Hagamos Ciencia Program Interim Report III

Presented to the Republic of Panama

May 26, 2009

The purpose of this report is to brief educators and government officials about the Pilot II study and the implications of the findings for the next evaluation study. The Pilot II study was carried out in order to evaluate the impact of the *Hagamos Ciencia* (HC) Program on student achievement. This report focuses on Phase 2 - Development of an Evaluation Instrument and Methodology. The report is an extension of the information provided in the previous two interim reports sent in November 2008.

Developing a Technically Sound Assessment Instrument

As described in the first interim report, developing a technically sound assessment instrument consists of several steps: (1) Item development; (2) Developing administration procedures; (3) Piloting the items to evaluate their technical quality; (4) Revising the items and developing new ones, if necessary; and (5) Piloting the items again to establish the item parameters.

The previous two reports focused on the first four steps: Item Development, Administration Procedures, Piloting the assessments, and Revising the items. The main topic of this report is the second pilot (Pilot II) conducted in November 2008, in which both the revised version of old items and newly developed items were pilot again with a larger sample. We also discuss another important topic–some methodological considerations for conducting case studies.

Overall, the second pilot was successful in providing more accurate information about the quality of the items, the quality of the administration materials, and more insights about the achievement of students in the HC program. Whereas large-scale test development of this type usually occurs over the course of many months, perhaps an entire year, it is important to point out that this project only began nine months ago and we are reporting about the implementation of the second pilot test.

Pilot II Module Assessment Booklets

As in Pilot I two booklets per module were compiled and tested. Informed by Pilot I, the items for Pilot II were modified based on findings from the item analysis and talk aloud protocols. The booklets for each of the three modules include exactly the same items but in different orders. Due to the perceived difficulty by students in switching back and forth from questions about specific modules and more general questions about national curriculum topics, booklets for Pilot II were designed such that all module items were placed sequentially and all national curriculum items were placed sequentially. Within each set (module or national curriculum), the sequence of items was randomly assigned after ensuring that the first two items were easy, based on the p values and/or on the easiness of the newly developed items. Booklet 1 includes as the first set the module items, followed by the set of national curriculum items, and Booklet 2 has first the national curriculum set followed by the module items set. This design allows us to ascertain whether it makes a difference which set of items is presented to

students first on the test. The two-booklet design complicates the test administration process in that it requires a spiral distribution of booklets within each classroom, as was carried out for the first pilot administration.

For the Soils booklets (third grade), there were 25 multiple-choice test items for each booklets and one short-answer question. For the Circuits booklets (fourth grade) and the Ecosystems booklets (sixth grade) there were 30 multiple-choice test items on each booklet and one short-answer question. In this report, we only report about the multiple-choice items.

Booklets were sent to SENACYT on November 19th, 2008. The Pequeño Científico Group conducted a final revision of the booklets.

Pilot II Sample Demographics

Two pieces of information obtained from Pilot I were critical for making some decisions for the Pilot II sampling: information about the effect size and information about the intra-class correlation. The magnitude of the effect size helps to decide a cost effective sample size.[1] Since effect size is population-specific; and no preliminary evidence was available about effect sizes with respect to the effect of HC, the analyses conducted in Pilot I were critical in gathering this particular piece of information. Effect sizes were calculated and analyzed to determine the impact of the HC program.

Another critical piece of information obtained in Pilot I that was useful to determining the Pilot II sample size was the intra-class correlation (ICC).[2] In Pilot I, ICC was calculated for different units of analysis, teachers and schools and within area, urban and rural (see Report II). Therefore, as the ICC increases from the teacher level to the school level, the focus then shifts to sampling more schools, and perhaps fewer students within schools.

Sample II Demographics. Considering the ICC found in Pilot I, the sample selected for Pilot II consisted of 136 schools with approximately 20,000 students. The selected sample was provided to Maria Heller on November 20th, 2008. Sampling was conducted proportionally to represent the number of schools by region. After administering the assessments, the resulting sample consisted of 118 schools from nine regions, 927 teachers, and 19,177 students. Information by grade and module is presented in Table 1.

---

[1] Effect size is a key piece of information to determine the *power* of a statistical test. Whereas a statistical significance level (e.g., p = .01) considers the null hypothesis to be rejected, power considers the researchers' hypothesis as true.

[2] ICC is a measure of homogeneity among analytical units; it provides a summary glance into the variability (variance) structure of the data. ICC helps to determine sample sizes for hierarchical populations like the one studied in HC (e.g., students nested within teachers and teachers nested within schools). For example, a low ICC (less than 0.25) indicates relatively small between-school variation. As ICC increases (greater than 0.25), between-school variation increases; that is some schools perform better than others. Therefore, the larger the ICC, which indicates variability between units of analysis (e.g., teachers or schools), the larger the sample required.

Table 1. *Pilot II Sample Demographics*

Participation Characteristic	Grade 3 Soils		Grade 4 Circuits		Grade 6 Ecosystems	
	Booklet 1	Booklet 2	Booklet 1	Booklet 2	Booklet 1	Booklet 2
Type of Group						
HC Program	2069	2038	2139	2186	2209	2215
Non-HC Program	1008	1035	1019	1022	1113	1124
Area						
Urban	2306	2305	2343	2367	2626	2624
Rural	771	768	815	841	696	715
Gender*						
Males	1611	1511	1610	1680	1668	1716
Females	1465	1461	1548	1528	1651	1622
Total	3077	3073	3158	3208	3322	3339

* Gender - The number of 'Total' was larger than the sum of 'Males' and 'Females' in some booklets since few students did not complete the demographic question about gender.

The numbers of male and female students were close to each other across all grades. The number of students who participated in the experimental group (HC program) was double of those who were part of the control group (non-HC program). The students from urban schools outnumbered those from rural schools in the ratio of three to one. Despite the effort to maintain both urban and rural groups of similar size within the HC or non-HC program, this was not possible since many schools had already enrolled in the HC program. On average, 91% of the students sampled in rural schools were part of the HC program. In contrast, only 60% of the students sampled in urban schools were part of the HC program (see Tables 6b and 7b for the break down counts).

As it was the case in Pilot I, for some students one or more demographic fields were left blank, resulting in the total number of responses occasionally being larger than the sum of the demographic subtotals (e.g., see note in Table 1 about gender). Likewise, in Pilot I, administration issues that were out of control of UC Denver (e.g., roads were closed or schools were flooded due to the intense rain season) partially contributed to this unbalanced design.

Booklets and Items. For any given grade level, half of students took Booklet I and another half took Booklet II, even though the split between the two booklets was more uneven at the class level. This artifact was probably due to the quality in the process of administrating the assessments. Overall, more students took the Ecosystems assessment (*n*=6661) than the Circuits assessment (*n*=6366) and the Soils assessment (*n*=6150).

Item Analyses

Two critical purposes of item analysis are to determine the *difficulty* and *discrimination* of each item. It is expected that important insights into the students' thinking and understanding of the content being assessed is found. We performed the item analysis with both the Classic Test Theory (CTT) and the Item Response Theory (IRT) approaches. Taking the CTT approach, difficulty is represented by the item difficulty index, named *p value*[3]; discrimination is represented by a discrimination index, D.[4] In the IRT approach, difficulty is represented by the *b* parameter, also called the item location[5]; discrimination is represented by the *a* parameter, also called the slope.[6] It is important to note that lower discrimination values are associated with lower test reliability. Table 2 provides information about the item analysis by booklet.

Item difficulty estimate was calculated for the following groups using the CTT approach: *Gender* - male and female groups, *Treatment group* - experimental or control groups students, *Alignment* – module and national curriculum, and *Area* - rural and urban groups. Due to the complexity of the IRT calibration process, the IRT analysis was only performed for the whole sample.

First, the level of difficulty of the items became easier along with the school grades. That is, the lowest *p* values were found for the Soils Assessment items, and highest for the Ecosystems Assessment items (from .37 for third grade, .42/.43 for fourth grade, to .49/.50 for sixth grade) compared to the counterparts of those (from .30 for third grade, .30 for fourth grade, to .36/.39 for six grade) in the Pilot I study. A similar pattern was also observed in the IRT analysis in which the *b* parameter values decreased from lower grades to sixth grade.

---

[3] Difficulty (p value) is the proportion of students who answered the item correctly. Items with p values < 0.25 are considered relatively difficult and items with p values > 0.75 relatively easy; items with p values around .050 are preferred.

[4] Discrimination measure the extent to which a test item discriminates or differentiates between students who do well on the overall test and those who do not do well on the overall test. The higher the value, the better. Items with discrimination of .40 or higher are considered to have an excellent discrimination. Items with a value from .30 to .39 are considered with good discrimination. Items with values below .29 are considered with undesirable discrimination.

[5] Item location refers to the point on the latent trait scale at which the probability of correctly answering an item is .50. The estimated values mostly fall into the range of -3 to 3. The greater the value of item location, b, the harder the item because the greater the ability that is required for a student to have 50% chance of getting the item right. Items with negative values of the b parameters are considered relatively easier; items with positive values are considered difficult; and items with the b parameters around 0 are preferred.

[6] Item slope indicates the degree to which an item discriminates between individuals varying in the latent trait. This parameter characterizes the slope of the item characteristics curve where the slope is at its maximum. The estimates mostly fall into the range of -2.80 to 2.80 as a higher positive value suggests a higher discrimination.

Table 2. *Item Characteristics Summary*

Participation Characteristic	Grade 3 Soils		Grade 4 Circuits		Grade 6 Ecosystems	
	Booklet 1	Booklet 2	Booklet 1	Booklet 2	Booklet 1	Booklet 2
Cronbach's Alpha	0.678	0.656	0.732	0.757	0.726	0.756
General						
Item Difficulty (CCT)	0.37	0.37	0.43	0.42	0.49	0.50
Discrimination (CCT)	0.34	0.33	0.34	0.35	0.33	0.35
Item Difficulty (IRT)	1.16	1.16	0.96	1.04	0.18	0.00
Discrimination (IRT)	0.41	0.38	0.42	0.44	0.42	0.45
Gender Item Difficulty (CCT)						
Female	0.37	0.38	0.44	0.43	0.50	0.51
Male	0.36	0.36	0.43	0.42	0.49	0.50
Treatment Item Difficulty (CCT)						
HC Group	0.38	0.38	0.44	0.42	0.50	0.50
Control Group	0.33	0.35	0.43	0.41	0.49	0.51
Alignment Item Difficulty (CCT)						
Module Based	0.35	0.33	0.46	0.42	0.51	0.49
Curriculum Based	0.39	0.43	0.40	0.42	0.48	0.52
Area item Difficulty (CCT)						
Urban	0.37	0.37	0.44	0.43	0.50	0.51
Rural	0.36	0.35	0.41	0.39	0.47	0.48

In order to further explore the characteristics of individual test items, student responses for each multiple-choice item within booklets were extensively analyzed to determine the following pieces of information: percentage of students who responded to each option (correct answers and the rest of distractors), percentage of students who did not respond, percentage of students whose responses were unclear (e.g., more than one option), correlation to adjusted total booklet score, contribution to reliability of the booklet score, response differences between boys and girls, and response differences between HC and control school students. Each of these pieces of information was compiled to represent the psychometric properties for individual test items and statistical analysis. Appendix A, an Excel file, provides all the item analysis results at the item level for all items across grades.

