ABEL 2

Advancing Basic
Education and Literacy
Phase 2

SAMPLE DESIGN FOR MALAWI ANNUAL PRIMARY SCHOOL SURVEY

JUNE 1995

Conducted by

The ABEL 2 Consortium
Academy for Educational Development with
Creative Associates International, Inc.,
Educational Development Center,
Florida State University,
Harvard Institute for International Development,
and Research Triangle Institute

Conducted for

Office of Field Support and Technical Assistance
Center for Human Capacity Development
Bureau for Global Programs, Field Research, and Support
U.S. Agency for International Development
Contract No. HNE-5832-C-00-4075-00

Trip Report

Consultant: Joseph George Caldwell

Purpose of trip: Develop sampling plan for Malaŵi annual primary school survey

Dates of trip: May 9, 1995 - June 30, 1995

Summary of trip:

I departed for Lilongwe, Malaŵi, on May 9, and checked in with Dr. Andrew Sisson of the United States Agency for International Development (USAID) on May 15, as planned. Dr. Sisson and Dr. Laurie Cameron (also of USAID) briefed me on the task, and a meeting was set (for 2:00 pm that day) up with members of the Ministry of Education Planning Unit.

I was assigned office space and provided access to the Primary Pupil Registration and School Survey data for 1993 and 1994, and began work on the task of designing a sample design for the 1995 survey. The report submitted to Dr. John Hatch describes the methodological approach used in this work, in detail. Throughout the assignment I worked very closely with Mr. John Khozi, Statistician, whose office I shared. During the assignment I had several meetings with Mr. Khozi and other members of the Planning Unit, to discuss sample design requirements and alternatives, and to select a preferred design. A variety of tables were prepared to describe the precision associated with sample design alternatives, and these were discussed in the various meetings.

A field trip to a number of schools in Ntchisi district was scheduled in early June, to identify records maintained at the school level and to assist evaluation of the feasibility of alternative sampling approaches. Mr. Khozi, Dr. Cameron, and I participated in this one-day trip. The result of this field visit was the conclusion that it was not feasible to consider sampling plans that were dependent on technical assistance support from local personnel, such as designs involving student subsampling at sample schools.

After a number of alternative designs had been identified and analyzed in a series of small meetings, a larger meeting was held in mid-June with Dr. Cameron and officials of the Planning Unit to discuss a recommended approach. After a consensus was reached, I prepared a draft report (dated June 19) describing the project approach, the various sample design alternatives that had been considered, and the recommended design. The draft report was circulated to members of USAID and the Planning Unit for review and comment, and a final revised version was completed and submitted on June 29.

A final meeting was held at USAID offices on June 29 with Drs. Sisson and Cameron to discuss the results of the assignment and plans for the future.

Dr. J. George Caldwell
4615 Schirra Court
Spartanburg, SC 29301 USA
1-803-576-8407

Memorandum

June 29, 1995

To: Distribution List
From: J. G. Caldwell
Subject: Transmittal of Report, Sample Design for Malawi Annual Primary School Survey, Revision of June 28

Transmitted with this memorandum is a copy of a report describing the recommended sample design for the Malawi annual primary school survey. This report describes sample designs that were discussed at meetings of USAID and MOE Planning Unit staff during the process of developing the recommended design, and the sample design that was selected as the preferred design as a result of those discussions.

The report transmitted herewith is a minor revision of the draft dated June 19. A few of the statistics cited in Part I were revised to reflect recent additions to the 1994 survey data base, some of the precision tables were revised to incorporate estimates of intrazone and intradistrict correlation coefficients, and two new sections were added to Part II (Section V, "Sample Selection in Future Years," and Section VI, "Implementation Schedule").

The report is divided into two parts: Part I (Sample Design Alternatives) describes sampling plan alternatives that were identified and analyzed during the course of the sample design effort, and the rationale for preferring the recommended design over the other candidates; and Part II (Description of Recommended Sample Design), which describes procedures for selecting a probability sample in accordance with the recommended design and for analyzing data collected under that design. Part II also presents an actual sample selected according to this design, and a description of the characteristics of the selected sample.

Complementing the report is a computer diskette containing a copy of the report (in WordPerfect 5.1) and listings of all of the computer programs and data files used in the sample design effort (in dBASE IV v2.0).

b

This report constitutes the two deliverables required under the terms of my consulting agreement.

I wish to take this opportunity of expressing my appreciation for the opportunity of working with United States Agency for International Development and Ministry of Eduction Planning Unit staff on the sample design effort.

Transmitals:

Report, Sample Design for MalaŵI Annual Primary School Survey, Parts I and II, plus two computer diskettes

Distribution:

Dr. Andrew Sisson, United States Agency for International Development / MalaŵI (report and diskettes)
Dr. Laurie Cameron, USAID (report only)
Mr. John Khozi, Statistician, MalaŵI Ministry of Education Planning Unit (report and diskettes)
The Honorable Sam Mpasu, Minister of Education (report only)

Sample Design

for

Malaŵi Annual Primary School Survey

Part I: Sample Design Alternatives

June 19, 1995

Revised June 28, 1995

Prepared by:
J. George Caldwell, Ph.D.
4615 Schirra Court
Spartanburg, SC 29301 USA
Tel. 1-803-576-8407

Prepared for:
Academy for Educational Development
1875 Connecticut Avenue NW, 9th Floor
Washington, DC 20009-1202
Tel. 1-202-884-8000

Under contract to:
US Agency for International Development
Washington, DC

Table of Contents

Foreword v

I. Introduction 1

II. Requirements 3

III. Population Characteristics 5

IV. Alternative Data Collection Approaches 13

V. Alternative Sample Designs 23

VI. Recommended Sample Design 59

Reference 60

Appendix

A. Computer Programs Used in the Analysis 61

List of Tables

1. Advantages and Disadvantages of Alternative Data Collection Approaches 14

2. Potential Advantages and Disadvantages of Alternative Sample Designs 19

3. Sampling Errors of Estimated Population Mean or Total for Simple Random Sampling without Replacement, Standard Deviation of Variable of Interest = 500 27

4. Relative Sampling Errors of Estimated Proportions for Simple Random Sampling without Replacement 28

5. Relative Sampling Errors of Estimated Proportions for Single-Stage Cluster Sample Design, Clusters Selected Using Simple Random Sampling without Replacement, p=0.20 30

6. Relative Sampling Errors of Estimated Proportions for Two-Stage Sample Design Using Simple Random Sampling without Replacement at Both Stages, p=0.20 33

7. Relative Sampling Errors of Estimated Proportions for Two-Stage Sample Design Using Simple Random Sampling without Replacement at Both Stages, p=0.20, Using Optimal Sampling and Subsampling Fractions 34

8. Relative Sampling Errors of Estimated Proportions for Three-Stage Sample Design Using Simple Random Sampling without Replacement at All Stages for p=0.20 and First Stage Intraunit Correlation Coefficient (r1) = 0.54 38

9. Relative Sampling Errors of Estimated Proportions for Three-Stage Sample Design Using Simple Random Sampling without Replacement at All Stages for p=0.20 and First Stage Intraunit Correlation Coefficient (r1) = 0.54, Using Optimal Sampling and Subsampling Fractions 39

10. Relative Sampling Errors of Estimated Proportions for Three-Stage Sample Design Using Simple Random Sampling without Replacement at All Stages for p=0.20 and First Stage Intraunit Correlation Coefficient (r1) = 0.81 43

11. Approximate Sampling Errors of Estimated Cluster Mean or Population Total for Single-Stage Cluster Sample Design, Clusters Selected without Replacement with Probabilities Proportional to a Measure of Size, Standard Deviation of Variable of Interest = 500 46

12. Approximate Sampling Errors of Estimated Proportions for Single-Stage

Cluster Sample Design, Clusters Selected without Replacement with Probabilities Proportional to a Measure of Size, Correlation between Variable of Interest and Measure of Size = 0.9000 47

13. Approximate Sampling Errors of Estimated Cluster Mean or Population Total for Stratified Single-Stage Cluster Sample Design, Clusters Selected within Strata Using Simple Random Sampling without Replacement, Standard Deviation of Variable of Interest = 500, No. of Strata = 2, Optimal Construction of Strata and Optimal Allocation of Sample to Strata Assumed. 50

14. Frequency Distribution of School Enrolments, and Cumulative Sum of Square Root of Frequencies 51

15. Approximate Sampling Errors of Estimated Cluster Mean or Population Total for Stratified Single-Stage Cluster Sample Design, Clusters Selected within Strata Using Simple Random Sampling without Replacement, Standard Deviation of Variable of Interest = 500, No. of Strata from 1 to 10, Optimal Construction of Strata and Optimal Allocation of Sample to Strata Assumed. 55

16. Approximate Sampling Errors of Estimated Cluster Mean or Population Total for Single-Stage Cluster Sample Design, Clusters Selected without Replacement with Probabilities Proportional to a Measure of Size, Standard Deviation of Variable of Interest = 500 57

17. Approximate Sampling Errors of Estimated Proportions for Single-Stage Cluster Sample Design, Clusters Selected without Replacement with Probabilities Proportional to a Measure of Size, Correlation between Variable of Interest and Measure of Size = 0.9000 58

Foreword

This document was prepared by Dr. J. George Caldwell under contract to the Academy for Educational Development. The work was funded by the US Agency for International Development (USAID) under USAID Agreement No. 612-T-605A, Program No. 612-0240, "Girls' Attainment in Basic Literacy and Education (GABLE)." The USAID project officer was Dr. Andrew Sisson.

The sample design presented in this report is the product of the efforts of a number of individuals. During the effort to develop a sample design, the author worked with Dr. Sisson and Dr. Laurie Cameron of USAID and with several members of the Planning Unit of the Malaŵi Ministry of Education. The principal Planning Unit staff member who worked on the sample design effort was Mr. John Khozi, Statistician.

I. Introduction

This report presents a sampling plan to be used to collect data for the 1995 Malaŵi primary school survey. The purpose of the sampling plan is to enable certain pupil-related data to be collected efficiently from a modest sample of primary schools. The 1994 survey was a census of all primary schools (approximately 3,400) and all students (approximately 2.9 million). Because of the heavy data collection burden associated with the census approach, a significant level of school nonresponse (about 15%) resulted, and resources were not available to reduce the nonresponse level appreciably. Furthermore, the substantial effort required to process the large amount of census data has required a considerable amount of time. The large size of the student data base requires very large data files (stored in a number of different computer drives) and long data-processing running times. The current effort was undertaken to identify data collection methods that could produce the required information with significantly less data collection and processing effort. With a significantly reduced effort, it was hoped that the response rate could be improved and needed estimates could be available much sooner.

The methodology adopted in this effort was to identify a number of alternative approaches for the enrolment survey, to discuss their advantages and disadvantages, and to select a preferred approach. The details of the selected approach were then worked out. The alternative approaches were synthesized after identifying the survey information requirements, the resources available for the survey, and the general characteristics of the school and pupil populations. The various alternatives were presented to USAID and Planning Unit staff concerned with the survey for their review and comment. Out of these discussions a consensus was reached that the needed information could be obtained efficiently using a cluster sample design involving a sample of 500 primary schools, with all students at sample schools included in the sample.

This report describes the methodology by which the primary school sampling plan was developed. It then describes the sampling plan in detail, including a listing of the sample schools and the various sample design parameters required to analyze the data. The formulas required to analyze the survey data are described. A computer diskette has been prepared containing a dBASE IV version 2.0 program that performs the computations required by the analysis formulas, as well as all other computer programs used in the course of developing the sample design.

This report consists of two parts. This part, Part I, identifies, describes, analyzes and compares a number of sample design alternatives. These are designs that are a priori may considered as reasonable candidates for consideration for the annual school survey. Based on the analysis and comparison of the designs, one of them was selected as a preferred design, and recommended for use in the survey. Part II describes the recommended design in additional detail, selects a probability sample of schools in accordance with the design, presents design parameters needed in the data analysis, and describes the formulas to be used in analyzing the sample data collected in accordance with the design.

This part consists of six additional sections and an appendix. Section II, "Requirements," describes the data presentations and/or estimates required to be

produced as a result of the enrolment survey. Section III, "Population Characteristics," describes the general characteristics of the school and student populations from which the enrolment survey is collecting data and about which information is desired. It also describes the sample frame (list of schools). Section IV, "Alternative Data Collection Approaches," identifies the general data collection methodologies that were proposed as feasible approaches for collecting the needed data from the school and student populations. The various data-collection methodologies are described, and the preferred methodology (selected after discussions among Planning Unit and USAID staff) is identified. The preferred methodology is based on probability sampling. Section V, "Alternative Sample Designs," examines a number of different probability sampling plans for collecting the data. Section VI, "Recommended Sample Design," describes the cluster sampling plan recommended for the annual primary school survey. Appendix A, "Computer Programs," identifies the computer programs used to evaluate the various sample design alternatives, to generate a probability sample in accordance with the recommended sampling plan, and to analyze the survey data collected for the sample.

II. Requirements

The first step in sample design is to obtain a good understanding of what data presentations and estimates are required or desired to be obtained from the collected data. The primary requirement is that it be possible to produce future editions of the annual publication, Basic Education Statistics, Malaŵi. For the most recent year (1993), that publication contained 64 pages, of which 28 were dedicated to primary education. The publication presents a variety of tables showing characteristics of schools and students for the nation, for the three major political regions of the country, and for the 28 school districts. Most of the tables deal with school characteristics: total enrollment, equipment, number of classes, books, teaching staff, teaching houses, and pupil/teacher ratios. A few tables deal with pupil characteristics: distribution of students by age, sex, and standard, number of repeaters (students who are in a standard for the second (or later) time, number and percentage of students passing the Primary School Leaving Certificate of Education (PSLCE) examination.

The tables dealing with pupil characteristics were the following:

- o A table showing student counts by district, standard, and sex
- o A (national) table showing student counts by age for girls and total (boys plus girls)
- o A (national) table showing counts of repeaters by standard and sex
- o A (national) table showing counts of students who entered and students who passed the PSLCE, the percentage passing, the number of government Form 1 places, and the transition rate (number of Form 1 places divided by number of PSLCE passers)

In addition to production of the Education Statistics each year, there are occasional ad-hoc surveys conducted each year to address specific issues for which the Education Statistics data are not adequate.

After discussing the requirements for data and estimates, it was agreed that next year's primary-school enrolment survey must be able to produce all of the primary-school-related data presentations and estimates presented in the 1993 edition of Education Statistics. In addition, it was desired to be able to present data on dropout rates and average number of years in school for students in standard 8. National-level estimates are satisfactory for the repetition, dropout and years-in-school estimates, i.e., estimates are not required by district.

It was desired that any sampling plan developed for the annual school survey would be appropriate for use at other times in the year for ad-hoc surveys, but this was not taken as a requirement (since the specific requirements of future ad-hoc surveys, by their nature, is not known).

The issue of whether a sampling plan that sampled students would be used in future years to collect data on the same schools and possibly the same students

sampled in earlier years was discussed. (Such plans can be useful for measuring change over time.) No requirement was imposed in this regard.

If possible, it was desired to develop a survey design that could be used for several years, without the need for major a major redesign effort each year.

III. Population Characteristics

A. The Structure of the Population

The populations of interest in the primary school survey are the primary schools and primary school students of Malaŵi. The first step in the survey design process was to obtain basic information about the basic structure of the population -- how students are assembled into schools and how the schools are organized. This information is used in the survey design process to assist the development of sampling plans that are highly efficient, i.e., provide relatively high levels of precision for the amount of sampling effort expended.

The structure of the population may be taken into account in a variety of ways. A survey design may take advantage of the organizational structure of a population (e.g., the organization of schools into districts and zones) to facilitate distribution of questionnaires, monitoring of the data collection effort, and collection of the data. Substantial reductions in cost may be achieved by taking advantage of the fact that students are organized in schools -- collection of data from a sample of 1,000 students in four schools is obviously less costly than collection of data from 1,000 students selected at random and located in 400 schools.

Basic information about primary schools is presented in Education Statistics. Although the latest edition of Education Statistics is for the year 1993, school and student data are available in computer data files for the 1994 school year. The 1994 data are quite different from the 1993 data with respect to total student enrolment, since 1994 was the first year of free primary education in Malaŵi. The following is a summary of the salient characteristics of the school and student populations for 1994.

The country is divided into three political regions and 28 school districts. The school districts are further divided into zones, and a zone inspector is assigned to each zone. There are 120 zones in the country.

The total number of schools is 3,424, and the total number of students is approximately 2.9 million. The mean and standard deviation of the number of students per school ("school size") were 841 and 665, calculated from the 1994 data. From the 1994 data (not yet edited), the mean and standard deviation of the number of schools per district were 109 and 58. This mean is a little less than the current value (3,424/28 = 122). From the 1994 data, the mean and standard deviation of the number of schools per zone were 25 and 11. This mean is a little less than the current value (3,424/120 = 28.5).

The mean and standard deviation of school size differs considerably from district to district; the mean and standard deviation of school size are substantially larger for "urban" districts than for "rural" districts.

From the 1994 school survey and primary pupil registration system, a variety of tables and crosstabulations can be constructed, showing the numbers of students in various categories of interest (e.g., tables of numbers of students by age, sex, standard; crosstabulations of numbers of students by sex and standard or by age and standard; tables of repeaters by age, sex, or standard). Since none of

the proposed sample designs presented in this report control the sizes of these subpopulations, the numbers of students falling in these categories in the selected samples will be approximately proportional to their representation in the full student population. While a number of tables and crosstabulations were examined during the sample design process, they are not included in this report (since the data are not yet edited). The principal fact to keep in mind is that if students were to be sampled at sample schools, the numbers of students in the higher standards would be very small, precluding the production of estimates of meaningful precision for these categories.

B. Previously-Known Information about Variables of Interest in the Survey

Cluster and Multistage Sampling The preceding paragraphs described information about the size and organization of the population units (schools, students) of interest in the annual primary school survey. The schools and students are referred to as primary sampling units (or primary units, or first-stage units) and secondary sampling units (or second-stage units, or subunits), respectively. (Note that the word "primary" in the term "primary sampling unit" has no relation to the same word in the term "primary school.") The survey is concerned with making observations (taking measurements) on these units -- measurement of school characteristics and measurement of student characteristics. The characteristics being measured are usually referred to as the variables of the survey. In some instances a characteristic of interest may be measured either as a school variable or as a student variable. For example, the age distribution of the students at a school may be observed as a school variable by collecting data from school registration records, or as a student variable by collecting the date of birth from a sample of students. The reason for preferring to take a measurements on one unit or the other depends on the relative cost of making the measurements and the relative magnitudes of the measurement errors (e.g., more accurate data on student ages may be obtained by asking the students directly than by relying on school records, but it is more costly to do so).

A sampling plan in which a sample of primary sampling units (e.g., schools) is selected and all of the secondary sampling units (e.g., students) are included in the sample is called a "cluster" sample design, or a single-stage sample design without subsampling. A sampling plan in which a sample of primary sampling units is selected and samples of secondary sampling units are selected from each selected primary unit is called a two-stage sample design. Sampling may be extended to additional stages (e.g., a three-stage sample design might consist of a first-stage sample of zones, a second-stage sample of schools, and a third-stage sample of students). Whether cluster or multistage sampling is desirable depends on what information is available about the variance structure of the population and the relative costs of sampling the units at the various stages. When cluster or multistage sampling is done, it may be advantageous to select the primary sample units with unequal probabilities, or to select different fractions of sample units from certain categories (e.g., large and small schools). The decision concerning the use of selection with unequal probabilities depends on what is known about the relationship of variables of interest in the survey to variables that are already known about the sample units.

Selection of Primary Sampling Units with Probabilities Proportional to a Measure of Size Just as the precision of survey estimates may be increased by taking into account the structure of the population, it may be possible to realize large increases in precision (for a given level of survey cost) by taking account of information that is known to be related to the variables of interest in the survey. In the case of the primary school survey, it is known that the variables of interest (e.g., numbers of repeaters) are related to school size (enrolment), and, since the school enrolment in one year is related to the enrolment in the previous year, this means that the variables of interest are also related to the enrolment in the previous year.

Based on the strength of the relationship of the enrolment in one year to the enrolment in the preceding year, it may be advantageous to group schools into size categories, or strata, prior to sampling, and to allocate the sampling effort to the strata in ways that increase precision. Or, it may be advantageous to select schools with probabilities that are proportional to their last-year enrolment. The approach of selecting primary sampling units with probabilities proportional to a known variable related to survey variables of interest is called "sampling with probabilities proportional to a measure of size." The "size" in this case is the school (total) enrolment, and the "measure of size" is the previous year's enrolment.

The standard measure of the relationship of two variables, such as enrolment in the current year and enrolment in the preceding year, is the correlation coefficient. The correlation coefficient, which measures the extent to which two random variables are linearly related, varies between the values -1 and +1. A value of +1 indicates that two variables are perfectly linearly related; a value of 0 indicates that there is no linear relationship between them. The correlation between the total school enrolment for 1993 and 1994 is .84. This is a high degree of correlation. With this high a degree of correlation, it should definitely be possible to improve the efficiency of the primary school survey by taking into account the enrolments in previous years (either by stratifying schools by the preceding year's enrolment or by selecting them with probabilities proportional to the preceding year's enrolment).

Note that since the enrolment increased dramatically from 1993 to 1994, the correlation of the 1993 enrolment to the 1994 enrolment is likely to be somewhat less than that between years prior to 1994 and between years after 1993. For this reason, the correlation of the 1994 enrolment to the 1995 enrolment is expected to be somewhat larger than the correlation between the 1993 enrolment and the 1994 enrolment, e.g., .90 instead of .84.

Information about school size is useful in the survey design process because it is related to the values of the variables to be measured in the survey. The extent to which school size is taken into account in designing the survey depends on the extent to which each of the variables to be measured in the survey is related to school size. A sample design that is efficient for obtaining information (producing estimates) about one variable may not be efficient, however, for obtaining information about another variable. The precision of estimates for variables that are highly correlated (positively or negatively) with enrolment (e.g., number of desks, books, repeaters) is increased by sampling schools with probabilities proportional to last year's enrolment, but the

precision of estimates for variables that bear little relation to enrolment will not be increased, and may be substantially decreased (compared to that associated with a sample design in which the school selection probabilities are not proportional to school size).

In order to assess the value of sampling schools with probabilities proportional to previous enrolment, it is necessary to assess the extent to which the estimates of interest -- both school estimates and student estimates -- are correlated with enrolment. All of the variables of interest in the annual school survey are in fact highly correlated with enrolment, so that selecting schools with probabilities proportional to previous enrolment will surely be desirable (i.e., it will result in a design that yields a higher level of precision than a design having the same school sample size that does not select schools with probabilities proportional to school enrolment).

