the schools was +0.45 and +0.25 points (out of 100 points) for the flip chart and comparison groups in 1997, respectively. In terms of standard deviations, this corresponds to average individual test differences of 0.6 and 0.4% of a standard deviation, respectively. In 1996, excluding the schools with missing 1997 data would therefore have lead to an overestimate of the relative scores of the flip-chart schools on the average individual test by 0.2% of a standard deviation. Assuming that any effect on the 1997 and 1998 results would be roughly of this magnitude, there is little reason to think that the inclusion of the missing schools would materially affect the results.

7. Conclusion

There are two possible explanations for the discrepancy between retrospective and prospective estimates of the effect of flip charts. The first is that students, teachers, or parents cut back on other inputs in response to the provision of flip charts and that they did so sufficiently to almost fully offset the impact of flip charts themselves. Under this interpretation, the retrospective results accurately estimate the partial derivative of test scores with respect to flip-chart provision, while the prospective estimate accurately measures the total derivative. The second interpretation is that the retrospective estimates are subject to omitted variable bias and schools with flip charts have other unobserved characteristics that increase test scores, such as good headmasters. It would be possible to distinguish between these hypotheses using the results from the prospective study if we had been able to collect data on other inputs, such as the time students spent on homework. We conducted such analyses in another study of textbook provision, but it was not possible in this case.

However, for a variety of reasons we feel that it is unlikely that schools and parents cut back dramatically on other inputs in response to flip charts. First, some inputs would be difficult to change rapidly in response to flipcharts. For example, classrooms are durable and even textbooks are often handed down from one sibling to another within a family. Second, many other inputs affect not just one subject, but all subjects together, so, for example, if pupils cut back on school attendance in response to flip-chart provision, test scores would go up in flip-chart subjects relative to non-flip-chart-subjects. Yet we do not observe this. Third, it would be difficult to reconcile the substitution hypothesis with rationality; charts are so cheap that almost all schools should use them if they are effective yet most schools do not have flip charts. Fourth, in the retrospective analysis flip charts were uncorrelated with other inputs; hence, the estimated effect of flip charts in the retrospective study was almost identical whether or not one controlled for other inputs. This is consistent with the hypothesis that flip charts have little effect and that their provision stimulates little endogenous response in other inputs. There is no evidence that schools cut back on other inputs in response to provision of flip charts.

Fifth, the retrospective difference-in-differences results comparing performance of schools with flip charts in subjects with and without flip charts also suggest a modest or negligible structural impact of flip charts on test scores. In particular, the absence of substantially higher test scores in the subjects in which flip charts were present could only be reconciled with a large positive effect if flip charts substitute for subject-specific inputs, rather than inputs such as the quality of classrooms or school attendance which would operate across all subjects. However, teacher time is not likely to be substituted across subjects, given that a fixed time period is allocated to each subject. Moreover, insofar as the flip charts are visually attractive, and represent a break from traditional teaching methods, it seems likely that student effort is a complement rather than a substitute for flip charts.

For all these reasons, we conclude that the second hypothesis is much more likely, i.e., that schools with flip charts have other unmeasured advantages that allow them to achieve higher test scores, but that flip charts have little impact, direct or indirect, on test scores in the Kenyan environment we examined.[9] The view that retrospective regressions would overestimate the partial derivative (structural impact) makes sense in a developing country context, where compensatory programs (which could lead to underestimation) are rare. More subtle retrospective analyses that compare test scores across subjects yield lower estimates of the impact, which may indicate that this approach is less likely to yield biased estimates. Yet such techniques are not applicable for inputs that affect all subjects, such as school buildings or smaller pupil–teacher ratios, and they could easily go astray in other contexts, given the ordinal nature of test score data.

Note finally that our conclusion that retrospective estimates overestimate reduced form impacts, if robust, is a sobering one; it suggests that Hanushek's (1995) pessimistic conclusions about the effects of school inputs in developing countries based on retrospective studies will be strengthened, rather than weakened, by prospective studies.

Acknowledgements

This paper is the outcome of a project that is a joint undertaking of many people. Assistance in data collection and analysis was provided by Chris Asoka, Irene Majale, Robert Namunyu, Caroline Nekesa, Stacy Nemeroff, Michael Wambetsa, Polycarp Waswa, and Maureen Wechuli. Invaluable assistance was provided by the staff of International Christelijk Steunfonds: Damary Alembi, Chip Bury, Jos Huizinga, Japheth Mahagwa, Pamela Oranja, and Susan Walji. We thank our discussant at Stanford, Ed Vytlacil, and two anonymous referees. 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. Financial support was provided 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.

[9] If the schools in the prospective sample already had some flip charts, the negligible impact of the prospective estimates could be explained by diminishing marginal impacts, as suggested by one referee. Yet flip charts are rare in rural Kenyan schools. While we have no data for the prospective sample, we do have data on the schools in the retrospective sample. About four-fifths of those schools had no flip charts at all. In the one-fifth that did have charts, the mean number of charts is 4.1, which is comparable to the set of charts provided to schools in the prospective study. Thus, even if there are diminishing marginal effects of flip charts, the "dose" of flip charts for the schools that had them is about the same in the retrospective and prospective analyses.

They do not necessarily represent the views of the National Science Foundation, the World Bank, its Executive Directors, or the countries they represent.

References

Davis Jr., O.L., 1968. Effectiveness of Using Graphic Illustrations with Social Studies Textual Materials. Final Report Kent State University, Ohio.

Dunn, R., et al., 1989. Learning Styles Inventory. National Association of Secondary Principles, Reston, VA, USA.

Dwyer Jr., F.M., 1970. Exploratory studies in the effectiveness of visual illustrations. AV Communication Review, 235–249.

Glewwe, P., Kremer, M., Moulin, S., 2004. Textbooks and Test Scores: Evidence from a Prospective Evaluation in Kenya. Policy Research Group. The World Bank.

Hanushek, E., 1995. Interpreting recent research on schooling in developing countries. World Bank Research Observer, 227–246 (August).

Holliday, W.G., 1973. A Study of the Effects of Verbal and Adjunct Pictorial Information in Science Instruction. Mimeo, Ohio State University.

Holliday, W., Benson, G., 1991. Enhancing learning using questions adjunct to science charts. Journal of Research in Science Teaching, 97–108 (January).

Krueger, A.B.,Whitmore, D.M. 2000. The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Results: Evidence from Project STAR. NBER Working Paper 7656.

LaLonde, R.J., 1986. Evaluating the econometric evaluations of training programs with experimental data. American Economic Review 76 (4), 604–620 (September).

Levin, J.R., et al., 1976. Pictures, repetition, and young children's oral prose learning. AV Communication Review 24 (4), 367–380 (Winter).

Lookatch, R.P., 1995. The strange but true story of multimedia and the type I error. Technos, 10–13.

Samuels, S.J., 1970. Effects of pictures on learning-to-read, comprehension, and attitude. Review of Educational Research 16, 397–407.

Shepard, R., 1967. Recognition Memory for Words, Sentences, and Pictures. Journal of Verbal Learning and Verbal Behavior 6.

Wallace, J., 1995. Accommodating elementary students' learning styles. Reading Improvement, 38–41.