In general, results of the item analysis indicated that the assessment items in the Pilot II study were easier to answer correctly compared with those in the Pilot I study. In Pilot II, the p values were closer to .50 compared to in Pilot I, allowing for larger variation of total test scores which contributes to higher reliability. For most testing purposes, items closer to .50 difficulty level are preferable and often appropriate for item selection. In addition, the test items were free of gender bias across all grades with a finding of a maximum .02 difference (ranging from zero to .02) of difficulty index between males and females as well as free of group bias across all grades with only a maximum .05 difference (ranging from .01 to .05) between the HC program and the control group.

Second, discrimination index (*D*), which differentiates higher ability/ more knowledgeable students from lower ability/less knowledgeable students, was also higher in Pilot II than in Pilot I. Item analysis showed that all of the indices were over .30 (ranging from .33 to .35) across all booklets and grades. The test items have a good discrimination (and some do have excellent discrimination, around .5) to detect differences of student performance compared with those indices (ranging from .20 to .28) of the test items in Pilot I study.

Finally, the differences of item difficulty indices between module-based items and national curriculum-based items remained small for all grades/booklets (ranging from zero to .06) except for Soils booklet 2 on third grade with a index of .1 (.43 for national curriculum-based items and .33 module-based one). Overall, curriculum-based items were easier for students to answer correctly than module-based ones for third graders. However, for fourth and sixth graders, no consistent patterns were found with respect to which type of items was easier to get right answers.

Item Selection for Data Analysis

The findings of item analysis across booklets guided the selection of items that were used for data analyses. That is, those items with inappropriate technical characteristics were excluded from the booklets (both 1 and 2 within a module). Table 3 provides information about the items that were kept by module and grade. Hereafter the reported results in this report only involve these selected items.

Table 3. *Items Discarded by Module and Grade*

	Grade 3 Soils	Grade 4 Circuits	Grade 6 Ecosystems
Original Number of Items	25	30	30
No. of Items Discarded	3	5	2
No. of Items Used for Analysis	22	25	28
Module-Based			
No. of Items	14	15	16
Averaged *p* Values	0.345	0.462	0.513
Curriculum-Based			
No. of Items	8	10	12
Averaged *p* Values	0.438	0.440	0.523

Table 4 provides the summary of the item analysis only for the selected items. The Cronbach's alpha coefficients (which will be discussed in detail later) are slightly higher when those poorly-performing items were dropped. Regarding item difficulty indices, the module items were relatively more difficult compared to the national curriculum items at Grade 3 only. At fourth and sixth grades, the two clusters of items exhibited similar difficulties when averaged across the two booklets.

Table 4. *Item Characteristics Summary of the Selected Items Only*

Participation Characteristic	Grade 3 Soils		Grade 4 Circuits		Grade 6 Ecosystems	
	Booklet 1	Booklet 2	Booklet 1	Booklet 2	Booklet 1	Booklet 2
New Cronbach's Alpha	0.688	0.665	0.748	0.767	0.742	0.766
General						
Item Difficulty (CCT)	0.38	0.38	0.46	0.45	0.51	0.52
Discrimination (CCT)	0.36	0.35	0.37	0.38	0.35	0.37
Item Difficulty (IRT)	1.09	1.04	0.51	0.82	0.19	-0.23
Discrimination (IRT)	0.76	0.71	0.63	0.62	0.54	0.63
Gender Item Difficulty (CCT)						
Female	0.38	0.39	0.46	0.45	0.52	0.53
Male	0.37	0.37	0.46	0.44	0.50	0.52
Treatment Item Difficulty (CCT)						
HC Group	0.39	0.35	0.45	0.44	0.51	0.53
Control Group	0.34	0.39	0.46	0.45	0.52	0.52
Alignment Item Difficulty (CCT)						
Module Based	0.35	0.34	0.48	0.44	0.52	0.50
Curriculum Based	0.42	0.46	0.42	0.45	0.50	0.55
Area item Difficulty (CCT)						
Urban	0.38	0.39	0.47	0.46	0.52	0.53
Rural	0.37	0.36	0.43	0.41	0.49	0.50

Preliminary Reliability Indicators

An initial indicator of the technical quality is the internal consistency (Cronbach's Alpha) of the items selected for each booklet. Cronbach's alpha coefficient represents the magnitude of homogeneity of the test items in an assessment. The more interrelated the test items selected for an assessment are, the higher the internal consistency, and the higher the Cronbach's Alpha is. When no test items were eliminated from the original booklets, the Cronbach's alpha coefficients ranged from .656 to .757 across the six booklets (see Table 2). After the item selection, the Cronbach's Alpha improved and the range was from .665 to .767. Although these magnitudes do not meet the requirement (.80), they are far higher than those observed in Pilot I (Alpha = .460 to .636). Furthermore, it has been argued that for large-scale studies a coefficient equal or higher than .70 can be considered appropriate (Davis, 2006).

This information needs to be carefully interpreted. First, it is important to remember that the magnitude of the p values and D values contributes to the internal consistency of the items. When no variability is observed (either because most of the students responded correctly or incorrectly) which often leads to extremely high or low p values and lack of discrimination, the reliability is affected. Second, since several new items were developed for the Pilot II study it should be expected that in order to increase the

reliability of the booklets, another round of revising and piloting the items should be carried out. However, due to the cost that this would imply, doing so was not feasible.

Preliminary Results on the Impact of "Hagamos Ciencia"

All Sample

As mentioned, to conduct the statistical analyses, only those items that had the appropriate technical quality were considered. Tables 5a and 5b summarize the descriptive by booklet within grade and group (Hagamos Ciencia and Control).

Table 5a. *Descriptives by Grade and Booklet for the Complete Sample*

Type of Group	Grade 3 - Soils		Grade 4 - Circuits		Grade 6 - Ecosystems	
	Booklet Max = 22		Booklet Max = 25		Booklet Max = 28	
	1	2	1	2	1	2
Hagamos Ciencia Group						
n	2069	2038	2139	2186	2209	2215
Mean	8.65	8.64	11.59	11.24	14.43	14.55
S.D.	3.97	3.80	4.70	4.81	4.85	5.06
S.E. of the Mean	0.087	0.084	0.102	0.103	0.103	0.107
Control Group						
n	1008	1035	1019	1022	1113	1124
Mean	7.58	7.80	11.32	10.97	14.22	14.72
S.D.	3.21	3.25	4.19	4.48	4.62	4.83
S.E. of the Mean	0.101	0.101	0.131	0.140	0.139	0.144
Statistical Analyses						
t Independent	7.99	6.38	1.65	1.56	1.20	-0.94
p	0.00	0.00	0.10	0.12	0.23	0.35
Effect Size	0.33	0.26	0.07	0.06	0.04	-0.04

Table 5b. *Descriptives by Grade Only for the Complete Sample*

Type of Group	Grade 3 - Soils	Grade 4 - Circuits	Grade 6 - Ecosystems
	Max = 22	Max = 25	Max = 28
Hagamos Ciencia Group			
n	4107	4325	4424
Mean	8.64	11.41	14.49
S.D.	3.88	4.76	4.95
S.E. of the Mean	0.061	0.072	0.074
Control Group			
n	2043	2041	2237
Mean	7.69	11.14	14.47
S.D.	3.23	4.34	4.73
S.E. of the Mean	0.072	0.096	0.100
Statistical Analyses			
t Independent	10.16	2.25	0.14
p	0.00	0.025	0.89
Effect Size	0.26	0.06	0.00

A new piece of information in these tables is the *standard error of the mean* (S.E.). The S.E. is an estimated value of the precision of a selected sample; that is, the closeness with which it can be expected to approximate the relevant population value by using the mean of a sample, since this value is generally unknown. The greater the value of the S.E., the greater the degree to which means of different samples vary among themselves and the less any of them can be relied upon. In sum, S.E. reflects the precision of the sample value at hand. For example, a S.E. of .10 (see Control Group at Grade 6 in Table 5b) indicates that sample means based on 2237 students can be expected to have a variability of .1 assessment score units, as measured by their own standard deviation. The results in Table 5a and 5b indicate that the magnitudes of the S.E. estimates are very small. Therefore, it can be argued that the samples used in Pilot II are quite precise and reliable.

A significant difference between the HC and the control groups was observed only in third grade, but not in fourth or sixth grade. The HC third grade students performed, on average, significantly better than those in the control group ($p < 0.001$). Such a difference was observed in both booklets with an averaged effect size of .30 (a small effect). It is important to remember that an effect size measures the strength of the effect of the HC-program on student performances compared with the non-HC program of students. The larger the value of an effect size, the greater the degree to which the phenomenon (e.g., effect of a treatment) under study is manifested.

On the other hand, no statistical significance of students' test scores was found between the HC-program and the non-HC program for either of booklets at fourth and sixth grades, indicating no effect of the treatment (HC-program). Furthermore, in Booklet 2 for sixth graders, student performance of the control group (non-HC program) outperformed their peers in the experimental group (HC-program), though not statistically significant. Further investigation is needed to explore the possible reasons resulting in this finding.

Figure 1 presents the averaged effect sizes by module (linked to grade) and clusters of items: all items, module items, and national curriculum items. Apparently the highest effect sizes are observed in the module-based items and the lowest in the national curriculum items. The students in the control group yielded similar or even better scores in the national curriculum items than their counterparts in the HC program. Indeed the six graders in the control group had a fairly better grasp of the national curriculum materials as compared to their peers' learning of the Ecosystems module. We interpret the results as evidence that suggests that the positive impact of the HC program was salient only in the module items. Its beneficial effect has not been transferred by teachers or students to the learning of the national curriculum materials.

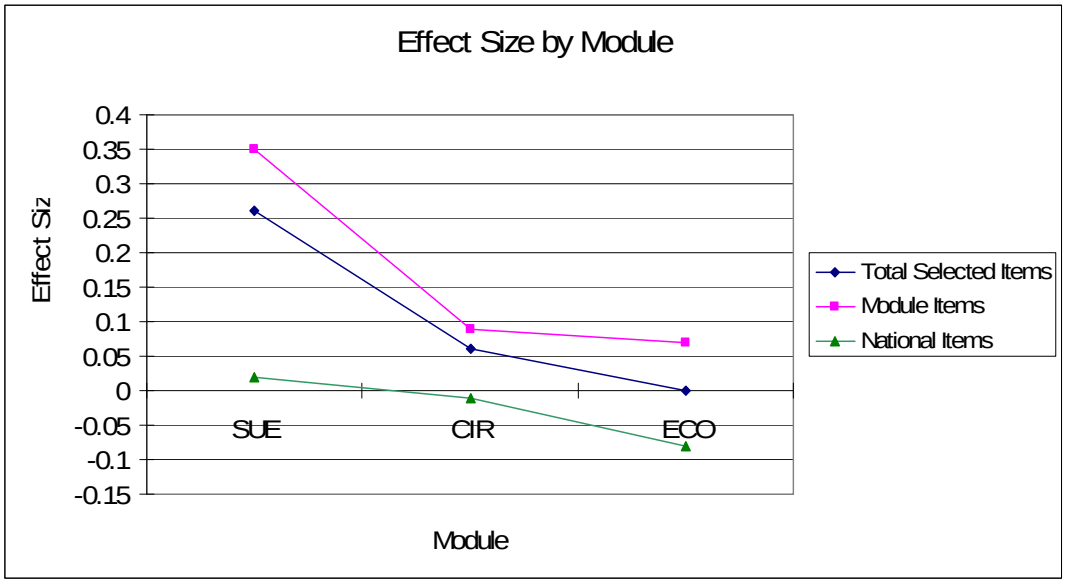


Figure 1. Effect sizes by cluster of items across modules.