It should be recognized, however, that if it is desired to collect information in an ad-hoc survey about a school variable that is unrelated to school size (e.g., perhaps, the use of double-shifting, or the method of recording attendance), then a sample design based on selection of schools with probabilities proportional to enrolment may be substantially less efficient than another type of design (e.g., a simple random sample of schools, regardless of size). For estimation of student-related characteristics, however, and even for estimation of many school-related characteristics (viz., those related to school size), the selection of schools with probabilities proportional to enrolment is likely to enhance precision substantially.

Sampling of Units in Clusters May Increase or Decrease Precision Although sampling of primary units with probabilities proportional to a measure of size may increase the precision of estimates related to enrolment (for a given level of sampling effort), the estimates for second-stage units are generally of substantially lower precision for a cluster or multistage sample than for a simple random sample having the same number of second-stage units. The reason why a multistage design is nevertheless often preferred is that the cost of sampling second-stage units from a sample of first-stage units is often much less than the cost of sampling second-stage units at random from the population. In order to assess the desirability of using cluster or multistage sampling, and in order to determine the optimal sizes of the second-stage samples (or third- or lower-stage samples, if additional stages of sampling are involved), it is necessary to know something about the variance structure of the clusters and the relative costs of sampling first-stage vs. second-stage (or lower-stage) units.

Stratification May be a Useful Means of Controlling Precision or Increasing the Efficiency of a Design Stratification may be beneficial in a survey for several reasons -- to take advantage of administrative or organizational features of the population, to insure an adequate sample size for subpopulations of interest, to increase the precision of sample estimates by varying the sampling rates in different strata, and to take into account known information about the stratum sizes. The decision to employ stratification depends on what information is available about potential variables of stratification before and after the survey, with respect to the variance structure of the strata and the costs of conducting stratified sampling.

The following sections describes what information is available about the variance structure of the primary school and student populations, and the cost of implementing cluster, multistage, or stratified sampling. Note that a useful method of characterizing the variance structure of a population is in terms of intraunit correlation coefficients. The relationship between the variance of the among-unit means and the variance of the elements within units is:

$$(1 - r)/r = S_2^2/[(N-1)/N)S_1^2 - S_2^2/M],$$

where r denotes the intraunit correlation coefficient (correlation coefficient between second-stage units in the same primary unit), N denotes the number of primary units in the population, M denotes the number of second-stage units in each primary unit, S_1^2 denotes the variance among primary unit means, and S_2^2 denotes the variance among subunits within primary units. This is approximately equal to S_2^2/S_u^2, where $S_u^2 = S_1^2 - S_2^2/M$. Solving for r yields the approximation

$$r = (S_1^2 - S_2^2/M)/(S_1^2 + S_2^2(1 - 1/M) .$$

In terms of S_1^2 and S^2 (the population variance), the expression for r is

$$r = [(N-1)M^2S_1^2 - (NM-1)S^2]/[(NM-1)(M-1)S^2],$$

or, approximately (if N is large),

$$r = (S_1^2 - S^2/M)/[S^2(1-1/M)] .$$

(The relation of S^2 to S_1^2 and S_2^2 is

$$S^2 = [(N-1)MS_1^2 + N(M-1)S_2^2]/(NM-1) ,$$

which for large N and M is approximately $S^2 = S_1^2 + S_2^2$.)

C. Information about the Variance Structure of the Population

Since the student variables of interest in this survey are highly related to school enrolment, it is desirable to consider the variance structure of the population with respect to school enrolment. Such information was available when this sample design effort began, from the 1993 and 1994 surveys. The following estimates were obtained from the 1994 survey data, using data for the 3,109 schools that had responded as of June 26, 1995.

Means and Standard Deviations of Total School Enrolments (1994 survey data)

Nation

Mean total enrolment per school = 840.25
Standard deviation of total enrolment per school = 664.34

Districts

Mean of district means (of total enrolment per school) = 975.21

Standard deviation of district means = 599.03
Mean of district standard deviations = 566.93

Zones

Mean of zone means (of total enrolment per school) = 893.28
Standard deviation of zone means = 501.54
Mean of zone standard deviations = 499.14

There is a wide variation in the mean school enrolment by district and zone. This is manifested in the fact that the standard deviations of the district and zone means are not small compared to the standard deviation of all schools in the nation. (Otherwise, we would expect the standard deviation of the unit means to be approximately equal to the standard deviation of school enrolments divided by the square root of the number of schools in the unit.) This means that the intradistrict and intrazone correlations are substantial. Using the formula given above for the correlation coefficient (r) in terms of S_1^2 (which is 599^2 for district means and 502^2 for zone means) and S^2 (which is 664^2 for the population of schools), we obtain estimates of .81 and .54 for the intradistrict and intrazone correlation coefficients, respectively.

Because of Malaŵi's explosive population growth, it is expected that the mean school enrolment for 1995 year will be greater than for 1994. In the analysis presented in this report, we shall use the value 900 for the mean student enrolment per school. The number of schools per district will be taken as 120, and the number of schools per zone will be taken as 30.

Although information on school enrolments is very useful in assisting the development of a sampling plan, it is highly desirable to analyze the variance structure of student variables of interest, such as repetition rates and dropout rates. Unfortunately, although the 1994 primary pupil registration system data were available when work began on this survey design effort, work had not been completed on editing and documenting these data, and neither the time nor the resources were available to edit the data and perform a variance analysis of the student data. For this reason, data-based estimates of the variances of student variables within and among schools are not known. It is conjectured that the intraschool correlation coefficient of these characteristics is positive, but not particularly high, e.g., in the range .1-.3. In the analysis of the efficiency of alternative designs, it will generally be assumed that the intraschool correlation coefficient has the value .2, although tables will be presented containing a range of values of the correlation coefficient.

The sample design analysis presented in this report is concerned to a considerable extent with assessment of the likely magnitudes of standard errors of estimates associated with alternative sample designs. These standard errors depend not only on the sample design, but also on the particular estimate (i.e., estimation formula) used. The analysis assumes that "standard" estimates are used. In the course of the data analysis, alternative estimators may be considered and selected (e.g., because they may have substantially lower standard errors).

Note that in the present application the school variables for which intraunit correlations are of interest are totals of students, not proportions of students. The reason for this is that the national estimates involve weighted sums of these totals.

D. Information about Sampling Costs

The objective of a sample design effort is to produce a sample design that returns a high level of precision for the cost expended in collecting and processing the data. In order to develop a design that is optimal, or at least highly efficient, it is necessary to specify the relative cost of sampling schools versus students. From previous primary school surveys, much information is available about the relative cost of sampling. While the sampling costs for future surveys may differ to some extent (e.g., because of changes in the questionnaires or data collection procedure (e.g., mail vs. personal visits)) this information may be used to suggest efficient survey designs.

The 1994 primary school survey cost about MK200,000 for questionnaire production, mailing, and fees to headmasters for recording of data about students. The fees paid to headmasters were independent of school size. While the MK200,000 may accurately reflect the cost to the Ministry of Education budget for the survey, it does not reflect the total cost of the survey, since it does not reflect salaries of district, zone, and school staff who were involved in the survey effort.

In searching for an efficient survey design, it will be attempted to use a cost function that reflects the actual total cost of the survey, not just the MOE "out-of-pocket" costs. The cost function will include components that address the cost of contacting schools, recording school and student information, following up nonrespondents, and data processing. Because not all of the costs are known with certainty, a "parametric" approach will be adopted, in which optimal designs will be developed under a number of different assumptions about the sampling costs. While the total sample size depends on the total amount available for the survey, the relative allocation of sampling effort between the first-stage units (schools) and second-stage units (students) depends mainly on the relative cost of sampling the first- and second-stage units, i.e., on the ratio of the cost of sampling first-stage units to second-stage units. Sample designs will be developed for a range of values of the sampling cost ratio, and a design selected that is appropriate for a reasonable range of sampling cost ratios.

For a survey administered similar to the last one, but with aggressive follow-up of nonrespondents, the ratio of the cost of sampling schools to the cost of sampling students is estimated to be in the range of 100-1000. The ratio of the cost of sampling zones to the cost of sampling schools is assumed to be in the range 5-10, and the ratio of the cost of sampling districts to the cost of sampling schools is assumed to be in the range 10-20.

E. Sample Frame

A probability sample is one in which each item of the population under study has a known, nonzero probability of selection (being included in the sample). In order to select a probability sample, it is necessary to have a complete, accurate, unduplicated list of the items to be selected. Such a list is called a sample frame. It is not necessary to have a list of all of the items to be included in the survey, prior to beginning the sample selection process. It is quite satisfactory to have a list of the first-stage sample units (e.g., schools), to select a probability sample of first-stage units, and to construct a sample frame for the second-stage units (e.g., students) only for those first-stage units that are selected in the sample.

During the course of the sample design process, effort was expended in the task of developing a school sample frame. A relatively complete list of schools was available from the 1994 school survey, but it was known that this list was now somewhat out-of-date. In March, a letter was sent to all 28 Malaŵi school districts, requesting an up-to-date school list. Complete school lists had been received from all districts prior to the initiation of the current sample design effort. During the first few weeks of the effort, these lists were used to update the list used in the 1994 school survey.

The lists received from the schools grouped all schools according to school zone. Since it was possible that it might be desirable to take school zone into account in the survey design process, the zone information was added to the school name file. (A dBASE program, ZONENTER.PRG, was written to facilitate the entry of the zone data.) The school name file was stored in the form of a dBASE IV version 2.0 data base file. The name of the 1994 school name file was SMASTER.DBF. The name of the updated school name file is SMAST2.DBF.

IV. Alternative Data Collection Approaches

The following is a list of the alternative methods that were identified for collecting the school and student data in the annual primary school survey:

1. Current method: census of schools and students (data are recorded for each school on a school data collection form, and data are recorded for each student (name, age, date of birth)

2. Census of schools, with no individual-student measurement (i.e., no recording of data for individual students -- student characteristics (age, sex, standard) are recorded on school-level data collection forms as tallies (counts))

3. Census of schools, with "sample" schools identified for intensive follow-up to reduce nonresponse; no individual-student measurement

4. Census of schools with "sample" schools identified for intensive follow-up to reduce nonresponse; individual-student measurement (either at sample schools or all schools), but collect all desired student data at both the school level (by means of tallies of students on a school data collection form) and the student level (by recording data for individual students). Measurements may be made on all students at a sample school (a "cluster" sample design), or on a sample of students at a sample school (a "two-stage" sample design).

The advantages and disadvantages of each of the preceding data collection methods were discussed in a series of meetings. A summary of the advantages and disadvantages is presented in Table 1.

As a result of the discussions of the alternative data collection methods, it was decided to use method 4. The significant advantage of method 4 over the other three methods is that it possesses their best advantages but not their worst disadvantages. It enables all of the data presentations of method 1, since schools are censused. It avoids the massive burden associated with method 1, by allowing for individual-student measurements on a sample of students. It overcomes the vulnerability of method 2 to a high level of nonresponse by permitting estimates to be based on a small (sampled) portion of the population for which it is feasible to expend special efforts to reduce the level of nonresponse. It allows for individual-student measurement (with the potential for smaller measurement errors), which neither method 2 nor method 3 does.

Table 1. Advantages and Disadvantages of Alternative Data Collection Approaches

1. Current method (census of schools and students)

Advantages:

- o If had 100% response, no sampling errors
- o Estimation straightforward
- o Can disaggregate as finely as desired (to school, standard, age, sex), tables and crosstabulations by all variables
- o Can use data for enforcement and resource allocation as well as policy analysis and planning
- o The use of the student sampling unit may reduce measurement errors
- o Can use SPSS (or even dBASE) to analyze the survey data

Disadvantages:

- o Heavy burden at Ministry of Education (MOE) headquarters (HQ)
- o Because can't afford to follow up all nonresponding schools (15% x 3400 = 510), have a high level of nonresponse, and therefore a possible high bias. Can't use statistical theory to estimate errors.
- o Long delays in processing data
- o Difficult to clean up data (because of massive size of data base) -- errors remain

2. Census of schools, no individual-student measurement

Advantages:

- o If had 100% response, no sampling errors
- o The level of nonresponse may be lower than for method 1 since the schools do not have to record individual-student data
- o Estimation straightforward
- o Can disaggregate to school level. Tables to any level of disaggregation, crosstabulations to school level.
- o Can use data for resource allocation, as well as policy analysis and planning
- o Can use SPSS or dBASE to analyze the survey data

Disadvantages:

- o Possible high level of nonresponse; possible high bias
- o Can't use data for enforcement at student level
- o Measurement of student characteristics may be subject to greater error
- o Can't disaggregate crosstabulations past school level (unless the crosstabulation was collected on the school survey form)

Table 1. Advantages and Disadvantages of Alternative Data Collection Approaches (continued)

3. Census of schools, with "sample" schools identified for intensive follow-up to reduce nonresponse; no individual-student measurement

Advantages:

- o With intensive follow-up of a small or modest sample of schools, can obtain estimates of known precision, very likely of higher accuracy than current census having 15 percent nonresponse
- o By end of year, may obtain results for rest of population
- o Can assess accuracy of census approach (for schools)
- o Otherwise, advantages and disadvantages are the same as for Alternative 2.

4. Census of schools with "sample" schools identified for intensive follow-up to reduce nonresponse; individual-student measurement (either at sample schools or all schools), but collect all desired student data at both the school level (by means of tallies of students on a school data collection form) and the student level (by recording data for individual students). Measurements may be made on all students at a sample school, or on a sample of students at a sample school.

Advantages:

- o Same as Alternative 3, plus can disaggregate tables to student level
- o May have less measurement error for student characteristics
- o Can assess the accuracy of the no-student-sampling approach

Having decided upon method 4 as the preferred data collection method, work began on the task of identifying alternative sample designs for that method. Before proceeding further, it is emphasized that method 4 is a "hybrid" method, that includes both a census and a sample of schools. The data from the census of schools will be used to produce all of the tables -- district-level as well as national-level -- of Education Statistics. The data from the sample of schools will be used to reduce the effect of nonresponse, and will enable use of the individual-student measurement approach for student-level estimates (e.g., repetition rates, dropout rates). The important thing to keep in mind in the discussion that follows, which concentrates on sampling of schools, is that the school sampling is done in conjunction with the school census, not exclusive of it.

Having decided on the hybrid census/sample approach, it remains to identify and evaluate alternative sample designs for this approach. The following designs were identified. The designs are divided into two categories -- those intended to provide national-level estimates of a reasonable level of precision, and a design intended to provide district-level estimates of a reasonable level of precision.

In the designs that follow, the selection of schools is done either with replacement or without replacement. In sampling without replacement, a sample item is excluded from the population after it is selected, so that it cannot be selected more than once. In sampling with replacement, selected items are not excluded after selection, so they may be selected more than once. The estimation formulas are simpler for selection with replacement, but the precision of the sample estimates is less.

The precision of sample estimates (such as population means, proportions, or totals) is measured by the variance of the estimate, or by the standard error of the estimate (which is the square root of the variance). The variance of estimates obtained using without-replacement sampling is smaller than that of estimates obtained using with-replacement sampling by a factor called the "finite population correction," or FPC. For simple random sampling, this factor is $1-n/N$, where n denotes the sample size and N denotes the population size. For situations in which the sample size is a substantial proportion of the total population, the FPC can result in a dramatic reduction in the variance. As the sample size approaches the population size, the variance (of an estimated mean) approaches zero for without-replacement sampling, but is proportional to 1/(square root of n) for with-replacement sampling, independent of the population size.

Since the sample sizes in the current application can be an appreciable portion of the total population (e.g., a sample of n=500 schools out of a total population of N=3,400 schools), the reduction associated with using without-replacement sampling is appreciable, and far outweighs any advantage associated with simplicity of the estimation formulas. Hence, without-replacement selection is assumed for all of the designs considered below.

Note that it is not possible to use SPSS or dBASE to analyze the student sample data for any of the probability-sampling designs considered (because the student sample is not a simple random sample). SPSS or dBASE may be used to assist

analysis of the school sample data for those designs in which a simple random sample of schools is selected, as long as allowance is made for the finite population correction (i.e., the estimates produced by SPSS or dBASE will be correct, but the estimated variances must be reduced by the FPC).

Designs for National-Level Estimates

1. Single-stage sample (or "cluster" sample) of schools: a simple random sample of schools, with inclusion of all students in each sample school. (As discussed, the sample is selected using sampling without replacement.)

2. Two-stage sample of schools and students: a simple random sample of schools, with a simple random sample of students selected from each sample school. The schools are the first-stage sample units and the students are the second-stage sample units.

3. Three-stage sample of zones, schools, and students: a simple random sample of zones, with a simple random sample of schools selected from each sample zone, and a simple random sample of students selected from each sample school. (All of the students or a sample of students at a sample school may be included in the sample.)

4. Three-stage sample of districts, schools, and students: a simple random sample of districts, with a simple random sample of schools selected from each sample zone, and a simple random sample of students selected from each sample school. (All of the students or a sample of students at a sample school may be included in the sample.)

5. Single-stage sample of schools, with selection of schools with probabilities proportional to a measure of size. The measure of size is the 1994 (total) enrolment. (All of the students at a sample school are included in the sample.)

6. Single-stage sample of schools, with schools stratified by a measure of size, and schools selected from within each stratum using selection with equal probabilities without replacement. The measure of size is the 1994 enrolment.

Design for District-Level Estimates

7. Single-stage sample of schools, with schools stratified by district and selected with probabilities proportional to a measure of size within each district.

Table 2 summarizes the major potential advantages and disadvantages associated with each of the preceding sample designs. The table describes potential advantages and disadvantages. Whether a particular design is advantageous depends on a number of factors, such as the relative cost of sampling units at the various stages, the intraunit (e.g., intraschool) correlation coefficient of the variables under study, and the ratio of the within-stratum variance to the

between-stratum variance for the variables under study. The following section will examine each type of design in detail, prior to selecting a preferred design.

Table 2. Potential Advantages and Disadvantages of Alternative Sample Designs

Designs for National-Level Estimates

1. Single-stage sample (or "cluster" sample) of schools: a simple random sample of schools, with inclusion of all students in each sample school.

Advantages:

- o Easy to select sample, and to delete items or add items to the sample
- o Relatively simple estimation formulas
- o The efficiency of the design is not affected adversely by a high value of the intraunit (intracluster) correlation coefficient
- o Relatively easy to analyze the data: SPSS or dBASE may be used to compute school estimates (but the estimated standard errors must be reduced by the FPC)

Disadvantages:

- o The design may be very inefficient, if the relative costs of first-stage and second-stage sampling vary substantially ("efficiency" refers to the return of precision for the level of sampling effort expended). It may be possible to realize substantial cost savings by sampling zones or districts, or by selecting a sample of students from each sampled school.
- o The design may be very inefficient, if the first-stage sample units have a high intraunit correlation coefficient (i.e., units are internally homogeneous -- units within the cluster are much more similar to each other than to units in the general population)
- o Will have sample schools in most or all zones, precluding concentration of follow-up effort in a small number of zones; may therefore have high nonresponse

2. Two-stage sample of schools and students: a simple random sample of schools, with a simple random sample of students selected from each sample school.

Advantages:

- o For variables for which the first-stage sample units have a high intraunit correlation coefficient, the efficiency of the design can be increased dramatically by selecting a sample of second-stage items, rather than by including all of them in the sample

Table 2. Potential Advantages and Disadvantages of Alternative Sample Designs (continued)

Disadvantages:

o The drawing of a student sample at each sample school is somewhat complicated, and must be carefully supervised. Since a probability sample must be selected at the second stage (e.g., a sample of students at each sample school), the design is more difficult to administer

o If the second-stage sample is small, it may not be possible to obtain estimates of adequate precision for "small" categories (e.g., female students in standard 8)

o Analysis may be somewhat complicated, depending on size and manner of selection of student samples

o Precision will be low for estimates of characteristics for which the intraunit correlation coefficient (intraschool correlation coefficient) is high

3. Three-stage sample of zones, schools, and students: a simple random sample of zones, with a simple random sample of schools selected from each sample zone, and a simple random sample of students selected from each sample school.

Advantages:

o If the number of zones is small, it may be possible to achieve a very high response rate by suitable training and monitoring of zone inspectors. (Each zone inspector is responsible for a small number of schools (e.g., six). It may be desirable to use a design with a limited number of zones, in order to take advantage of this administrative feature of the population, e.g., by training a smaller number of zone inspectors to assist in the data collection process, or by concentrating nonresponse follow-up efforts in a smaller number of zones. The survey-related workload of each zone inspector could be controlled by using the same-size sample of schools within each zone.)

o The design may be tailored to be highly efficient

Disadvantages:

o A design having a small number of zones might be substantially less efficient than a design having a larger number of zones. Precision will be low for estimates for which either the intrazone correlation coefficient or the intraschool correlation coefficient is high

Table 2. Potential Advantages and Disadvantages of Alternative Sample Designs (continued)

4. Three-stage sample of districts, schools, and students: a simple random sample of districts, with a simple random sample of schools selected from each sample zone, and a simple random sample of students selected from each sample school.

Advantages and disadvantages similar to (3)

5. Single-stage sample of schools, with selection of schools with probabilities proportional to a measure of size. The measure of size is the 1994 (total) enrolment.

Advantages:

- o Since the variables of interest in the present survey are correlated with enrolment, a substantial increase in design efficiency (return of precision for cost expended) may be realized by selecting primary sample units with probabilities proportional to a variable (1994 enrolment) related to current enrolment

Disadvantages:

- o Formulas for the estimates and their standard errors are available only if complicated sample-selection procedures are strictly followed
- o The sample, once selected, cannot easily be modified (the number of first-stage sample units changed). The sample must be completely redrawn.
- o The sample design will be efficient for the estimates presented in Education Statistics, but will not necessarily be efficient for ad-hoc surveys in which the variables of interest are unrelated to school size.
- o The analysis (computation of estimates and their estimated standard errors) must be done in accordance with specific formulas. It is in general not possible to use the SPSS statistical program package to analyze the data. The SPSS package may be used to compute the values of the estimates in some instances (e.g., a "self-weighting" sample), but even in these cases it cannot be used to compute the estimated standard errors of the estimates).

6. Single-stage sample of schools, with schools stratified by a measure of size, and schools selected from within each stratum using selection with equal probabilities without replacement. The measure of size is the 1994 enrolment.

Advantages and disadvantages similar to (5).

Table 2. Potential Advantages and Disadvantages of Alternative Sample Designs (continued)

Design for District-Level Estimates

7. Single-stage sample of schools, with schools stratified by district and selected with probabilities proportional to a measure of size within each district.

