

UNIVERSITY OF CALIFORNIA, LOS ANGELES

UCLA

BERKELEY · DAVIS · IRVINE · LOS ANGELES · RIVERSIDE · SAN DIEGO · SAN FRANCISCO



SANTA BARBARA · SANTA CRUZ

SCHOOL OF PUBLIC HEALTH
LOS ANGELES, CALIFORNIA 90024

September 28, 1983

TO: Dr. Charlotte Neumann

✓ FROM: Anne H. Coulson *Anne H. Coulson*

RE: Trip Report - Nutrition CRSP Kenya Project, August 1983

I left for Kenya August 10 and returned September 1. My primary activities were related to aspects of the data gathering and management for the second phase of the study. Following are brief comments regarding the various concerns I addressed in conjunction with the field staff. Attached are appendices with more complete information on the respective methodologies developed during my stay.

Data gathering forms: With respect to the numbering system of the data forms I have expanded the code for the "schooler" age group to incorporate columns for age and sex, thus pushing the "toddlers" into columns 41 and 42, and the new infants into columns 51 and 52. See Appendix N for details.

The physical exam form seemed to be lacking questions on some important issues. For example, we do not have any life style questions, such as "smoking." Smoking seems to be prevalent in the Embu area, and I think we should include room for data on this - especially among the lead males. Material relevant to these aspects of data gathering is in Appendices F and H.

Data management: Mr. Njeru and I worked out a list of his responsibilities in managing the data as well as a timetable for sending materials to me. Regarding the latter, I asked him to check with DHL Courier Service about the weight of the least expensive shipment so we can review the DHL vs. airmail costs. The DHL may be more economical as well as being otherwise more preferable. If necessary, shipments of data could be cut down to one every three weeks. One tape can easily accommodate three weeks of data.

By the end of September we should place an ad in the Nairobi paper for the data key-entry positions. Qualifications should include facility with English and mathematics, and three years experience. I suggest that Mr. Njeru, Prof. Scott, and Mr. Kinyanjui (the key-entry supervisors) all interview the finalists of the application process, and that the final choice be unanimous. See Appendix DM for more information on data management.

Sampling: With regard to the material on sampling methods which you already have, one further thought is that we might preferentially include Kiambu speakers, which compose about 96% of the population; and exclude those who do not speak Kiambu.

Remaining to be addressed is the problem of training aids and reference manuals for the data collection instruments.

During my visit I met several times with Professor Scott, head of the Department of Computer Sciences, University of Nairobi. We discussed problems of computer capability, personnel and application of funding algorithms to the project computer use. In this connection, I also met with the Finance Officer of the University of Nairobi to discuss the treatment of the project, for funding purposes, as an "outside user" rather than as a research project.

I visited Embu for three days during which I worked with Mr. Njeru on preparation of the reduced sample for the pregnancy and newborn (1983) infants survey necessary before households can be recruited into the study and with morbidity and nutrition staffs regarding preparation of forms. During this time I talked to the assembled enumerators about their role in the data collecting activities in terms of the use and usefulness of the ultimate data in meeting the objectives of the overall study.

AHC/gk
Atts.

Numbering system:

The identification system will be based on household numbers (5 digits) and personal numbers (2 digits). The combination of the seven digits will uniquely identify an individual. Use of the household number with a "personal" number of 99 will indicate that the data in the record relate to the whole household and not specifically to any member of it. Similarly, use of a "household" number of 00001, 00002 or 00003 will indicate that the data in that record relate to the sublocation 1, 2 or 3. If a fourth sublocation is added, the number 00004 will be used, and so on. Use of 00009 will indicate that the data in the record relate to the whole community. (The leading zeroes shown here are included only as place holders; the single digit will appear in the 5th position in the household number.)