Urban and Rural Samples

Results by area, urban or rural, show similar patterns as those described for the complete sample, although the effects sizes clearly varied by area. Tables 6 (a and b) and 7 (a and b) present the results by area.

All Selected Items. In urban area (Tables 6a and 6b), students in the HC program significantly outperformed those in the control group across booklets in third and fourth grade ($p < .001$), but their performance became evened out at sixth grade, showing no significant difference (p=.26 for Booklet 1, p=.91 for Booklet 2, and p=.49 when two booklets were analyzed together). Effect sizes decrease as school grade increases with an averaged .33 of effect size for third grade ($p < 0.01$), an averaged .18 of effect size for fourth grade ($p < 0.01$), and .02 for sixth grade (no statistical significance).

Table 6a. *Descriptive by Grade and Booklet in Urban Areas*

Type of Group	Grade 3 Soils		Grade 4 Circuits		Grade 6 Ecosystems	
	Booklet Max = 22		Booklet Max = 25		Booklet Max = 28	
	1	2	1	2	1	2
Urban Hagamos Ciencia Group						
n	1363	1330	1386	1408	1600	1596
Mean	8.87	9.00	12.14	11.80	14.59	14.80
S.D.	3.94	3.83	4.64	4.80	4.77	5.02
S.E. for Mean	0.107	0.105	0.125	0.128	0.119	0.126
Urban Control Group						
n	943	975	957	959	1026	1028
Mean	7.65	7.84	11.29	10.97	14.38	14.82
S.D.	3.23	3.24	4.14	4.46	4.61	4.83
S.E. of the Mean	0.105	0.104	0.134	0.144	0.144	0.151
Statistical Analyses						
t Independent	8.12	7.87	4.65	4.24	1.12	-0.12
p	0.00	0.00	0.00	0.00	0.26	0.91
Effect Size	0.33	0.32	0.19	0.18	0.04	0.00

Table 6b. *Descriptive by Grade in Urban Areas*

Type of Group	Grade 3 Soils	Grade 4 Circuits	Grade 6 Ecosystems
	Max = 22	Max = 25	Max = 28
Urban Hagamos Ciencia Group			
n	2693	2794	3196
Mean	8.93	11.96	14.69
S.D.	3.88	4.72	4.89
S.E. of the Mean	0.075	0.089	0.087
Urban Control Group			
n	1918	1916	2054
Mean	7.75	11.13	14.60
S.D.	3.24	4.31	4.73
S.E. of the Mean	0.074	0.098	0.104
Statistical Analyses			
t Independent	11.29	6.27	0.69
p	0.00	0.00	0.49
Effect Size	0.33	0.18	0.02

In the rural areas (Tables 7a and 7b), the differences of averaged assessment scores between the two groups, Hagamos Ciencia group and control Group, varied from grade to grade and from booklet to booklet. Because of the small sample size in each booklet, we mainly discuss the interpretation based on Table 7b instead of Table 7a in this report. Still, the sample of students in the control group is too small compared to the HC group. Therefore, the results must be interpreted with extra caution.

Table 7a. *Descriptive by Grade and Booklet in Rural Areas*

Type of Group	Grade 3 Soils		Grade 4 Circuits		Grade 6 Ecosystems	
	Booklet Max=22		Booklet Max=25		Booklet Max=28	
	1	2	1	2	1	2
Rural Hagamos Ciencia Group						
n	706	708	753	778	609	619
Mean	8.24	7.95	10.59	10.24	14.00	13.91
S.D.	3.99	3.64	4.65	4.65	5.01	5.11
S.E. of the Mean	0.150	0.137	0.170	0.167	0.203	0.205
Rural Control Group						
n	65	60	62	63	87	96
Mean	6.62	7.12	11.82	10.89	12.36	13.66
S.D.	2.84	3.34	4.96	4.88	4.40	4.72
S.E. of the Mean	0.352	0.431	0.630	0.615	0.472	0.482
Statistical Analyses						
t Independent	4.24	1.71	-1.99	-1.07	2.89	0.47
p	0.00	0.09	0.05	0.29	0.00	0.64
Effect Size	0.41	0.23	-.26	-0.14	0.33	0.05

Unlike the students in the urban setting, the rural HC students outperformed their rural counterparts in the control group at third and sixth grade with a small effect size (effect size=.33 for third grade and effect size=.18 for sixth grade). However, students in the rural HC program were out-performed by those in the control group with small effect size at Grade 4 (p=0.03, effect size=-.20). Although the standard error of the means are higher in the rural areas than in the urban areas, it is important to note that the magnitude of the values are still small and not close to even one standard deviation.

Taking together, the findings of urban and rural schools confirm the positive effects of the HC program in third grade for the Soils module. However, findings are less conclusive for the HC modules at upper grades. The school setting, urban or rural, may play an unexpected important role in the implementation of the HC program mediated by characteristics of students and their teachers, resources, and the training/assignment of the facilitators.

Table 7b. *Descriptive by Grade in Rural Areas*

Type of Group	Grade 3 Soils	Grade 4 Circuits	Grade 6 Ecosystems
	Max = 22	Max = 25	Max = 28
Rural Hagamos Ciencia Group			
n	1414	1531	1228
Mean	8.09	10.41	13.96
S.D.	3.82	4.65	5.06
S.E. for Mean	0.102	0.119	0.144
Rural Control Group			
n	125	125	183
Mean	6.86	11.35	13.04
S.D.	3.09	4.93	4.61
S.E. for Mean	0.276	0.441	0.340
Statistical Analyses			
t Independent	4.20	-2.16	2.31
p	0.00	0.03	0.02
Effect Size	0.33	-0.20	0.18

Module-Based Items. Urban students in the HC-program outperformed urban students in the control group across grades and booklets ($p < 0.01$) except for booklet 2 of sixth grade ($p > 0.05$). In the rural area, students performed better than those in the control group on both booklets 1 and 2 for third grade ($p < 0.01$) and on booklet 1 only for sixth grade ($p < 0.01$), but no significant difference on booklet 2 for sixth grade. However, for both booklets for fourth grade, students from the control group performed better than the students in the HC program.

National Curriculum-Based Items. In the urban area, when examining the curriculum-based items, the urban students of the HC program outperformed the control group students only on booklet 2 for third grade ($p < .05$) and on booklet 1 for 4^{th} grade ($p < .05$). In the rural area, the students from the HC program outperformed the students from the control group on booklet 2 for sixth grade ($p < .05$). Overall, the comparison between the two groups did not statistically favor the HC program in urban or rural for other booklets.

Regions

Nine regions participated in the Pilot II study: Chiriquí, Coclé, Colón, Herrera, Los Santos, Panamá Centro, Panamá Oeste, San Miguelito, and Veraguas. Figure 2 shows the magnitude of the effect sizes by region and Table 8 provides the detailed information by region and module. Appendix B, an excel file attached to the report, provides the descriptives for each teacher by region and module.

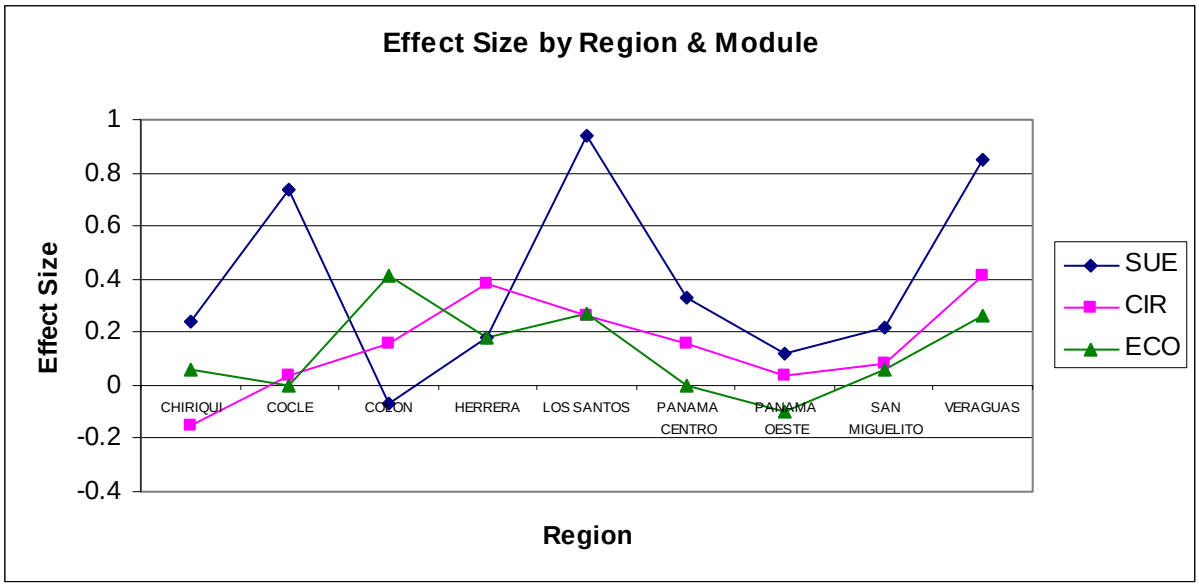


Figure 2. Effect sizes by region and module.

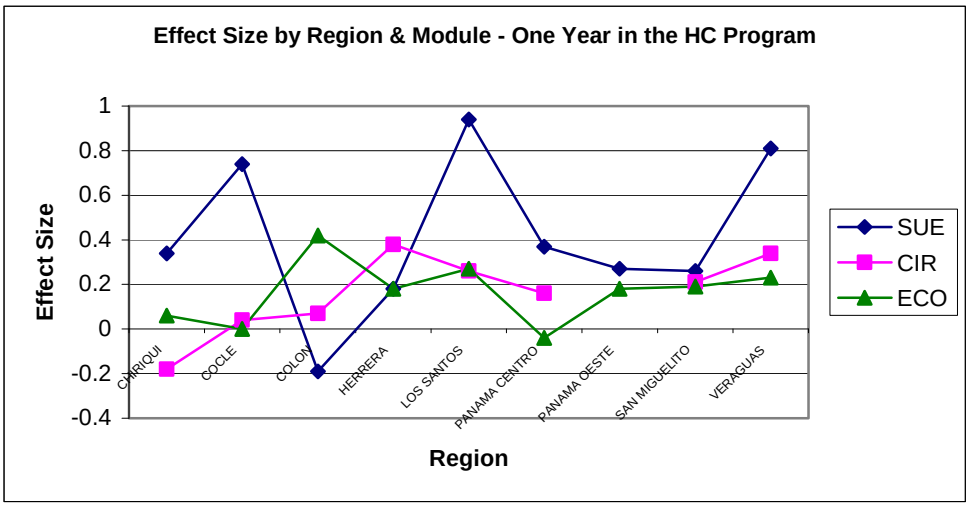
The effect sizes within region were not consistent across modules. That is, within the same region, the effect sizes may be high in third grade (Soils module) but low in sixth grade (Ecosystems module), such as Region Coclé. Overall, Region Veraguas had the highest averaged effect size across the three modules (averaged effect size = .51). Veraguas was followed by Los Santos (averaged effect size = .45). Both regions had a medium effect size, which means that in the two regions the effect of the Hagamos Ciencia program was the strongest among all regions. In addition, four regions had small effect size: Coclé (averaged effect size =.26), Herrera (averaged effect size = .25), Colón (averaged effect size = .17), and Panamá Centro (averaged effect size = .16). Finally, the smallest effect sizes were observed in two regions, Chiriquí (averaged effect size = .05) and Panamá Oeste (averaged effect size = .02). In sum, the HC program was mostly associated with positive effect on student learning across the regions except a few cases, such as the third graders in Colón, the fourth graders in Chiriquí, and the sixth graders in Panamá Oeste.