Advantages:

- o At some future time, it may be useful to have district-level statistical estimates, in addition to those available from the census-of-schools data

Disadvantages:

- o The (school) sample size in each district must be quite large relative to the total (school) population size. The total number of schools for all districts will likely be very large, approaching the total population size. There would likely be little advantage to expending the effort require to use probability sampling (the complications of sample selection and data analysis), because the total sample size would approach the population size

V. Alternative Sample Designs

A. Approach Used to Assess the Adequacy of Alternative Probability Sampling Plans

The preceding section identified a number of sample designs based on probability sampling, any one of which might a priori be satisfactory for the annual primary school survey. In order to select a preferred design from the list of candidate alternatives, it is necessary to analyze the cost, precision, and other characteristics of each design. Then, by comparing the various designs on these characteristics, a preferred design may be selected. The advantages and disadvantages of each design were discussed among Planning Unit and USAID staff, as part of the process of identifying the preferred design.

The preceding section described the general reasons why each of the various alternative sample designs is used -- it discussed potential advantages and disadvantages of each design, over a variety of applications. In order to recommend a design or select a preferred design from among these alternatives, it is necessary to assess or evaluate the performance of each alternative relative to the particular application at hand, that is, to "optimize," or "tailor," the design for the Malaŵi annual primary school survey.

The process of optimizing each of seven alternative sampling plans requires more effort than the short-cut process of selecting a specific design based on its general characteristics, but it increases the likelihood of obtaining a final design that is best for the survey. The process is facilitated by conducting a parametric evaluation of each design, i.e., by determining the optimal configuration of each design for a range of values of key design parameters. These parameters include sampling cost ratios, variances (or standard deviations) of the total population and subpopulations of interest, and intraunit correlation coefficients.

Specification of Variances In the analysis that follows, estimates will be presented for the standard error of the estimate for estimated means and estimated proportions, such as the proportion of students who are repeaters. In the case of estimating proportions, the analysis is simplified somewhat, since the variance of the underlying (binomial) distribution is functionally dependent on the mean (through the relation $var(X) = p(1-p)$, where X is a random variable having a binomial distribution with mean value $E(X) = p$, i.e., the probability that X equals 1 is p and the probability that X equals 0 is 1-p). In this case, the variance is determined simply by specifying the value of p.

In the case of estimating proportions, the precision of the sample estimates will be measured in terms of the "relative sampling error," which is defined as 1.96 times the coefficient of variation of the estimate, or 1.96 times the standard error of the estimate (in percent) divided by the mean (i.e., divided by p). The relative sampling error may be interpreted as follows. For reasonably large samples (e.g., more than 30 schools), the probability is approximately 95% in repeated sampling that the interval formed by the estimate plus or minus the sampling error will include the true (population) value. (Note that it is the interval that is the random variable -- the population value is considered fixed.)

In the general case in which the variance is not determined by the mean, it is necessary to specify both the mean and the variance (or standard deviation), in order to specify the relative sampling error of an estimate. An example is the total enrolment (i.e., the estimate is the total enrolment for the nation, or for a specified subpopulation of interest). If the precision is measured by the sampling error rather than the relative sampling error, it is necessary only to specify the variance, not the mean.

Specification of Intraunit Correlation Intraunit correlation may be specified in two different ways -- in terms of the intraunit correlation coefficient or in terms of the variance among unit means (S_1^2) and the variance (S_2^2) among elements within units. For two-stage designs, intraunit correlation will be measured by the intraunit correlation coefficient, r, which is the correlation coefficient between two units within the same first-stage unit. Although a correlation coefficient may in general range from -1 to +1, in the present application the intracluster correlation coefficient is expected to be positive for all variables of interest.

For three-stage designs, there are two intraunit correlation coefficients of interest -- the correlation coefficient, r1, among second-stage units within a first-stage unit, and the correlation coefficient, r2.1, among third-stage elements within a second-stage unit, conditional on the mean of the second-stage units (within a first-stage unit). Alternatively, the degree of intraunit correlation may be specified in terms of the variance (S_1^2) among first-stage unit means, the variance (S_2^2)among second-stage unit means with first-stage units, and the variance (S_3^2) among elements within second-stage units.

The presentations that follow will represent intraunit correlation in terms of the intraunit correlation coefficients. (The computer program used to estimate sampling errors for multistage designs presents a table of values of the among-unit and within-unit variance (S's) in the case of estimation of proportions.)

Specification of Sampling Costs For two-stage surveys, a sampling cost function of the following form will be assumed:

$$C = c_1 n + c_2 nm$$

where C denotes the total cost, c_1 denotes the cost per first-stage unit included in the sample, c_2 denotes the cost per second-stage unit included in the sample, n denotes the number of first-stage units included in the sample, and m denotes the mean number of second-stage units selected per first-stage unit.

For three-stage surveys, a sampling cost function of the following form will be assumed:

$$C = c_1 n + c_2 nm + c_3 nmk$$

where C denotes the total cost, c_1 denotes the cost per first-stage unit included in the sample, c_2 denotes the cost per second-stage unit included in the sample, c_3 denotes the cost per third-stage unit included in the sample, n denotes the number of first-stage units included in the sample, m denotes the mean number of

second-stage units selected per first-stage unit and k denotes the mean number of third-stage units selected per second-stage unit.

In determining an optimal design, it is possible to minimize the total cost subject to a constraint on the variance (i.e., on the standard error of the estimate), or to minimize the variance subject to a constraint on the total cost. With the specified sampling cost functions, either problem produces the same formulas for estimating the optimal second- and third-stage sample sizes. In the present application, the total cost, C, is not so much a concern as the minimum sample size required to produce estimates having an adequate level of precision. That is, the problem is to minimize cost with respect to a constraint on the variance.

With the above cost functions, it turns out that the optimal numbers of second- and third-stage sample units are independent of n, the number of first stage sample units. Because of this, the total cost for the optimal design is of the form C = n * constant, where the constant involves the cost parameters (the c's) and the optimal second- and third-stage sample sizes. Furthermore, it also turns out that the optimal values of the second- and third-stage sample sizes depend only on the sampling cost ratios, c_1/c_2 and c_2/c_3.

Because of these facts, it is satisfactory in the present analysis to present tables of standard errors simply in terms of the sampling cost ratio(s) and the total number of first-stage units included in the sample. The design to be selected is the one for which the value of n produces a sufficiently small sampling error. Since there is no need to compute C, there is no need to specify all of the c's individually -- only the sampling cost ratios need be specified. This simplifies the analysis considerably, since it is significantly easier to specify reasonable ranges for the sampling cost ratios than to specify ranges for the individual sampling costs.

The tables that follow calculate the theoretical values of standard errors of estimate means or proportions and optimal sample sizes in accordance with formulas presented, in W. G. Cochran, Sampling Techniques, 3rd edition, John Wiley & Sons, 1977 (Reference 1). Exact page references are provided in Appendix A, and also in the computer program listings (included on the diskette transmitted with this report).

In one case, the theoretical formula presented in Reference 1 was modified. The variance formula presented in Reference 1 for stratified sampling assumes that the strata are optimally constructed and that the sample is optimally allocated to the strata. The formula used in the computer program used to construct the tables that follow use an approximate formula that accounts for errors in the construction of the strata resulting from imperfect knowledge of the unit sizes. The approximate formula is specified in the computer program listing for program SAMPDES6.PRG.

In the case of single-stage cluster sampling, no formula was given in Reference 1 for the theoretical standard error of the estimated mean or proportion. For this case, an approximation was derived; it is specified in the computer program listing for program SAMPDES5.PRG.

The preceding paragraphs describe the methodology for assessing and describing the precision of alternative probability sampling plans. In some instances, other factors, in addition to precision, were also important. For example, a design might be efficient from a theoretical viewpoint, but present practical difficulties that raise the risk of nonsampling errors to unacceptable levels. Such factors will be discussed along with the consideration of precision.

Note that each sample design considered here can be used to produce estimates of means (or totals) both for schools and for students. The precision associated with the two types of estimates can be quite different for a particular design. Since both types of estimates are of interest, the precision of both types of estimates will be discussed. In the present survey design effort, however, it is the student estimates that are of primary concern. In order for a probability sampling approach to be useful for the primary school survey, it is necessary that the school sample size be a relatively small portion of the total school population. With relatively small school sample sizes, however, the precision of school estimates will not be very high. This is not a great concern, since all of the school estimates will be available for the school census -- the primary motivation for probability sampling is to obtain better and more efficient student estimates.

On the other hand, since the average school size is about 900 students, even a modest-sized sample of schools contains a large number of students, so that (since the intraschool correlation coefficients are low for the variables of interest) the precision of student estimates will be acceptable. The evaluation of the precision characteristics of each design will address student estimates, followed by school estimates, since if the precision of the student estimates is not adequate, there is little point to considering the precision of the school estimates.

Most of the discussion will address estimation of proportions, with some discussion of estimation of total enrolment. In the case of estimation of proportions, attention will center on the case in which the true value of the proportion is $p = .2$. This corresponds approximately to the case of repetition rate.

B. Evaluation of Alternative Sampling Plans

1. Design 1. Single-stage sample (or "cluster" sample) of schools: A simple random sample of schools (without replacement), with inclusion of all students in each sample school.

Design 1 is a nationwide simple random sample of schools. The schools are selected without replacement, using equal probabilities of selection. This design is about the simplest probability design possible. It is easy to select the sample and to conduct the analysis.

Table 3 describes the precision associated with this design, for estimates of school means and totals. In order to achieve a sampling error of less than 100,000 for estimating the total (national) enrolment, a sample of over 800 schools is needed.

Table 4 describes the precision of this design for estimating the proportion of schools possessing a specified characteristic. In order to achieve a relative sampling error of approximately 10% for estimates of the proportion of schools possessing some characteristic in the case in which the true value of the proportion is $p = .2$, a sample size of approximately 1,000 schools is required (relative sampling error = 10.415 for $n = 1{,}000$ when $p = .2$).

Table 5 presents the precision for estimating the proportion of students having a specified characteristic, in the case in which the true value of the proportion is $p = .2$. In order to estimate the precision, it is necessary to specify the approximate value of the intraschool correlation coefficient. It is expected that the intraschool correlation coefficient is low for most variables of interest in the annual school survey, e.g., .1 or .2. Assuming a value of .2, a sample of 250 schools is adequate to estimate the proportion with a relative sampling error of approximately 10% (relative sampling error = 10.695 for $p = .2$, $r = .2$, and $n = 250$).

Table 3. Sampling Errors of Estimated Population Mean or Total for Simple Random Sampling without Replacement, Standard Deviation of Variable of Interest = 500

Program: SAMPDES7.PRG (dBASE IV v2.0) Date: 06/21/95 Time: 15:04:02

Population size (N) = 3400

Sample Size (n)	Sampling Error of Estimate (1.96 x standard error of estimate)	
250	Mean:	59.658
	Total:	202839.
500	Mean:	40.476
	Total:	137619.
600	Mean:	36.307
	Total:	123444.
700	Mean:	33.008
	Total:	112227.
800	Mean:	30.299
	Total:	103017.
900	Mean:	28.011
	Total:	95239.
1000	Mean:	26.037
	Total:	88526.
2000	Mean:	14.062
	Total:	47810.
3000	Mean:	6.137
	Total:	20866.
3400	Mean:	0.000
	Total:	0.

Table 4. Relative Sampling Errors of Estimated Proportions for Simple Random Sampling without Replacement

Program: SAMPDES1.PRG (dBASE IV v2.0) Date: 06/21/95 Time: 15:04:49

Population size (N) = 3400

Sample Size (n)	Relative Sampling Error (1.96CV, in percent) for Estimated Total-Population Proportion (p est) when True Value = p (CV = Coef. of Var. = 100 SE(est p)/p %) True Population Proportion (p) .1	.2	.3	.5	.7	.9
250	35.795	23.863	18.226	11.932	7.811	3.977
500	24.286	16.191	12.366	8.095	5.300	2.698
600	21.784	14.523	11.092	7.261	4.754	2.420
700	19.805	13.203	10.084	6.602	4.322	2.201
800	18.179	12.120	9.256	6.060	3.967	2.020
900	16.807	11.205	8.558	5.602	3.668	1.867
1000	15.622	10.415	7.954	5.207	3.409	1.736
2000	8.437	5.625	4.296	2.812	1.841	0.937
3000	3.682	2.455	1.875	1.227	0.804	0.409
3400	0.000	0.000	0.000	0.000	0.000	0.000

Table 5.- Relative Sampling Errors of Estimated Proportions for Single-Stage Cluster Sample Design, Clusters Selected Using Simple Random Sampling without Replacement, p=0.20

Program: SAMPDES2.PRG (dBASE IV v2.0) Date: 06/21/95 Time: 15:06:22

Total no. of clusters (primary sample units) in population (N) = 3400
Cluster size (no. of subunits (elements) per primary unit (M)) = 900

Cluster Sample Size (n)	Total No. of Subunits in Sample (nM)	Relative Sampling Error (1.96CV, in percent) for Estimated Total-Population Proportion (est p) when True Value (p) = 0.20 (CV = Coef. of Var. = 100 SE(est p)/p %) Intracluster Correlation Coefficient (r) 0	.1	.2	.5	.9	1.0
250	225000	0.796	7.585	10.697	16.886	22.644	23.867
500	450000	0.540	5.146	7.258	11.456	15.363	16.193
600	540000	0.484	4.616	6.510	10.276	13.780	14.525
700	630000	0.440	4.197	5.919	9.343	12.528	13.205
800	720000	0.404	3.852	5.433	8.576	11.500	12.121
900	810000	0.374	3.561	5.023	7.928	10.632	11.206
1000	900000	0.347	3.310	4.669	7.370	9.882	10.416
2000	1800000	0.188	1.788	2.521	3.980	5.337	5.625
3000	2700000	0.082	0.780	1.100	1.737	2.329	2.455
3400	3060000	0.000	0.000	0.000	0.000	0.000	0.000

2. Design 2. Two-stage sample of schools and students: a simple random sample of schools, with a simple random sample of students selected from each sample school (sampling at both stages without replacement).

To the extent that the characteristics of students within the same school are correlated, it may be more efficient to include just a sample of students from the same school and to sample a larger number of schools, than to include all of the students at each sample school. The average school size in Malaŵi is about 900 students. Table 6 shows that if students are sampled from the sample schools, reducing the number of students selected from the average of 900 to as low as 100 causes a negligible increase in the sampling error (which increases from 10.697 to 10.900 for p = .2 and r = .2).

Table 7 shows the optimal number of students to select at a sample school, for two different values of the sampling cost ratio (c_1/c_2). This table shows that quite small numbers of students can produce adequate levels of precision for national-level estimates (e.g., m_{opt} = 20 for cost ratio = 100, and m_{opt} = 63 for m_{opt} = 1,000). This result suggests that, after data have been collected from about 60 students at a school, it would be more cost-effective (yield a higher level of precision for a given sampling cost) to include another school in the sample rather than to continue sampling students at the same school.

While Table 6 and Table 7 suggest that student sampling at sample schools is an efficient approach, it was decided not to consider this approach, for two reasons. First, based on a field visit to a number of schools in the district of Ntchisi, it was observed that the local (district-level) resources available to assist the survey are extremely limited. Zone inspectors may make no more than one visit per term to remote schools. It was decided that it was not feasible to implement an approach that would require a student sample to be selected at each sample school. Resources are not available for MOE headquarters staff to visit each sample school, and a large amount of resources would be required to adequately train the zone inspectors in correct procedures for selecting a student sample at a school, and to provide them with the wherewithal to visit the schools and select a student sample. Apart from the issue of resources, the risk of nonsampling errors caused by procedural errors in the selection of the student sample was considered high, since this had not been done before.

A second reason by student sampling at student schools was rejected was that it precluded the ability to construct estimates (or tables) at a high level of disaggregation. For example, it might be desired at some point to construct a table showing repetition rate by standard, or by standard and sex. The numbers of students falling in the higher standards is quite small, so that estimates for these categories would be of very low precision. Although the estimation of means and proportions for small categories was not stated as a requirement, it is reasonable to believe that this might change. If student sampling were undertaken, exercising this option at a later date would be precluded.

The issue of the efficiency of sampling students at sample schools arises designs other than Design 2. The same rationale will apply for not recommending student sampling at sample schools. It should be remembered, however, that the practice of including all of the students at a sample school is inefficient (from the

point of view of the precision of national estimates). In order not to lose sight of this fact, tables of sampling errors will generally be presented that include student-sampling options in addition to the no-student-sampling option.

Table 6. Relative Sampling Errors of Estimated Proportions for Two-Stage Sample Design Using Simple Random Sampling without Replacement at Both Stages, p=0.20

Program: SAMPDES3.PRG (dBASE IV v2.0) Date: 06/28/95 Time: 08:19:54

Total no. of primary sample units in population (N) = 3400
Average no. of secondary sample units per primary unit (M) = 900

1st Stage Sampl Size (n)	2nd Stage Sampl Size (m)	Total 2nd Stage Sample Units (nm)	Relative Sampling Error (1.96CV, in percent) for Estimated Total-Population Proportion (est p) when True Value (p) = 0.20 (CV = Coef. of Var. = 100 SE(est p)/p %) Intraunit Correlation Coefficient (r) 0	.1	.2	.5	.9	1.0
250	50	12500	3.499	8.245	11.123	17.057	22.669	23.867
250	100	25000	2.469	7.903	10.900	16.967	22.656	23.867
250	900	225000	0.796	7.585	10.697	16.886	22.644	23.867
500	50	25000	2.469	5.631	7.571	11.582	15.382	16.193
500	100	50000	1.739	5.380	7.407	11.516	15.372	16.193
500	900	450000	0.540	5.146	7.258	11.456	15.363	16.193
1000	50	50000	1.739	3.684	4.911	7.467	9.897	10.416
1000	100	100000	1.219	3.491	4.784	7.416	9.889	10.416
1000	900	900000	0.347	3.310	4.669	7.370	9.882	10.416
2000	50	100000	1.219	2.122	2.742	4.070	5.351	5.625
2000	100	200000	0.847	1.952	2.627	4.023	5.344	5.625
2000	900	1800000	0.188	1.788	2.521	3.980	5.337	5.625
3000	50	150000	0.987	1.216	1.409	1.871	2.350	2.455
3000	100	300000	0.680	1.009	1.255	1.801	2.339	2.455
3000	900	2700000	0.082	0.780	1.100	1.737	2.329	2.455
3400	50	170000	0.924	0.877	0.826	0.653	0.292	0.000
3400	100	340000	0.634	0.601	0.567	0.448	0.200	0.000
3400	900	3060000	0.000	0.000	0.000	0.000	0.000	0.000

Table 7. Relative Sampling Errors of Estimated Proportions for Two-Stage Sample Design Using Simple Random Sampling without Replacement at Both Stages, p=0.20, Using Optimal Sampling and Subsampling Fractions

Program: SAMPDES3.PRG (dBASE IV v2.0) Date: 06/28/95 Time: 08:20:07

Total no. of primary sample units in population (N) = 3400
Average no. of secondary sample units per primary unit (M) = 900

1st Stage Sample Size (n)	Sampling Cost Ratio (c1/c2)		Relative Sampling Error (1.96CV, in percent) for Estimated Total-Population Proportion (est p) when True Value (p) = 0.20 (CV = Coef. of Var. = 100 SE(est p)/p %) Intraunit Correlation Coefficient (r) 0	.1	.2	.5	.9	1.0
250	100.00	mopt:	900.0	30.0	20.0	10.0	3.3	0.0
		mopt+:	900	30	20	10	3	2
		n*mopt+:	225000	7499	4999	2500	833	500
		RSE:	0.80	8.68	11.77	17.76	23.05	23.87
250	1000.00	mopt:	900.0	94.9	63.2	31.6	10.5	0.0
		mopt+:	900	95	63	32	11	2
		n*mopt+:	225000	23714	15809	7905	2635	500
		RSE:	0.80	7.92	11.03	17.16	22.77	23.87
500	100.00	mopt:	900.0	30.0	20.0	10.0	3.3	0.0
		mopt+:	900	30	20	10	3	2
		n*mopt+:	450000	14998	9999	4999	1666	1000
		RSE:	0.54	5.95	8.04	12.10	15.66	16.19
500	1000.00	mopt:	900.0	94.9	63.2	31.6	10.5	0.0
		mopt+:	900	95	63	32	11	2
		n*mopt+:	450000	47427	31618	15809	5270	1000
		RSE:	0.54	5.39	7.50	11.66	15.46	16.19
1000	100.00	mopt:	900.0	30.0	20.0	10.0	3.3	0.0
		mopt+:	900	30	20	10	3	2
		n*mopt+:	900000	29996	19997	9999	3333	2000
		RSE:	0.35	3.93	5.27	7.87	10.11	10.42
1000	1000.00	mopt:	900.0	94.9	63.2	31.6	10.5	0.0
		mopt+:	900	95	63	32	11	2
		n*mopt+:	900000	94854	63236	31618	10539	2000
		RSE:	0.35	3.50	4.86	7.53	9.96	10.42
2000	100.00	mopt:	900.0	30.0	20.0	10.0	3.3	0.0
		mopt+:	900	30	20	10	3	2
		n*mopt+:	1800000	59991	39994	19997	6666	4000
		RSE:	0.19	2.33	3.06	4.43	5.55	5.63
2000	1000.00	mopt:	900.0	94.9	63.2	31.6	10.5	0.0
		mopt+:	900	95	63	32	11	2
		n*mopt+:	1800000	189708	126472	63236	21079	4000
		RSE:	0.19	1.96	2.69	4.12	5.40	5.63

Table 7.. Relative Sampling Errors of Estimated Proportions for Two-Stage Sample Design Using Simple Random Sampling without Replacement at Both Stages, p=0.20, Using Optimal Sampling and Subsampling Fractions (continued)

3000	100.00	mopt:	900.0	30.0	20.0	10.0	3.3	0.0
		mopt+:	900	30	20	10	3	2
		n*mopt+:	2700000	89987	59991	29996	9999	6000
		RSE:	0.08	1.45	1.79	2.36	2.64	2.46
3000	1000.00	mopt:	900.0	94.9	63.2	31.6	10.5	0.0
		mopt+:	900	95	63	32	11	2
		n*mopt+:	2700000	284563	189709	94854	31618	6000
		RSE:	0.08	1.02	1.35	1.95	2.43	2.46
3400	100.00	mopt:	900.0	30.0	20.0	10.0	3.3	0.0
		mopt+:	900	30	20	10	3	2
		n*mopt+:	3060000	101985	67990	33995	11332	6800
		RSE:	0.00	1.14	1.33	1.49	1.16	0.00
3400	1000.00	mopt:	900.0	94.9	63.2	31.6	10.5	0.0
		mopt+:	900	95	63	32	11	2
		n*mopt+:	3060000	322504	215003	107502	35834	6800
		RSE:	0.00	0.62	0.73	0.83	0.65	0.00

3. Design 3. Three-stage sample of zones, schools, and students: a simple random sample of zones, with a simple random sample of schools selected from each sample zone, and a simple random sample of students selected from each sample school (sampling without replacement at all stages).