The household numbers will be the five digit numbers assigned in the preliminary survey. All household numbers which did not use a fifth digit, will have the number zero added as the fifth digit. The first digit presently indicates the sublocation and will continue to do so. The reason for this continuation is threefold: 1) The houses are already identified by that number -- on survey code sheets, marked on the house and on the maps. Thus, retaining the number will preclude confusion of numbers. 2) The enumerators are familiar with the numbers and the numbering system. 3) The relatively large number of available numbers and the small number of households can be utilized to reduce the possibility of error -- a wrongly copied number has a substantially smaller probability of being a valid study number.

(Approximately 1650 household numbers were assigned, 821 of these resulted in households eligible for the preliminary study. About 300 will participate in the final study. Thus, less than twenty percent of the possible household numbers will be valid numbers in the final study.)

The individuals already surveyed will retain their original numbers (01, 02, 03, etc.) with the following exceptions:

If the lead male, head of household is not numbered 01, the numbering system will be adjusted so that he is so numbered.

If the lead female, spouse of head of household is not number 02, the numbering will be similarly adjusted so that she is so numbered.

The schooler selected for study, if any, in the household will be renumbered to indicate age and sex. Male children to be followed as schoolers between ages 7 and 8 will be numbered 27, between ages 8 and 9, 28, and nine year olds will be numbered 29. Similarly, female schoolers selected for the special study will be numbered 37, 38, and 39.

The toddlers selected for study will be numbered 41 if male and 42 if female.

Infants born during the study who are part of mother-infant pairs will be numbered 51 if male and 52 if female.

Note that all 01's and 02's (lead males and females) in participating households are target individuals and no other adults are. The only other target individuals are those with personal numbers in the 20's, 30's, 40's and 50's. These individuals are target only during their period of study; one year for the schoolers and toddlers and six months for the infant who is part of the mother-infant and pair.

Numbers can be assigned to toddlers and schoolers at the time of enrollment, as the decision on which toddler or schooler will have been made

in the sampling, and toddlers and schoolers to be studied must have been members of the household at the time of the preliminary survey. Numbers will be assigned to infants in mother-infant pairs as they are born.

The problem of twinning is managed by the exclusion of twins from any of the studies of children. Members of twin pairs will not be selected for follow up as schoolers or toddlers; twin births will not be followed as mother-infant pairs, and the data collected on the mother during the last two trimesters of pregnancy will be deleted from the mother-infant pair data set.

5

Data Management - Forms to computer entry.

Forms

Forms for data capture will be designed to allow the collection of data in the most effective and efficient manner. Given the subject matter, the field situation, the skills of the data collector or recorder, and the requirements of the data management system.

All forms will have a common heading, specifying the form, identifiers and other necessary information. This is described below under the section "Header." Each sublocation, household and individual will have a unique identifier. This is described under "Numbering System," below.

The forms for field use, in which the enumerator must ask specific questions of the household respondent or target individual will be written in Kiambu so that all the enumerators will have the same translation. Forms for the use of the enumerators in making observations or measurements may be written in English, unless this is practical for the enumerators. It may be desirable to include Kiambu words or phrases to assist in the observations (e.g., local terms for signs or symptoms of illness).

Coding

To the extent possible, the forms will be self-coding. This means that the enumerator or recorder so designates a response or so enters a measurement that the data processing (key entry) can be done without further copying or coding of the information. A self-coding measurement might look like this:

Blood pressure	syst/dias	<u>24</u> <u>25</u> <u>26/27</u> <u>28</u> <u>29</u>	
or			
Blood pressure	syst/dias	-- -- --/-- -- --	(24-29)

Both allow the observer to record the reading in the correct place from which the key entry operator will enter into columns 24-29. (Most numerical measures are self-coding).

Where self-coding is not feasible, precoding will be utilized to the extent possible. For precoded items, a code exists but is not printed on the form or given to the respondent. For example, the International Classification of Diseases may be used to code disease entities reported or diagnosed. A coder would look up the appropriate code after the information had been collected.

Occasionally, as in the morbidity data, an expert will have to code using his/her expert judgement (e.g., do these symptoms warrant a diagnosis of clinical malaria). The codes have been developed; the expert decides which of the codes applies to the information collected. Usually a number of items of information are taken together in making the judgement.