It is important to note that the magnitudes of the S.E. of the mean by region and module/grade are larger than those found for the complete sample. The highest S.E. was found in Los Santos, followed by Coclé, and Herrera (averaged S.E. of .43, .40, and .39 respectively), and the lowest in Panamá Oeste, San Miguelito, and Chiriquí (averaged S.E. of .18, .19, and .20 respectively). Panamá Centro, Veraguas, and Colón showed a slightly higher S.E. (averaged S.E. of .21, .24 and .25 respectively). Still, none of them had an S.E. close to one standard deviation.

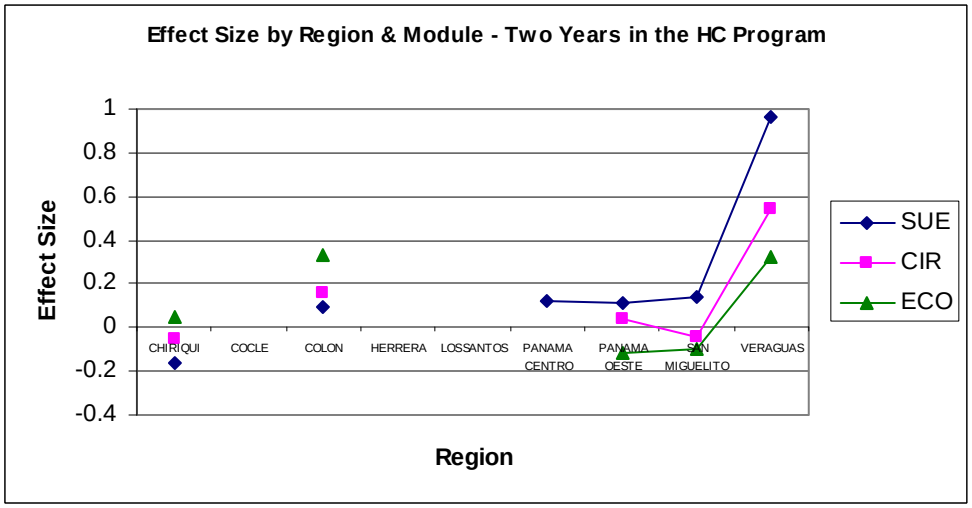
Table 8. *Descriptives by Group, Region, and Module*

	Chiriquí	Coclé	Colón	Herrera	Los Santos	Panamá Centro	Panamá Oeste	San Miguelito	Veraguas
SOILS Max =22									
Effect size	0.24	0.74	-0.07	0.18	0.94	0.33	0.12	0.22	0.85
Hagamos Ciencia Group									
n Schools	12	2	13	2	2	9	6	8	10
n of Stu.	679	84	802	121	168	485	655	490	623
Mean	9.25	10.04	6.83	9.68	9.43	8.67	7.74	8.64	10.66
SD	3.85	4.29	3.18	4.77	3.99	3.69	3.35	3.67	4.01
SE of Mean	0.15	0.47	0.11	0.43	0.31	0.17	0.13	0.17	0.16
Control Group									
n Schools	12	4	2	2	2	6	4	6	4
n of Stu.	365	216	121	130	45	312	396	268	190
Mean	8.34	7.43	7.05	8.95	5.82	7.57	7.35	7.89	7.37
SD	3.49	3.19	3.44	3.42	3.17	2.79	3.01	3.13	3.31
SE of Mean	0.18	0.22	0.31	0.30	0.47	0.16	0.15	0.19	0.24
CIRCUITS Max =25									
Effect size	-0.15	0.04	0.16	0.38	0.26	0.16	0.04	0.08	0.41
Hagamos Ciencia Group									
n Schools	12	2	12	2	2	9	5	10	11
n of Stu.	620	95	901	136	217	439	632	616	669
Mean	11.72	11.57	10.27	13.24	11.60	11.65	11.15	11.20	12.51
SD	4.74	5.02	4.41	4.72	4.56	4.68	4.93	4.80	4.72
SE of Mean	0.19	0.52	0.15	0.40	0.31	0.22	0.20	0.19	0.18
Control Group									
n Schools	11	4	2	3	3	7	4	7	5
n of Stu.	354	169	104	122	58	352	386	357	139
Mean	12.42	11.38	9.58	11.53	10.40	10.95	10.95	10.82	10.61
SD	4.54	4.45	3.32	4.26	4.74	4.31	4.40	4.03	4.29
SE of Mean	0.24	0.34	0.33	0.39	0.62	0.23	0.22	0.21	0.36
ECOSYSTEMS Max = 28									
Effect size	0.06	0.00	0.41	0.18	0.27	0.00	-0.10	0.06	0.26
Hagamos Ciencia Group									
n Schools	12	2	14	2	2	9	6	10	11
n of Stu.	694	90	714	128	185	506	859	623	625
Mean	15.34	15.36	12.83	15.66	14.95	13.93	14.02	14.97	15.57
SD	5.36	5.17	4.52	4.66	4.50	4.41	5.20	4.87	4.64
SE of Mean	0.20	0.54	0.17	0.41	0.33	0.20	0.18	0.19	0.19
Control Group									
n Schools	12	4	2	5	2	6	4	7	5
n of Stu.	372	216	86	141	85	294	482	365	196
Mean	15.03	15.36	10.99	14.79	13.72	13.95	14.52	14.67	14.36
SD	4.74	4.66	4.38	5.10	4.74	4.47	4.92	4.48	4.31
SE of Mean	0.25	0.32	0.47	0.43	0.51	0.26	0.22	0.23	0.31

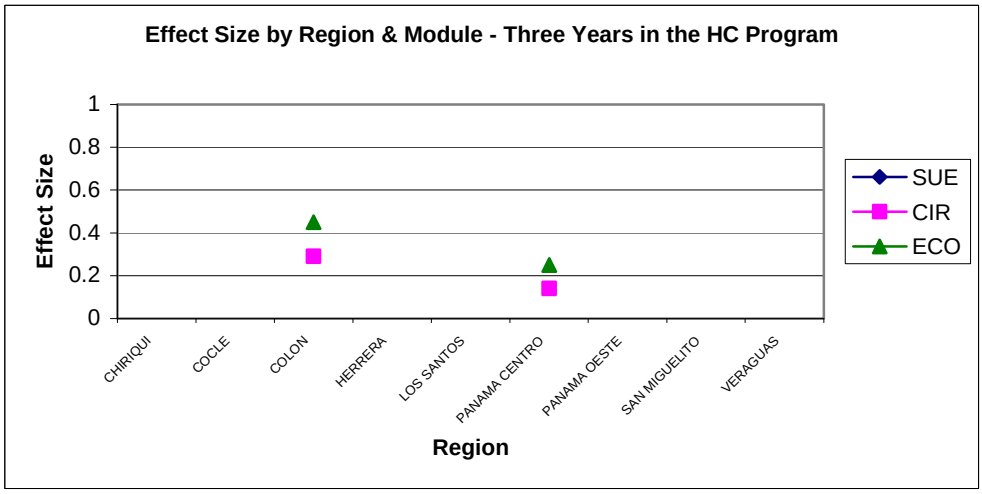
Figures 3 present similar results but controlling for the number of years of the Hagamos Ciencia program in schools.



(a)



(b)



(c)

Figure 3. Effect sizes by Region and Module controlling for year in the HC program.

Effects sizes by module and years in the program are presented in Table 9. For those schools with only one year as part of the program (Figure 3a), the pattern was similar to one described before. On average across modules, the region with the highest effect size is Los Santos (.49) followed by Veraguas (.46). In both cases the highest effect size was observed for third grade (Soils module: .94 and .81 respectively). These two regions were followed by Coclé (.26), Herrera (.25), Panamá Oeste (.23), and San Miguelito (.22). The lowest effect size was found in Chiriquí (.07) mainly due to the negative effect size found in fourth grade (Circuits module effect size = -0.18). It is important to mention that the region of Panamá Oeste did not have schools in fourth grade which implemented the HC program for only one year.

When the focus was on schools with two years in the program, the averaged effect sizes were highly reduced in most of the regions, except for Veraguas. Veraguas was the one with the highest averaged effect size (.61), followed by Colón (.19). The region of Chiriquí showed a negative effect size (-0.05). Panamá Centro has data for only one module (Soils module effect size = .12). Coclé, Herrera, and Los Santos did not have any schools with two years of implementing the HC program.

Only two regions have schools with three years' involvement of the HC program, Colón and Panamá Centro. In both regions information from only two grades was available (Circuits module and Ecosystems module) with an averaged effect size of .37 and .20 respectively.

Table 9. *Effect Sizes by Year, Module, and Region*

	Chiriquí	Coclé	Colón	Herrera	Los Santos	Panamá Centro	Panamá Oeste	San Miguelito	Veraguas
Third Grade - SOILS									
1 Year	0.34	0.74	-0.19	0.18	0.94	0.37	0.27	0.26	0.81
2 Years	-0.16	NA	0.09	NA	NA	0.12	0.11	0.14	0.96
3 Years	NA	NA	NA	NA	NA	NA	NA	NA	NA
Fourth Grade - CIRCUITS									
1 Year	-0.18	0.04	0.07	0.38	0.26	0.16	NA	0.21	0.34
2 Years	-0.05	NA	0.16	NA	NA	NA	0.04	-0.04	0.54
3 Years	NA	NA	0.29	NA	NA	0.14	NA	NA	NA
Sixth Grade - ECOSYSTEMS									
1 Year	0.06	0.00	0.42	0.18	0.27	-0.04	0.18	0.19	0.23
2 Years	0.05	NA	0.33	NA	NA	NA	-0.12	-0.10	0.32
3 Years	NA	NA	0.45	NA	NA	0.25	NA	NA	NA

Note: NA-not data available for the analysis.

Overall, Veraguas was the region with highest effect size across the three grades. Only in the region of Chiriquí and Colón was a negative effect was observed in the first year of implementation of the program: in Chiriquí in the Circuits module (-0.18, a small negative effect) and in Colón in the Soils module (-0.19). Contrary to what should be expected, effect sizes of the HC program do not increase as

the years of Hagamos Ciencia increase. Instead, the initially observed positive effect tends to either stay almost the same or drift away (e.g., Veraguas went from .96 to .81 in third grade and from .54 to .34 in fourth grade). A similar pattern can be observed in fourth grade in the region of Colón (Circuits module). It would be important to focus on the possible causes for this decline in order to sustain the effects of the HC program.

Variability at the Teacher and School Level: Intra-Class Correlation

Because of a two-stage sampling design adopted for this study, the sampling unit was no longer individual students; Instead, the sampling unit was school first and then teacher/class and student, as students are nested within teachers who are then nested within schools. The intra-class correlation (ICC) permits to learn the proportion of total variance of test scores that exists in higher level of sampling units (school, teacher, or class). In other words, ICC is the degree to which individual students share common experiences due to closeness in schools, teachers, or classes. A larger ICC coefficient represents greater heterogeneity of student test scores among higher-level sampling units of class/teacher level or school level. The findings of intra-class correlation indices among teachers and schools for booklets within grade are summarized on Table 10.