This design includes both the case in which there is no student sampling at a sample school, and the case in which a sample of students is selected at each sample school. For the reasons discussed above, sampling of students at a sample school is not recommended (but this was not decided at the time when this design was first selected for consideration).

This design was considered because it was conjectured that it might be desirable to rely heavily on zone inspectors for implementation of the survey, particularly when following up to reduce nonresponse levels. The advantage of such a design is that it might be possible to include a relatively small number of zones in the sample, thereby enabling concentration of resources.

Table 8 shows the precision associated with this design. The first-stage (zone) intraunit correlation coefficient (r1) is assumed to be equal to .54 (i.e., the value estimated from the 1994 data). As discussed earlier, the second-stage (school) intraunit correlation coefficient is expected to be low for the survey variables of interest (e.g., r2.1 = .2). Even if it is assumed that no student sampling is done, it is necessary to include a large percentage of the zones in the sample, in order to achieve adequate precision (e.g., relative sampling error = 12.727 for 41 = .54, r2.1 = .2, n = 100 zones, n = 5 schools per zone, k = 900 students per school; or relative sampling error = 4.441 for r1 = .54, r2.1 = .2, n = 120 zones, m = 5 schools per zone, k = 900 students per school). The number of schools must be increased considerably when zone sampling is done, to compensate for the loss in precision associated with zone sampling (e.g., in these examples from about 250 schools to 500 or 600 schools).

Table 8 shows that, from the point of view of increasing precision, there is little point to sampling all or even a large number of students in each school. Table 9 shows the optimal first- and second-stage sample sizes, for ranges of the sampling cost ratios. Note that there are 3,424 schools and 120 zones, for an average of about 30 schools per zone. The minimal sample size for any stage of unit two, to enable estimation of the variance. The optimal number of schools to sample per zone when r1 = .54 and r2.1 = .2 is two. If optimal school and student sample sizes are used, it is necessary to include all 120 zones in the sample, in order to achieve a relative sampling error not greater than 10% (c_1/c_2 = 5, c_2/c_3 = 100, n = 120 zones, m_{opt} = 2 schools per zone, k_{opt} = 20 students per school, relative sampling error = 8.17%). (Note: Even if the value of r1 is assumed to be much smaller than .54, e.g., r1 = .1, essentially the same conclusion is reached -- in the case of r1=.1, for example, it is necessary to select 80-100 zones to achieve adequate precision.) If all 120 zones must be included, the optimal design is in fact not a three-stage design at all, but a two-stage design (first stage schools, second stage students).

In summary, Tables 8 and 9 show that if zone sampling is used, it is necessary to include a large proportion or all of the 120 zones in order to achieve adequate precision. No design having a small number of zones provides adequate precision. The precision associated with a small zone sample size is too low,

and if a large number of zones is selected, the advantage of being able to concentrate effort in a small number of zones is lost. It is hence concluded that zone sampling is not advisable.

Table 8. Relative Sampling Errors of Estimated Proportions for Three-Stage Sample Design Using Simple Random Sampling without Replacement at All Stages for p=0.20 and First Stage Intraunit Correlation Coefficient (r1) = 0.54

Program: SAMPDES4.PRG (dBASE IV v2.0) Date: 06/28/95 Time: 08:22:02

Total no. of primary sample units in population (N) = 120
Average no. of secondary sample units per primary unit (M) = 30
Average no. of tertiary sample units per secondary unit (K) = 900

1st Stage Sampl Size (n)	2nd Stage Sampl Size (m)	3rd Stage Sample Size (k)	Total 3rd Stage Sample Units (nmk)	Relative Sampling Error (1.96CV, in percent) for Estimated Total-Population Proportion (est p) when True Value (p) = 0.20 (CV = Coef. of Var. = 100 SE(est p)/p %) 2nd Stage Intraunit Corr. Coefficient (r2.1) 0	.1	.2	.5	.9	1.0
60	5	50	15000	26.385	26.746	27.102	28.143	29.473	29.797
60	5	100	30000	26.340	26.706	27.067	28.122	29.469	29.797
60	5	900	270000	26.301	26.671	27.036	28.103	29.466	29.797
60	10	50	30000	26.340	26.485	26.628	27.054	27.612	27.750
60	10	100	60000	26.318	26.465	26.611	27.043	27.610	27.750
60	10	900	540000	26.298	26.447	26.595	27.034	27.608	27.750
80	5	50	20000	18.688	19.069	19.442	20.521	21.878	22.204
80	5	100	40000	18.641	19.027	19.406	20.500	21.874	22.204
80	5	900	360000	18.599	18.990	19.374	20.481	21.870	22.204
80	10	50	40000	18.641	18.794	18.945	19.393	19.973	20.116
80	10	100	80000	18.617	18.773	18.927	19.381	19.971	20.116
80	10	900	720000	18.596	18.754	18.910	19.371	19.969	20.116
100	5	50	25000	11.879	12.353	12.810	14.093	15.640	16.003
100	5	100	50000	11.819	12.302	12.766	14.068	15.635	16.003
100	5	900	450000	11.766	12.256	12.727	14.045	15.631	16.003
100	10	50	50000	11.819	12.011	12.200	12.750	13.448	13.617
100	10	100	100000	11.789	11.984	12.177	12.736	13.445	13.617
100	10	900	900000	11.762	11.961	12.156	12.724	13.443	13.617
120	5	50	30000	1.527	3.452	4.637	7.089	9.412	9.908
120	5	100	60000	1.075	3.295	4.534	7.047	9.406	9.908
120	5	900	540000	0.330	3.149	4.441	7.010	9.400	9.908
120	10	50	60000	1.075	2.229	2.963	4.496	5.955	6.266
120	10	100	120000	0.753	2.106	2.882	4.463	5.950	6.266
120	10	900	1080000	0.209	1.992	2.809	4.434	5.945	6.266

Table 9. Relative Sampling Errors of Estimated Proportions for Three-Stage Sample Design Using Simple Random Sampling without Replacement at All Stages for p=0.20 and First Stage Intraunit Correlation Coefficient (r1) = 0.54, Using Optimal Sampling and Subsampling Fractions

Total no. of primary sample units in population (N) = 120
Average no. of secondary sample units per primary unit (M) = 30
Average no. of tertiary sample units per secondary unit (K) = 900

1st Stage Sample Size (n)	Smplng Cost Ratio (c1/c2)	Smplng Cost Ratio (c2/c3)		Relative Sampling Error (1.96CV, in percent) for Estimated Total-Population Proportion (est p) when True Value (p) = 0.20 (CV = Coef. of Var. = 100 SE(est p)/p %) 2nd Stage Intraunit Corr. Coefficient (r2.1) 0	.1	.2	.5	.9	1.0
80	5.00	100.00	mopt:	***.**	0.7	0.9	1.5	2.0	2.1
			mopt+:	2	2	2	2	2	2
			n*mopt+:	160	160	160	160	160	167
			kopt:	900.0	30.0	20.0	10.0	3.3	0.0
			kopt+:	900	30	20	10	3	2
			n*m+*n+:	144000	4797	3198	1599	533	335
			RSE:	18.61	20.01	21.12	23.96	27.02	27.17
80	5.00	1000.00	mopt:	***.**	0.7	0.9	1.5	2.0	2.1
			mopt+:	2	2	2	2	2	2
			n*mopt+:	160	160	160	160	160	167
			kopt:	900.0	94.8	63.2	31.6	10.5	0.0
			kopt+:	900	95	63	32	11	2
			n*m+*n+:	144000	15170	10114	5057	1686	335
			RSE:	18.61	19.78	20.83	23.64	26.85	27.17
80	10.00	100.00	mopt:	***.**	0.9	1.3	2.1	2.8	3.0
			mopt+:	2	2	2	2	3	3
			n*mopt+:	160	160	160	166	224	237
			kopt:	900.0	30.0	20.0	10.0	3.3	0.0
			kopt+:	900	30	20	10	3	2
			n*m+*n+:	144000	4797	3198	1662	748	474
			RSE:	18.61	20.01	21.12	23.77	24.74	24.79
80	10.00	1000.00	mopt:	***.**	0.9	1.3	2.1	2.8	3.0
			mopt+:	2	2	2	2	3	3
			n*mopt+:	160	160	160	166	224	237
			kopt:	900.0	94.8	63.2	31.6	10.5	0.0
			kopt+:	900	95	63	32	11	2
			n*m+*n+:	144000	15170	10114	5255	2364	474
			RSE:	18.61	19.78	20.83	23.46	24.61	24.79

Table 9. Relative Sampling Errors of Estimated Proportions for Three-Stage Sample Design Using Simple Random Sampling without Replacement at All Stages for p=0.20 and First Stage Intraunit Correlation Coefficient (r1) = 0.54, Using Optimal Sampling and Subsampling Fractions (continued)

100	5.00	100.00	mopt:	***.**	0.7	0.9	1.5	2.0	2.1
			mopt+:	2	2	2	2	2	2
			n*mopt+:	200	200	200	200	200	209
			kopt:	900.0	30.0	20.0	10.0	3.3	0.0
			kopt+:	900	30	20	10	3	2
			n*m+*n+:	180000	5997	3998	1999	666	419
			RSE:	11.78	13.49	14.78	17.91	21.11	21.27
100	5.00	1000.00	mopt:	***.**	0.7	0.9	1.5	2.0	2.1
			mopt+:	2	2	2	2	2	2
			n*mopt+:	200	200	200	200	200	209
			kopt:	900.0	94.8	63.2	31.6	10.5	0.0
			kopt+:	900	95	63	32	11	2
			n*m+*n+:	180000	18963	12642	6321	2107	419
			RSE:	11.78	13.21	14.45	17.57	20.94	21.27
100	10.00	100.00	mopt:	***.**	0.9	1.3	2.1	2.8	3.0
			mopt+:	2	2	2	2	3	3
			n*mopt+:	200	200	200	208	280	296
			kopt:	900.0	30.0	20.0	10.0	3.3	0.0
			kopt+:	900	30	20	10	3	2
			n*m+*n+:	180000	5997	3998	2077	934	592
			RSE:	11.78	13.49	14.78	17.71	18.75	18.80
100	10.00	1000.00	mopt:	***.**	0.9	1.3	2.1	2.8	3.0
			mopt+:	2	2	2	2	3	3
			n*mopt+:	200	200	200	208	280	296
			kopt:	900.0	94.8	63.2	31.6	10.5	0.0
			kopt+:	900	95	63	32	11	2
			n*m+*n+:	180000	18963	12642	6569	2955	592
			RSE:	11.78	13.21	14.45	17.38	18.61	18.80

Table 9. Relative Sampling Errors of Estimated Proportions for Three-Stage Sample Design Using Simple Random Sampling without Replacement at All Stages for p=0.20 and First Stage Intraunit Correlation Coefficient (r1) = 0.54, Using Optimal Sampling and Subsampling Fractions (continued)

120	5.00	100.00	mopt:	***.**	0.7	0.9	1.5	2.0	2.1
			mopt+:	2	2	2	2	2	2
			n*mopt+:	240	240	240	240	240	251
			kopt:	900.0	30.0	20.0	10.0	3.3	0.0
			kopt+:	900	30	20	10	3	2
			n*m+*n+:	216000	7196	4797	2399	800	502
			RSE:	0.55	6.02	8.17	12.33	16.01	16.18
120	5.00	1000.00	mopt:	***.**	0.7	0.9	1.5	2.0	2.1
			mopt+:	2	2	2	2	2	2
			n*mopt+:	240	240	240	240	240	251
			kopt:	900.0	94.8	63.2	31.6	10.5	0.0
			kopt+:	900	95	63	32	11	2
			n*m+*n+:	216000	22756	15170	7585	2528	502
			RSE:	0.55	5.50	7.66	11.92	15.82	16.18
120	10.00	100.00	mopt:	***.**	0.9	1.3	2.1	2.8	3.0
			mopt+:	2	2	2	2	3	3
			n*mopt+:	240	240	240	249	337	355
			kopt:	900.0	30.0	20.0	10.0	3.3	0.0
			kopt+:	900	30	20	10	3	2
			n*m+*n+:	216000	7196	4797	2493	1121	711
			RSE:	0.55	6.02	8.17	12.08	13.33	13.39
120	10.00	1000.00	mopt:	***.**	0.9	1.3	2.1	2.8	3.0
			mopt+:	2	2	2	2	3	3
			n*mopt+:	240	240	240	249	337	355
			kopt:	900.0	94.8	63.2	31.6	10.5	0.0
			kopt+:	900	95	63	32	11	2
			n*m+*n+:	216000	22756	15170	7883	3546	711
			RSE:	0.55	5.50	7.66	11.68	13.17	13.39

4. Design 4. Three-stage sample of districts, schools, and students: a simple random sample of districts, with a simple random sample of schools selected from each sample zone, and a simple random sample of students selected from each sample school (sampling without replacement at all levels).

As was the case of Design 3, this design may or may not include sampling of students at sample schools.

This design was conjectured that it might be desirable to restrict sampling to a modest number of school districts. The advantage of this design is (as in the preceding case of zone sampling) that if sampling can be restricted to a small number of districts it is possible to concentrate survey resources, thereby promoting a higher response rate.

Table 10 shows the precision associated with this design. The first-stage (district) intraunit correlation coefficient (r1) is assumed equal to .84 (the value estimated from the 1994 data). As discussed, it is expected that the second-stage (school) intraunit correlation coefficient is low (e.g., r2.1 = .2). It is seen from Table 10 that even if it is assumed that no student sampling is done, it is necessary to include all of the districts in the sample, in order to achieve adequate precision. In fact, any design with less than all 28 districts included is highly inefficient -- no amount of school and student sampling can compensate for the loss of precision caused by omitting a district. The optimal design corresponds to n = 28 districts, i.e., the optimal design is a two-stage design involving schools and students. (Note: This same conclusion is still reached even if it is assumed that r1 is much smaller than .84, e.g., r1 = .1.)

In summary, Table 10 shows that district sampling should not be done.

Table 10. Relative Sampling Errors of Estimated Proportions for Three-Stage Sample Design Using Simple Random Sampling without Replacement at All Stages for p=0.20 and First Stage Intraunit Correlation Coefficient (r1) = 0.81

Program: SAMPDES4.PRG (dBASE IV v2.0) Date: 06/27/95 Time: 13:59:57

Total no. of primary sample units in population (N) = 28
Average no. of secondary sample units per primary unit (M) = 120
Average no. of tertiary sample units per secondary unit (K) = 900

1st Stage Sampl Size (n)	2nd Stage Sampl Size (m)	3rd Stage Sample Size (k)	Total 3rd Stage Sample Units (nmk)	Relative Sampling Error (1.96CV, in percent) for Estimated Total-Population Proportion (est p) when True Value (p) = 0.20 (CV = Coef. of Var. = 100 SE(est p)/p %) 2nd Stage Intraunit Corr. Coefficient (r2.1) 0	.1	.2	.5	.9	1.0
16	5	50	4000	57.803	58.099	58.393	59.266	60.411	60.694
16	5	100	8000	57.772	58.071	58.368	59.251	60.408	60.694
16	5	900	72000	57.744	58.046	58.346	59.237	60.405	60.694
16	10	50	8000	57.772	57.913	58.054	58.476	59.033	59.171
16	10	100	16000	57.756	57.899	58.042	58.468	59.031	59.171
16	10	900	144000	57.742	57.887	58.031	58.461	59.030	59.171
16	50	50	40000	57.747	57.764	57.782	57.836	57.907	57.925
16	50	100	80000	57.744	57.762	57.780	57.834	57.906	57.925
16	50	900	720000	57.741	57.759	57.778	57.833	57.906	57.925
16	120	50	96000	57.743	57.743	57.743	57.742	57.741	57.741
16	120	100	192000	57.742	57.742	57.741	57.741	57.741	57.741
16	120	900	1728000	57.741	57.741	57.741	57.741	57.741	57.741
20	5	50	5000	42.237	42.560	42.880	43.828	45.061	45.364
20	5	100	10000	42.202	42.529	42.853	43.812	45.058	45.364
20	5	900	90000	42.171	42.501	42.829	43.797	45.055	45.364
20	10	50	10000	42.202	42.357	42.511	42.971	43.576	43.726
20	10	100	20000	42.185	42.341	42.497	42.962	43.574	43.726
20	10	900	180000	42.169	42.328	42.485	42.955	43.573	43.726
20	50	50	50000	42.174	42.194	42.213	42.272	42.350	42.369
20	50	100	100000	42.171	42.191	42.211	42.270	42.349	42.369
20	50	900	900000	42.168	42.188	42.208	42.269	42.349	42.369
20	120	50	120000	42.170	42.170	42.170	42.169	42.168	42.168
20	120	100	240000	42.169	42.169	42.169	42.168	42.168	42.168
20	120	900	2160000	42.168	42.168	42.168	42.168	42.168	42.168

Table 10: Relative Sampling Errors of Estimated Proportions for Three-Stage Sample Design Using Simple Random Sampling without Replacement at All Stages for p=0.20 and First Stage Intraunit Correlation Coefficient (r1) = 0.81 (continued)

24	5	50	6000	27.308	27.723	28.132	29.324	30.842	31.210
24	5	100	12000	27.264	27.683	28.097	29.303	30.838	31.210
24	5	900	108000	27.224	27.648	28.066	29.285	30.834	31.210
24	10	50	12000	27.264	27.463	27.661	28.246	29.008	29.196
24	10	100	24000	27.241	27.443	27.643	28.235	29.006	29.196
24	10	900	216000	27.221	27.425	27.628	28.226	29.004	29.196
24	50	50	60000	27.228	27.253	27.278	27.354	27.454	27.479
24	50	100	120000	27.223	27.249	27.275	27.351	27.453	27.479
24	50	900	1080000	27.219	27.245	27.271	27.349	27.453	27.479
24	120	50	144000	27.223	27.222	27.222	27.221	27.219	27.219
24	120	100	288000	27.221	27.221	27.220	27.220	27.219	27.219
24	120	900	2592000	27.219	27.219	27.219	27.219	27.219	27.219
28	5	50	7000	2.039	4.871	6.580	10.100	13.427	14.137
28	5	100	14000	1.440	4.675	6.452	10.048	13.419	14.137
28	5	900	126000	0.471	4.493	6.336	10.002	13.412	14.137
28	10	50	14000	1.440	3.380	4.558	6.988	9.286	9.777
28	10	100	28000	1.016	3.238	4.466	6.950	9.280	9.777
28	10	900	252000	0.326	3.107	4.382	6.917	9.275	9.777
28	50	50	70000	0.638	1.258	1.661	2.507	3.315	3.488
28	50	100	140000	0.446	1.181	1.610	2.486	3.312	3.488
28	50	900	1260000	0.116	1.108	1.563	2.468	3.309	3.488
28	120	50	168000	0.405	0.384	0.362	0.286	0.128	0.000
28	120	100	336000	0.278	0.264	0.248	0.196	0.088	0.000
28	120	900	3024000	0.000	0.000	0.000	0.000	0.000	0.000

5. Design 5. Single-stage sample of schools, with selection of schools with probabilities proportional to a measure of size. The measure of size is the 1994 (total) enrolment.

This design is of interest because all of the student characteristics of interest in the school survey are correlated with enrolment. Table 11 shows the sampling error of the estimated total enrolment for the nation and for the mean enrolment per school. In order to construct this table, it was necessary to specify the standard deviation of school enrolment. Based on the 1994 data, this was observed to be about 500 students. To construct the table it is also necessary to specify the mean value of the reciprocals of the selection probabilities, divided by the square of the number of primary units in the population (N = 3,400). This is equal to the mean value of the reciprocals of the measures of size (1994 enrolments) multiplied by the total measure of size of the population (3,400 x 900 = 3,060,000) and divided by the square of the number of primary units (N = 3,400) in the population. From the 1994 data, this is observed to be approximately 2.0. Finally, it is necessary to specify the correlation of the variable of interest (e.g., 1995 enrolment) with the measure of size (1994 enrolment). This is estimated, as discussed earlier, to be at least as large as the correlation between the 1993 and 1994 enrolments, which is .85. We shall assume a value of .90.

Table 11 shows that for a sample of 500, the total national primary school enrolment may be estimated with a sampling error of approximately 85,000 (84,834 in the table). This means that a 95% confidence interval for the true total enrolment will be approximately $\pm$ 85,000.

Note that when schools are selected with probabilities proportional to a measure of size, the exact number of students falling the sample is rather uncertain, until the sample is selected. Furthermore, the expected number of students falling in the sample is expected to be somewhat greater than the number of sample schools (n = 500) times the mean number of students per school (M = 900, approximately) because, when schools are selected with probabilities proportional to a measure of size, it is the larger schools that tend to be selected. Hence, for a sample of 500, the number of students in the sample is expected to be somewhat larger than 500 x 900 = 450,000.

Table 12 shows the relative sampling errors of estimated proportions for Design 5, corresponding to the value of .9 for the correlation of the variable of interest with the measure of size. This table shows that if the true value of the proportion is p = .2, then a sample of size 500 schools or more is required to achieve a relative sampling error of 10%.

Recall that the efficiency of this design could be improved by allowing for sampling of students at each sample school rather than inclusion of all of the students at each sample school in the student sample. For the reasons discussed earlier, however, sampling of students from sample schools was rejected.

Table 11. Approximate Sampling Errors of Estimated Cluster Mean or Population Total for Single-Stage Cluster Sample Design, Clusters Selected without Replacement with Probabilities Proportional to a Measure of Size, Standard Deviation of Variable of Interest = 500

Program: SAMPDES5.PRG (dBASE IV v2.0) Date: 06/27/95 Time: 14:12:59

Total no. of clusters (primary sample units) in population (N) = 3400
Average no. of subunits (elements) per primary unit (M) = 900
(SUM(1/zi))/(N**2) = (Mo'SUM(1/Mi'))/(N**2) = 2.0
zi = probability of selection for i-th primary unit
Mi' = measure of size for i-th primary unit
Mo' = total measure of size for population

Cluster Sample Size (n)	Approx. Lower Bound to Expected No. of Subunits in Sample (nM)		Approximate Sampling Error of Estimate (1.96 x approx. standard error of estimate) Correlation between Variable of Interest (e.g., True Size) and Measure of Size 0	.2	.5	.9	.95	1.0
250	225000	Mean:	84.370	82.665	73.066	36.776	26.344	0.000
		Total:	286857.	281062.	248426.	125038.	89571.	0.
500	450000	Mean:	57.242	56.086	49.573	24.951	17.874	0.000
		Total:	194623.	190691.	168549.	84834.	60771.	0.
600	540000	Mean:	51.346	50.308	44.467	22.381	16.033	0.000
		Total:	174576.	171049.	151187.	76096.	54511.	0.
700	630000	Mean:	46.680	45.737	40.426	20.348	14.576	0.000
		Total:	158713.	155507.	137450.	69182.	49558.	0.
800	720000	Mean:	42.849	41.984	37.109	18.678	13.380	0.000
		Total:	145687.	142744.	126169.	63504.	45491.	0.
900	810000	Mean:	39.614	38.814	34.307	17.267	12.370	0.000
		Total:	134688.	131967.	116643.	58709.	42056.	0.
1000	900000	Mean:	36.822	36.078	31.889	16.050	11.498	0.000
		Total:	125195.	122665.	108422.	54571.	39092.	0.
2000	1800000	Mean:	19.886	19.484	17.222	8.668	6.209	0.000
		Total:	67613.	66247.	58554.	29472.	21112.	0.
3000	2700000	Mean:	8.679	8.504	7.516	3.783	2.710	0.000
		Total:	29509.	28912.	25555.	12863.	9214.	0.
3400	3060000	Mean:	0.000	0.000	0.000	0.000	0.000	0.000
		Total:	0.	0.	0.	0.	0.	0.

