PD-ABW-439

Performance Monitoring and Evaluation Project
The Mitchell Group Inc. with Harvard Institute for International Development

Quality Improvement in Primary Schools Program (QUIPS)
USAID and the Government of Ghana

DEVELOPING ACHIEVEMENT INSTRUMENTS TO ASSESS THE IMPACT OF EDUCATIONAL INTERVENTIONS IN GHANA

Trip Report for a Test Development Assignment in Accra, Ghana
October 25 to December 18, 1999

By

Richard S. Sandman

Submitted to the

United States Agency for International Development
Accra, Ghana

January 8, 2000

A

Developing Achievement Instruments to Assess the Impact
of Educational Interventions in Ghana

Richard S. Sandman
January 8, 2000

Background

USAID's Quality Improvement in Primary Schools Program (QUIPS) seeks to improve the quality of education in Ghana's primary schools through carrying out a prescribed set of interventions. The Performance Monitoring and Evaluation Project (PME) is a part of QUIPS that, among other things, is charged with determining the impact of these interventions. The impact of educational interventions on student achievement is, of course, of primary importance. Thus I was asked by PME to come to Ghana to help in developing achievement tests for measuring the impact on student achievement of the QUIPS interventions. The period of my assignment was October 25 to December 18, 1999.

It had been decided beforehand that the new tests were to be in English and mathematics, and that for each of these subjects one test was to be developed for the Primary 3 level, and another for the Primary 5 level. Tests had earlier been prepared in English and mathematics to measure QUIPS impact for the first two cohorts of children involved in the project. However, these instruments were seen to be unsatisfactory, as the baseline scores obtained with them were very low. In addition, the measurement of students had been cross-sectional in nature, with different students being measured on different occasions to ascertain achievement growth.

The new instruments to be developed were intended for Cohort 3. It was planned to use a longitudinal approach for measurement, with the same students being tested in English and mathematics on three occasions – January 2000, July 2000, and July 2001. Students were to be measured in all 108 project (Partnership) schools, as well as in 54 comparison schools. Comparison of test results would show the relative growth in achievement in the two sets of schools, and hence the impact of the project interventions.

Eddie Williams, a second language literacy specialist, was enlisted from the United Kingdom to assist in the preparation of the English tests. He was present in Ghana during the first three-and-a-half weeks of my consultancy.

Initial Planning Activities

Work on the test development began on October 25, 1999. The first task was to form two teams of local educators, one to work on the mathematics tests, and the other to work on the English tests. Such teams were hastily put together, assembling for the first time on October 27. These teams consisted mainly of university lecturers, along with

staff of one of the component QUIPS programs, Improving Learning through Partnerships (ILP). During the period October 27 to 29, there were three math team members working with me, and five English team members working with Eddie Williams, to develop some basic structures for the needed achievement tests.

In mathematics, the subject area was divided into four major topics: basic operations, basic number concepts, geometry and graphs, and story problems. For each topic, the principal objectives to be examined were specified for each grade level to be included in the tests. For the P3 test, it was decided to include P2, P3, and P4 items. For the P5 test, P4, P5, and P6 items were to be included. It is to be remembered that although the initial administration of the tests was planned for students near the beginning of P3 and P5, these same students will be at the end of P4 and P6 when the third administration of the tests is carried out. Thus items at these higher grade levels need to be included in the tests in order to accurately assess achievement growth.

In English it was decided to include sections on vocabulary, grammar, dictation, punctuation and capitalization, listening comprehension, reading comprehension, and, for the P5 test only, composition. In addition, a small number of students were to be pulled out separately and measured in oral facility on a one-to-one basis. The English tests were to include items from the same grade levels as included in the mathematics tests, above. The subject content of the included items from the various grade levels was also specified for each of the above sections.