Table 10. *Intra-class Correlation by Two Units of Analysis, Grade and Booklet*

Type of Group	Grade 3 Soils		Grade 4 Circuits		Grade 6 Ecosystems	
	Booklet		Booklet		Booklet	
	1	2	1	2	1	2
Both Groups						
ICC Teachers	0.78	0.76	0.73	0.71	0.73	0.72
ICC Schools	0.87	0.84	0.79	0.74	0.75	0.76
Hagamos Ciencia						
ICC Teachers	0.79	0.78	0.75	0.74	0.75	0.74
ICC Schools	0.88	0.87	0.81	0.79	0.77	0.78
Control Group						
ICC Teachers	0.71	0.66	0.69	0.65	0.67	0.66
ICC Schools	0.73	0.65	0.75	0.63	0.71	0.73

Overall, high ICC estimates are found across grades at both levels teachers and schools. At the teacher level without considering type of group (i.e., HC program and non-HC program), the ICC coefficients ranged from .71 to .78. The range was similar for both the teachers within the Hagamos Ciencia group and within the control group (range = .05 and .06 respectively). This information indicates that student assessment scores are quite different in terms of teacher-level comparisons. Interestingly, lower ICCs were consistently found for the control group. The ICC indices for each booklet and grade in the HC program were relatively higher as compared to corresponding counterparts in the control group. This

comparative pattern justified the need to sample more schools (therefore more teachers and students) in the HC group than the control group in this evaluation study and any follow up work.

At the school level sampling, the ICC coefficient varied from .74 to .87 regardless of type of group (HC or non-HC group). Likewise, a similar pattern was observed at the teacher level as well for the whole sample, with small ranges (.13, .11, and .12 respectively). The ICC indicators for school comparisons were larger than those of counterparts for teacher comparisons across all 18 analyses with two exceptions (.65 smaller than .66 in the control group of third grade with booklet 1, and .63 smaller than .65 in the control group of the 4th grade with booklet 2). This result indicates that student assessment scores are far more different at the school-level comparisons than variations between teachers or students.

Facilitator Analysis

In addition to teachers and schools, facilitators play a vital role in offering and disseminating the HC program. This section discusses the results related to the facilitators. Table 11 presents the results of intra-class correlation indices among facilitators for booklets within grade. The analysis was executed for students in the HC group only as teachers in the control group did not receive any training from the facilitators.

Similar to the results at the levels of teachers and schools, the ICC coefficients were consistently high across grades at the facilitator level, ranging from .75 to .87. This indicates that student scores are different at facilitator-level comparisons. In other words, students' scores were somehow more similar when they shared a same facilitator though they may be taught by different teachers. Based on the results, the facilitator influence needs to be considered in the effect size.

Table 11. *Intra-class Correlation for Facilitators by Grade and Booklet*

	Grade 3 Soils		Grade 4 Circuits		Grade 6 Ecosystems	
	Booklet		Booklet		Booklet	
	1	2	1	2	1	2
ICC Facilitators	0.87	0.84	0.81	0.75	0.75	0.79

Before comparing students' performance across facilitators, we first report the counts of cases with missing values. As shown in Table 12, more missing data are involved for facilitators in the year of 2007, therefore we chose to only use the facilitator information in 2008 in the remaining part of this section.

Table 12. *Summary of the Missing Cases for Facilitator Information by Year, Grade, and Level*

		Student Level			Teacher Level			School Level		
Year	Grade	*n* of missing cases	*n* of total cases	%	*n* of missing cases	*n* of total cases	%	*n* of missing cases [a]	*n* of total cases	%
2008	3	982	4107	23.91	41	195	21.03	11	63	17.46
	4	343	3982	8.61	16	196	8.16	6	63	9.52
	6	510	4424	11.53	24	199	12.06	10	67	14.93
2007	3	1037	1591	65.18	49	76	64.47	18	23	78.26
	4	1189	1986	59.87	53	88	60.23	21	25	84.00
	6	955	1939	49.25	44	85	51.76	18	26	69.23

a - A few schools may have the facilitators for some teachers but other teachers. Those schools were considered as cases with missing information.

Descriptive results for facilitators across grades are summarized in Table 13. (See Appendix C, an Excel file attached to the report, for information on facilitators at the teacher level). The means and standard deviations for the control group are included as well as a reference. Facilitators whose students had significantly higher or lower scores compared to the grade mean in the control group are highlighted in green or in pink, respectively. Those means of student scores which were lower than the grade mean for the control group are also flagged with an asterisk and the green cells refer to the statistical significance. Still, when the sample size is too small, the facilitators may just work with only one or two teachers.[7] With the insufficient sampling at the teacher level for some facilitators, the interpretation of the table below needs to be extremely cautious.

[7] In the Pilot II sample, 15% of the facilitators worked with only one or two schools in 2008 and another 15% worked with three schools across the grades.

Table 13. *Means and SDs for Individual Facilitators in 2008 by Grade*

Name of Facilitator	Type	Grade 3				Grade 4				Grade 6			
		Mean	SD	*n* Students	*n* Teachers	Mean	SD	*n* Students	*n* Teachers	Mean	SD	*n* Students	*n* Teachers
Non-HC Program		7.69	3.23	2043	102	11.14	4.34	2041	99	14.47	4.73	2237	103
ABAD AIZPRUA	F	11.81	3.98	88	4					15.00	3.91	36	2
ALMA CHEN	F	10.23	4.16	150	7					16.35	4.84	68	3
AMINTA ROBLES	F	12.71	3.65	41	2	11.63	5.20	32	2	15.68	4.49	284	13
ANA ROSA AVILA	F	9.94	4.51	52	3								
ANAYANSI RAMOS	FF					12.07	3.93	14	1	*13.85	5.44	537	20
ANERIS RIOS	M	*6.45	3.07	98	5								
ARACELY GONZALEZ	-[a]	9.55	3.34	40	3	13.00	3.97	10	1	*13.47	4.83	51	2
ARCADIO DE LEON	M									15.69	4.48	91	3
ARISTIDES BEITIA	FF	10.57	3.76	46	3	12.10	4.18	115	6				
BENITO CASTILLO	F	9.90	3.85	49	2	12.22	4.68	182	7	16.84	5.28	58	2
BERTHA PEÑA	F	*7.25	2.67	20	1	*10.92	4.11	164	6	*13.12	4.43	224	8
CARLOS CAMAÑO	F					*11.12	4.96	427	16	*14.24	5.17	21	1
CARLOS GOMEZ	F	8.34	3.74	41	2	*10.22	4.48	78	5	14.82	4.91	34	2
CHRISTIAN GARRIDO	M	*6.69	3.47	55	3								
CIBELES ZOMARRIBA	F	9.67	3.72	167	9	13.16	4.61	108	6	*12.97	5.27	31	2
DIANA ARAUZ	F	9.14	3.84	124	5	14.03	4.42	86	5	16.35	4.83	97	5
DIGNA CAICEDO	F					12.35	3.94	20	1	*12.23	3.74	48	1
DIMIA RUJANO	F	9.83	4.28	117	6								
DORIS QUEZADA	FF	7.92	3.74	76	4	11.56	4.95	88	4	15.06	3.98	62	3
EDIL GARCIA	F									15.68	4.77	144	6
EDUARDO ZAMORANO	F					14.28	5.10	50	3				
EDWARD MONTENEGRO	F					*10.44	4.87	72	3				
ELISA MORGAN	F	*7.51	3.11	197	8	*10.58	4.55	167	6	*14.17	4.60	155	6

Name of Facilitator	Type	Grade 3				Grade 4				Grade 6			
		Mean	SD	*n* Students	*n* Teachers	Mean	SD	*n* Students	*n* Teachers	Mean	SD	*n* Students	*n* Teachers
Non-HC Program		7.69	3.23	2043	102	11.14	4.34	2041	99	14.47	4.73	2237	103
ELIZABETH LAZARO	FF					11.30	4.73	151	6				
ERIC PEREZ	F	9.43	3.99	168	8	12.17	4.31	145	6	14.95	4.50	185	8
EYRA GERRA	FF					11.55	4.55	62	3				
FERNANDO LOPEZ	F	7.74	3.51	140	7	*10.92	4.71	91	4	14.61	4.40	169	7
FLORENTINA GUTIERREZ	F	13.25	3.78	44	2	13.90	3.89	125	5				
GLORIA RODRIGEZ	F	10.76	3.68	25	2								
ILSA AUSTIN	F	8.20	3.52	130	5					*13.35	3.69	26	1
IRMA SALDAÑA	F					15.12	4.48	52	3	16.40	4.86	72	3
ISAAC QUINTERO	F	8.34	3.20	50	3								
JESSICA APARICIO	M									*13.90	5.17	132	5
JORGE DIMAS	-[a]					14.00	3.28	22	1	*14.42	3.49	24	1
JOSE LUIS RODRIGUEZ	F	9.86	3.77	113	5	11.99	4.37	166	7	*14.04	5.81	192	7
JUAN MENDIETA	FF	7.78	3.14	40	3	11.15	5.07	26	2	*12.98	4.37	46	3
JUDITH CARRION	M					11.42	5.34	125	5				
JULIO MENOCAL	F	10.58	4.70	52	2	12.88	4.46	76	4	16.82	5.28	74	5
KATHIA HERNANDEZ	M	8.94	3.72	68	3								
KENIA MATHIUS	-[a]	8.95	2.56	20	1								
LIDIA QUIROS	F	7.86	3.91	50	3	*9.89	4.21	38	2	*11.55	4.41	40	2
LINETH CAMPOS	M	*7.52	3.37	21	1	*9.93	4.36	126	6	*11.15	4.59	46	4
LORENA LOPEZ	M	*6.37	2.12	49	2								
LUIS CAMARENA	-[b]	9.07	3.58	110	5					*14.11	5.12	186	7
LUIS MORENO	M	13.00	3.43	19	1								
LUIS PEÑALOSSA	F	8.58	3.00	26	1								
LUIS TUÑON	F	*7.68	3.64	92	4	13.33	4.85	55	2	15.88	4.61	59	2
MARIA COBA	F					13.24	4.72	136	6	15.66	4.66	128	6
MARIELA BATISTA	F	8.43	4.29	68	3								

Name of Facilitator	Type	Grade 3				Grade 4				Grade 6			
		Mean	SD	*n* Students	*n* Teachers	Mean	SD	*n* Students	*n* Teachers	Mean	SD	*n* Students	*n* Teachers
Non-HC Program		7.69	3.23	2043	102	11.14	4.34	2041	99	14.47	4.73	2237	103
MIRENA MENDOZA	F	9.47	3.98	117	5	*10.93	4.30	107	5	*14.21	4.12	94	4
NIEVE VILLAREAL	F					12.84	4.89	62	2				
NITZIA SUAREZ	M					*9.40	3.81	35	2	*11.72	4.44	43	3
OMAYRA JAEN	F	8.92	3.67	64	3	12.03	4.33	129	7	15.47	5.25	53	2
PEDRO ADAMES	FF	9.58	3.62	62	3	11.42	5.14	156	9	15.90	5.24	30	2
RAMON GOMEZ	-[a]									*12.71	4.25	17	1
RENAN GONZALEZ	F	8.12	2.53	26	2	*10.30	4.03	123	7	*11.04	4.82	53	4
RIGO PINZON	M					*10.22	4.95	187	7				
ROBERTO GARRIDO	F					*11.04	4.94	101	4				
ROBERTO MELILLO	FF	10.67	3.80	64	4	*9.73	4.48	52	3	14.93	5.00	59	4
SUSAN CALDERON	M	8.08	3.10	26	2	*8.90	3.93	20	1	*13.17	4.30	66	3
TERESA PEREZ	-[a]									15.11	5.47	45	2
VANESSA CENTOLLA	F									16.61	4.49	49	2
VERONICA CASTRO	F	9.48	4.98	69	3								
YANETH RODRIGUEZ	F					11.19	4.86	21	1	15.60	4.50	91	5
YENIVETH URBINA	F	10.04	4.29	84	4	*9.18	4.42	33	2	16.35	5.86	31	2
YIRA PHILLYPS	F					11.64	4.78	50	2	14.52	4.67	31	2
ZULEIKA HERNANDEZ	F	8.46	3.97	41	2	12.64	4.70	45	2	13.51	3.77	53	2

Note: (1) F refers to facilitators, FF refers to Facilitator FFs, and M refers to monitors.