Table 12: Approximate Sampling Errors of Estimated Proportions for Single-Stage Cluster Sample Design, Clusters Selected without Replacement with Probabilities Proportional to a Measure of Size, Correlation between Variable of Interest and Measure of Size = 0.9000

Program: SAMPDES5.PRG (dBASE IV v2.0) Date: 06/27/95 Time: 14:13:08

Total no. of clusters (primary sample units) in population (N) = 3400
Average no. of subunits (elements) per primary unit (M) = 900
(SUM(1/zi))/(N**2) = (Mo'SUM(1/Mi'))/(N**2) = 2.0

Cluster Sample Size (n)	Approx. Lower Bound to Expected No. of Subunits in Sample (nM)	Approx. Rel. Sampling Error (1.96CV, in %) for Estimated Total-Population Proportion (p est) when True Value = p (CV = Coef. of Var. = 100 SE(est p)/p %) True Population Proportion (p) .1	.2	.3	.5	.7	.9
250	225000	22.066	14.710	11.235	7.355	4.815	2.452
500	450000	14.971	9.981	7.623	4.990	3.267	1.663
600	540000	13.429	8.952	6.838	4.476	2.930	1.492
700	630000	12.209	8.139	6.216	4.070	2.664	1.357
800	720000	11.207	7.471	5.706	3.736	2.445	1.245
900	810000	10.360	6.907	5.275	3.453	2.261	1.151
1000	900000	9.630	6.420	4.903	3.210	2.101	1.070
2000	1800000	5.201	3.467	2.648	1.734	1.135	0.578
3000	2700000	2.270	1.513	1.156	0.757	0.495	0.252
3400	3060000	0.000	0.000	0.000	0.000	0.000	0.000

6. Design 6. Single-stage sample of schools, with schools stratified by a measure of size, and schools selected from within each stratum using selection with equal probabilities without replacement. The measure of size is the 1994 enrolment.

Instead of selecting schools with probabilities proportional to a measure of size, a similar design is one in which schools are categorized, or stratified, by size, and the sample is allocated to the categories, or strata, in such a way as to minimize the sampling error. When schools are stratified by size, a sample of schools may be selected from each stratum using selection with equal probabilities. There is no need to select the sample within a stratum with probabilities proportional to the measure of size, since all of the schools in a stratum are of similar size.

An issue that arises with stratified sampling is how to construct, or define, the strata, i.e., how many strata should there be, and what should the endpoints (specified as a number of students) of each stratum be. Statistical theory (Reference 1, pp. 127 -135) provides a solution to this problem. The procedure is to construct a frequency distribution (histogram) of schools by size, and break the distribution at points that are equidistant on the cumulative sum of the square roots of the frequencies.

The number of strata to use is determined by assuming a cost function in terms of the number of strata, and to select the number of strata that minimizes the variance for of the estimates for specified total survey cost or minimizes the total sampling cost for a specified variance. The following cost function will be assumed:

$$C = L\ C_s + n\ C_n,$$

where L denotes the number of strata, n denotes the total sample size, C_s denotes the cost per stratum, and C_n denotes the cost per observation. For the same reasons presented in the discussion of optimal sample sizes in multistage sampling, the optimum depends on the ratio C_s/C_n.

Table 13 shows the standard error of the estimated total enrolment, for various numbers of strata, assuming optimal construction of the strata using the measure of size. In order to construct this table, it is necessary to specify a quantity that is equivalent to the mean of the reciprocals of the measures of size, multiplied by the total measure of size for the population and divided by the square of the number of primary units in the population. As discussed earlier, this quantity is equal to 2.0. Also, it is necessary to specify the correlation of the variable of interest with the variable of stratification (1994 enrolment). As discussed earlier, this value will be taken as .9.

Table 13 shows that the sampling error corresponding to a sample of 500 is about 100,000 students (100,660 in the table), if two strata are used. This design is not quite as precise as the preceding one because the probabilities of selection are not quite proportional to the measure of size -- the schools within a stratum do vary a little in size, and this variation is ignored by the sample selection process within each stratum (i.e., selection with equal probabilities).

In order to determine the stratum boundaries, it is necessary to examine the frequency distribution of school sizes for the 1994 school survey. Table 14 shows this frequency distribution, as well as the cumulative sums of the square roots of the frequencies. This table shows that the following stratum definitions should be used:

For two strata: 0-1,000, 1,001+

For three strata: 0-650, 651-1,400, 1,401+

For four strata: 0-500, 501-1,000, 1,001-1,700, 1,701+

For five strata: 0-400, 401-800, 801-1,200, 1,201-1,900, 1,901+

Table 15 shows the sampling errors for samples of n = 500 schools, where the number of strata is varied from two to ten (the case of a single "stratum" corresponds to no stratification at all -- the case of Design 1). From Table 14, it is seen (the column with the correlation between the variable of interest and the variable of stratification equal to .9) that there is little point to using more than two size categories (strata), or perhaps three, from the point of view of an optimal return of precision. For L = 2 strata the sampling error is 100,660 students. For L = 3 strata the sampling error is 92,204, and for L = 4 strata the sampling error is 89,055.

Since enrolment will change less from year to year in future years than in 1993-94 (when free primary education was introduced), the correlation between the enrolment in successive years may increase, e.g., to .95. In this case, there would be an advantage to having a larger number of size strata (e.g., 3 or 4).

Table 13. Approximate Sampling Errors of Estimated Cluster Mean or Population Total for Stratified Single-Stage Cluster Sample Design, Clusters Selected within Strata Using Simple Random Sampling without Replacement, Standard Deviation of Variable of Interest = 500, No. of Strata = 2, Optimal Construction of Strata and Optimal Allocation of Sample to Strata Assumed.

Program: SAMPDES6.PRG (dBASE IV v2.0) Date: 06/27/95 Time: 14:15:43

Total no. of clusters (primary sample units) in population (N) = 3400
Average no. of subunits (elements) per primary unit (M) = 900
(Mo'SUM(1/Mi'))/(N**2) = 2.0
Mi' = measure of size for i-th primary unit
Mo' = total measure of size for population

Cluster Sample Size (n)	Approx. Lower Bound to Expected No. of Subunits in Sample (nM)		Approximate Sampling Error of Estimate (1.96 x approx. standard error of estimate) Correlation between Variable of Interest (e.g., True Size) and Variable of Stratification 0	.2	.5	.9	.95	1.0
250	225000	Mean:	78.921	77.556	69.956	43.636	37.554	29.829
		Total:	268330.	263690.	237849.	148364.	127684.	101419.
500	450000	Mean:	53.545	52.619	47.463	29.606	25.479	20.238
		Total:	182053.	178905.	161373.	100660.	86629.	68810.
600	540000	Mean:	48.030	47.199	42.574	26.556	22.855	18.153
		Total:	163301.	160477.	144751.	90291.	77706.	61722.
700	630000	Mean:	43.666	42.910	38.705	24.143	20.778	16.504
		Total:	148463.	145895.	131598.	82087.	70645.	56114.
800	720000	Mean:	40.082	39.389	35.529	22.162	19.073	15.150
		Total:	136278.	133922.	120798.	75350.	64847.	51508.
900	810000	Mean:	37.056	36.415	32.846	20.489	17.633	14.006
		Total:	125989.	123811.	111677.	69661.	59951.	47619.
1000	900000	Mean:	34.444	33.848	30.531	19.044	16.390	13.019
		Total:	117109.	115084.	103806.	64751.	55726.	44263.
2000	1800000	Mean:	18.602	18.280	16.489	10.285	8.852	7.031
		Total:	63246.	62152.	56062.	34970.	30095.	23905.
3000	2700000	Mean:	8.118	7.978	7.196	4.489	3.863	3.068
		Total:	27603.	27126.	24467.	15262.	13135.	10433.
3400	3060000	Mean:	0.000	0.000	0.000	0.000	0.000	0.000
		Total:	0.	0.	0.	0.	0.	0.

Table 14. Frequency Distribution of School Enrolments, and Cumulative Sum of Square Root of Frequencies

VALUE	FREQUENCY	ROOTFREQ	CUMROOTF	CUMFREQ	CUMFREQF	CUMFREQS	CUMFREQSF
50	10	3.16	3.16	10	0.0031	89	******
100	33	5.74	8.91	43	0.0134	2772	0.0010
150	74	8.60	17.51	117	0.0364	12226	0.0046
200	136	11.66	29.17	253	0.0786	36437	0.0137
250	148	12.17	41.34	401	0.1246	69663	0.0262
300	155	12.45	53.79	556	0.1728	112464	0.0422
350	160	12.65	66.44	716	0.2225	164620	0.0618
400	121	11.00	77.44	837	0.2601	210044	0.0789
450	172	13.11	90.55	1009	0.3135	283033	0.1063
500	130	11.40	101.95	1139	0.3539	344713	0.1294
550	130	11.40	113.35	1269	0.3943	412982	0.1551
600	132	11.49	124.84	1401	0.4354	489073	0.1836
650	136	11.66	136.51	1537	0.4776	573830	0.2155
700	122	11.05	147.55	1659	0.5155	656445	0.2465
750	114	10.68	158.23	1773	0.5510	738992	0.2775
800	94	9.70	167.92	1867	0.5802	812082	0.3049
850	132	11.49	179.41	1999	0.6212	921265	0.3459
900	104	10.20	189.61	2103	0.6535	1012164	0.3800
950	92	9.59	199.20	2195	0.6821	1097162	0.4119
1000	78	8.83	208.03	2273	0.7063	1173136	0.4405
1050	104	10.20	218.23	2377	0.7387	1279527	0.4804
1100	69	8.31	226.54	2446	0.7601	1353834	0.5083
1150	66	8.12	234.66	2512	0.7806	1428091	0.5362
1200	72	8.49	243.15	2584	0.8030	1512567	0.5679
1250	49	7.00	250.15	2633	0.8182	1572611	0.5905
1300	47	6.86	257.00	2680	0.8328	1632426	0.6129
1350	49	7.00	264.00	2729	0.8480	1697413	0.6373
1400	47	6.86	270.86	2776	0.8626	1761959	0.6615
1450	42	6.48	277.34	2818	0.8757	1821891	0.6840
1500	32	5.66	283.00	2850	0.8856	1869096	0.7018
1550	43	6.56	289.55	2893	0.8990	1934584	0.7264
1600	35	5.92	295.47	2928	0.9099	1989594	0.7470
1650	28	5.29	300.76	2956	0.9186	2035099	0.7641
1700	24	4.90	305.66	2980	0.9260	2075306	0.7792
1750	24	4.90	310.56	3004	0.9335	2116691	0.7947
1800	15	3.87	314.43	3019	0.9382	2143341	0.8047
1850	12	3.46	317.90	3031	0.9419	2165282	0.8130
1900	16	4.00	321.90	3047	0.9469	2195300	0.8242
1950	21	4.58	326.48	3068	0.9534	2235641	0.8394
2000	8	2.83	329.31	3076	0.9559	2251406	0.8453
2050	10	3.16	332.47	3086	0.9590	2271717	0.8529
2100	13	3.61	336.08	3099	0.9630	2298692	0.8631
2150	10	3.16	339.24	3109	0.9661	2319930	0.8710
2200	8	2.83	342.07	3117	0.9686	2337303	0.8776

Table 14. Frequency Distribution of School Enrolments, and Cumulative Sum of Square Root of Frequencies (continued)

2250	6	2.45	344.52	3123	0.9705	2350619	0.8826
2300	1	1.00	345.52	3124	0.9708	2352885	0.8834
2350	6	2.45	347.96	3130	0.9727	2366856	0.8887
2400	4	2.00	349.96	3134	0.9739	2376390	0.8922
2450	4	2.00	351.96	3138	0.9751	2386099	0.8959
2500	7	2.65	354.61	3145	0.9773	2403438	0.9024
2550	3	1.73	356.34	3148	0.9782	2411021	0.9052
2600	3	1.73	358.07	3151	0.9792	2418705	0.9081
2650	2	1.41	359.49	3153	0.9798	2423952	0.9101
2700	2	1.41	360.90	3155	0.9804	2429292	0.9121
2750	4	2.00	362.90	3159	0.9817	2440217	0.9162
2800	3	1.73	364.64	3162	0.9826	2448593	0.9193
2850	4	2.00	366.64	3166	0.9838	2459918	0.9236
2900	4	2.00	368.64	3170	0.9851	2471401	0.9279
2950	0	0.00	368.64	3170	0.9851	2471401	0.9279
3000	0	0.00	368.64	3170	0.9851	2471401	0.9279
3050	5	2.24	370.87	3175	0.9866	2486499	0.9336
3100	2	1.41	372.29	3177	0.9873	2492634	0.9359
3150	4	2.00	374.29	3181	0.9885	2505149	0.9406
3200	0	0.00	374.29	3181	0.9885	2505149	0.9406
3250	1	1.00	375.29	3182	0.9888	2508370	0.9418
3300	0	0.00	375.29	3182	0.9888	2508370	0.9418
3350	4	2.00	377.29	3186	0.9901	2521667	0.9468
3400	4	2.00	379.29	3190	0.9913	2535200	0.9519
3450	2	1.41	380.70	3192	0.9919	2542055	0.9544
3500	2	1.41	382.11	3194	0.9925	2548991	0.9570
3550	0	0.00	382.11	3194	0.9925	2548991	0.9570
3600	2	1.41	383.53	3196	0.9932	2556157	0.9597
3650	1	1.00	384.53	3197	0.9935	2559791	0.9611
3700	2	1.41	385.94	3199	0.9941	2567105	0.9638
3750	1	1.00	386.94	3200	0.9944	2570822	0.9652
3800	1	1.00	387.94	3201	0.9947	2574615	0.9667
3850	0	0.00	387.94	3201	0.9947	2574615	0.9667
3900	0	0.00	387.94	3201	0.9947	2574615	0.9667
3950	0	0.00	387.94	3201	0.9947	2574615	0.9667
4000	0	0.00	387.94	3201	0.9947	2574615	0.9667
4050	1	1.00	388.94	3202	0.9950	2578630	0.9682
4100	0	0.00	388.94	3202	0.9950	2578630	0.9682
4150	1	1.00	389.94	3203	0.9953	2582771	0.9697
4200	1	1.00	390.94	3204	0.9956	2586964	0.9713
4250	0	0.00	390.94	3204	0.9956	2586964	0.9713
4300	0	0.00	390.94	3204	0.9956	2586964	0.9713
4350	1	1.00	391.94	3205	0.9960	2591291	0.9729
4400	0	0.00	391.94	3205	0.9960	2591291	0.9729
4450	0	0.00	391.94	3205	0.9960	2591291	0.9729
4500	1	1.00	392.94	3206	0.9963	2595757	0.9746
4550	0	0.00	392.94	3206	0.9963	2595757	0.9746

Table 14: Frequency Distribution of School Enrolments, and Cumulative Sum of Square Root of Frequencies (continued)

4600	1	1.00	393.94	3207	0.9966	2600324	0.9763
4650	2	1.41	395.36	3209	0.9972	2609593	0.9798
4700	1	1.00	396.36	3210	0.9975	2614293	0.9816
4750	0	0.00	396.36	3210	0.9975	2614293	0.9816
4800	0	0.00	396.36	3210	0.9975	2614293	0.9816
4850	0	0.00	396.36	3210	0.9975	2614293	0.9816
4900	0	0.00	396.36	3210	0.9975	2614293	0.9816
4950	0	0.00	396.36	3210	0.9975	2614293	0.9816
5000	1	1.00	397.36	3211	0.9978	2619276	0.9834
5050	0	0.00	397.36	3211	0.9978	2619276	0.9834
5100	1	1.00	398.36	3212	0.9981	2624375	0.9853
5150	0	0.00	398.36	3212	0.9981	2624375	0.9853
5200	0	0.00	398.36	3212	0.9981	2624375	0.9853
5250	1	1.00	399.36	3213	0.9984	2629614	0.9873
5300	0	0.00	399.36	3213	0.9984	2629614	0.9873
5350	0	0.00	399.36	3213	0.9984	2629614	0.9873
5400	0	0.00	399.36	3213	0.9984	2629614	0.9873
5450	1	1.00	400.36	3214	0.9988	2635017	0.9893
5500	1	1.00	401.36	3215	0.9991	2640502	0.9914
5550	0	0.00	401.36	3215	0.9991	2640502	0.9914
5600	0	0.00	401.36	3215	0.9991	2640502	0.9914
5650	0	0.00	401.36	3215	0.9991	2640502	0.9914
5700	0	0.00	401.36	3215	0.9991	2640502	0.9914
5750	0	0.00	401.36	3215	0.9991	2640502	0.9914
5800	0	0.00	401.36	3215	0.9991	2640502	0.9914
5850	0	0.00	401.36	3215	0.9991	2640502	0.9914
5900	0	0.00	401.36	3215	0.9991	2640502	0.9914
5950	0	0.00	401.36	3215	0.9991	2640502	0.9914
6000	0	0.00	401.36	3215	0.9991	2640502	0.9914
6050	0	0.00	401.36	3215	0.9991	2640502	0.9914
6100	0	0.00	401.36	3215	0.9991	2640502	0.9914
6150	0	0.00	401.36	3215	0.9991	2640502	0.9914
6200	0	0.00	401.36	3215	0.9991	2640502	0.9914
6250	0	0.00	401.36	3215	0.9991	2640502	0.9914
6300	0	0.00	401.36	3215	0.9991	2640502	0.9914
6350	0	0.00	401.36	3215	0.9991	2640502	0.9914
6400	0	0.00	401.36	3215	0.9991	2640502	0.9914
6450	0	0.00	401.36	3215	0.9991	2640502	0.9914
6500	0	0.00	401.36	3215	0.9991	2640502	0.9914
6550	0	0.00	401.36	3215	0.9991	2640502	0.9914
6600	0	0.00	401.36	3215	0.9991	2640502	0.9914
6650	0	0.00	401.36	3215	0.9991	2640502	0.9914
6700	0	0.00	401.36	3215	0.9991	2640502	0.9914
6750	0	0.00	401.36	3215	0.9991	2640502	0.9914
6800	0	0.00	401.36	3215	0.9991	2640502	0.9914
6850	1	1.00	402.36	3216	0.9994	2647343	0.9940
6900	0	0.00	402.36	3216	0.9994	2647343	0.9940

Table 14. Frequency Distribution of School Enrolments, and Cumulative Sum of Square Root of Frequencies (continued)

6950	0	0.00	402.36	3216	0.9994	2647343	0.9940
7000	0	0.00	402.36	3216	0.9994	2647343	0.9940
7050	0	0.00	402.36	3216	0.9994	2647343	0.9940
7100	0	0.00	402.36	3216	0.9994	2647343	0.9940
7150	1	1.00	403.36	3217	0.9997	2654455	0.9966
7200	0	0.00	403.36	3217	0.9997	2654455	0.9966
7250	0	0.00	403.36	3217	0.9997	2654455	0.9966
7300	0	0.00	403.36	3217	0.9997	2654455	0.9966
7350	0	0.00	403.36	3217	0.9997	2654455	0.9966
7400	0	0.00	403.36	3217	0.9997	2654455	0.9966
7450	0	0.00	403.36	3217	0.9997	2654455	0.9966
7500	0	0.00	403.36	3217	0.9997	2654455	0.9966
7550	0	0.00	403.36	3217	0.9997	2654455	0.9966
7600	0	0.00	403.36	3217	0.9997	2654455	0.9966
7650	0	0.00	403.36	3217	0.9997	2654455	0.9966
7700	0	0.00	403.36	3217	0.9997	2654455	0.9966
7750	0	0.00	403.36	3217	0.9997	2654455	0.9966
7800	0	0.00	403.36	3217	0.9997	2654455	0.9966
7850	0	0.00	403.36	3217	0.9997	2654455	0.9966
7900	0	0.00	403.36	3217	0.9997	2654455	0.9966
7950	0	0.00	403.36	3217	0.9997	2654455	0.9966
8000	0	0.00	403.36	3217	0.9997	2654455	0.9966
8050	0	0.00	403.36	3217	0.9997	2654455	0.9966
8100	0	0.00	403.36	3217	0.9997	2654455	0.9966
8150	0	0.00	403.36	3217	0.9997	2654455	0.9966
8200	0	0.00	403.36	3217	0.9997	2654455	0.9966
8250	0	0.00	403.36	3217	0.9997	2654455	0.9966
8300	0	0.00	403.36	3217	0.9997	2654455	0.9966
8350	0	0.00	403.36	3217	0.9997	2654455	0.9966
8400	0	0.00	403.36	3217	0.9997	2654455	0.9966
8450	0	0.00	403.36	3217	0.9997	2654455	0.9966
8500	0	0.00	403.36	3217	0.9997	2654455	0.9966
8550	0	0.00	403.36	3217	0.9997	2654455	0.9966
8600	0	0.00	403.36	3217	0.9997	2654455	0.9966
8650	0	0.00	403.36	3217	0.9997	2654455	0.9966
8700	0	0.00	403.36	3217	0.9997	2654455	0.9966
8750	0	0.00	403.36	3217	0.9997	2654455	0.9966
8800	0	0.00	403.36	3217	0.9997	2654455	0.9966
8850	0	0.00	403.36	3217	0.9997	2654455	0.9966
8900	0	0.00	403.36	3217	0.9997	2654455	0.9966
8950	1	1.00	404.36	3218	1.0000	2663401	1.0000

Table 15. Approximate Sampling Errors of Estimated Cluster Mean or Population Total for Stratified Single-Stage Cluster Sample Design, Clusters Selected within Strata Using Simple Random Sampling without Replacement, Standard Deviation of Variable of Interest = 500, No. of Strata from 1 to 10, Optimal Construction of Strata and Optimal Allocation of Sample to Strata Assumed.