Item Development

Item development began on Monday, November 1. At that point, the three members of my math team remained, but the English team had been reduced to two members plus Eddie Williams. Later that week one of the math team members withdrew, and the remaining group of team members – two in math and two in English – has remained unchanged up to this point. This group consists of Kafui Etsey and Charles Duedu for math, and Isaac Amuah and Lawrence Owusu-Ansah for English. All four are lecturers at the University of Cape Coast. It is to be noted that the small size of the English group, considering the complexity of its task, has always constituted a problem in the development of the English tests.

For mathematics, item development consisted of seeing how each objective to be included at each grade level was treated in the grade-level textbook, and by devising items that were similar in content and difficulty to those found in the textbook. Three items were prepared for each objective. Ultimately, the best two of these were chosen for field testing.

For English, this curriculum-based item-development process was supplemented by the use of new types of items introduced by Eddie Williams. Notable among these is a reading comprehension passage where the student fills in gaps in the passage by selecting appropriate words from those included in a box, the same box of words being used for all

gaps in the passage. It should be noted that in the development of items for the English tests, it was decided to extend downward the grade-level range of included items. Thus the P3 test was to include items from P1 through P4, and the P5 test was to include items from P1 through P6.

Field Testing of New Items

After a sufficient number of new items had been prepared, these items were field-tested in a number of schools in the Accra area. The field testing took place over the period November 8 to 21. The field testing was carried out by the four members of the test development teams, with the help of support staff from PME. Four schools were visited in the mathematics field testing, and about an equal number for English. An attempt was made to choose schools of varying characteristics for the field testing, so that the items could be tried out on different types of students. For the English field testing, in particular, some rural schools were visited. In each school the needed P3 and P5 classes were selected at random, as much as possible, from those available in the school. In each class, five boys and five girls were chosen randomly for the field testing.

In mathematics, two parallel versions of each field test were prepared and tried out, each containing one of the two selected items for each objective, mentioned above. Some of the English sections also had two versions.

For the mathematics tests, all instructions and most items were read to the students and, where necessary, translated into local language. Students proceeded through the items one-by-one at a pace determined by the oral presentation of the test administrator. The exception to this was the basic operations section, where the reading needed was very limited and students were allowed to proceed at their own pace. Each section of the test began with one or more examples of the format used, also orally presented.

For the English tests, all instructions and most items were also read to the students. Instructions were translated into local language, but, of course, the items were not. The only items the students had to read for themselves were in picture vocabulary and in reading comprehension. One or more orally-presented examples also preceded each section of the test.

For mathematics the results of the field testing varied considerably from school to school, and sometimes from class to class within schools. Overall class performance ranged from 20.4% to 40.4% on the P3 test, and from 6.2% to 41.1% on the P5 test. The P5 test was also given to a JSS1 class in three of the four schools as a validation check. In two of these schools the JSS1 class did considerably better than the P5 class, while in the other school it did considerably worse. It turned out that 14 P4 items were common to the P3 and P5 field-tests. For almost every available comparison, the P5 children in a given school performed better on these items than the P3 children in the same school.

The English field testing was supervised by Eddie Williams. Thus I don't have precise statistics as to the results obtained, results which may be included in his trip report. It was reported to me, however, that as a result of this field testing it was decided to drop the punctuation and capitalization, and dictation, portions of the test. In addition, the picture vocabulary items proved to be too easy, so some more-challenging word completion vocabulary items were developed and field-tested. It was also decided to rewrite the grammar sections of the tests. An additional story was added to listening comprehension for P5, so as to make this section go beyond just the picking out of discrete bits of information. I am told also that the reading comprehension passages proved to be rather difficult for many of the students.