(2) -[a] refer to cases where the information for facilitator type was not provided. -[b] refers to the case which was identified as both F and FF.

(3) * denotes that the mean score is lower than the Control group grade mean.

(4) Cells highlighted in green indicate the students from those facilitators statistically significantly outperformed the control group after controlling the multiple comparison errors. Pink highlights those facilitators whose students had scores significantly lower than the control group.

In the academic year of 2008, the Pilot II sample involved 67 facilitators of three types. Among them, there are 41 F facilitators, six FF facilitators, and 12 monitors who worked with the teachers across the three grades (See Table 13). A preliminary analysis was also performed to compare these three types of facilitators. Table 14 summarizes the results of the descriptive analysis by grade and area. In previous section, the effect of HC was found to have differentiated effects on student learning depending on the areas schools are located. The statistical analysis was also conducted for each grade to determine whether type of facilitator and area influence the learning outcomes of students. In urban schools, students with the F facilitators appeared to take the lead compared to those with the FF facilitators. The difference between the F facilitators and FF facilitators was statistically significant at third and sixth grades ($p<.01$). In the rural schools, an opposite pattern emerged. Students with the FF and F facilitators statistically outperformed their counterparts with the monitors consistently across the three grades ($p<.01$). Only in third grade, the FF facilitators were associated with student performance higher than their F colleagues ($p<.01$).

Based on students' performance on the assessments, the F facilitators were associated with the highest student performance in the urban schools but the FF facilitators were related to the highest student performance in the rural schools. Due to the small sample size for some types of facilitators (e.g., 89 students were taught by teachers who worked with the monitors), the results are mainly suggestive and exploratory in nature. Additional work with larger sample sizes is needed before drawing any conclusions about types of facilitators.

Table 14. *Descriptive by Type of Facilitators, Area, and Grade for Facilitators in 2008*

Grade	Type of Facilitator	Area							
		Urban				Rural			
		n of students	Mean	S.D.	S.E. of Mean	*n* of students	Mean	S.D.	S.E. of Mean
3	F	1733	9.50	4.10	0.10	672	8.62	3.82	0.15
	FF	116[a]	7.87	3.53	0.33	172	10.25	3.74	0.28
	M	89[a]	8.61	3.67	0.39	247	7.16	3.49	0.22
4	F	2192	12.19	4.71	0.10	749	10.77	4.57	0.17
	FF	400	11.78	4.60	0.23	264	10.85	4.92	0.30
	M	233	11.34	5.01	0.33	260	9.44	4.50	0.28
6	F	1828	15.09	4.70	0.11	772	14.27	5.20	0.19
	FF	645	13.90	5.26	0.11	89[a]	15.26	5.08	0.54
	M	223[a]	14.63	4.97	0.33	155[a]	12.17	4.49	0.36

Note: (1) Facilitators denoted with the superscript of *a* only worked with 3 or fewer schools.
(2) Facilitators shadowed in grey had students' average scores lower than that of the control group (also see Tables 4b and 5b).

Case Studies

Sampling

Using findings from Pilot II, we apply a set of criteria to propose a list of teachers from which SENACYT will further select a subsample to be used in a *multiple-case embedded design* (Yin, 2003). This first round of sampling only considers four regions suggested by Maria Heller in May, 2009 for the case studies: Chiriquí, Panamá Centro, San Miguelito, and Veraguas. The sample is provided in Appendix D, an Excel file attached to the report. The file provides detailed information related to the number of teachers per school, years in the HC program, and facilitator (whenever information was available in the data set) and some facilitator characteristics available (e.g., whether the facilitators was selected as with teachers of highly performance consistently). The sampling considered the following selection criteria:

(1) *Performance of students on the selected items*. Within each region, high- and low-performing teachers were identified after their class mean scores were ranked (i.e., the highest 10 teachers and the lowest 10 teachers, respectively). Generally those teachers' class mean performance significantly differed from the grade mean score. However in one region, some selected teachers were labeled as high-performing within that region even though their class level performance was actually at medium level based on *t* test results when compared to all other schools at that grade.

(2) *Facilitator information*. At least one teacher from the high and low performing facilitators was selected whenever applied. The decisions of high and low facilitators were made based on summarizing the information from Table 13. That is, the mean of student performance at facilitator level is higher or lower than the grade mean. We were also interested in identifying four types of facilitators across the three grades, those who were always associated with high or low averaged student scores and those who were less consistent or extremely inconsistent across grades (see Appendix E, an Excel file attached to the report, provides information available for all facilitators that were part of the data base; facilitators coded by color after facilitator average scores were sorted at each grade). Finally, we attempted to take the number of teachers into account, assuming student scores are less biased estimates if facilitators worked with several teachers (except two low facilitators at Grade 6).

(3) *Class Size.* Only teachers with 10 or more students were included in the recommended list. This decision was influenced by the reasoning that for a later large-scale study we will mostly focus on a reasonably large class size. Such focus takes into account two issues: the required minimal sample size at class level and the logistics concern when testing with schools with too few students.

(4) *Other Relevant Factors*. During this preliminary sampling, we also sought a balance between HC vs. control group, urban vs. rural settings, and one year only vs. more than one year of implementing HC when appropriate. This balance ensures that the case study will be conducted with a purposive sample that provides valuable information about those relevant variables that may mediate the impact of the HC program. To achieve this goal, more weight was given to teachers from the control

group, rural schools, or two to three years of HC implementation when individual teachers had similar average scores.

The suggested list includes 96 teachers, four with the highest mean performance within the region and four of the lowest performance within the region, considering type of group (HC and control) and area (urban or rural). Table 15 provides general information about the sample selected.

Table 15. *Distribution of the Selected Sample by Level of Performance, Area, and Type of Group by Region.*

		HC				NON-HC				Both
		Urban		Rural		Urban		Rural		Total
Region	Grade	High	Low	High	Low	High	Low	High	Low	
Chiriquí	3	1	3	2	0	1	0	0	1	8
	4	1	2	0	2	2	0	1	0	8
	6	2	1	1	2	1	0	0	1	8
Panamá Centro	3	3	2	0	0	1	2	0	0	8
	4	3	2	0	0	1	2	0	0	8
	6	3	3	0	0	1	1	0	0	8
San Miguelito	3	4	1	0	1	0	2	0	0	8
	4	3	1	0	2	1	1	0	0	8
	6	4	3	0	0	0	0	0	1	8
Veraguas	3	4	0	0	1	0	3	0	0	8
	4	2	0	2	2	0	2	0	0	8
	6	1	0	2	4	1	0	0	0	8
Total	All	31	18	7	14	9	13	1	3	96

As it could be expected, the number of teachers varies by area and program. However, to remain the same numbers of teachers in all the cells would increase the sample considerably. It would be even impossible for some regions. Not surprisingly, it is important to point out that the highest number of teachers can be found in the column of high performing teachers within urban areas and the HC program despite the effort to maximize teachers from rural or non-HC group. It is also necessary to consider that, if more than one teacher could be observed within a selected school, those teachers should be included in the sample whenever possible. This would allow a systematic examination on the school contextual factors related to teachers' implementation. Thus, the final sampling will involve a selection of a sub-sample of teachers and addition of their colleagues from the same schools, if applicable.

Proposed Characteristics for the Case Studies

In this section we propose some aspects to be considered in the *multiple-case embedded design,* a research strategy that has been regarded as an appropriate method for tapping research questions about "how" and "why" (Yin, 2003). This type of design follows the logic of *replication*. Evidence

collected from multiple cases (that can be translated as sites) is more compelling and, if consistent across cases, more robust and worthy of continued investigation. Of course, cases should be carefully selected; they should show similar or contrasting results, as it is the case in the sample selected.

An important characteristic of case studies is the analysis of the contextual conditions of the "case" since it can be expected that contexts of these multiple cases differ to some extent. Therefore, if common conclusions are drawn across cases, it is more likely to ascertain critical characteristic related to the appropriate implementation and the effect of the HC program. We believe that a multiple-case embedded design will provide an appropriate description of how teachers implement the HC program and how students are impacted by these reform practices.

In the context of case studies, *embedded* implies to focus on more than one unit of analysis within each case. In this project, there are four relevant units of analysis: students, teachers, facilitators, and principals. Therefore, *embedded* gets translated as following different students, teachers, and facilitators within the same school whenever possible and applicable. Although principals are considered another unit of analysis, it is assumed that there is only one principal per school. We acknowledge that it is possible to have two principals per school, one per shift, morning and afternoon. However, this would only apply if the sample includes teachers in different shifts within the same school.

Below we articulate the rationale for the case studies by elaborating the five components that characterize this type of study (Yin, 2003): research questions, propositions, unit(s) of analysis, logic linking the data to the propositions, and criteria for interpreting the findings.

Research Questions

Knowing the student performance from Pilot II, the research questions should address more the "how". We hope that knowing "how" would help later explain "why". In case studies it is important to also define propositions that help to direct attention about what needs to be examined. Below are some possible research questions that can be explored in the multiple-case embedded design and some propositions linked to each of them:

1. Does the HC program have an impact on teachers' science instructional practices as observed in the classroom?
    A. HC teachers' instructional practices are different to Non-HC teachers' instructional practices.
    B. HC teachers' instructional practices should be more aligned to the science instructional practices promoted by the HC program than the Non-HC teachers' instructional practices.
2. If the HC program is having an impact on teachers' science instructional practices, is the degree of alignment of the teachers instructional practices to what HC envisioned related to the student performance in the science assessments?

C. Teachers' understanding about scientific inquiry and about the goals and characteristics of the HC program are linked to the degree of alignment to the HC instructional practices expected.

D. Teachers' instructional practices more aligned with the science instructional practices promoted by the HC program should be linked to higher student performance.

3. What setting factors may be affecting the effectiveness of HC in changing teachers' instructional practices?

E. Facilitators' understanding about scientific inquiry, about their role with teachers, and about the goals and characteristics of the HC program are linked to the degree of alignment of the HC teachers' actual instructional practices to those expected by HC.

F. Facilitators who have received direct training from the HC program (F) tend to train teachers better, measured by the aligned instructional practice of the trained teachers, than those who have received training from other facilitators (FF or M).

G. Schools that are more supportive of teachers' implementing the HC program have higher levels of student performance.

Framework

The case studies will be guided by a theoretical framework for understanding social settings proposed by Tseng and Seidman (2007), which includes: *resources* (i.e., human, economic, physical), *organization of resources* (how resources are arranged or located), *social processes* (i.e., transaction between groups), and *setting level outcomes* (e.g., student learning). We believe that focusing on these components will help develop understanding as to how the HC program is implemented and what factors may be affecting its effectiveness. The cases will involve schools implementing the HC program and schools not implementing the HC program. Figure 4 presents a graphic representation of the framework.

