ADMINISTRATION OF EDUCATIONAL AND COGNITIVE TESTS IN THE CONTEXT OF THE GHANA LIVING STANDARDS SURVEY

Chris Scott, World Bank Consultant

Accra, July 1988

DN 2477

Background

A research proposal by Paul Glewwe and Peter Moock has been accepted for funding by the World Bank's Research Committee which involves the administration of cognitive and educational tests in the field on the basis of a representative household sample. The researchers have approached (among others) the Ghana Statistical Service with the suggestion that these tests be carried out on the sample of the Ghana Living Standards Survey (GLSS).

The Government Statistician was sufficiently receptive of the proposal to allow a pilot test to be conducted for one month in 5 of the 10 teams currently engaged in field work for the GLSS. This testing period coincided with my visit to Ghana in connection with other aspects of the GLSS, and the Government Statistician asked me to assess the feasibility of the proposal and any potentially negative impact it might have on the ongoing survey work.

I was able to visit two areas in southern Ghana* and to observe 4 testing sessions. I also talked to the staff concerned.

This report assesses the following issues:

- Is the project desirable?
- Is the administration of the tests in field conditions feasible?
- Do the tests impose an unacceptable burden on respondents?
- Do the tests impose an unacceptable burden on the field team?
- Can the respondent burden be reduced? How?
- Can the burden on team members be reduced? How?
- Sample size
- The schools questionnaire.

The report ends with the consultants recommendations. Some notes on detailed problems with the test materials appear in the appendix.

---

* Ashiaman, a dormitory town near the industrial port of Tema and Atuakrom, a remote village near Nsawam in the Eastern Region.

Is the project desirable?

Comparisons of survey data on years of schooling with reported literacy suggest that the quality of primary schooling in Ghana may be exceptionally low by African norms. It seems obvious that further information on this important issue would be of great concern to the Ghana Government, and indeed the Ministry of Education has expressed a positive interest in the research project under consideration.

In the opinion of this consultant the above consideration leaves no room for doubt about the importance of collecting "output" measures of educational attainment in Ghana, for comparison with the standard input measure, years of schooling.

I am less convinced by some of the other arguments advanced on behalf of the project in the research proposal. These revolve around the question of the efficiency of modelling of education in relation to economic achievement. It is not obvious that failure to measure an educational output variable will lead to biased conclusions regarding the effect of educational inputs on socio-economic behaviour. Doubts also arise as to use of the Raven's Progressive Matrices test as a supposed measure of innate ability. The hypothesis that such tests are culture-free and education-free has been widely (and rightly, in my view) questioned; it becomes particularly unconvincing where the range of educational levels is very great.*

Despite these reservations it is certainly true that a knowledge of the relationship between the educational input and output, and between the latter and socio-economic performance, will improve our understanding of the socio-economic impact of education. At the very least such data would throw light on the crucial and urgent question whether future inputs would be more productively directed towards quantity or quality.

Thus on any basis the project seems certain to provide valuable information. Moreover there seems to be no way of collecting such information other than from a general population survey and clearly this can be done more efficiently in the framework of an already existing survey.

I therefore strongly support the desirability of the project.

Is the administration of the tests in field conditions technically feasible?

This consultant has experience of administering tests of the kind involved in this project to many thousands of subjects in strict classroom conditions. Inevitably the rigour that can be imposed in the classroom is not obtainable in the conditions of a household survey. However, from the small sample of observations which I was able to make it appeared to me that the defects of the testing environment in a Ghanaian

* One only has to watch illiterate or semi-literate subjects struggling with the problem of completing the answer sheet to appreciate at least one of the learned skills measured by Raven's.

household survey, though real, are unlikely to cause serious distortion of the test scores. The significant problems encountered are discussed below.

First there is the problem of subjects helping each other. I saw only one attempt to seek such help and it came to nothing. It should be noted that the group normally consists of the eligible members of the selected household only, typically 2 - 4 people at the 1st round and 1 or 2 at the 2nd round. The session seems to be typically carried out in the open air - if only because indoors the light is too poor, even during the day - and generally in the courtyard of the household's dwelling or perhaps the village square. Here there is normally no difficulty in seating the subjects at least 2m. apart. In these circumstances S's cannot see each others' work and could only help each other by asking and answering questions aloud, and such questions would have to be very specific, such as: "What's the answer to question A4?". I never heard such questions and one feels that they would rarely be formulated - or answered if they were. In any case an only moderately alert tester can easily intervene to put a stop to them.

A second potential problem is that of spectators. These can be numerous. A tester may find it difficult to disperse the crowd but it is easy enough to make them stand back - at 3m. or more, say, from the S's. At this distance they cannot learn anything useful nor give any help, and in fact they very soon lose interest and drift away.