Pilot Testing of New Test Forms

Based on the field-test results, pilot test forms were developed for P3 math, P5 math, P3 English, and P5 English. For the math tests this was a fairly straightforward procedure. Almost all of the items had behaved well in the field testing, and only a few small corrections to items were made. The main task was to choose between the two items field-tested for each objective. In most cases, either item would have served, but a slight preference for one or the other resulted in its being selected for the pilot test. In most cases, the easier of the two items was chosen. The P3 math pilot test ended up with 27 items, 2 being multiple-choice, while the P5 math pilot test had 45 items, 3 of these being multiple-choice. For the English tests, there were fewer items to select from, and less than universal satisfaction with the items that were available. However, it was felt that the two pilot tests that were ultimately formed were reasonably sound. The P3 English pilot test had 34 items, 18 of these being multiple-choice, while the P5 English pilot test had 45 items, also 18 being multiple-choice.

We were requested by the Chief of Party, Elizabeth Barcikowski, to include some demographic questions with the test, so that the demographic information collected could later be related to student achievement. Thus we constructed a short survey to go in front of the test booklet. Questions asked concerned the student's gender and age, the language used at home, the occupation of father and mother, and whether the student owned his or her textbook, and could use the textbook at home.

Test administration manuals were also developed by members of the two teams, with my assistance. Separate manuals were developed for the math tests and the English tests.

The pilot testing was carried out over the period November 29 through December 7. Again, the testing was implemented by the four team members, with the help of PME support staff. Nine schools were selected for the pilot testing, three from Northern Region, three from Central Region, and three from Greater Accra. In each region, one school was supposed to be a good school, one a medium school, and one a poor school. Thus the ratings A, B, and C, provided by the Ministry of Education, were used in school selection. While names of schools selected in Central Region and Greater Accra were

provided in advance of the testing, the schools selected in Northern Region were not made available to us until arrival of the testing team on site. This team reported that the schools visited in this region were of lower quality than expected, given the specifications that were to be used in guiding the selection.

Test administrators were instructed to administer the P3 and P5 mathematics tests in a school on one day, and to give the P3 and P5 English tests to the same classes on a different day. One randomly-selected class from each grade level was to be used, with a maximum of 40 students being tested in a class. The oral testing in English was to follow the regular English testing, but was to use only 5 boys and 5 girls randomly selected from those participating in the regular testing.

I observed the pilot testing only in the Greater Accra region, and what I saw went relatively smoothly, with few problems in administration. The survey questions on father's and mother's occupation, however, took a long time for students to complete, adding considerably to the total time for the testing. These questions should probably be omitted when the general testing takes place.

Reports from the testing team in the Northern Region suggest that students there have great difficulty with the English language, some not even being able to write their names. This, of course, also added to the time needed for test administration. This administration time for a test thus varied considerably, ranging from less than one hour to close to two hours, depending upon location, grade level, and subject being tested.

Near the end of the pilot testing, it was suggested by the Chief of Party that we also look at some P4 and P6 classes, to see if they perform better on the tests than the respective P3 and P5 classes in the same school. This would be to serve as a validation check for the tests. Thus we administered the P3 and P5 math tests to P4 and P6 classes in two of our pilot schools, one in Greater Accra and one in Central Region. We did the same thing with the P3 and P5 English tests, but for logistical reasons used different pilot schools in the two regions from those used for the math P4 and P6 testing.

Results of the Pilot Testing

The response data from the pilot testing were entered into the computer by PME staff using a coding scheme that I devised. The data files were then analyzed using the ITEMAN program. I analyzed the data myself, although I provided some instruction and exercises to PME staff on the use of this program. The overall results of the pilot testing are summarized in the following table. It should be noted that these results do not include the P4 and P6 testing that was described above:

	P3 Math	P5 Math	P3 English	P5 English
N of Items	27	45	34	45
N of Examinees	256	265	250	254
Mean Score	7.8	10.0	11.1	17.4
Mean % Correct	28.7%	22.3%	32.6%	38.8%
Highest Score	22	27	31	42
Lowest Score	0	0	2	3
Alpha Reliability	.84	.83	.87	.92
Mean Item-Tot. Cor.	.45	.32	.44	.48

