

APPENDIX B

GRANT HENNING

LANGUAGE TESTING CONSULTANT

TRIP REPORT

AUGUST 28, to SEPTEMBER 10, 1981

BEST AVAILABLE

CONSULTATION IMPRESSIONS AND COMMENTS

1. The number of schools and sampling method seem adequate.
2. The primary difficulties will be:
  - a. preparing simple enough items .
  - b. devising tasks that are valid for measuring reading comprehension and listening comprehension.
  - c. preparing enough items per task to ensure reliable measurement without firing students.
  - d. standardizing procedure so that identical instructions and practice activities are available to all language groups.
  - e. piloting the test in order to reject inappropriate items, provide tentative reliability and validity measures, and develop possible equated forms before the November administration.
  - f. devising a procedure to identify nonvalid participants (e.g. impulsive markers, guessers, cheaters, hearing impaired, anxiety crippled, non-attentive) so as to prevent them from contaminating the data.
  - g. reducing administration time by streamlining classroom procedures as much as possible.
3. One possible measurement problem depending on one's philosophy of instruction is that the tests will be developed and administered to control before the broadcasts. This will permit foreknowledge of control group weaknesses and design of lessons to the advantage of the experimental group, in terms of the instrumentation employed. This need not occur if lesson developers are not shown control group results.

4. Further on sampling procedure, the stratification of schools on general examination results is probably a superior approach to that of the Nicaragua experiment which stratified on urban-rural location. Ensuring the greatest possible ability spread will probably enhance reliability of measurement.
  
5. The Nicaragua project achieved .82 KR20 reliability on their \*28-item first-year math pretest, achieving a standard error of measurement of 1.91, indicating a test standard deviation of 4.50 on their Spanish revision of the TOBE. By Spearman Brown Prophecy formula, this suggests a similar test of 55-item length would have produced .90 KR20 reliability; 35 items would have produced .85 reliability. This suggests that, although content and sample are different, it would be inadvisable for the present test to contain fewer than 35 items and unnecessary to exceed 60 items. Flexibility in that range will depend on ease of the tasks involved, total time of test administration including instructions, distribution of forms, etc., and also upon the level of reliability desired for subscales of the test. Thirty items of listening comprehension and thirty reading items would possibly permit a minimum acceptable level of confidence in the respective subtests, provided the test is piloted and revised on a sample of 100 or more representative children.

\* (This was a pretest achieving a mean of around 20 out of 28. Presumably more items could be used on a post-test because of greater maturity of the children.)

6. Test administration will need to be conducted by trained project staff rather than teachers with a vested interest in the comparative success of their classes. This is a serious matter --otherwise the entire procedure may fail.
7. Of course children will need to be tested before they leave for summer or term break. It is important to be sure that testing dates preempt any premature exodus from schools. In Egypt boys begin dropping out in March for various reasons, although teachers are paid to teach until May.
8. If it proves undesirable to employ a minimum of 60 items (30 LC and 30 Rdg) (preferably 35/35 = 70), a matrix sampling procedure may be followed similar to the Nicaragua project. Here 120 items would be required (60 LC and 60 Rdg), but any given student would receive no more than 30 items (15 LC and 15 Rdg) or 40 items (20 LC 20 Rdg) depending on the matrix pattern. The matrix sampling procedure has advantages:
  1. students are not overburdened in the testing situation: the task is reduced.
  2. a variety of forms could be available in each class to prevent cheating.
  3. more items, hence, achievement objectives are tested: more diagnostic information.
  4. an element of random assignment of treatment form to subject may be introduced.

It also has profound disadvantages:

1. the sample is effectively reduced to a third or a fourth the size for statistical inference; i.e., fewer completed tests are available.
  2. analysis of results becomes vastly more complex.
  3. simultaneous or group testing (LC) becomes complicated and confusing to the children if forms are individualized; i.e. if there is variation within class.
  4. coordination of administration procedures is greatly complicated.
9. In the situation where three groups are employed in the design (21 -school experimental and control groups for summative evaluation, and the observation formative evaluation group), the summative posttest could be given to all three groups separately to test treatment effects and treatment plus formative evaluation effects.
10. It is crucial that not one test form applied to control group children before the treatment is administered to the experimentals should pass into the hands of the teachers or the children. Otherwise experimental teachers may teach to the test.
11. In light of difficulty 2.f. discussed above, I recommend that the summative test data be subjected to Rasch person fit analysis in addition to the traditional analyses necessary. This could be done using American University in Cairo's BICAL software at very minimal cost. The advantages would be (1) that nonfitting persons (e.g. cheaters, hearing impaired, non-attentive, etc) could be identified and prevented from contaminating the data, (2) reliability of measurement might be substantially increased, and (3) all items could be calibrated and tested for fit to a latent trait model.