A third problem arises from the illiteracy of some test subjects in the Raven's test. It is certainly too optimistic to state (as the tester's manual does) that Raven's "can be given to groups of people up to a maximum of 10 persons". Any group larger than 3 in Ghana has a substantial chance of including more than one illiterate. Testers are instructed to deal with an illiterate S by sitting beside S and having him/her point to the answer; the tester then completes the answer sheet himself. Clearly he can only do this with one S at a time. The supervisor and the anthropometrist might perhaps be brought in to help with this task, given a minimum of initial training. However some bias seems inevitable. Imagine two illiterates, one of whom really cannot handle a pencil or write numbers, while the other can just barely manage this but is totally inexperienced with any kind of paper work. The clerical task of indicating the chosen answers in the right place in the answer sheet must be solved by the second S without help, while the first S, who has less ability, is excused this task altogether.

Only the last of these three problems appears serious. The Raven's test has a procedure for recording answers which assumes significant familiarity with paper-and-pencil clerical work. (For example, S needs to know that each response-pattern is denoted by the number which appears above it. He can, of course, use intelligence to infer this, by checking that not all patterns have a number below, while all have one above. But the amount of intelligence required to conceive of this check may exceed that required to answer the first few test items.) If we are going

to help some S's to solve this problem but not others, we have a bias from the start. Perhaps a few more "worked examples" would be helpful before the test proper begins?

A related problem that is likely to cause bias is that the 2nd round mathematics test assumes a significant knowledge of English. For example, the words "represent" and "shaded" will not be understood by all who claim to read English. This problem might be met by simplifying the English used. (See notes in the Appendix.)

The general conclusion is that, with some modifications in the questions and instructions, it should be possible to administer the tests in the course of a household survey without serious distortion in the scores.

Do the tests impose an unacceptable burden on respondents?

In the 5 interviews I witnessed in Ghana only one respondent expressed overt impatience. (There was also only one household among the 4 whose interviews I watched in Ivory Coast which showed impatience.) These are of course very small samples. Certainly the GLSS interviews are extremely long and must quite frequently cause irritation.

However, like the anthropometric measurements, the tests represent a break in the monotony of the long series of questions. They seem to come as a welcome diversion rather than an added burden. Moreover in the villages the cleverer S's seemed proud to show off their greater skills in front of less educated friends. However this conclusion, itself only a subjective impression, is based on observation of households which are not sufficiently modernized to feel that their time is valuable: they are pleased to be offered something new to do, rather than irritated to see their time purloined for an activity which interrupts their chosen pursuits. It might well be that wealthier or urban households would show greater resistance. This can only be determined by observing reactions over a larger sample. If no adverse reactions are found during the current pilot tests it will be safe to assume that the problem can be ignored.

Meanwhile it may be noted that by far the greatest burden, the two 2nd round tests, takes place after the final interview. If these cause a negative reaction at least it will not reflect on any of the GLSS results.

Do the tests impose an unacceptable burden on the field team?

We assume first that the test work is done by a tester attached as an additional member of the team. In these circumstances additional burdens might arise as follows:

(a) One more person must be carried in the team's vehicle. This does not appear to constitute any appreciable burden.

(b) Lodging must be found for one more person. Occasionally this may cause difficulty. If the tester is female she should be attached to one of the teams which has a female

member so that accommodation can be shared if necessary. With this precaution the extra burden should be negligible.

(c) Testing is normally carried out at the end of the interview. This means that when the team has finished in the cluster there will be an additional delay of up to ½ hour for the 1st round or ¾ hour for the 2nd round before they can leave together in the car. Normally this delay will occur only at the end of the week's work, but in a rural cluster with several villages it is liable to occur after each village. Even so, such additional delays cannot be regarded as a great hardship.

If, secondly, we assume that the testing work is to be done by a member of the existing team, we seem to have two possible alternatives: the supervisor or the anthropometrist. If the job were given to the supervisor there is no doubt that the work of supervision would suffer; this does not seem to be a good solution. If the work is done by the anthropometrist the problems mentioned above under (a) and (b) disappear, though problem (c) remains. In general the anthropometrist seems to be currently under-employed. In the 2nd round, in particular, his main job* is to re-measure and re-weigh a 50% sample of the persons measured and weighed 2 weeks before. Unfortunately this work also comes after completion of the interviews (at least in the 1st round), so that the problem (c) is actually aggravated when we use the anthropometrist for testing. (But see below for possible strategies to reduce the problem.)

How can the respondent burden be reduced?

A great deal has already been done to reduce the respondent burden. Marginal sub-populations are eliminated from the testing and ingenious filtering arrangements have been introduced which take advantage of the 2-stage survey structure.