Program: SAMPDES6.PRG (dBASE IV v2.0) Date: 06/27/95 Time: 14:16:14

Total no. of clusters (primary sample units) in population (N) = 3400
Average no. of subunits (elements) per primary unit (M) = 900
(Mo'SUM(1/Mi'))/(N**2) = 2.0

Cluster Sample Size (n)	Approx. Lower Bound to Expected No. of Subunits in Sample (nM)	No. of Strata (L)		Approximate Sampling Error of Estimate (1.96 x approx. standard error of estimate) Correlation between Variable of Interest (e.g., True Size) and Variable of Stratification 0	.5	.9	.95	1.0
500	450000	1	Mean:	40.476	40.476	40.476	40.476	40.476
			Total:	137619.	137619.	137619.	137619.	137619.
500	450000	2	Mean:	53.545	47.463	29.606	25.479	20.238
			Total:	182053.	161373.	100660.	86629.	68810.
500	450000	3	Mean:	55.629	48.646	27.119	21.587	13.492
			Total:	189140.	165398.	92204.	73397.	45873.
500	450000	4	Mean:	56.341	49.054	26.193	20.048	10.119
			Total:	191558.	166784.	89055.	68162.	34405.
500	450000	5	Mean:	56.667	49.242	25.753	19.293	8.095
			Total:	192667.	167421.	87559.	65597.	27524.
500	450000	6	Mean:	56.843	49.343	25.510	18.871	6.746
			Total:	193267.	167766.	86735.	64161.	22937.
500	450000	7	Mean:	56.949	49.404	25.363	18.612	5.782
			Total:	193628.	167974.	86235.	63279.	19660.
500	450000	8	Mean:	57.018	49.444	25.267	18.441	5.060
			Total:	193861.	168109.	85909.	62700.	17202.
500	450000	9	Mean:	57.065	49.471	25.201	18.324	4.497
			Total:	194022.	168201.	85684.	62300.	15291.
500	450000	10	Mean:	57.099	49.490	25.154	18.239	4.048
			Total:	194136.	168267.	85524.	62013.	13762.

7. Design 7. (Design for District-Level Estimates) Single-stage sample of schools, with schools stratified by district and selected with probabilities proportional to a measure of size within each district.

This design is of interest if it is desired to produce estimates by district. In this case, a comparable sample size is needed in each district. There are 3,424 schools nationwide in 28 districts, for an average of approximately 120 schools per district.

Table 16 shows the sample size needed to yield various levels of precision for estimated school means and district totals, assuming that each district contains 120 schools and that the standard error of school sizes is 500. To produce a sampling error of 4,000 students for the estimated district total enrolment requires (looking in the .9 column) a sample of at least 90 schools per district, for a total sample of 2,520 schools.

Table 17 shows the relative sampling error for estimated proportions. To produce a relative sampling error of about 10% within each stratum requires a sample of about 100 schools per district if the true value of the proportion is p = .2. This corresponds to a nationwide sample of 2,800 schools.

In summary, Tables 16 and 17 show that it is not feasible to use probability sampling to produce estimates of reasonable precision for districts -- the total sample size required is almost the entire school population (3,400).

Table 16: Approximate Sampling Errors of Estimated Cluster Mean or Population Total for Single-Stage Cluster Sample Design, Clusters Selected without Replacement with Probabilities Proportional to a Measure of Size, Standard Deviation of Variable of Interest = 500

Program: SAMPDES5.PRG (dBASE IV v2.0) Date: 06/21/95 Time: 16:00:06

Total no. of clusters (primary sample units) in population (N) = 120
Average no. of subunits (elements) per primary unit (M) = 900
(SUM(1/zi))/(N**2) = (Mo'SUM(1/Mi'))/(N**2) = 2.0

Cluster Sample Size (n)	Approx. Lower Bound to Expected No. of Subunits in Sample (nM)		Approximate Sampling Error of Estimate (1.96 x approx. standard error of estimate) Correlation between Variable of Interest (e.g., True Size) and Measure of Size 0	.2	.5	.9	.95	1.0
30	27000	Mean:	219.135	214.707	189.776	95.519	68.425	0.000
		Total:	26296.	25765.	22773.	11462.	8211.	0.
40	36000	Mean:	178.923	175.308	154.952	77.991	55.869	0.000
		Total:	21471.	21037.	18594.	9359.	6704.	0.
50	45000	Mean:	149.697	146.673	129.642	65.252	46.743	0.000
		Total:	17964.	17601.	15557.	7830.	5609.	0.
60	54000	Mean:	126.517	123.961	109.567	55.148	39.505	0.000
		Total:	15182.	14875.	13148.	6618.	4741.	0.
70	63000	Mean:	106.927	104.766	92.601	46.608	33.388	0.000
		Total:	12831.	12572.	11112.	5593.	4007.	0.
80	72000	Mean:	89.461	87.654	77.476	38.995	27.934	0.000
		Total:	10735.	10518.	9297.	4679.	3352.	0.
90	81000	Mean:	73.045	71.569	63.259	31.840	22.808	0.000
		Total:	8765.	8588.	7591.	3821.	2737.	0.
100	90000	Mean:	56.580	55.437	49.000	24.663	17.667	0.000
		Total:	6790.	6652.	5880.	2960.	2120.	0.
110	99000	Mean:	38.146	37.376	33.036	16.628	11.911	0.000
		Total:	4578.	4485.	3964.	1995.	1429.	0.
120	108000	Mean:	0.000	0.000	0.000	0.000	0.000	0.000
		Total:	0.	0.	0.	0.	0.	0.

Table 17. Approximate Sampling Errors of Estimated Proportions for Single-Stage Cluster Sample Design, Clusters Selected without Replacement with Probabilities Proportional to a Measure of Size, Correlation between Variable of Interest and Measure of Size = 0.9000

Program: SAMPDES5.PRG (dBASE IV v2.0) Date: 06/21/95 Time: 16:00:09

Total no. of clusters (primary sample units) in population (N) = 120
Average no. of subunits (elements) per primary unit (M) = 900
(SUM(1/zi))/(N**2) = (Mo'SUM(1/Mi'))/(N**2) = 2.0

Cluster Sample Size (n)	Approx. Lower Bound to Expected No. of Subunits in Sample (nM)	Approx. Rel. Sampling Error (1.96CV, in %) for Estimated Total-Population Proportion (p est) when True Value = p (CV = Coef. of Var. = 100 SE(est p)/p %) True Population Proportion (p) .1	.2	.3	.5	.7	.9
30	27000	57.311	38.207	29.181	19.104	12.506	6.368
40	36000	46.794	31.196	23.827	15.598	10.211	5.199
50	45000	39.151	26.101	19.935	13.050	8.543	4.350
60	54000	33.089	22.059	16.848	11.030	7.221	3.677
70	63000	27.965	18.643	14.239	9.322	6.102	3.107
80	72000	23.397	15.598	11.913	7.799	5.106	2.600
90	81000	19.104	12.736	9.727	6.368	4.169	2.123
100	90000	14.798	9.865	7.535	4.933	3.229	1.644
110	99000	9.977	6.651	5.080	3.326	2.177	1.109
120	108000	0.000	0.000	0.000	0.000	0.000	0.000

VI. Recommended Sample Design

Based on analysis and discussion of the alternative designs presented in the preceding section, it is recommended that Design 5, in which a single-stage sample of 500 schools is selected with probabilities proportional to a measure of size (1994 enrolment), be used for the annual Malaŵi primary school survey. (Recall that with a single-stage sample, all of the students in a sample school are included in the sample -- there is no subsampling of students at each sample school.) The following comments summarize the reasons for preferring this design.

Design 5 is preferred to Designs 1 and 2 (selection of schools with equal probabilities) because it is substantially more efficient. The greater degree of complexity of Design 5 is justified by the higher level of precision.

Design 5 is preferred to Designs 3 and 4 (zone and district sampling) also because it is substantially more efficient -- the same level of precision can be obtained with a significantly smaller size sample of schools. Also, a reasonable level of precision can be achieved with a sample of 500 schools.

Design 5 is preferred to Design 6 (stratification of schools by size) for several reasons. First, it is more efficient -- the sampling errors for a sample of 500 are somewhat smaller. Second, it is considered somewhat simpler to implement. Third, the design changes less each year. With the stratified design, the stratum boundaries and number of strata must be reexamined each year to make sure that they are still optimal. With Design 5, the measure of size (the most current year's enrolment) is simply updated, and the sample-selection program is reexecuted.

Design 7 is rejected because the total sample size is unacceptably large.

In the recommended design, there is no sampling of students from the sample schools -- all students at every sample school are included in the student sample. Although sampling of students at sample schools is more efficient (i.e., a higher level of precision can be obtained by selecting a larger sample of schools and a modest sample of students from each sample school than by selecting a smaller sample of schools and taking all students from each sample school), the risk associated with sampling of students was considered unacceptably high.

Reference

1. Cochran, W. G., Sampling Techniques, 3rd edition, John Wiley & Sons, New York, 1977

Appendix A. Computer Programs Used in the Analysis

The following computer programs were used in the analysis of the sample designs discussed in this report. The programs process 1993 and 1994 survey data to assist the development and analysis of the various designs, calculate the theoretical sampling errors of the various designs, select a sample in accordance with the recommended design, and analyze data collected in accordance with that design. The programs are written in the dBASE IV version 2.0 programming language. Listings of the programs are provided on the computer diskette that accompanies this report.

SAMPDES1.PRG. This program calculates the theoretical relative sampling error for estimated proportions in the case of simple random sampling without replacement. The formula is presented on p. 23 of Reference 1.

SAMPDES2.PRG. This program calculates the theoretical relative sampling errors for a single-stage cluster sample design, for the case in which the clusters are selected using simple random sampling without replacement. The formula is presented on p. 241 of Reference 1.

SAMPDES3.PRG. This program calculates the theoretical relative sampling errors for a two-stage sample design using simple random sampling without replacement at both stages. The formula for the variance is presented on p. 277 of Reference 1, and the formula for the optimal second-stage sample size is given on p. 281.

SAMPDES4.PGR. This program calculates the theoretical relative sampling errors for a three-stage sample design using simple random sampling without replacement at all stages. The formula for the variance is presented on p. 286 of Reference 1, and the formulas for the optimal second- and third-stage sample sizes are given on p. 288.

SAMPDES5.PRG. This program calculates approximate sampling errors of estimated cluster means or population totals for a single-stage cluster sample design, with clusters selected without replacement with probabilities proportional to a measure of size. The formula for the approximate variance is presented in the program listing.

SAMPDES6.PRG. This program calculates approximate sampling errors of the estimated cluster mean or population total for a single-stage cluster sample design, with clusters selected without replacement with probabilities proportional to a measure of size. The formula for the approximate variance is presented in the program listing.

SAMPDES7.PRG. This program estimates sampling errors for estimated means and totals in the case of simple random sampling without replacement.

AGGREG1.PRG. This program aggregates observations in a data file, by summing values for a specified variable (field) of aggregation. It was used to construct a file containing total 1993 enrolment for each school from a file containing enrolment by age, sex, and standard.

AGGREG1B.PRG. This program aggregates observations in a data file, by summing over different fields. It was used to construct a file containing total 1994 enrolment for each school from a file containing enrolment by age and standard.

AGGREG2.PRG. This program computes means and variances of a selected variable (field) of a data file, for subsets of the data file. It was used to construct estimates for school districts.

AGGREG2B.PRG. This program computes means and variances of a selected variable (field) of a data file, for subsets of the data file. It was used to construct estimates for school zones.

AGGREG3.PRG. This program generates a frequency distribution for a sample of observations, and the cumulative sum of the square root of the frequencies.

AGGREG4.PRG. This program selects a probability sample using the Rao-Hartley-Cochran (RHC) method of selecting a single-stage (cluster) sample of units without replacement, with probabilities proportional to a measure of size. The sample units in this application are primary schools and the measure of size is the 1994 total school enrolment. The RHC procedure for cluster sampling is described on pp. 266-67 of Reference 1.

AGGREG5.PRG. This program calculates characteristics of the selected sample.

CORR1.PRG. This program estimates the correlation coefficient between two variables.

ESTIM1.PRG. This program computes a sample estimate and its standard error, for a sample selected with the RHC procedure. It also simulates two random variables, for use as examples of the estimation procedure. The formula for the estimate is presented on p. 266 of Reference 1, and the formula for the estimated variance of the estimate is presented on p. 267.

UPDAT2.PRG. This program updates the value of a single field in one file with the value from a single field in another file. It was used to add enrolment data to the file, SMAST2.DBF, containing school ID numbers and names.

ZONENTER.PRG. This program assisted the entry of zone data into the school master file, SMAST2.DBF.

Procedures for Updating the Enrolments in file SMAST2.DBF, after Making Changes to the Enrolment Data File SUMENR.DBF. The following is a summary of the dBASE interactive commands required to update the total enrolments in the SMAST2.DBF data file, to take into account changes that are made to the SUMENR.DBF data file (e.g., data for another school, or changes to existing records).

Sign on the network.

Get in drive C: (command c:).

cd c:\edreport

copy s:\edreport\sumenr.dbf /v

Reboot (ctrl-alt-del) the computer, so the system will use the PATH statement in the local CONFIG.SYS file.

cd edreport

copy sumenr.dbf sumenrs.dbf /v

erase sumenr.mdx

dbase (when dBASE starts in the ASSIST mode, move to the dot-prompt mode by pressing Esc followed by Y)

use sumenr2

Index on SCHOOLID, with the command modi struct, and then enter a Y in the SCHOOLID index column, and then press ctrl-end.

disp stat (will now show that the SUMENR2.DBF file is now indexed on SCHOOLID).

use enragg94

zap

do aggreg1b (to aggregate the SUMENR2 records by school)

use enragg94

Update the field DZCODE in ENRAGG94 from SMAST2, with the command do updat2, with parameters enragg94,smast2,schoolid,schoolid,dzcode,dzcode.

Update the enrolments in SMAST2 from ENRAGG94, with the command do updat2, with parameters smast2,enragg94,schoolid,schoolid,totalenrol,total.

Save the new SMAST2.DBF file on a diskette, with the command !copy smast2.dbf a:/v from dBASE environment or copy smast2.dbf a:/v from MS-DOS environment.

Procedures for Using Programs AGGREG2.PRG, AGGREG2B.PRG, and AGGREG3.PRG to Compute District and Zone Statistics and an Enrolment Histogram

use enragdis

zap

do aggreg2

use enragzon

zap

do aggreg2b

use enraghis

zap

do aggreg3 (with parameters 0.0, 50.0)

Procedures for Using Program AGGREG4.PRG to Draw a New Sample and Summarize It

do aggreg4 (enter sample size, e.g., 500)

do aggreg5

To Print a Listing of Primary Schools by District and Zone, with Schools Alphabetically Arranged within Zones

use smast2

set order to order1 (Note: The index order1 was created with the command index on substr(dzcode,1,4)+substr(sname,1,20) tag order1.)

assist , then select REPORT1

To Print a Listing of Nonrespondent Schools (enrolment=0) by District and Zone

use smast2

set order to order1

set filter to totalenrol=0

assist , then select REPORT1

To Print a Listing of Primary Schools, Ordered Alphabetically

use smast2

set order to sname

assist , then select REPORT2

To Print a Listing of Primary Schools, Ordered by SCHOOLID, with Enrolments for Standards 3, 4, and 5

use steve1

set order to schoolid

assist , then select STEVREP2

Sample Design

for

Malaŵi Annual Primary School Survey

Part II: Description of Recommended Sample Design

June 19, 1995

Revised June 28, 1995

Prepared by:
J. George Caldwell, Ph.D.
4615 Schirra Court
Spartanburg, SC 29301 USA
Tel. 1-803-576-8407

Prepared for:
Academy for Educational Development
1875 Connecticut Avenue NW, 9th Floor
Washington, DC 20009-1202
Tel. 1-202-884-8000

Under contract to:
US Agency for International Development
Washington, DC

65

Table of Contents

Foreword iv

I. Introduction; Summary Description of the Annual Primary School Survey Sample Design 1

II. Sample Selection Procedure 2

III. Data Analysis Formulas 2

IV. Characteristics of the Selected Sample 3

V. Sample Selection in Future Years 5

VI. Implementation Schedule 5

Reference 9

Appendix

A. Listing of Sample Units 10

List of Tables

1. Characteristics of the Selected Sample 8

Foreword

This document was prepared by Dr. J. George Caldwell under contract to the Academy for Educational Development. The work was funded by the US Agency for International Development (USAID) under USAID Agreement No. 612-T-605A, Program No. 612-0240, "Girls' Attainment in Basic Literacy and Education (GABLE)." The USAID project officer was Dr. Andrew Sisson.

The sample design presented in this report is the product of the efforts of a number of individuals. During the effort to develop a sample design, the author worked with Dr. Sisson and Dr. Laurie Cameron of USAID and with several members of the Planning Unit of the Malaŵi Ministry of Education. The principal Planning Unit staff member who worked on the sample design effort was Mr. John Khozi, Statistician.

I. Introduction; Summary Description of the Annual Primary School Survey Sample Design

This report describes a sampling plan to be used to collect data for the Malaŵi annual primary school survey. This sample design was selected from a set of alternative designs described, evaluated, and compared in Part I of this report. The selected design consists of a nationwide sample of 500 schools and an estimated 524,093 students, intended to enable the production of national estimates of student characteristics with relative sampling errors of about ± 10% for estimated proportions, and an estimate of total (national) enrolment with a sampling error of approximately ± 85,000. (The "sampling error" is the maximum size of a 95% confidence interval for the estimate. The "relative sampling error" is the sampling error expressed as a percentage of the proportion or mean being estimated.)

The selected design was described briefly in Part I of this report, which described the rationale for recommending this design. This part of the report describes the selected design in greater detail, including procedures for selecting a sample in accordance with the design and procedures for analyzing sample data collected using the design. In addition, a sample is selected in accordance with the design. The sample schools are identified, along with information required to analyze the sample data.

The survey design is a single-stage (cluster) sampling plan in which the primary sample units -- schools -- are selected without replacement with probabilities proportional to a measure of size (ppms). The measure of size is the most recent known enrolment for each school. For the survey to be conducted this year, the measure of size is the 1994 total school enrolment, if known, and the average school enrolment, if the 1994 enrolment is not known.

All students at sample schools (i.e., schools falling in the sample) will be included in the sample. With this design, it will be possible to estimate tables and crosstabulations of a variables of interest according to student characteristics, such as age, sex, and standard. For small categories such as the higher standards, however, the precision of the estimated means and proportions will not be as high as for the overall (all-student) estimates (i.e., the relative sampling errors will be larger).

This part of this report consists of five additional sections and an appendix. Section II describes procedures for drawing the sample (i.e., selecting the sample schools). A sample is selected in accordance with the design, and the sample items are identified in Appendix A. Section III describes procedures for analyzing the sample data; specifically, it describes the formulas to be used to calculate the estimated population total from data collected using the recommended sample design, and to calculate its estimated standard error. Two random variables are simulated, and analyzed in accordance with the design. Section IV presents a description of the selected sample, in terms of the characteristics of the sample schools (locations and sizes of schools, total number of students, and expected sample sizes in categories ("domains") of interest). Section V describes sample design options in future years. Section VI presents an implementation plan. Appendix A presents a listing of a sample of 500 schools selected using the cluster sample design, including the values of

sample design parameters required (for each sample school) to analyze the survey data.

II. Sample Selection Procedure

The proposed sample design is a complex design, and the design complexity affects both the sample selection process and the data analysis procedures. The design features that introduce complications are the selection of schools with probabilities proportional to a measure of size and the selection without replacement. These features were introduced because they significantly increase the precision of the sample estimates, but they are not without an associated cost in terms of sampling and analysis complexity.

The significant problem associated with the sample selection is to insure that formulas are available (known) for producing good (high precision, low bias) estimates of the proportions, means, and totals of interest, and for estimating the precision (standard errors) of those estimates. A good procedure for selecting a ppms-without-replacement is the Rao-Hartley-Cochran (RHC)procedure. With this procedure, the school population is randomly divided into a number of groups equal to the desired sample size (500), such that the number of schools in each group is as similar as possible. In this case, there are 500 random groups, each consisting of 6 or 7 schools (3,400 schools in the population divided by 500 schools in the sample = 6.8 schools per random group). A single school is selected from each random group, with probability proportional to the measure of size.

A computer program, AGGREG4.PRG, was written to select a probability sample of 500 schools in accordance with the RHC procedure. The sample selected by the program is presented in Appendix A.

III. Data Analysis Formulas

Let N_g denote the number of schools in the g-th random group, z_i denote the probability of selection of the i-th school in the g-th random group, Z_g denote the sum of the z_i's for the g-th random group, and y_i denote the value of the random variable observed on the i-th school in the g-th random group. Let z_g denote the selection probability for the school selected from the g-th random group, and let y_g denote the value of the random variable observed on the school selected from the g-th random group.

In terms of this notation, the following formula is used to estimate the population total for the variable y:

$$Y_{RHC} = \text{Sum} \; (\; Z_g \; y_g \; / \; z_g \;),$$

where the sum is over all n = 500 random groups (i.e., over the 500 sample observations, one of which is in each group).

The following formula is used to estimate the standard error of the estimated total:

$$v(Y_{RHC}) = [(\text{Sum } N_g^2 - N)/(N^2 - \text{Sum } (N_g^2))] \ / \ \text{Sum } (Z_g \ (y_g/z_g - Y_{RHC})),$$

where the sums are over all n = 500 random groups.

In addition to selecting a sample in accordance with the RHC procedure, the program AGGREG4.PRG generates two random variables -- a simulated total enrolment for each school and a simulated proportion that is uncorrelated with enrolment. These two random variables represent two extreme types of variables -- one highly correlated with the measure of size and the other uncorrelated with it. A computer program, ESTIM1.PRG, was written to calculate the estimate and its standard error according to the preceding formulas. Table 18 illustrates the program output for these two simulated random variables.

The simulated data are stored in the file SMAST2.DBF, in the fields SMALLY1 and SMALLY2. The ESTIM1.DBF program is set up to also analyze whatever data are contained in the field SMALLY in SMAST2.DBF.

The estimate presented above is the standard one associated with RHC sampling, for estimation of totals. The data analysis will surely include additional estimators, as well. For example, the proportion of repeaters is estimated as a ratio estimate (the ratio of the estimated total number of repeaters to the estimated total enrolment). As another example of a ratio estimator, it may be desirable to use estimates involving ratios of the current year's value of a variable to the previous year's value. In addition, the sample mean will be examined -- it is biased for RHC selection, but may have high precision. In addition, consideration may be given to several other estimation methods (e.g., the jackknife method, balanced repeated replications). These estimation procedures are described in Reference 1. In cases in which alternative estimators are available (e.g., two different estimates of a proportion), the standard errors of the estimates will be compared to assist a selection of a preferred estimator.

It is noted that the results of the analysis of data collected from the school probability sample may be used to improve the estimates obtained from the school census. If student-based estimates are seen to be better than school-based estimates for some variables, a ratio estimate involving the ratio of the student-based estimate to the school-based estimate may be applied to the census (school-based) data to obtain improved estimates from the census data.