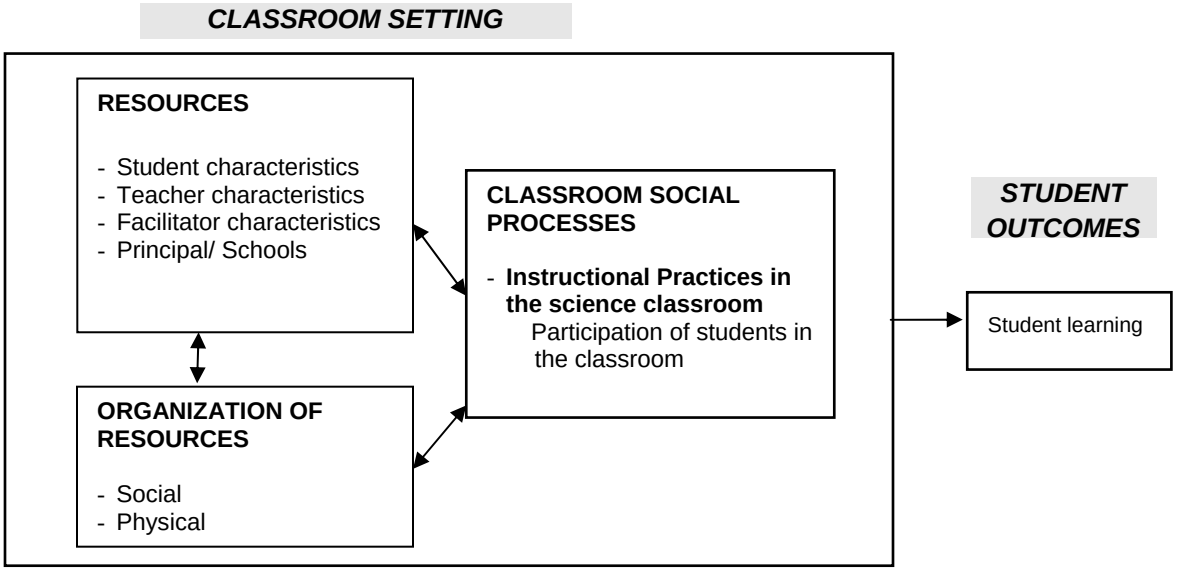


Figure 4. Conceptual framework to study the implementation of the Hagamos Ciencia Program in the school context.

Following the sequence of the research questions, the description of the framework starts with the third component proposed by Tseng and Seidman (2007), the *social processes* in the context of the science classrooms. It is the classroom where the HC program is expected to have its major impact by modifying the type of transactions observed between teachers and students when implanting science curricula. It is expected that the quality of these transactions can directly impact student learning.

Classroom Social Processes. In the context of Tseng's and Seidman's (2007) framework, social processes are defined as the transactions between one or more people or one or more groups of people. These transactions are expected to be modified if teachers' science instructional practices change as a result of the implementation of the HC program in the schools. To approach the study of the transactions between teachers and students in the classroom, it is proposed to focus on the six aspects that are expected to be the focus of the HC teachers. These aspects were discussed in Panama (in October, 2008) and were developed based on the document, *Towards a Practical Conceptual Framework of Scientific Inquiry*: (a) students' understanding of the learning targets, (b) helping students to construct conceptual schemes, (c) encouraging discussion and argumentation, (d) developing and improving scientific processes, (e) developing understanding about the role of data and evidence in explanations, and (f) developing communication in science. Appendix F presents the observation protocol as discussed in October, 2008.[8]

Resources: This component focuses on the presence, characteristics, and quality of resources in the settings (i.e., schools). It refers to how people, space, time, and money are organized or allocated in each setting. Although the availability and quality of resources matter in achieving the proposed outcomes, it is how these resources are used that makes a difference. Human resources refers to the individuals who are part of the setting and includes the characteristics of the students, teachers, facilitators (as they are embedded in schools and may adapt to the setting in which they work), and staff in the school. Physical resources refer to the availability of space, facilities, and materials. Temporal resources will focus on available time to science instruction and the planning time available. Economic resources will be inferred by the poverty level of the schools. We propose to focus mainly in human resources.

Organization of resources: This component refers to ways in which resources (e.g., people, space) are allocated in the setting. For example, students may be grouped by their teacher according to student achievement level, more time may be spent for some instructional activities than others, and money may be allocated for professional development of the teachers or to buy new tables for the classrooms. It has been argued (Tseng & Seidman, 2007) that changing the organization of the resources may help change the setting outcomes (e.g., student achievement). For example, changing the physical

[8] It is very likely that this protocol has been used and modified by the PCien in Colombia. It is important to find out the last version of the instrument.

organization of the classrooms can impact the social organization and the type of transactions between teachers and students. The organization of time in the school can also affect the setting (e.g., the school schedule may allocate, or not, time for teachers to get together and discuss issues related to their students and instructional practices).

Linking Propositions to Sources of Evidence

To link the data to the propositions, the fourth component of case studies, we propose to use different sources of evidence. Table 16 provides information that links the units of analysis, the propositions, the sources of evidence, and the proposed time to collect the data during the school year. In the next sections we describe the instruments, the procedures we will use to collect the information, and the criteria to interpret the data collected.

Table 16. *Linking Unit of Analysis, Propositions, Sources of Evidence, and Timing*.

Unit of Analysis	Propositions	Sources of Evidence	Module Instruction		
			Before (Pre-Test)	During	After (Post-Test)
Students	D	Assessment Performance	x		x
		Questionnaire			x
		Interview		x	x
Teachers	A, B, C,D	Classroom Videotapes		x	
		Evaluator's Direct Observation		x	
		Questionnaire			x
		Interview			x
		Artifacts (e.g., notebooks)			x
Facilitators	E, F	Questionnaire	x		
		Interview			x
		Classroom Videotapes		x	
		Teacher-Facilitator Videotapes		x	
School Setting	G	Evaluator's Direct Observations	x	x	x
		Principal Questionnaire		x	

Units of Analysis and Sources of Evidence. We propose to focus on four units of analysis within each school although the sampling will be primarily started at the teacher level: the students, the teachers and the facilitators linked to them, and the school setting. The first three units of analysis become the sources of information about themselves, as well as for other units of analysis (e.g., students can inform about their teachers instructional practices). For the school setting we propose to use the principal as the main source of information, along with the teacher. In the following discussion, we describe each unit of analysis and the proposed sources of evidence. Table 17 provides the different sources of evidence linked to the setting components.

Table 17. *Setting Components and Sources of Evidence*

Setting Components	Sources of Evidence								
	Student Science Assessment Performance	Students Questionnaire & Interview	Teacher Questionnaire & Interview	Facilitator Questionnaire & Interview	Classroom Artifacts	Classroom Videotapes	Classroom Direct Observation	Principal Questionnaire	School Visit (School Direct Observation)
Classroom Social Processes		X	X	X	X	X	X		
Resources			X	X		X	X	X	X
Organization of Resources			X	X		X	X	X	X
Student Outcomes	X				X	X			

Students. Students can be administered three measures: the science assessment, a questionnaire, and a follow-up interview. The science assessment will be administered before and after the HC modules are implemented. The science assessment characteristics are presented at the beginning of this report.

The student questionnaire (see Appendix G) has 31 Likert-type questions and was designed to collect information on factors that may influence student achievement. The questionnaire gathers information on two aspects: (a) about the students and their families (e.g., information about the educational resources do students have in their homes), and (b) activities conducted in their science class (e.g., information about the students' perceptions about what they do in their science class?).[9] The questionnaire has some highlights that will have to be adapted according to the grade (module) of the students to whom it will be administered (e.g., the topic of question 13 of the student questionnaire will need to be changed to make soils in third grade and Ecosystems in sixth grade). It is important to note that most of the questions can be presented in a matrix from. If there is an agreement that students can respond to this type of questions in a matrix, the number of question will be reduced considerably.[10]

[9] The questions are only possibilities to be considered. Maria Heller with her team should decide which ones are appropriate in the Panama context and the schools. At least it is expected that they help to elicit other questions that may be of interest for the HC program.

[10] Please not that there are some question number with no question written. They are highlighted in pink. They are there in case more questions are written. If they are used, the numbering of the questions will automatically be changed. Do not delete them until the final version is achieved.

A follow-up nine-question interview (see Appendix H and footnote 10) can be conducted with a random sample of students. The interview will focus on few critical questions around the instruction and learning. One set of questions asks students to recall the instructional activities they have received (e.g., Can you describe a typical science class?). Another set of questions probe students' understanding about of inquiry (e.g., What do scientists do? What did you learn in your science class that you have not learned in other class?). Similar questions can also be asked of the students in the control program.

Teachers. As with students, we propose four sources of data: Classroom videotapes, classroom direct observation, questionnaire, and interview. To collect information on the classroom transactions, teacher and students (i.e., classrooms) will be videotaped. There are different options to approach this task. Teachers and students (classroom) can be videotaped:

1) from the beginning of the school year. Collecting this type of data over time will provide insights of how the practices evolve over time within and between classroom settings. Teacher can be videotaped once a week or every two weeks

2) from the beginning of the module (for HC teachers) or one unit (for control teachers). We propose videotaping everyday or at least once a week during the duration of the implementation of the module/unit at hand.

3) during one purposefully selected lesson from the module everyday. The same can apply for the control teachers; select one lesson of one topic based on the national curriculum.

Videotaping needs to capture teachers' and students' talking, as well as any intervention of the facilitator, if appropriate and if there is one on the day of the videotaping. Therefore, it is important that at least two microphones be set on each classroom, one for the teacher (lapel) and a rod microphone to capture student participation. Depending on the type of video camera used, either videotapes or storage devices (USBs) should be collected and properly organized.

Direct observation will be conducted by an evaluator (or more than one if necessary). The evaluator should visit the school (and classroom) at least twice during the implementation of the HC modules and/or during the school year. Direct observation should focus on capturing information that is difficult to capture in the videotapes, questionnaires, and interview. For example, direct observation in the classroom should provide a detailed description of the classroom environment (e.g., arrangements of the tables or desks, posters in the walls, and space for the students). Photos or maps (drawings) should be considered as part of the data collection. Classroom visits should be paired with the days in which videotaping will occur. This strategy will allow for the collection of information that will be useful for inter-rater agreement.

The teacher questionnaire should be administered to all teachers in the sample. The questionnaire has 28 questions and focuses on three aspects that have been assumed to be related to student achievement (see Appendix I and note 10 and 11): (a) teacher background (i.e., questions focusing on the characteristics of the science teachers; this section includes questions about the HC program), (b)

teacher's science class (i.e., questions focusing on the activities that students do in their science class), and (c) teacher's perception about their school (e.g., What are the roles and responsibilities of the teachers and staff?).

A follow-up semi-structured interview (see Appendix J) should be conducted with the whole sample if at all possible. The 25 question protocol focuses on teachers' conceptions about their role and the student role in the science class, their conception about the HC program and science inquiry. Questions about their successes and problems in implementing the HC program are included as well as the factors that have contributed more or less to those success and/or problems (e.g., What role do you think the facilitator plays in your success implementing the modules?). These interviews can be conducted during the implementation of the modules or at the end of the academic year. However, any decision should be consistent across cases. Some examples of the questions to be included in the interviews include: What do you think are the goals of the HC program? What is its importance? Questions will be adapted and/or ignored when interviewing teachers in the control group.