One further step seems desirable: the short mathematics and reading tests could be timed at 5 minutes instead of 8. Observation (and consultation with the field workers) confirms that almost invariably all work has stopped by the end of the 4th minute. Consideration should also be given to shortening the time limits for the 2nd round tests, perhaps to 15 minutes. These tests are in the true sense timed tests - S's are working against the clock. In these circumstances it is actually an advantage to fix the time limit so that no S finishes the test: this maximises the variance of scores and hence the discriminatory power.

How can the burden on the team be reduced?

Training the anthropometrist seems the best solution. In addition it seems reasonable, now that a year's data have been collected, to eliminate the re-measuring and re-weighing programme of the 2nd round. (If this is not acceptable to analysts, one might eliminate re-measuring but retain the re-weighing, perhaps with a reduced subsampling rate, currently 50%.)

---

* His only other duty is price collection.

This would give the anthropometrist more time to spare at the 2nd round for conducting the relatively long tests of that round and hence reduce waiting time for the rest of the team.

Note that a possible mode of organization would be to do all the testing at the 2nd round. One could either do the short tests at the beginning of the week, even before the interviews, and the long tests later, or each household could do the short tests and then move straight on to the long tests (after a brief interval for marking and eliminating persons who do not qualify). It is possible one or other of these procedures may be found more convenient: teams should be encouraged to experiment to find their own preferred solution. One possible advantage of confining the testing to the 2nd round would be to reduce the refusal rate if, as has been suggested, households are currently refusing the 2nd round altogether on the grounds that the 1st round is too burdensome.

Sample size and subsampling

I was not asked to comment on sample size as such, but I was asked to consider the question whether all GLSS households in the selected cluster should be included in the testing programme or only a proportion of them (one half?), or whether possibly there might be a sampling of persons within households.

Since the testing programme takes about 20 minutes (1st round) or 45 (2nd round) per household while the interviews average about 2 hours per round, there is enough time to test all households during the interviews (assuming 1 tester and 2 interviewers) and little or nothing would be gained by sub-sampling households for the tests. There might be some saving of time at the end of the 1st round in each cluster, or each village, and this could be ensured if interviewers could be persuaded to leave until the end the households not selected for the test, so that the last testing session in an area could be conducted simultaneously with the last interviews. On the whole the advantage does not seem worth the loss of half the sample.

Since all eligible members of each household are normally tested together there is negligible saving in sampling persons within households and there would be considerable added complexity. This strategy is not recommended.

The school questionnaires

I do not have full information on the objectives of these questionnaires. There is one for primary schools and one for middle/junior secondary schools, though only the former is mentioned (very briefly) in the project proposal. These are lengthy questionnaires and the sample does not seem to represent any defined population of schools. The burden of data collection must be considerable.

On the information available to me I cannot comment on these questionnaires except to say that this part of the study seems to require fuller specification and justification.

Conclusions and recommendations

1. The proposed testing programme is important and should yield very valuable data.

2. It is also feasible to conduct in the context of the GLSS.

3. The additional burden on respondents is likely to be a problem only among urban and wealthy respondents. Pilot results from Accra should be evaluated to see whether there were negative reactions in higher class areas.

4. There is an additional burden on the team but, again, this will be minor in most cases.

5. The following modifications to procedures are suggested to reduce these burdens:

   (a) Reduce the time limit for the 1st round mathematics and reading to 5 minutes (possibly 4?).

   (b) Reduce the time limit for the 2nd round mathematics and reading to 15 minutes each.

   (c) Use the anthropometrist as tester. He will need at least 1 week's training. All other team members should be given a day's training to enable them to assist illiterates in the Raven's test and to stress the importance of not revealing the correct answers.

   (d) To facilitate (c) above, reduce the anthropometric work required at the 2nd round. Preferably eliminate it altogether. If this is not acceptable eliminate re-measurement but retain re-weighing, preferably with a lower (1-in-5?) subsampling rate.

   (e) Whether or not the tester is the anthropometrist, there is no need for <u>2nd</u> round testing to wait until after the 2nd round interview. If the later-interviewed households are tested before their 2nd round interview the problem of the tests delaying the team's departure from the cluster is eliminated. (The problem remains for the 1st round, but here the delay caused by testing is shorter.)

6. It appears desirable to include in the testing programme <u>all</u> households, and all eligible persons, covered by the GLSS in a given cluster. Whether the total of 200 clusters should be covered or only a subsample of them is outside my terms of reference. It depends on objectives and resources.

7. The tests and instructions could be improved at many points. Suggestions are listed in the Appendix.

A P P E N D I X : Notes on the test materials

1. The answer sheets all need re-designing. There should be provision for an ID linking S to the household interviewed. With this, sex and age are presumably not needed? Are the times of starting and ending needed? If so, on the Round 1 mathematics and English answer sheet the times should be shown separately for these two parts.