12. Unlike procedures in the Nicaragua project, measures of test validity should be included. This is particularly true since both listening comprehension and reading comprehension are being measured and generalizations are being made from these measures. This procedure could take the form of multitrait-multimethod validation (Campbell & Fiske, 1959) where two methods of assessment (say, recognition and production tasks) may be employed for each of the two traits (listening comprehension and reading). This would permit inferences about construct validity from a simple 4 x 4 matrix of correlation coefficients.
  
13. The formative evaluation should include some of the following components:
  - (a) an observation rating form for the observer to note on a Likert scale the extent to which students are attending, responding, following instructions, etc., for each segment daily.
  - (b) an affective questionnaire for teachers, observers and possibly students to indicate on a Likert scale the extent students enjoy the different components of each lesson or series of lessons;
  - (c) a cognitive criterion test to measure the extent to which children have mastered objectives of instruction for each lesson. These may be administered on a daily, weekly, or monthly basis, depending on the feedback needs for formative evaluation;
  - (d) a free-response component for the observer to make general comments on each lesson after it has been presented. Here it will be useful to have several observers simultaneously observing several different classrooms to compare responses on a and d above.

14. It was noted in group discussion that there are at least three basic ways to design the summative test.

(a) A 60-item instrument would be written and piloted with 30 items of listening comprehension and 30 items of reading comprehension. This would probably satisfy reliability needs, but might prove exhausting for the children, even if total administration time could be reduced to 45 minutes, which would be a useful target time.

(b) A matrix sampling procedure could be followed using 60 items of listening and 60 items of reading. By this procedure each student would receive no more than 30 (i.e., 15/15) or 40 (i.e., 20/20) of the total 120 items, depending on the matrix variation. One variation is illustrated as follows:

	Listening				Reading			
Group 1	1A				1B			
2		2A				2B		
3			3A				3B	
4				4A				4B

15 items per cell

Another variation could be:

	Listening				Reading			
Group 1	1A				1B			
2		2A				2B		
3			3A				3B	

20 items per cell

63

Advantages and disadvantages of these approaches were noted in comment number eight above.

(c) A Rasch Model approach would permit selection of an appropriate number of items arranged along a difficulty continuum for each subtest. Children could be encouraged to stop when items become too difficult for them. Thus also some administration time might be reduced in this way and frustrations minimized. While this overall approach has much to offer, it is probably too radically innovative to introduce at this point in the plan.

From the viewpoint of the children the approaches would probably rank in the following order of preference: b1, b2, c, a. From the perspective of measurement theory the prioritizing would probably be: c, a, b2, b1. From the standpoint of the statisticians who have to analyze the data and make sense of it all the preferred order may be either a, b2, b1, c or b2, b1, a, c, depending on whether they desire more information on a greater range of objectives with less statistical power (latter priority order) or greater statistical power with less specific information tested (former priority order).

By weighting the four priority orderings equally and averaging, option b2 appears the strongest. By this option children would receive 40 items according to the second matrix above. This would also permit construct validation if 10 items are tested in each of two modes for both skills measured.

15. In staff discussion five objective categories were identified for testing of listening and five for testing of reading. Each of the five persons present agreed to prepare five items in each of two objective categories, one under reading and one under listening. After editing, this will give us fifty items for tryout on a mini sample of a maximum of 30 children of standard one this week. By correlation of objective category scores with total scores for reading or listening it should be possible to identify the most promising two objective categories within each of the two general skills. Hopefully common differentiating qualities will be found for the two sets of objective categories (e.g. production, recognition) to permit construct validation. The plan after the mini-pilot would be to prepare about 160 items over all structures and vocabulary in the syllabus in two general skills, two objective categories; and two administrative modes -- about 20 items in each of eight developmental cells. These would be piloted and analysed next week for about 100 children.

The five within-skill objective areas mentioned above were:

Listening Comprehension

1. responding to instructions or implications
2. recognizing sound contrasts.
3. word recognition (meaning/form) with pictures
4. sentence comprehension (dictation)
5. answering questions.

65

Reading Comprehension

1. naming upper and lower case letters of the alphabet, ordering letters.
2. reading vocabulary, matching written to written or spoken words or structures to a picture.
3. analysing structure to read new words including plurals, inflections,
- 4 matching pictures with words and sentence options
5. cloze recognition with auditory stimuli.