For the sake of comparison, these results were also derived for the three Northern Region schools alone:

	P3 Math - North	P5 Math – North	P3 English – North	P5 English – North
N of Items	27	45	34	45
N of Examinees	66	61	66	60
Mean Score	4.7	8.4	6.7	11.0
Mean % Correct	17.6%	18.6%	19.6%	24.5%
Highest Score	13	24	12	38
Lowest Score	0	0	2	3
Alpha Reliability	.74	.80	.26	.86
Mean Item-Tot. Cor.	.41	.36	.20	.43

It may be noted that the English scores are somewhat higher than the math scores. This, however, may be due to the larger number of multiple-choice items in the English tests, producing a guessing factor which may inflate scores to some degree. The performance of the Northern schools is somewhat lower than the general average, particularly in regard to English. Thus the Northern mean of 19.6% on the P3 English test is not much above chance, probably contributing to the low alpha reliability obtained on the test with this group of students.

It must be remembered, though, that these tests are considered to be pretests for the purpose of collecting baseline data. They contain much material that has not yet been covered by the students in their classes, resulting in scores of zero or near zero on some items. It is hoped that when students in the general administration are given the tests for a second and third time, they will show improved performance over the baseline results, based on the learning that has taken place in the intervening period.

It remains to give the results on the oral portion of the English test. For both P3 and P5 tests, the maximum score possible is 10. For P3 the mean score for all children tested was 3.19, while in the North it was 2.60. For P5 the general mean score was 4.87, while the mean for the North was 3.87. Here again, the participating Northern schools showed somewhat less oral facility in English than the average of all participating schools.

Some Validity Information

It will be recalled that we administered the pilot tests to P4 and P6 classes in some of our pilot schools, two schools for math and two schools for English. The objective was to see if the tests could pick up the hoped-for difference in knowledge between P3 and P4 students, and between P5 and P6 students. The results of these comparisons are as follows:

	Mathematics							
	A School – Greater Accra				B School – Central Region			
	P3 Test		P5 Test		P3 Test		P5 Test	
	P3	P4	P5	P6	P3	P4	P5	P6
N of Items	27	27	45	45	27	27	45	45
N of Examinees	22	26	34	27	40	22	27	25
Mean Score	9.6	14.6	13.9	18.9	4.4	11.8	7.7	12.5
Mean % Correct	35.5%	54.0%	30.9%	42.0%	16.2%	43.8%	17.1%	27.8%

	English							
	C School – Greater Accra				C School – Central Region			
	P3 Test		P5 Test		P3 Test		P5 Test	
	P3	P4	P5	P6	P3	P4	P5	P6
N of Items	34	34	45	45	34	34	45	45
N of Examinees	38	36	32	41	32	32	38	36
Mean Score	15.0	18.0	25.2	26.0	10.0	7.2	12.8	11.6
Mean % Correct	44.2%	52.9%	56.0%	57.8%	29.4%	21.1%	28.4%	25.7%

Here it can be seen that in both schools in which mathematics tests were given to P4 and P6 classes, P4 students did better than P3 students on the P3 test, and P6 students did better than P5 students on the P5 test. Of course, for a given school and test, there were some individual items on which the lower grade level did as well as or better than the upper grade level. A variety of factors accounts for such a result, including the grade level of the item in question, the time period when the material contained in the item was taught, the quality of the instruction given, and the extent and nature of any later review of the material. Differences in the students included in the two classes also play a role. In addition, as the pilot testing took place in the early part of the school year, there was material on each test that had not yet been covered by either of the grade levels tested – P4 material on the P3 test, and P6 material on the P5 test.