IV. Characteristics of the Selected Sample

Appendix A includes a listing of all of the primary sample units (schools) of the sample, ordered by district, zone, and school name. To assist understanding of the proposed sample design, a descriptive analysis will be presented for the sample. The description will include frequency distributions of the numbers of schools and students falling in different districts and school size categories. These counts are useful in planning for the printing and distribution of survey data collection forms, and for understanding the level of precision that will be associated with estimates associated with small categories.

The proposed sample design is appropriate for producing estimates of characteristics (attributes) of the total school and student population. For the proposed design, the sampling error of (national) estimates of proportions is about ± .05 (±5%), and the relative sampling error is about ± 10%. (As mentioned, the "sampling error" is the maximum size of a 95% confidence interval for an estimate, and the "relative sampling error" is the sampling error divided by the proportion.) For simple random sampling, the sampling error is given by the formula ±1/(square root of sample size). The proposed sample design contains two types of sample units -- schools and students, and it is not a simple random sample of 500 schools or 524,093 students -- it is a complex design. Because schools are selected with probabilities proportional to a measure of school size (previous enrolment) that is highly correlated with current school size and other variables of interest in the survey, the precision of the recommended design for estimating school characteristics of interest will be substantially higher than the precision associated with simple random sampling (i.e., than the precision associated with a simple random sample of 500 schools). Because the student characteristics being measured are somewhat more similar for students in the same school than for students in the general population, however, the precision of the proposed design for estimating student characteristics will be substantially less than that associated with a simple random sample of 524,093 students. (The precision of a complex sample design is described in terms of the "design effect," which is the ratio of the variance of an estimator for the design to the variance of an estimator based on simple random sampling. For estimates of school characteristics, the design effect for the proposed design is substantially greater than 1.0; for estimates of student characteristics, the design effect for the proposed design is substantially less than 1.0.)

Since the sample design imposes no controls on the numbers of sample units falling in various subpopulation categories (i.e., no stratification is involved), the proportion of sample units falling into various population categories ("domains," such as categories of age, sex, or standard, or combinations of these characteristics) is approximately the same as the proportion of the total population falling into those categories. The only exception to this is school size; since schools are selected with probabilities proportional to a measure of size, there will be a disproportionate representation of large schools in the sample.

As discussed, the precision of estimates for subpopulations is determined by the number of sample items in the subpopulations. Estimates of useful precision may be computed for large subpopulations, e.g., the characteristics of students in standard 1 could be estimated with a reasonable level of precision, because there is a large proportion of such students in the population (and hence in the sample, which will contain students in approximate proportion to their representation in the population). For small subpopulations (e.g., girls in standard 8), the level of precision of the sample estimates will be low. (If additional information is desired about the precision of estimates for small subpopulations is desired, the computer programs described in Part I may be used to determine precision estimates. The proportion of the total population falling in a category of interest may be determined by applying the SPSS statistical analysis program to the 1994 survey / registration system data. The number of sample items falling in the category will be approximately the same proportion of the total sample. The dBASE program SAMPDES5.PRG may be executed using this

sample size, to obtain an approximate estimate of the precision of estimates associated with the subpopulation.)

The characteristics of the sample units are summarized in Table 1, which presents a frequency distribution of the number of sample schools falling in the various school districts, and a frequency distribution by school size (total enrolment).

V. Sample Selection in Future Years

The issue of whether to use the same sample for several years was not decided at the present time. This issue may be addressed during the coming year. The precision of estimates of change from year to year is higher if the same sample is used than if a new, independent, sample is selected each year. If the same sample is used for several years, all new schools coming into existence should be included in the sample, and their selection probabilities set equal to one. The sample design effort did not assess the precision associated with estimates of change obtained by using the same school sample each year.

A disadvantage of using the same sample each year is that the response rate for nonsample schools may decline over time (since they are subjected to less follow-up than sample schools).

A "hybrid" approach is to employ a "panel" sample design, in which a portion of the sample is replaced every year. Additional analysis would be required to estimate the precision associated with alternative panel sample designs, and to make recommendations concerning the portion of the sample to be replaced each year.

VI. Implementation Schedule

It is recommended that the school census and school sample data collection efforts be conducted at the same times as last year (over the period September-December), by the Planning Unit. Further, it is recommended that Mr. John Khozi, Statistician, be in charge of the data collection and analysis effort.

During the period between now and September, it is recommended that a high level of attention be given to the matter of design of the survey instruments (data collection forms and instructions). The recommended sample design will enable the production of estimates of high efficiency, but the advantage of the design will be lost if the data collection instruments are flawed.

The proposed design is premised on the measurement of all student concepts of interest (e.g., repetition rate) on both the school and student sample units. In order for this approach to be successful, high-validity means must be developed for collecting repetition-rate, dropout-rate, and years-to-completion data from both sources.

It is recommended that the survey instruments (forms, logs, and instructions for data collection, transfer, and editing) be drafted, pretested at MOE

72

headquarters, and pilot-tested in the field. The importance of conducting the instrumentation pilot test cannot be underemphasized.

It is recommended that at least six weeks be allowed for the instrument development, pretesting, pilot testing, and revision. The instruments should be pilot tested in at least two districts and at least ten schools. The pilot-test schools should include schools of a range of sizes and distances from the district headquarters.

In the full-scale data collection effort, it is highly recommended that data on the amount of follow-up (e.g., numbers of follow-up contacts) associated with each response be recorded, so that improved methods may be developed for adjusting census results (which will be based on less intensive follow-up than the survey results) based on the sample data results.

The proposed sampling plan will dramatically reduce the burden of data collection and processing associated with the annual primary school survey, and should be instrumental in increasing the school response rate (because more intensive follow-up measures can be undertaken for the sample schools than for the population of all schools). The advantage of this approach is not, however, without a cost. The use of probability sampling requires that strict attention be given to the sampling and data collection procedures, and that the proper procedures be employed in the statistical analysis. If the proper procedures are not employed, the estimates may be seriously biased or of unacceptably high precision.

This year will be the first year in which the Malaŵi annual primary school survey will be based on use of a probability sampling plan. The survey design is a complex design (i.e., it is not based on simple random sampling), and specialized formulas must be used to compute estimates and their standard errors (estimates of precision). The selection of the sample has been completed in accordance with the correct procedures, but the data analysis must also be done in strict compliance with the correct procedures, in order for the results to be useful. Some estimates of interest may be based on small sample sizes, and have high standard errors. If such estimates are presented in reports without discussion of their accuracy, the result will be embarrassment at least and wrong conclusions at worst.

Under the previous approach (census of schools and students), the data analysis was straightforward -- although the massive size of the data base may have caused problems, the estimation procedures were simple. A standard statistical program package (such as SPSS, SAS, or BMDP) or even a standard database development package (such as dBASE IV) could be used to analyze the data. The data represented the entire population, and the data analysis involved simply the computation of totals or means for this entire population. With the probability sample design approach, the data cannot be analyzed directly using a standard statistical program package or database package. Instead, the estimates and their estimated standard errors must be computed using the formulas that are appropriate for the sample design and sample selection procedures. Furthermore, alternative estimators may be available, and analysis is required to select the best one. This cannot be done in advance of the data analysis.

It is possible that, after the data are collected, requirements may arise for estimates other than those already identified. It may be desired to compute estimates for certain subpopulations ("domains of study") of the total population, or to determine estimates based on "poststratification" of the data. The procedures for determining these estimates are complicated, and they must be executed carefully if the results are to be correct and of value.

With the previous census approach, it was not necessary to present or discuss the sampling errors of the estimates in the annual Basic Education Statistics publication. With the probability-sampling approach, it is necessary to discuss the sampling errors of the estimates. There are two standard ways of doing this: (1) presentation of standard errors with each estimate (this may be cumbersome); or (2) presentation of tables and graphs of generalized variances, which summarize the accuracy of similar types of estimates (as a function of school and student sample sizes). Since this will be the first time for Ministry of Education personnel to undertake this approach, it is highly recommended that serious consideration be given to obtaining qualified technical assistance at various times during the coming year, especially during the instrumentation development, data processing, statistical analysis, and data presentation phases of the survey.

The development of the data processing and statistical analysis procedures can be initiated during the September-December period, while data collection is underway.

A major area of concern that may be overlooked is data "cleaning." In any data collection activity, errors may occur for a variety of reasons -- mistakes in data recording or data entry, lost forms, data processing errors. In order for estimates based on the survey data to be of value, considerable effort must go into the development and application of procedures for editing the data. These procedures include the development of data entry screens that perform validity and consistency checks on each data item, and the application of data editing (consistency, correlation) procedures to the full data set and subsets.

For the database and statistical software to be of value in future years, it is essential that it be properly documented. It is recommended that an established software development standard be used for the software development and documentation (e.g., the principal software development standard in the United States, the DOD-STD-2167A standard).

The development of adequate data editing, processing, and presentation software and documentation is not a trivial undertaking. It can have a dramatic affect on the quality and usefulness of the end product and on the amount of effort required to repeat the survey and make use of previous-year data in future years. A substantial commitment of technically qualified effort is required if this operation is to be accomplished successfully.

Table 1. Characteristics of the Selected Sample

Number of schools (in sample) = 500
Number of students (in sample schools) = 524093
Number of schools with enrolment <= 1,000 = 276
Number of students in schools with enrolment <= 1,000 = 127421
Number of schools with enrolment > 1,000 = 224
Number of students in schools with enrolment > 1,000 = 396672

Sample Allocation to Districts:

District No.	No. of Schools	No. of Students
1	8	6370
2	11	9827
3	8	3991
4	49	17644
5	5	7666
6	9	6680
7	27	23027
8	15	12254
9	15	15930
10	20	20680
11	7	5020
12	11	30798
13	50	43833
14	20	12092
15	14	6205
16	32	25673
17	24	28700
18	26	27218
19	1	1478
20	21	22189
21	7	9737
22	20	72437
23	14	11351
24	16	19189
25	38	51918
26	9	5268
27	18	22694
28	5	4224

Reference

1. Cochran, W. G., Sampling Techniques, 3rd edition, John Wiley & Sons, New York, 1977

Appendix A

Listing of Sample Units

(Nationwide PPMS Sample of 500 Schools, Selected without Replacement)

SCHOOLID	TOTALENROL	NG	SMALLZG	BIGZG	DISTRICT	SNAME
01027	481	7	0.000165	0.002080	CHITIPA	IPULA
01029	1008	7	0.000346	0.001818	CHITIPA	ISYALIKILA
01042	686	7	0.000235	0.001878	CHITIPA	KAROPA
01046	565	7	0.000194	0.001457	CHITIPA	KASISI
01050	1674	7	0.000574	0.002543	CHITIPA	KAWALE
01063	575	7	0.000197	0.002077	CHITIPA	MPALE
01078	773	7	0.000265	0.001798	CHITIPA	NGOYA
01090	608	7	0.000208	0.001679	CHITIPA	ZAMBWE
02002	1564	6	0.000536	0.002354	KARONGA	BWIBA
02011	333	7	0.000114	0.002132	KARONGA	CHIWIRA
02014	522	7	0.000179	0.001059	KARONGA	FULIRWA
02047	1330	7	0.000456	0.002372	KARONGA	LUPEMBE
02061	1110	7	0.000381	0.001736	KARONGA	MLARE
02067	1128	7	0.000387	0.001526	KARONGA	MWENITETE
02083	1636	7	0.000561	0.001970	KARONGA	ST MARY'S GIRLS
02088	618	7	0.000212	0.001861	KARONGA	TILORA
02091	870	7	0.000298	0.001664	KARONGA	ULIWA
02094	502	6	0.000172	0.001247	KARONGA	WILIRO
02102	214	7	0.000073	0.002453	KARONGA	SABI
03025	795	7	0.000273	0.001602	RUMPHI	CHOZOLI
03049	506	6	0.000173	0.001768	RUMPHI	LUVIRI
03052	145	7	0.000050	0.001065	RUMPHI	MARA
03054	675	6	0.000231	0.001311	RUMPHI	MBULUNJI
03068	1071	6	0.000367	0.002957	RUMPHI	NKHOZO
03071		7	0.000292	0.001840	RUMPHI	PHWAMPHWA
03078	583	7	0.000200	0.002428	RUMPHI	THUNDA
03088	216	7	0.000074	0.001569	RUMPHI	VUVU J.P
04001	489	7	0.000292	0.001744	MZIMBA	BARA
04008	1192	7	0.000409	0.001384	MZIMBA	BWENGU
04013	608	7	0.000208	0.001791	MZIMBA	CHAMUNGUMA
04033	733	7	0.000251	0.001661	MZIMBA	CHIPATA
04043	702	7	0.000241	0.001786	MZIMBA	DIMI
04051	2076	7	0.000292	0.001367	MZIMBA	EKWENDENI
04058	1391	7	0.000477	0.002161	MZIMBA	EMBANGWENI
04060	1068	6	0.000292	0.001313	MZIMBA	EMFENI
04061	534	7	0.000183	0.001406	MZIMBA	EMONENI
04065		7	0.000292	0.001729	MZIMBA	EMTIYANI
04072	706	7	0.000292	0.001813	MZIMBA	ENKWELENI
04074	1043	7	0.000358	0.001504	MZIMBA	ENUKWENI
04078	407	7	0.000140	0.001801	MZIMBA	ETHULENI
04085	326	6	0.000292	0.002013	MZIMBA	JANGAVYA
04089	583	7	0.000292	0.001421	MZIMBA	KABENA
04103	534	6	0.000183	0.002050	MZIMBA	KALITUBI
04130	350	7	0.000120	0.001729	MZIMBA	KAPANGAMAWE
04135	702	6	0.000241	0.001393	MZIMBA	KAPOLI
04149		6	0.000292	0.001761	MZIMBA	KAVITENGO
04157	361	7	0.000292	0.001925	MZIMBA	KAVIYOMBO
04158		7	0.000292	0.001446	MZIMBA	KHUYU
04163		7	0.000292	0.002193	MZIMBA	LUNJIKA

04170	1009	6	0.000346	0.002001	MZIMBA	LUZI
04172	804	7	0.000276	0.001984	MZIMBA	MABIRI
04176	523	7	0.000179	0.001394	MZIMBA	MALANGAZI
04184	754	7	0.000258	0.001790	MZIMBA	MATHANDANI
04203	551	7	0.000189	0.002561	MZIMBA	MPENI
04205	737	7	0.000253	0.001614	MZIMBA	MPHONGO
04211	564	7	0.000193	0.001411	MZIMBA	MTELEMUKA
04212	850	7	0.000291	0.001757	MZIMBA	MTENDE
04213	783	7	0.000292	0.001999	MZIMBA	MTENTHE
04215	737	7	0.000253	0.002171	MZIMBA	MTUZUZU
04238		7	0.000292	0.002132	MZIMBA	PHUTULA
04254	821	6	0.000292	0.001849	MZIMBA	UNYOLO
04256	674	7	0.000231	0.001725	MZIMBA	VIBANGALALA
04259	398	7	0.000292	0.001702	MZIMBA	WATEREKA
04268	704	7	0.000241	0.001971	MZIMBA	KAVUULA
04271	885	7	0.000303	0.001178	MZIMBA	EMBOMBENI
04281		7	0.000292	0.001899	MZIMBA	CHANG'ANGA
04284	656	7	0.000225	0.002573	MZIMBA	MAKAZI
04285	298	6	0.000292	0.001319	MZIMBA	CHIKUKA
04297		7	0.000292	0.002346	MZIMBA	MTANGATANGA
04310	0	7	0.000292	0.001540	MZIMBA	MUKANAUMOZA
04313	0	7	0.000292	0.002813	MZIMBA	KAULIRA
04316	130	7	0.000292	0.001697	MZIMBA	KAMUWOLI J.P
04319		7	0.000292	0.002034	MZIMBA	KAMUZOMELA
04320	312	7	0.000292	0.001423	MZIMBA	KAMPHAMBE
04324	291	7	0.000292	0.002100	MZIMBA	MANGOMBERA
04327	485	7	0.000292	0.003851	MZIMBA	LUMEMO J.P.
05007	670	7	0.000230	0.001935	MZUZU CITY	LUPASO
05011	1186	6	0.000407	0.001921	MZUZU CITY	MSONGWE
05014	2109	7	0.000723	0.002383	MZUZU CITY	CHIPUTULA
05016	1239	7	0.000425	0.002141	MZUZU CITY	ST AUGUSTINE
05017	2462	7	0.000844	0.001922	MZUZU CITY	ST PETERS
06015	926	7	0.000317	0.001687	NKHATA BAY	CHIKWINA
06017	716	7	0.000245	0.002539	NKHATA BAY	CHIWISI
06023	444	7	0.000152	0.001961	NKHATA BAY	CHIPUZUMUMBA
06042	435	7	0.000149	0.001722	NKHATA BAY	KASANGAZI
06049	599	7	0.000205	0.001809	NKHATA BAY	KUWIRWI
06059	837	7	0.000287	0.001499	NKHATA BAY	MAULA
06066	1219	7	0.000418	0.002499	NKHATA BAY	MPAMBA
06084	802	7	0.000275	0.001558	NKHATA BAY	SANGA
06087	702	7	0.000241	0.001383	NKHATA BAY	ST PETERS
07019	418	7	0.000143	0.001558	KASUNGU	CHIDAOLA
07041		7	0.000292	0.001114	KASUNGU	MWEZI
07042	1044	7	0.000358	0.001817	KASUNGU	CHITENJE
07046	1154	7	0.000396	0.001722	KASUNGU	CHULU
07049	816	7	0.000280	0.002457	KASUNGU	DWANKHWALI
07058	593	7	0.000203	0.002196	KASUNGU	KALIMANYUNGU
07065	1319	7	0.000452	0.002241	KASUNGU	KAMCHOCHO
07066	1038	7	0.000292	0.001672	KASUNGU	KAMDIDI
07070	1458	6	0.000500	0.001679	KASUNGU	KAMUZU
07079	305	7	0.000105	0.001457	KASUNGU	KASIKIDZI
07083	862	6	0.000296	0.001326	KASUNGU	KATOZI

07087	512	7	0.000176	0.002016	KASUNGU	KAVIKULA
07090	1261	7	0.000432	0.002174	KASUNGU	KAWIYA
07095	1280	7	0.000439	0.002171	KASUNGU	LINYANGWA
07097	1294	7	0.000444	0.001975	KASUNGU	LISASADZI
07101	636	7	0.000218	0.001549	KASUNGU	LONGWE
07106	1076	7	0.000369	0.001689	KASUNGU	MANKHAKA
07107	1501	7	0.000515	0.002135	KASUNGU	MANYANI
07111	779	7	0.000267	0.001497	KASUNGU	MILENJE
07114	671	7	0.000230	0.001895	KASUNGU	MKANAKHOTI
07125	855	7	0.000293	0.001862	KASUNGU	NKHAMENYA BOYS
07126	712	7	0.000244	0.002490	KASUNGU	NKHAMENYA GIRLS
07149	1038	7	0.000356	0.001979	KASUNGU	ZYALAMBE
07168	410	7	0.000141	0.001098	KASUNGU	KAVUMBIRI
07183	395	6	0.000135	0.000941	KASUNGU	CHIKOKO J.P.
07186	2638	7	0.000904	0.002377	KASUNGU	MTUNTHAMA
07191	417	7	0.000292	0.003284	KASUNGU	DWALALA
08003	871	7	0.000299	0.001951	NKHOTAKOTA	CHAMALIRE
08007	561	7	0.000192	0.002236	NKHOTAKOTA	CHIGUMUKIRE
08012	1002	7	0.000343	0.002898	NKHOTAKOTA	CHIONDO
08016	1211	7	0.000415	0.002291	NKHOTAKOTA	CHISOTI
08018	812	7	0.000278	0.001787	NKHOTAKOTA	CHOMBO
08019	961	7	0.000329	0.002390	NKHOTAKOTA	CHONGOLE
08021	1348	7	0.000462	0.002982	NKHOTAKOTA	DWAMBAZI
08026	1305	6	0.000447	0.001822	NKHOTAKOTA	KANYENDA
08028	915	7	0.000314	0.001768	NKHOTAKOTA	KASAMBA
08038	536	7	0.000184	0.001977	NKHOTAKOTA	LIUDZI
08062	417	6	0.000143	0.001432	NKHOTAKOTA	MWALAWATONGOLE
08070	1650	7	0.000566	0.001888	NKHOTAKOTA	ST PAULS
08072	665	7	0.000228	0.001706	NKHOTAKOTA	WALEMERA
08095	289	7	0.000292	0.002448	NKHOTAKOTA	WOZI J.P.
08098		7	0.000292	0.001399	NKHOTAKOTA	KAUYE
09002	596	6	0.000204	0.001529	MCHINJI	BENJAMIN
09003	3010	7	0.001032	0.002428	MCHINJI	BUA
09007	379	6	0.000130	0.001825	MCHINJI	CHIMKOMENI
09012	1645	7	0.000564	0.002317	MCHINJI	CHIMTEKA
09013	1085	7	0.000372	0.001736	MCHINJI	CHIMWAMKANGO
09022	949	7	0.000325	0.002099	MCHINJI	GUILLEME GIRLS'
09027	854	7	0.000293	0.001865	MCHINJI	KAFULAMA
09034	2414	7	0.000828	0.002652	MCHINJI	KAMWENDO
09049	966	7	0.000331	0.001329	MCHINJI	LUDZI BOYS'
09053	1076	7	0.000369	0.002668	MCHINJI	MATUAMBA
09057	1165	7	0.000292	0.001667	MCHINJI	MCHINJI
09066	650	7	0.000223	0.001930	MCHINJI	MSUKWALA
09077	1116	7	0.000383	0.001600	MCHINJI	NYALUBWE
09079	441	6	0.000151	0.001002	MCHINJI	PITALA
09082	749	7	0.000257	0.002059	MCHINJI	SITORO
10005	689	7	0.000236	0.000986	DOWA	BWEYA
10015	1418	7	0.000486	0.001977	DOWA	CHIDOTHI
10018	664	7	0.000228	0.002499	DOWA	CHIKHOBWE
10020	1416	7	0.000485	0.001650	DOWA	CHIMBALU
10021	900	7	0.000309	0.001567	DOWA	CHIMBULI
10022	1019	7	0.000349	0.001577	DOWA	CHIMKOKA