Copies of student work (e.g., notebooks) can be collected at the end of the implementation of the HC module or the school year (for the Non-HC schools). This will allow us to analyze teachers' interaction with students with respect to assignments and feedback practices as well as teachers' organizations of classroom resources.

Facilitator. This unit of analysis only applies to HC schools. Four instruments are considered to focus on the facilitator as a source of information: questionnaires, interviews, classroom videotapes, and teacher-facilitator videotapes. The facilitator questionnaire has 18 questions classified in three aspects (see Appendix K, see footnote 10 and 11): (a) facilitator background (e.g., questions focusing on the characteristics of the facilitators), and (b) facilitator's perception about the training in the HC program (e.g., how useful are the courses taken in the HC program?), and (c) facilitators' school activities with teachers (e.g., interaction with the teachers, activities carried out when working with science teachers, etcetera).

The semi-structured interview protocol (see Appendix L) has 24 questions and focuses on both the training received (either by HC directly or by other facilitator) and in-school experiences with the teachers. They will be designed to obtain information about how facilitators experienced the learning activities (e.g., what, if anything, they would change to improve the course and training sessions; how much they were helped to improve their understanding about HC and their role wit the teachers). Some examples of the questions to be included in the interviews are: How do you conceptualize the HC program? Why do you think the HC program is important? What is the purpose of HC? What is your role in the HC program?

The classroom videotapes were described in the teacher section. If facilitators are in the classroom the day assigned to videotape, it is important to capture their interactions with the teachers and the students. It would be also important to collect videotapes of meetings that the facilitator and the teacher have before or after the implementation of the module in any given day. Videos will be analyzed

focusing on the interaction of the facilitator and the teacher (e.g., how much the facilitators help teachers to improve their instructional practices).

School. Different sources of information will be used to collect data about the school: the principal, the teachers, and the evaluator. A school questionnaire, to be completed by the school principal, collects information concerning some of the major factors thought to influence student achievement (see Appendix M). The questionnaire focuses on four major aspects: (a) school characteristics (e.g., enrollment), (b) staff characteristics (e.g., teacher professional development), (c) school climate (e.g., conflicts among staff members), and (d) support to the HC program. The teacher questionnaire has questions that focus on the school climate, as does the facilitator questionnaire. This questionnaire can be used as a guide for an interview if the questionnaire is not suitable and the interviews can be conducted during the academic year.

Direct observations at the school should be conducted on at least two occasions during the academic year. As mentioned previously, direct observation by the evaluator should focus on those aspects that are difficult to capture on the questionnaires and videotapes. The evaluator should collect information about the physical condition of the school and the resources, as well as the transactions (see teacher section). For example, a detailed description of school physical setting (e.g., if there is library, if it has enough books, if the classrooms have computers. whether those computers are old or new, etcetera), transactions of the teacher with other teachers, staff, and principal in the school.

Procedures

The technical qualities of all the instruments need to be evaluated. Reliability and validity will be evaluated according to the purpose and characteristics of the instruments. We will provide evidence of the inter-rater agreement (Cohen's Kappa) and inter-rater reliability whenever it is appropriate. Validity evidence will be provided for every instrument. For example, information collected through the videotapes can be used not only to focus on teacher instructional practices but also as a benchmark to validate other instruments (e.g., Are teacher's self-reported classroom activities consistent with what is observed in videos? Do the videotapes show evidence that teachers are implementing HC modules as expected?). Single instruments will be used as much as possible to collect different types of information. Information will be collected in units of analysis for both types of program HC and Non-HC.

Data Preparation

Pilot II involved a large sample size with 19,177 students within nine regions and 118 schools. In order to conduct a more accurate and comprehensive data analysis with such a large amount of data, data preparation is the first critical step.

Data preparation involves entering data into the computer files, checking data for accuracy, reading data from students' response sheets, and developing a database structure that integrates all the variables needed for the diverse analyses. Preparing a complete and accurate data set reduces time and cost invested in the project, and produces more reliable results. We acknowledge that, even though

most part of the data entering and preparation was done by a subcontractor, it is critical to consider some general guidelines of data preparation for the purposes of overseeing and facilitating the project.

Checking Data for Accuracy

Data accuracy is critical for saving time, avoiding errors, and having a more efficient way to approach data analyses. The data entered in a data files should be checked by randomly selecting about 2%-5% (depending on the sample size) to verify that the information was correctly entered. We suggest that Panama check the data before sending the files to UCD. For example, making sure that the values entered in variables such as "Years in HC Program" are appropriate (e.g., 0 to 3) since any other values will be a flag of an error. It is crucial to make sure that the control Groups does not have any value in Years in HC Program, or any facilitator assigned. Another important issue is to make sure that information is consistent across different sources. For example, is the region information of the schools consistent in different data sources? Checking for this type of inconsistency will save many hours of otherwise wasted time and will help find and correct these errors at an early stage of data preparation to reduce the need for corrections later on during the data analysis.

Excel files allow organizing the data in a way that makes it easier to check for inconsistencies in the names of the school or the teachers. However, there are some simple statistical procedures to check the data. For example, a simple frequency sometimes offers insights to identify possible typographical errors in data entry. For instance, if there are far more students in one booklet than another, special attention needs to be paid to check the variable of booklet. Joint-distributions across two related variables, such as "HAGAMOS_CIENCIA" and "Years in HC Program" also help to identify errors. The non-HC group should only have 0 as the value of "Years in HC Program". Reviewing the "facilitators in 2007" and "Years in HC Program" together, teachers with only 1 year implementation should not have any facilitators in the year of 2007.

Every time an error is found, the procedures need to be re-run, sometimes starting again with data transformation (recoding) and/or other steps to develop a database structure that is appropriate for the data analyses.

Reading Data from Students' Response Sheets

During the process of data-reading from test response sheet, the presence of missing data is an important issue. Whether this is an administration problem or it is due to a lack of time for students to fill in the booklet completely is an important issue to address. A large percentage of missing values in a data file may cause biased outcomes and result in misinterpretation of the results. Thus, one needs to plan how to ensure the data completeness even before collecting the data.

Additionally, finding and correcting the main source of this problem is important. Needless to say that one possibility is reducing the amount of information students need to fill in on the response sheet, which also avoids human errors. How much can it be pre filled? For example, information like Test ID or Booklet Number can be printed in advance on the test response sheets. We acknowledge that this was

not possible in Pilot II due to time constraints, but it seems critical in order to avoid as much missing data as possible.

During the process of data preparation for Pilot II data, we documented every inconsistency and/or inaccuracy found in the data sources, which enables us to trace the corrections made in the data files and provide some guidelines for subsequent data preparation. Having a workable data file took far longer than expected. We suggest that for the next round of data analysis, Panama review first the data file for accuracy related to names of teachers, schools, HC years, and all the other variables. Only when the information in the data has been checked for accuracy should be passed to the UCD team. Table 18 presents several important issues we found in the data files provided that were the main cause of the delay in processing the data and conducting the statistical analyses.

Table 18*. Data Inconsistencies and Errors in the Panama data Files.*

Variable Name	Issue	Solution Taken	Suggestion
Data Source: (a) Student-level data file – PILOT II-COMPLETE DATA-ALL GRADES-1-23-09-REVISED			
REGION	Some region names were similar but not identical.	Similar region names were corrected manually into the same one in SPSS syntax.	- Provide a more careful instruction for students to fill in the region info. - Panama carefully reviews the data before Passing it to the UCD team.
STUDENTID	STUDENTID was not totally unique for each student.	STUDENTID was not used in the analysis.	- Print a test ID on the test response sheet.
SCHOOL	1.Some school names were similar but not identical with those in the school-level data file 2.The small number of students found in several schools may be due to the school name problem in (1) 3.Several schools have errors in region info.	Those school names were corrected in SPSS syntax **manually**.	- Assign a unique school ID for each participating school in advance. - Panama carefully reviews the data before Passing it to the UCD team.
BOOKLET	There were missing data (88, 99, sys) for some students.	Those students with missing booklet info were dropped from the further analyses.	- Print the Booklet info on the test response sheet.
TEACHER_FULL NAME	1.Some teacher names are similar but not identical with those in other data files; 2.Two same teacher names are found in different schools.	Those teacher names were corrected in SPSS syntax manually.	- Assign a unique school-teacher ID identifier for each teacher in each participating school. - Panama carefully reviews the data before Passing it to the UCD team.

Variable Name	Issue	Solution Taken	Suggestion
YEARS_IN_HCPROGRAM	1.There were odd values including 4-15, 33, 1998, 2006-2008; 2.Multiple numbers of years were found in some same schools.	YEARS_IN_HCPROGRAM was corrected based on school info on Years in HC Program provided by Panama	- Have a separate data file that particularly deals with information at the school level, including school ID, YEARS_IN_HCPROGRAM. - Panama carefully reviews the data before Passing it to the UCD team.
HAGAMOS_CIENCIA	1.There were missing data (88, 99, sys) for some students; 2.Some schools had students both in HC and Control Group; 3.Inconsistency was found between the two variables -- HAGAMOS_CIENCIA & YEARS_IN_HCPROGRAM.	HAGAMOS_CIENCIA was corrected based on school info on HC and Control group provided by Panama.	- Have a separate data file with information at the school level, including HAGAMOS_CIENCIA . Do not ask students to fill in the information. - Panama carefully reviews the data before Passing it to the UCD team.
Q1 to Q30	1.Several students only responded to a few questions. This may be due to motivation issues; 2.System missing or odd values (i.e., 94) were found.	1.Students with a low response rate were dropped: 5 items for 3rd grade (about 1% deleted), 10 items for 4th (0.7% deleted), & 15 items for 6th (0.3% deleted); (2) System missing or other odd values (i.e., 94) were recoded as 99.	
All variables	There was still system missing for all variables.	The system missing was treated as 99.	
Data Source: (b) School-level data file – BASE_DE_DATOS_DENVER_0927_03122009			
SCHOOL NAME	Some schools included in Pilot II cannot be found in this data file.		
FACILITATOR NAME	1.Some facilitator info was missing; 2.Similar, but not identical, facilitator names were found; 3.Several facilitators were found in the control group; 4.Year info for facilitators was not consistent with another data file 'ANALISIS DATOS 090410 mvh' .	The facilitator info in the data source – 'ANALISIS DATOS 090410 mvh' was adapted in the facilitator analysis.	- Have a separate master teacher file with information at the teacher level, including facilitator. - Panama carefully reviews the data before Passing it to the UCD team.

Variable Name	Issue	Solution Taken	Suggestion
TEACHER NAME & TEACHER LAST NAME	Some teacher names and teacher last names were not consistent with other data files.	The teacher info in the data source – 'ANALISIS DATOS 090410 mvh' was adapted in the facilitator analysis.	Have a separate master teacher file with information at the school level, including teacher name and teacher last name.

Appendices List

A. List of the item analysis results for all items at item level: Excel file attached

B. Means and SDs for each teacher by region and module: Excel file attached

C. Facilitator information at teacher level: Excel file attached

D. Selected sample for the case study: Excel file attached

E. Facilitator information: Excel file attached

F. Observation Protocol: Word document

G. Student Questionnaire: Word document

H. Student Protocol Interview: Word document

I. Teacher Questionnaire: Word document

J. Teacher Protocol Interview: Word document

K. Facilitator Questionnaire: Word document

L. Facilitator Protocol Interview: Word document

M. School Questionnaire: Word document