In some situations, it may not be possible to develop the code at the time data collection starts, as the character and range of responses may not be known and may need to be developed after the first few responses. (This number is usually set at 50 or 100 responses, assuming that the first respondents are likely to have the same range of responses as the later respondents.)

In some cases, it may be useful to enter the whole observation or response, in an alphabetic form, for later print out or computer manipulation of Alpha information. (Observations of child activities or interpersonal interactions may fall in this category.)

Coding responsibility will vary with the level of coding, but will be fixed for each piece of information identified. Enumerators will be responsible for appropriate completion of self-coding forms, and may be

assigned responsibility for coding precoded items. Decision coding will be done by the expert in the area. Senior staff and investigators will work out codes for non-precoded items and will assign coding responsibility as appropriate, full entry information is not coded. Supervisors, calculators, senior staff and data entry operators will be assigned coding responsibilities as appropriate.

Editing

Editing (manual editing) will be performed at 4 levels. The enumerator will review the completed form before turning it over to the supervisor. The supervisor will review the form for completion and correctness. The data entry personnel will do final pre-entry editing. After key entry, but before the information is refiled in the household record, it will be reviewed one final time by senior staff.

There will be two stages to computer editing of key entered data. The first will comprise range checks and consistency checks within the key entered record. In range checking, each item of information is checked to make sure that the value on the computer record is a valid one, e.g., that sex is coded as 1 or 2. This checking does not ensure that the number is correct, but only that it is in the range of correct numbers. Consistency checking involves the identification of pieces of information that should exhibit a particular relationship, e.g., if an individual reports that he does not smoke, then the information about cigarette consumption per day should not be present. If it is, one of the two items of information is wrong. If information is found to be incorrect through range or consistency checks, the identified items and the record heading will be printed out.

Errors detected in manual editing will be corrected before key entry, except for the final review before filing. Questions raised by key entry operators will be marked for recheck. If possible, key entry will be completed, with a note made as to the actual key entry so that correction can be made, if necessary. If the question precludes continuation of key entry of the record, the form will be returned to Embu for correction.

Errors detected in range and consistency checks will be printed out and checked with the forms at Chirromo. If correction can be made from the forms, it will be done then, and a correction form prepared for the next computer run. Errors which cannot be resolved at Chirromo will be carried to Embu and distributed to the supervisor concerned who will be responsible for obtaining the correct information. This may require information from the enumerator from the household or from the individual. The information may have to be obtained again.

The second level of computer editing will occur when records for an individual or household collected at different times and/or including different information are linked in the computer to form the basic data set. This will be an ongoing process throughout the study. At the time of the linkage, further consistency checks will be made. For example, we would expect that the weight collected in the metabolism studies should be approximately the same as the weight collected at the anthropometry measure closest in time. Also, we would not expect that height at the second measurement is significantly shorter than at the first measurement. Record headings and the questionable information will be printed out on detection. The printouts will be returned to Embu and given to the supervisors to correct. These procedures cannot be started until the linkage routines are operational.

Correction forms will be completed for all errors. These forms will include the record heading, the variable number or numbers for the questioned information and the true values. These forms will be entered into the computer by key entry or by terminal entry. The program for this entry of corrections will reapply the range check and will print out the new value of the variable changed. (Note that the new value is read from the corrected data set and not from the correction entered.) Correction forms should be completed if at all possible within the week to be returned to Chirromo with the next batch of forms.

Key entry. All completed forms will be sent to Chirromo once each week (Friday). The key entry staff will perform such coding and editing functions as necessary for key entry and on Saturday and Sunday all completed forms will be key entered. Key entry will be 100% verified by a second operator. The key entry staff will continue with other data work (coding, editing, checking) on Monday and Tuesday. (Days off will be Wednesday and Thursday for the key entry operators and Njeru.)

Computer entry. On Monday, Njeru will use the computer to correct earlier entries and to enter all material key entered on Saturday and Sunday. Range and consistency checks will be run and forms annotated for checking in Embu. Every other week, a copy of all corrected records not yet sent to UCLA will be created for shipment to UCLA. Njeru will return to Embu late Monday (early Tuesday when absolutely necessary) and will work Tuesday to distribute checking needed to supervisors, return forms to Embu, write reports and check with study administration and coordinator on progress, needs, etc.