16. On September 3, 1981, the Min-pilot Summative Evaluation Test (MSET) consisting of 50 items in two skill areas and 10 testing formats was administered to 30 standard one children of Kahuho Elementary School. The school was selected because of its proximity to Nairobi, cooperation of its headmaster, and presumed median ability of the children. Muitungu, a project staff member who is a native speaker of Kikuyu, the language of the children, administered the test following a day of rehearsal at the project centre. The other staff members present assisted in distribution and collection of materials and in timing of the segments of the test.
  
17. The purpose of the MSET administration was to determine two listening and two reading item formats which would be best from among the ten formats described in no. 15 in terms of probable reliability and validity of the Final Summative Evaluation Test (FSET). It would be possible to compute measures of reliability (KR-20 and KR-21) and validity (predictive and construct) for all sub-scales of the MSET, and based on these estimates decisions could be made about the characteristics of the Pilot Summative Evaluation Test (PSET) to be administered to a larger sample of children (about 100) on September 9, 1981.
  
18. Results of the administration of the MSET may be summarized in the following tables:

TABLE I  
MEANS, STANDARD DEVIATIONS, CORRELATION COEFFICIENTS AND PREDICTIVE  
VALIDITY COEFFICIENTS FOR FIVE SUBSCALES OF LISTENING COMPREHENSION  
AND FIVE SUBSCALES OF READING COMPREHENSION.

(N = 30)

L I S T E N I N G

R E A D I N G

	1	2	3	4	5	T	1	2	3	4	5	T	GT
M	2.400	2.400	3.733	1.633	1.433	11.600	4.500	3.067	3.033	1.433	1.567	13.567	25.167
S	1.793	1.499	1.413	.850	.935	3.874	.820	1.484	.850	1.104	1.406	3.821	6.716
r	.784	.622	.881	.027	.287		.424	.814	.376	.716	.811		
r <sub>c</sub>	.460	.288	.737	-.189	.048		.225	.591	.163	.522	.604		

The correlation coefficients of Table 1 reflect the correlations of the individual subscale formats with the total scores for the skill area (listening or reading) which they represent. The bottom row presents these coefficients after correction for part-whole overlap to remove the contribution of the item format score to the skill area score. As such the bottom row reflects an estimate of predictive validity; i.e.

89

the extent to which each item format subscale predicts a more general measure of the skill in question. Based on the magnitudes of these coefficients, the decision was made to employ listening formats 1 (responding to instructions) and 3 (word recognition with pictures), and reading formats 2 (matching words to a picture) and 5 (cloze recognition with auditory stimuli). See parts one and two of the MSET.

TABLE 2  
 KR-21 RELIABILITY ESTIMATES OF MSET  
 BEFORE AND AFTER SELECTION OF PREFERRED  
 ITEM FORMATS (N=30)

	N of Items	original Reliability		N of Items	final Reliability	
			KR-21		KR-21	KR-20
Listening	25		.610	10	.805	.853
Reading	25		.599	10	.735	.807
Total	50		.738	20	.838	.880

BEST AVAILABLE COPY

It is important to note that KR Formula 21 provides a slightly more conservative estimate of reliability than KR Formula 20. The estimates above indicate that the reliability achieved with MSET with only 20 items is already considerably higher than that of the TOBE, with 28 items, employed in the Nicaragua Project. And the MSET is only a trial instrument. Some comment is warranted about the construct validity of the MSET. Initially, the reading and listening total scores were correlated for the original 50-item instrument (.524) and for the edited 20-item instrument (.531). This indicated that the skills of listening and reading comprehension were distinct as measured. The correlations were <sup>low</sup> enough to warrant a conclusion that something different was being measured in the two skill areas of the MSET. It is also important to note that individual formatting subscales correlated more highly with their own skill area than with the other skill area employed. This is indicated by the results in Table 3

TABLE 3  
CONSTRUCT VALIDITY OF THE REVISED  
MSET (N=30)

	L1	L3	R2	R5
Listening Total r	.784	.881	.539	.568
r <sub>c</sub>	.460	.737	-	-
Reading Total r	.394	.515	.814	.811
r <sub>c</sub>	-	-	.591	.604

Note that even after correction for part-whole overlap, the chosen subscale formats were clearly more highly related to their own skill area than to the other skill area tested. This was particularly important with R5 subscale format since for this task category an auditory stimulus was employed with a predominantly reading task.