Ravens

2. Answer sheet. The additional line at the bottom of each column is confusing and should be deleted. Probably the commonest error is getting out of step. Possible solutions to this: print the name of the colour of the test item beside each answer box? Give more emphasis to page numbers in the test booklet, and use a simple numbering system from 1 to 36. (The numbers could be changed by hand in all the booklets before field work begins.) The problem is rightly emphasized in the training manual for testers, but they still did not seem to be checking often enough.

Round 1 mathematics

3. The multiple choice version is needlessly complex: open response seems simpler for S and just as easy to mark.

4. Answer sheet. In one version I saw the lines for entering responses were too short and too crowded. S did not understand she was expected to squeeze in the response on the line. (But in most copies the spacing seems OK. Care should be taken if re-typing is planned.)

5. Answer to item 8: Official answer is 5 R6 but the following should be counted correct also:

$5\frac{6}{7}$ 5.9 5.8 5.86 5.85

6. Interpretation. In interpretative comments on findings it should be borne in mind that when marks are low (say 4 or less) we are almost certainly testing knowledge of the notation almost entirely, rather than mathematical skill. It would be interesting to try the following experiment on those who fail to get item 1 right: show S a ₵100 and a ₵200 note together and ask how much money that is. This would be a more realistic test of those who "really" know that 1 and 2 make 3.

Round 1 reading

7. Would be better entitled "English reading test".

8. Item 8. "best title" is surely an excessively subjective concept for an objective test. Best for what purpose? I presume answer A is right, yet there is only one sentence about John's learning to read while there are 4 about schools: perhaps if you want to attract Ghanaian readers

D would be a good title. I would suggest replacing this item by something more objective (hence less culture-bound).

Round 2 mathematics

9. Item 2. I did not know this notation (I thought it was a square root, badly made). Ghanaians all seem to know it but what about persons educated in a neighbouring French-speaking country?

Item 3. The word "represent" is difficult. Replace 2nd sentence by: "How do you find out the number of chairs?"

Item 4. Similar change.

Item 5. For "shaded" read "dark"?

Item 9. Read: "... which decimal shows the amount that is dark?"

Item 14. Delete "AB represents" and delete ", which"

Item 16. I suggest inserting short vertical marker lines along the x-axis at each unit value. Add s after "Thousand". Delete the horizontal line below this.

Items 18 and 19. Some Ghanaians have learnt the imperial system, others the metric system. Perhaps it would be better to drop these two questions?

Item 21. Mainly tests knowledge of the word "perimeter". Is this a useful objective?

Item 24. Illogical. The question cannot be answered.

Item 25. Eldoret - spelling.

Item 27. For "if" read "of".

Item 35. Correct response (C): insert "on" before $\overline{AC}$.

10. The test is surely unnecessarily long. Drop 31 - 34? Does anyone ever get that far?

Round 2 reading

11. Item 2. "best title" again. Most journalists would say B!

Item 13. Few Ghanaians have ever heard the screech of a train's wheels (I'm not sure whether I have). Why not change to a car? (Incidentally "railroad" is US English: I think Ghanaians say "railway".)

Item 17. Seems too culture-specific. Items A and B will be unknown to many. "Honk" may be unknown.

Item 21. To expect knowledge of the phrase "trading posts" seems too demanding.

Testers' Manual

12. The following points are all trivial and would only be worth changing if it is planned to re-type or re-run the manual.

Page 1. For "fill out" read "fill in" 5 times on this page. (Ghanaians more often use British English than US English.)

Page 2, line 5. For "nine" read "9" for greater clarity.

Page 5, 3rd line from end. Add after "answer": "after the first two items."

Page 7, middle. For "math" read "maths". (Brit English again.)

13. Page 2. "who have had at least 3 years of schooling" may be ambiguous. Better to insert "completed" before "years". This appears in the 1st para. on lines 7, 9 and 14.

14. Page 2. Last sentence of 1st para. As noted in my report, time could be saved if the rule about testing always after the interview were changed for the 2nd round. I would recommend this. I do not think it will lead to any refusals of GLSS 2nd round interviews.

14. Page 3. 2nd para. Is there any point in giving this explanation?

15. Page 7, middle. For "village" read "EA", since the teams also work in towns.

16. Page 7, 5th and 7th lines from end. The idea of tests of this kind "measuring" something may be a little too abstract. I suggest replacing "measure" by "find out".

17. Page 8, 2nd para., 2nd sentence. Reading the questions aloud to S. This will rarely be of any help. Anyone who can understand this kind of English when read aloud is likely to be able to read it himself. The real problem is with those who do not understand English. Is the tester supposed to translate the questions into Twi? Surely this is not practicable for the long maths test, and not necessary for the short.

18. Page 8, 1st word. For "village" read "cluster".

19. Page 11. "Fill in" for "fill out" again twice here.