Floppy (flexible) disks will be retained for six months after the last key entry on the disk as a back up to computer files. Two copies of the most recent complete tape will be kept, one in the computer facility and one in

Embu (under refrigeration) during weather 32°C or over. Working copies will be backed up on the computer disk or on tape kept at Chirromo, as appropriate.

Data flow. The data flow will be organized as follows:

Data capture forms will be prepared in the office and inserted in the "household" folder (community, household, individual), which will be filed by "household" number. The supervisor will remove the needed forms for the day or week -- depending on supervisor visits to the office. They will be distributed to the enumerators who will use them in the field and return the completed forms to the supervisor. The supervisor will do the necessary editing and turn in the completed forms. Forms will be coded and edited as appropriate at Embu and prepared for Njeru to transport to Chirromo on Friday of each week.

Data coding and editing, as appropriate, will be completed at Chirromo. Data key entry, with 100% verification, will be done Saturday and Sunday at Chirromo.

Computer entry of key entered data will be done on Monday, and range and consistency checks done. Errors will be printed out for checking. Alternate weeks, a copy of data will be sent to UCLA.

Error subroutine -- Problems encountered at Chirromo in editing, key entry or computer entry will be identified and returned to Embu. Responsibility for checking and correction rests with the supervisor of the enumerator who collected the data. The supervisor will refer the problem to the enumerator, household or individual as appropriate

11

for correction. The supervisor will return the completed correction form to the office for inclusion in Njeru's next Friday pick up. If at all possible, corrections should be made between Tuesday and Friday so that computer correction can be made at the next computer run.

The subsequent steps in data management including linkage, variable distribution and analysis will be worked out when the data entry and error programs have been worked out and the data to be collected have been fully defined and described.

Header for all data forms for the study:

Each record will be headed by a standard set of designators that will identify the record type uniquely and the community, household or individual to which the data in the record relate. Thus, for example, the household food intake record for the first day of collection in the 14th month of the study will be identifiable as such, as will the household on which the data were collected.

A record is defined as the information collected on a given unit -- community, household or individual -- concerning a given topic (e.g., morbidity, food intake) at a specific time. Thus, the household food intake data collected for one day would constitute a record as would each of the personal food intakes collected on target individuals. The immunology data collected every three months would constitute three records: though the information collected is not a large amount, it is collected from three sources, the field, the Kenyan laboratory and the Newfoundland laboratory. (The only exception to this general principle of record identification may occur if the information constituting the "record" is sufficiently massive to make data management as one record difficult. Such records will be divided into manageable lengths.)

The following "topics" are identified.

1. Food intake
2. Anthropometry
3. Morbidity
4. Metabolic adaptation
5. Cognitive function

6. Household
7. Case studies
8. Non core studies

The digit listed here will form the first digit of the three digit record designator.

Each of these major topics may be further subdivided:

1. Food Intake
 1. Household
 2. Target individual intake
 3. Intake measures during illness

2. Anthropometry
 1. Regular measures on target persons
 2. One time measures of all household members
 3. Special anthropometry during illness
 4. Special anthropometry month 5 of pregnancy
 5. Ditto for month 8
 6. Special anthropometry 8 days postpartum -- mother
 7. Special anthropometry 8 days postpartum -- infant

3. Morbidity
 1. Weekly morbidity checks
 2. Physical exam -- adults
 3. Physical exam -- children
 4. Immunology
 1. Field observations
 2. Kenya laboratory

3. Newfoundland laboratory
5. Special morbidity visits for illness
4. Metabolic adaptation
 1. Measurements of adults
 2. Pregnant women 5 mos.
 3. Pregnant women 8 mos.
5. Psychological evaluation
[This subdivision will depend on tests, observations]
6. Social sciences
 1. SES
 2. Sanitary conditions
 3. Cure giving
 4. Hygenic activities
7. Repro
8. Case studies
9. Non core studies