19. The high observed reliabilities of the MSET with comparatively few items lead me to revise downwards my original estimate of required numbers of items on the PSET and FSET. Probably 80 good items would be sufficient, i.e., 20 in each of four subscale formats.
20. Regarding administration time of the MSET, this is summarized below in minutes.

Test Segment		Distribution and Explanation	Administration	Total
Trial Sheet		8	-	8
Listening	1	4	6	10
	2	5.5	3	8.5
	3	2.5	3.5	6
	4	2.25	4.25	6.5
	5	2	6.5	8.5
a break of 13 minutes was allowed between sections				
Reading	1	5.25	2	7.25
	2	3	4.25	7.25
	3	2	2	4
	4	2	4.75	6.75
	5	1.75	5	6.75
grand total				79.5

Speed of administration probably increased as children

BEST AVAILABLE COPY

70

gained familiarity with tasks.

21. Considerations about summative evaluation procedures --
- (a) It would appear desirable that the summative evaluation schools be visited at least once a month to ensure that the radio broadcasts are being fully utilized, to check to see that the radio and other materials are fully operative -- supplying batteries, materials, or replacement radios where needed, to verify that control group students are not being exposed to the broadcasts or the supplementary materials, to gather useful anecdotal information from the headmasters about the application of the broadcasts, and to alert schools about summative testing dates. This would require at least one person visiting one summative school per school day throughout the academic year. A one-page report of each visit should be prepared on a form sheet prepared for this purpose.
- (b) The final summative evaluation test will probably require nearly two hours administration time, counting instructions, distribution, and a 15 minute recess in the middle. Minimally this would require three teams of two persons a period of seven consecutive school days (nine days total) to administer. An additional three days would be needed to include the formative schools in the FSET administration. The team should be chosen so that ideally one of the members could speak the native language of the children in each school visited. A formal schedule should be devised indicating who is travelling to which schools on which days, allowing adequate travel and hotel time each case. A one-page written account of the administration in each school should be prepared immediately afterwards. It should note any deviations, special problems, or irregularities observed in test administration and timing, as well as in the participants themselves. Form sheets should be prepared for this purpose.

(c) The test should consist of a half-page, familiarity exercise sheet, recorded native-language instructions on high quality portable cassette players, and two eight-page test booklets with five test items per page.

22.

Formative Evaluation Suggestions:

(a) Formative evaluation schools should be visited at least once per week by a team of two persons. Each visit should include (1) formal observation of a broadcast lesson in the classroom(s) using an observation form, (2) administration of a criterion-referenced test of 25 - item length, including five critical items from each lesson taught since the last visit to the school, (3) administration of a brief affective questionnaire concerning children's appreciation of the broadcast for that day, (4) collection of observation forms from resident field observers, (5) a brief prepared interview with the teacher(s) involved that week in the classroom(s), (6) a check on the radio equipment, ensuring that it is operational with sufficient batteries, and (7) distribution of materials to observers and schools as needed.

BEST AVAILABLE COPY

(b) This procedure would require at least two teams of two persons each visiting one different school each day. A visitation schedule might appear as follows:

Lesson day	1	2	3	4	5
team	1	1	1	1	1
school	1	2	3	4	5
team	2	2	2	2	2
school	6	7	8	9	10
Items/lessons	5/1	10/1-2	15/1-3	20/1-4	25/1-5
lesson day	6	7	8	9	10
team	1	1	1	1	1
school	6	7	8	9	10
team	2	2	2	2	2
school	1	2	3	4	5
items/lessons	25/2-6	25/3-7	25/4-8	25/5-9	25/6-10

12

In this way cognitive feedback should be available from every school for every lesson taught. Affective and interview feedback should be available from two schools for each lesson. Formal team observation should be available from two schools for each lesson. Additional resident observer observations could be available to increase the number of schools per lesson; however, great caution is necessary not to allow untrained, non-project-related persons to contaminate the procedures.

- (c) In all, formative and summative evaluation school visitation by this plan would require three project cars and five persons constantly on the move throughout the year. During summative evaluation three cars and six persons would be needed for a minimum of seven consecutive school days - or ten school days if formative schools are also post-tested, as they should be, including formative controls to permit formative/summative school comparisons.
- (d) During the year field persons should be debriefed once each week, when all their forms should be collected and filed against each lesson concerned, and the requisite sets of forms for the coming week could be supplied. This might take place on Friday afternoons. If changes occur in observation team personnel, it would be useful if this did not happen mid-week, but at the weekend.

**Best Available Document**