The results of the P4 and P6 testing for the English tests were less supportive of test validity than the math results. In the C pilot school in Greater Accra, the P4 class did somewhat better than the P3 class on the P3 test, while the P6 class did only marginally better than the P5 class on the P5 test. In the C pilot school in Central Region, these relationships were reversed, P3 doing better than P4, and P5 doing better than P6. A possible explanation for these results comes from the testing team in this school, which reported that the P4 class tested had not had a regular class teacher for the whole term, and that the P6 class tested had just completed a final term test in the moments prior to

the administration of our pilot test. Naturally, many of the English items showed a better performance by the lower level grade tested, particularly in the latter school.

Not wanting to base any validity conclusions on the results from only two schools, we went ahead and made some further comparisons involving all nine of the pilot schools. It turned out that there were seven items common to the P3 and P5 math tests, and also seven items common to the P3 and P5 English tests. For each of these 14 items, the P5 students did better overall than the P3 students. For the seven math items, the overall performance was 16.4% correct for P3 students and 42.7% correct for P5 students. For the seven English items, the overall performance was 37.9% correct for P3 students and 55.3% correct for P5 students. Thus these items, at least, appear to be sensitive to the different degrees of learning of different grade levels.

Issues Raised

At this point in test development work, it is usually time to review the items in the pilot tests one-by-one, using the item analysis results. The objective is to see which items can be accepted into the final tests as they are, and which need further revision and testing. In fact, we began going through this process, with the goal of having final copies of the tests ready for printing by Friday, December 17. However, certain outside factors impinged on this activity, and affected the whole further course of the test development effort:

1. Complaints from the North

The Catholic Relief Services (CRS) is responsible for the QUIPS interventions in the northern regions of the country. Apparently some of their staff had observed the pilot testing up North, and were concerned about the difficulty and appropriateness of the pilot tests for their children, particularly in regard to the English tests. It was stated that the level of English up North is very low, and that children there are unable to perform on many of the items included on the English tests. The fear is that the tests are based at too high a level to catch any improvements in learning that might occur as a result of the interventions offered. In addition, some of the situations portrayed in the items were thought to be foreign to the experience of Northern children, and that this negatively affected their chances of answering those items correctly. An example was offered in the picture of a beach scene, used in the P5 oral test to elicit a verbal description of the activities taking place. Northern children, it was said, have no experience with beach scenes.

Thus the possibility was raised of revising the tests to respond to the concerns of the CRS people. Further pilot testing of the revised instruments up North was also proposed. The possibility of having two sets of instruments, one for the North and one for the rest of the country, was also said to be under consideration.

2. Observations of the Chief of Party

Following the pilot testing, the Chief of Party indicated that she was unhappy with the item format or form of presentation for some of the sections in the English tests. In particular, the word completion vocabulary items came under criticism, and, indeed, students did not perform well on these items. In addition, she did not like the way the grammar section was orally presented, thought the listening comprehension story for P5 should not require a written response from students, and thought that we needed some simpler reading comprehension items to answer complaints from the North. With respect to the reading comprehension passages, Dr. Barcikowski preferred that students circle answers to fill in the gaps rather than writing in the words themselves, thus avoiding the writing problems that were reported in some classes in the North. She also suggested that the presented answer alternatives might differ for each gap, rather than having a common set of alternatives for all gaps. Finally, she didn't see why all students needed to do the composition for P5, and suggested that only the pull-out students do it, giving them as much time as necessary.

The Chief of Party was also concerned over the P3-P4 and P5-P6 comparison data presented above, and the fact that the English test did not invariably demonstrate greater learning for the higher grade level. She was also worried over the large number of English items which showed better performance by the lower grade level in one or both of the schools for which these comparison data were available.

Actions Taken

On December 9 and 10, the two members of the math team and I reviewed all the items in the math pilot tests using the item analysis information. Only minimal changes in items were made, including the use of a few more names from the North in story problems, the use of a bicycle rather than an automobile in a story problem, and the reduction in difficulty of one of the example items used. We also agreed not to ask questions on father's and mother's occupation, as too much time was consumed in having students answer these questions. On December 11 we made some minor revisions in the test administration manual for mathematics.