[Subdivisions will depend on lists, observations to be made]

Using these subdivisions and those still to be developed. The base record number will be three digits. To these will be added an additional two digits indicating the order, if any, of the data collection. For food intakes, for example, there may be as many as 48 visits to the household to collect the data. The two digits would be used to indicate that this record represents the *i*th visit in this potential series of 48. Morbidity visits are done weekly, with a potential of 104 visits - three digits. To save header space,

as it would only be needed for the few families still under follow-up in the 100-104th weeks, other computer arrangements will be made for those last 5 weeks, so only two digits will be used.

The household # and person # will be in the header. As described in the number section if the data relate to an individual, the seven digits used are:

5 digit household # + 2 digit personal #

If the data relate to a household:

5 digit household # + 99 in personal # digits

If the data relate to a sublocation:

0000 n in household digits + 00 in personal digits

n = 1 for sublocation 1, etc.

If the data relate to the whole study area:

00009 in household digits + 00 in personal digits

The data of the collection of the data also is incorporated as part of the header, in the system of day/month/year. To save space, only the last digit of the year will be used -- thus, August 29, 1983 would be coded 29083.

The name of the sublocation, the surname of the household or the full name of the target individual will also form part of the header, but will not be entered into the computer. The inclusion of the alphabetic information will allow cross-checking for ID accuracy, as well as being useful to the enumerator in sorting out the forms during household visits.

The header will be organized as follows:

col. 1-3	Major topics subdivision designator (unique to form)
4-5	Order of collection (ith)
6-10	Household number
11-12	Personal number

13-14 Day of data collection
15-16 Month of data collection
17 Yr of data collection

Name of (community), (household), (person).

This material should be placed accross the top of the data collection forms and should be as similar as possible for all forms.

17

KENYAN PART OF DATA MGMT

MUST

SHOULD

DESIRABLE

OK

INDIFFERENT

DATA COLL.

EDITING

KEY ENTRY

COMPUTER ENTRY

LINKAGE

DISTRIBUTIONS

ANALYSIS

On field completion of data collection, i.e., worker returns from field with completed forms.

WHO DOES

- EMBU 1. Worker must complete form for all self-coding and precoded questions for which code is on document.
- EMBU 2. Worker submits completed form to clerk who notes completion for management system.
- EMBU 3. Form goes immediately to worker's supervisor for review -- supervisory action if required: is form completed? is written material legible? has worker coding been done?
- EMBU or
CHIRROMO
(could be
combined) 4. Form goes to coder for completion of codes.
- EMBU or
CHIRROMO 5. Form goes to editor for final review including check of coder's work.
- EMBU 6. Form to Njeru for data entry.
- 109

- CHIRROMO 7. Data entry -- 100% verification questions to Njeru and back to Embu for resolution if necessary.
- CHIRROMO
and EMBU 8. Computer entry -- range and consistency checks problems to Njeru for resolution at Embu, if necessary.
- CHIRROMO 9. Corrections entry --
- CHIRROMO
or EMBU 10. UCLA copies (make or detached) sent -- do at Embu --
- CHIRROMO 11. Tape In to U.S.
- CHIRROMO
and UCLA 12. Linkage of In tape to previous to create In.
- EMBU (Apple)
and CHIRROMO
and UCLA 13. Notation of due but not present forms -- print request to clerk for check.

CHIRROMO, if
possible,

UCLA 14. Analysis to date -- counts, status of studied population
(loss to follow-up, etc.), preliminary distributions to
date.

UCLA 15. Analysis

UCLA Sample of data forms key entered, computerized, edited,
etc., compared with In tape.

read In tape, send ID #s and form #s on tape back to
Nairobi for checking.

linkage, etc. I_{nc} to F_{nc}

read F_{nk} tape, compare with F_{nc} tape -- list differences,
process and send to Nairobi

split data base into working data sets. Analysis.

check problems (unclean data) with forms/Nairobi.

set up "change" output form

21