On December 13 and 14, the two English team members and I reviewed the items in the English pilot tests. Based on the item analysis information, a few changes in item content and wording were made. However, the largest changes came from attempting to incorporate into the tests the suggestions of the Chief of Party, given above. Thus the word completion vocabulary items were changed to multiple-choice, with everything but the alternatives being read to the children. In the presentation of the grammar items, each sentence was now to be read four times, once for each alternative. The listening comprehension story in the P5 test was changed to multiple-choice. In addition, new simpler reading comprehension items were written to answer complaints from the North. Also, revised forms of the reading comprehension passages were developed to allow circling of responses from a unique set of alternatives for each gap. Finally, it was

planned to replace the picture of a beach scene in the P5 oral test with that of a playground scene.

It is clear that such major changes in item format and content as those described above require some additional field testing being carried out before the new and revised items can safely be included in the final forms of the tests being prepared. This additional field testing was planned for December 15 and 16.

However, on the afternoon of December 14 there was a meeting at USAID during which the tests were discussed. Present were the Chief of Party, the relevant USAID officers, and representatives of the organizations responsible for the QUIPS interventions. Apparently as a result of this meeting, it was announced on the morning of December 15 that the tests would not be given as scheduled in January. It was said that more time was needed to deal with the concerns over the tests described above, to decide what to do about the testing in the North, and to make revisions in the tests that would satisfy all parties concerned.

A rather subdued English testing team went ahead with the planned field testing of new and revised items on December 16. Two schools were visited. The testing seemed to go well, and the results suggest that most of the revisions were successful, but with still some question remaining over the most effective mode of presentation for the reading comprehension passages.

On that same day, December 16, I met with two USAID representatives for a debriefing on my activities in the test development effort. The following day there was a further meeting at USAID about the tests, at which I was not in attendance. During this meeting the concerned parties requested copies of the tests as they stood at that time. Presumably these parties will review the tests, and, in the New Year, will come up with suggestions as to what next should be done with them.

Some Final Observations

1. In my view, there are problems that can arise when those responsible for an intervention become too involved with the measurement instruments used to assess the intervention's impact. For one thing, it is conceivable that these intervention people, in their natural desire to demonstrate a positive impact for their intervention, may attempt to influence the form and content of the measurement instruments in such a way that this positive result is more likely to be obtained. For another thing, being intimately familiar with the measurement instruments to be used for evaluation purposes allows for the possibility of directing the intervention more particularly toward the material contained in these instruments, thus increasing the chances for a favorable result. While I am not saying that these factors are necessarily operating in the present instance, they do constitute considerations to be taken seriously where valid evaluation results are desired.

2. To guard somewhat against the second possibility mentioned above, the directing of interventions toward specific tests, it is suggested that new parallel tests, rather than the identical tests, be used for the second and third data collection points in the assessment design. This will also protect against the effects of possible leakage of the actual items used in the original tests. However, it should be noted that in the event that interventions are directed at material contained in the original tests, the later performance of students on tests that are parallel to the originals will also be affected.

3. While inputs into the test development activity are often helpful, it is best, where possible, that these be offered at an early stage in the test development process. When inputs on the form and content of items are provided after the pilot testing is completed, these, if taken seriously, are likely to set back the schedule of testing activities, and to discourage those involved in the testing enterprise. This, of course, is what has happened in the present instance.

Conclusion

During the assignment reported here, I have worked closely with Ghanaian colleagues in developing achievement tests to assess the impact of the QUIPS interventions – interventions supported by USAID to improve the primary education system in the country. I was very impressed with the commitment of these local colleagues to the test development effort – both the members of the test development teams and the PME support staff. These people were truly a joy to work with. I hope that our combined efforts will ultimately come to fruition in an accurate accounting of the effects of the interventions, so that the information obtained can be properly employed in planning further moves forward in Ghanaian primary education.