10025	881	7	0.000302	0.001603	DOWA	CHIMUNGU
10027	1037	7	0.000355	0.002032	DOWA	CHING'AMBA
10030	683	7	0.000234	0.001691	DOWA	CHITAKADZI
10040	972	7	0.000333	0.002374	DOWA	JAMES
10046	811	6	0.000278	0.001038	DOWA	KAFULU
10049	255	7	0.000087	0.002032	DOWA	KALEWA
10059	745	7	0.000255	0.001754	DOWA	KAPATAMOYO
10064	1724	6	0.000591	0.002008	DOWA	KATALIMA
10077	1306	7	0.000448	0.002444	DOWA	MADISI
10079	0	7	0.000292	0.001816	DOWA	MANONDO
10100	1013	7	0.000347	0.001793	DOWA	MPONELA 1
10103	3120	7	0.001070	0.003257	DOWA	MTANIRA
10107	1720	7	0.000590	0.002178	DOWA	MVERA D
10126	307	7	0.000105	0.001948	DOWA	THIWI
11020	257	7	0.000088	0.001315	SALIMA	CHITIPI
11024	451	7	0.000155	0.001808	SALIMA	KACHULU
11042	1264	7	0.000433	0.002145	SALIMA	LIFUWU
11072	589	6	0.000202	0.001835	SALIMA	NTHUMBO
11075	2348	7	0.000805	0.002277	SALIMA	SALIMA L E A
11094	111	7	0.000038	0.001391	SALIMA	TSAWALA J.P.
11098		7	0.000292	0.001557	SALIMA	MSAMBAMFUMU
12006	4141	7	0.001420	0.004173	LILONGWE U	CHILINDE
12008	2052	7	0.000703	0.001926	LILONGWE U	CHINKHUTI
12010	3011	7	0.001032	0.002137	LILONGWE U	CHIWOKO
12012	3568	7	0.001223	0.002672	LILONGWE U	KAFULU
12015	2090	7	0.000716	0.002370	LILONGWE U	LILONGWE BOYS
12019	2097	7	0.000719	0.001957	LILONGWE U	LILONGWE L E A
12020	1053	7	0.000361	0.001712	LILONGWE U	LIVIMBO
12021	2744	7	0.000941	0.002746	LILONGWE U	M'BINZI
12023	2873	7	0.000985	0.002783	LILONGWE U	MKOMACHI
12053	2526	7	0.000866	0.002818	LILONGWE U	CHINSAPO
12054	4643	7	0.001592	0.003060	LILONGWE U	KALIYEKA
13006	260	7	0.000292	0.001837	LILONGWE R	BUWA
13012	290	6	0.000099	0.001489	LILONGWE R	CHALUSA
13014	1833	7	0.000628	0.001576	LILONGWE R	CHANKHANDWE
13022	314	7	0.000108	0.002067	LILONGWE R	CHAZOZOMA
13034	846	7	0.000290	0.001833	LILONGWE R	CHILIKHANDA
13038	1051	6	0.000360	0.001606	LILONGWE R	CHIMANAZO
13039	599	7	0.000205	0.001926	LILONGWE R	CHIMBALANGA
13058	2467	7	0.000846	0.002365	LILONGWE R	CHITEDZE
13064	723	7	0.000248	0.000921	LILONGWE R	CHIWIRI
13072	929	7	0.000318	0.002053	LILONGWE R	DZAMA
13074	392	7	0.000134	0.001566	LILONGWE R	DZUWA
13078	880	7	0.000302	0.001589	LILONGWE R	KACHULE
13079	416	7	0.000143	0.001217	LILONGWE R	KADYAULENDO
13081	2071	7	0.000710	0.002282	LILONGWE R	KAFUTWE-CHIZUMBA
13091	1157	7	0.000397	0.002989	LILONGWE R	KAMBIRA
13093	549	7	0.000188	0.001644	LILONGWE R	KAMBWA
13094	266	6	0.000292	0.001109	LILONGWE R	CHIKULI
13095	815	7	0.000279	0.001835	LILONGWE R	KAMGANGA
13097	1329	7	0.000456	0.002076	LILONGWE R	KAMZIMBI
13102	691	7	0.000237	0.002016	LILONGWE R	KAPHALA

13112	1382	7	0.000474	0.002721	LILONGWE R	KASONJOLA
13116	1150	7	0.000394	0.002312	LILONGWE R	KAUNDAMA
13127	809	7	0.000277	0.002079	LILONGWE R	KUCHIPALA
13128	0	7	0.000292	0.002527	LILONGWE R	KUMAANI
13135	1544	7	0.000529	0.002238	LILONGWE R	MAKANYA
13137	1029	7	0.000353	0.002149	LILONGWE R	MALEMBO
13142	3032	7	0.001039	0.002396	LILONGWE R	MALINGUNDE
13147	1216	7	0.000417	0.001654	LILONGWE R	MATAPILA
13148	486	7	0.000167	0.002498	LILONGWE R	MATSIMBE
13149	1443	7	0.000495	0.002129	LILONGWE R	MATUNDULUZI
13152	680	7	0.000233	0.002263	LILONGWE R	MBABVU
13170	1126	6	0.000386	0.002719	LILONGWE R	MKANDA
13176		7	0.000292	0.003012	LILONGWE R	MLALE SEMINARY
13180	929	6	0.000318	0.001873	LILONGWE R	MLOMBWA
13199	1789	7	0.000613	0.001760	LILONGWE R	MSELERA
13207	1358	7	0.000466	0.001864	LILONGWE R	MTEMBE
13212	344	7	0.000118	0.001918	LILONGWE R	MWAKHUNDI
13214	653	7	0.000224	0.004504	LILONGWE R	MWASE
13223	1268	7	0.000435	0.002498	LILONGWE R	NAMITETE
13225	772	7	0.000265	0.002213	LILONGWE R	NANKHONDE
13226	648	7	0.000222	0.001887	LILONGWE R	NASALA
13228	1357	7	0.000465	0.001338	LILONGWE R	NDAULA
13247	1217	7	0.000417	0.001893	LILONGWE R	SINUMBE
13257	280	7	0.000096	0.001544	LILONGWE R	YEPA
13266	273	7	0.000292	0.002323	LILONGWE R	ZOKOTO
13269	547	7	0.000188	0.001412	LILONGWE R	GUMBI
13275	296	7	0.000101	0.001157	LILONGWE R	DZUNDI
13285	544	7	0.000186	0.001282	LILONGWE R	MBOBO J.P.
13291	582	6	0.000200	0.001194	LILONGWE R	NTHULU J.P.
13295	351	7	0.000292	0.002323	LILONGWE R	MILOMBWA J.P.
14008		7	0.000292	0.001678	DEDZA	CHIKUTA
14014		7	0.000292	0.001738	DEDZA	CHITHIMBA
14020		7	0.000292	0.002305	DEDZA	CHIPALUKWA
14027	476	7	0.000163	0.001660	DEDZA	CHULE
14029	1539	7	0.000528	0.002180	DEDZA	DEDZA L E A
14034	870	7	0.000298	0.001782	DEDZA	GUNDADZUWA
14046	1047	7	0.000359	0.002115	DEDZA	KAMTEDZA
14047	564	6	0.000193	0.001300	DEDZA	KANAMA
14069	1113	7	0.000382	0.002082	DEDZA	MADZUMBI
14072	1618	7	0.000555	0.001952	DEDZA	MAKOTA
14077	115	7	0.000039	0.001590	DEDZA	MAMTEGA
14085		6	0.000292	0.001993	DEDZA	MBIRIMA
14098	797	6	0.000273	0.001336	DEDZA	MLUNDUNI
14102		7	0.000292	0.001711	DEDZA	MSEKENI
14110	1064	7	0.000365	0.002055	DEDZA	MTEMWENDE
14122	439	7	0.000150	0.001706	DEDZA	NANKHANDE
14132	1379	7	0.000473	0.001810	DEDZA	THETE
14142	832	7	0.000285	0.002043	DEDZA	TCHETSA
14152	239	7	0.000082	0.001583	DEDZA	GOWOKE
14173		7	0.000292	0.002112	DEDZA	FUMBA
15003	749	7	0.000257	0.001984	NTCHISI	CHAKULANJALA
15014	880	7	0.000302	0.001337	NTCHISI	CHIKHOTA

15022	561	6	0.000192	0.001076	NTCHISI	GWANGWA
15029	826	7	0.000283	0.001489	NTCHISI	KANYULUNYULU
15030	443	7	0.000152	0.001592	NTCHISI	KAOMBE
15032	907	6	0.000311	0.001688	NTCHISI	KASAKULA
15033	519	6	0.000178	0.001463	NTCHISI	KAULATSITSI
15045	820	7	0.000281	0.002379	NTCHISI	MPALO
15072	209	7	0.000072	0.001730	NTCHISI	CHISANJA
15092	114	6	0.000039	0.000929	NTCHISI	VIKULA
15095	177	7	0.000061	0.001639	NTCHISI	NDENDERE
15104		7	0.000292	0.001498	NTCHISI	PHUYE II
15105		6	0.000292	0.001482	NTCHISI	CHAMDULU
15106		7	0.000292	0.002365	NTCHISI	MPHEZA
16002	1260	7	0.000432	0.001992	NTCHEU	BANGALA
16007	1301	7	0.000446	0.002030	NTCHEU	BILILA
16018	635	7	0.000218	0.001500	NTCHEU	CHIKUSE
16021	1873	7	0.000642	0.003050	NTCHEU	CHIOLE
16024	1065	7	0.000365	0.001746	NTCHEU	CHIPUSILE
16025	651	6	0.000223	0.000971	NTCHEU	CHITALE
16036	1948	7	0.000668	0.002343	NTCHEU	GUMBU
16040	1463	7	0.000502	0.001954	NTCHEU	KABWAZI
16049	1009	7	0.000346	0.001660	NTCHEU	KAMPANJE
16055	1632	7	0.000559	0.002271	NTCHEU	KATSEKERA
16062	462	7	0.000158	0.002154	NTCHEU	LAKE-VIEW
16065	930	7	0.000319	0.001797	NTCHEU	LIVALO
16067	1865	7	0.000639	0.002291	NTCHEU	LIZULU
16069	797	6	0.000273	0.002367	NTCHEU	LUDOVIKO
16075	315	7	0.000108	0.001639	NTCHEU	MALUZA
16080	948	7	0.000325	0.001547	NTCHEU	MATANDA
16083	1488	7	0.000510	0.001631	NTCHEU	MDEKA
16087	1356	7	0.000465	0.001828	NTCHEU	MLANDA
16093	1164	7	0.000399	0.002259	NTCHEU	MULUMA
16097	1162	6	0.000398	0.001264	NTCHEU	MWALAWOYERA
16106		6	0.000292	0.001925	NTCHEU	NSIYALUDZU
16120	1284	7	0.000440	0.001784	NTCHEU	TAMBALA
16122	878	7	0.000301	0.002588	NTCHEU	THAMBO
16129		7	0.000292	0.003185	NTCHEU	CHAMPITI
16164	187	7	0.000064	0.002299	NTCHEU	KAMPHEKO
16179		6	0.000292	0.001539	NTCHEU	PHEZA
16180		7	0.000292	0.001558	NTCHEU	MALIMWE
16184	450	7	0.000292	0.001748	NTCHEU	NKHONGWE
16193		7	0.000292	0.001777	NTCHEU	HAYISI
16207		7	0.000292	0.002711	NTCHEU	CHAUMBWI
16210		7	0.000292	0.002034	NTCHEU	JAMES
16218		7	0.000292	0.003280	NTCHEU	KHOLA 11
17002	1225	6	0.000420	0.001633	MANGOCHI	CHANGALI
17004	995	7	0.000341	0.001835	MANGOCHI	CHANTULO
17012	1091	7	0.000374	0.001264	MANGOCHI	CHIMVUU I
17016	1675	7	0.000574	0.001852	MANGOCHI	CHINGWENYA
17017	1131	7	0.000388	0.001936	MANGOCHI	CHIPALAMAWAMBA
17025	954	7	0.000327	0.002178	MANGOCHI	CHIWAULA MUSLIMS
17026	1127	7	0.000386	0.001851	MANGOCHI	CHOWE
17028	492	7	0.000169	0.002854	MANGOCHI	IPANI

83

17041	2400	7	0.000823	0.002179	MANGOCHI	KOCHE MODEL
17059	701	7	0.000240	0.002218	MANGOCHI	MAGANGA
17066	849	6	0.000291	0.001780	MANGOCHI	MALENGA
17073	882	7	0.000302	0.001203	MANGOCHI	MANGOMBO
17076	1344	7	0.000461	0.003111	MANGOCHI	MASUKU
17090	1159	7	0.000397	0.003254	MANGOCHI	MKOPE HILL
17095	1638	7	0.000562	0.002214	MANGOCHI	MPILIPILI
17096	1918	6	0.000658	0.001965	MANGOCHI	MPINGANJIRA
17097		7	0.000292	0.001695	MANGOCHI	MPITILIRA
17098	3316	7	0.001137	0.003030	MANGOCHI	MPONDAS
17108	1513	7	0.000519	0.002216	MANGOCHI	MTUWA
17110	593	7	0.000203	0.001278	MANGOCHI	NAKUNDU
17113	1110	7	0.000381	0.001870	MANGOCHI	NAMISI
17114	1128	7	0.000387	0.002060	MANGOCHI	NAMPINGUJA
17127	1237	7	0.000424	0.001582	MANGOCHI	ST JOSEPH
17148	222	7	0.000076	0.001161	MANGOCHI	MPAMILA J.P
18009	2182	7	0.000748	0.002628	MACHINGA	CHIKWEO
18010	1254	7	0.000430	0.002922	MACHINGA	CHILANGA
18021	711	7	0.000244	0.001751	MACHINGA	KALAMBO I
18025	1190	6	0.000408	0.002089	MACHINGA	KAOMBE
18027	1246	7	0.000427	0.002351	MACHINGA	KAWEYA
18028	838	7	0.000287	0.001408	MACHINGA	KAYUNI
18033	1162	7	0.000398	0.002221	MACHINGA	MACHINGA L E A
18036	1378	7	0.000472	0.002633	MACHINGA	MALUNDANI
18042	630	7	0.000216	0.001574	MACHINGA	MBAZA
18043	983	7	0.000337	0.001955	MACHINGA	MCHEMA
18059	866	7	0.000297	0.001834	MACHINGA	MPINIUMODZI
18060	2206	7	0.000756	0.002056	MACHINGA	MPIRI
18064	1021	7	0.000350	0.001756	MACHINGA	MSECHE
18074	1865	7	0.000639	0.001577	MACHINGA	NAMANOLO
18084	1746	7	0.000599	0.002211	MACHINGA	NKASAULO
18091	1144	7	0.000392	0.002023	MACHINGA	SOSOLA
18095	1543	7	0.000529	0.002346	MACHINGA	ULONGWE I
18100	1068	7	0.000366	0.002517	MACHINGA	VETHIWA
18101	983	7	0.000337	0.002055	MACHINGA	WATAKA
18104	757	6	0.000292	0.001787	MACHINGA	MISU
18105	746	7	0.000256	0.001626	MACHINGA	MITENGWE
18126	447	7	0.000153	0.001322	MACHINGA	NANJOKA
18128	409	7	0.000140	0.001993	MACHINGA	NJERENJE JUNIOR
18148	424	7	0.000145	0.001469	MACHINGA	CHINKHWAKWA
18151	1176	7	0.000403	0.002181	MACHINGA	CHIKOLEKA
18174		7	0.000292	0.003268	MACHINGA	KANGANKUNDI
19006	1478	7	0.000507	0.002173	ZOMBA URBA	SACRED HEART
20008	2509	7	0.000860	0.003157	ZOMBA RURA	CHIKUPIRA
20020		6	0.000292	0.001951	ZOMBA RURA	CHISONI
20025		7	0.000292	0.001610	ZOMBA RURA	JENALA
20034	973	7	0.000334	0.001848	ZOMBA RURA	LIFANI
20038	1149	7	0.000394	0.001468	ZOMBA RURA	MACHERENI
20042	1405	7	0.000482	0.001988	ZOMBA RURA	MAKOKA
20043	923	7	0.000316	0.001816	ZOMBA RURA	MAKONGWA
20057	940	7	0.000322	0.001767	ZOMBA RURA	MILOLA
20060	765	7	0.000262	0.001149	ZOMBA RURA	MPACHIKA

20063	1137	7	0.000390	0.001958	ZOMBA RURA MPYUPYU
20065	489	7	0.000168	0.001404	ZOMBA RURA MUHASUWA
20071		6	0.000292	0.001107	ZOMBA RURA NAKAMBA
20077	1373	7	0.000471	0.002048	ZOMBA RURA NAMATAPA
20082	2799	7	0.000960	0.002167	ZOMBA RURA NAMIWAWA
20084	1234	6	0.000423	0.001746	ZOMBA RURA NANJIRI
20092	1917	7	0.000657	0.002257	ZOMBA RURA NSWASWA
20102		6	0.000292	0.001216	ZOMBA RURA SONGANI
20106	1570	7	0.000538	0.001948	ZOMBA RURA ST MACHAELS
20107	2212	7	0.000758	0.002752	ZOMBA RURA ST PAULS
20129	794	7	0.000272	0.001715	ZOMBA RURA NAMITSITSI
20132		7	0.000292	0.001544	ZOMBA RURA NYAMBWE
21004	2020	7	0.000692	0.002182	CHIRADZULU CHIKANGULU
21009	792	7	0.000272	0.001595	CHIRADZULU CHINGALA
21021	965	6	0.000331	0.001359	CHIRADZULU MAFE
21028	1340	7	0.000459	0.001609	CHIRADZULU MAPESI
21045	1538	7	0.000527	0.002586	CHIRADZULU NAMAKA
21052	1980	7	0.000679	0.003222	CHIRADZULU NDUNDE C C A P
21066	1102	7	0.000378	0.001547	CHIRADZULU THOMBOWE
22002	7112	7	0.002438	0.003891	BLANTYRE U BANGWE CATHOLIC
22006	1951	6	0.000669	0.002367	BLANTYRE U CHIGUMULA
22008	2092	6	0.000717	0.001835	BLANTYRE U CHILOMONI L E A
22010	4626	7	0.001586	0.003189	BLANTYRE U CHIMWANKHUNDA
22011	8946	7	0.003067	0.005235	BLANTYRE U CHIRIMBA
22014	1936	7	0.000664	0.001852	BLANTYRE U KANJEDZA
22015	1526	7	0.000523	0.001837	BLANTYRE U KAPENI DEMONSTRATION
22016	3377	7	0.001158	0.003398	BLANTYRE U LIKHUBULA
22018	4015	7	0.001376	0.002908	BLANTYRE U LIMBE GIRLS
22019	2330	7	0.000799	0.002714	BLANTYRE U LUMBIRA
22020	5099	7	0.001748	0.002957	BLANTYRE U MAKATA
22023	1645	7	0.000564	0.002581	BLANTYRE U MLAMBALALA
22028	3793	7	0.001300	0.002412	BLANTYRE U NAMALIMWE
22029	2204	7	0.000756	0.002637	BLANTYRE U NAMAME
22031	2466	7	0.000845	0.002429	BLANTYRE U NAMILANGO L E A
22032	3124	7	0.001071	0.002885	BLANTYRE U NAMIWAWA
22034	3131	6	0.001073	0.002088	BLANTYRE U NAYIZI
22035	6841	6	0.002345	0.003552	BLANTYRE U NDIRANDE
22046	3221	6	0.001104	0.002961	BLANTYRE U ZINGWANGWA L E A
22051	3002	7	0.001029	0.002432	BLANTYRE U MULUNGUZI
23001	1131	7	0.000388	0.001887	BLANTYRE R ANDISENI
23011	1342	7	0.000460	0.002158	BLANTYRE R CHIKULI
23016	248	6	0.000085	0.001213	BLANTYRE R CHILEKA S D A
23031	730	7	0.000250	0.001510	BLANTYRE R KACHERE I
23035	1030	7	0.000353	0.001937	BLANTYRE R KHOLA
23042	1940	7	0.000665	0.002594	BLANTYRE R LIRANGWE
23050		7	0.000292	0.001000	BLANTYRE R MALAKA
23066	1358	7	0.000466	0.002271	BLANTYRE R MITSIDI
23075	768	6	0.000263	0.001975	BLANTYRE R MPUMBE
23080		7	0.000292	0.002172	BLANTYRE R NAMALONDWE
23084		7	0.000292	0.002351	BLANTYRE R NAMITSITSI
23094	1031	7	0.000353	0.001690	BLANTYRE R NG'ONGA
23102	1576	7	0.000540	0.002426	BLANTYRE R NTENJERA

23116	197	7	0.000068	0.001510	BLANTYRE R	DZUNGWI J.P. SCHOOL
24014	343	6	0.000118	0.001844	THYOLO	CHIPUNGU
24036	1086	6	0.000372	0.001538	THYOLO	KHAVE
24037	1665	7	0.000571	0.001802	THYOLO	KHONJENI
24038	912	7	0.000313	0.002059	THYOLO	KONZALENDO
24055	732	7	0.000251	0.000957	THYOLO	MALOSA
24060	1133	6	0.000388	0.001497	THYOLO	MAONGA
24081	2233	7	0.000765	0.002325	THYOLO	MITENGO
24098	1682	6	0.000577	0.002332	THYOLO	NAKAMBWE
24102	1077	7	0.000369	0.002682	THYOLO	NAMILEME
24104		6	0.000292	0.001456	THYOLO	NAMITETE
24122	1285	7	0.000441	0.002466	THYOLO	NKHWANGWA
24132	2216	7	0.000760	0.001973	THYOLO	NYAMBALO
24141	2039	7	0.000699	0.002395	THYOLO	THYOLO
24142	538	7	0.000184	0.001844	THYOLO	WILSON
24164	586	7	0.000201	0.002017	THYOLO	MPENDA
24167	1662	7	0.000570	0.002265	THYOLO	MATHIYA
25002	775	7	0.000266	0.002917	MULANJE	BONA
25003	2099	7	0.000720	0.002496	MULANJE	CHABUKA
25007	2159	7	0.000740	0.002673	MULANJE	CHINGOLI
25008	567	7	0.000194	0.001544	MULANJE	CHIGOMBE
25015	820	7	0.000281	0.001254	MULANJE	CHINJIKA
25016	1910	6	0.000655	0.001882	MULANJE	CHINYAMA
25018	788	7	0.000270	0.001232	MULANJE	CHISAMBO
25021	1196	7	0.000410	0.003184	MULANJE	DZENJE I
25022	953	7	0.000327	0.001847	MULANJE	DZENJE II
25025	1981	7	0.000679	0.001580	MULANJE	KADUYA
25032	1504	7	0.000516	0.002372	MULANJE	KHAYA
25036	2445	7	0.000838	0.002344	MULANJE	LAUDERDALE
25037	1398	7	0.000479	0.001492	MULANJE	LIHAKA
25039	1396	7	0.000479	0.002156	MULANJE	LIKANANI
25056	2463	7	0.000844	0.002329	MULANJE	MASUBI
25057	632	7	0.000217	0.002342	MULANJE	MAUNI
25059	721	6	0.000247	0.000901	MULANJE	MBEMBEMBE
25066	2071	6	0.000710	0.001735	MULANJE	MILEME
25071	1009	7	0.000346	0.002250	MULANJE	MISANJO
25072	1695	6	0.000581	0.003011	MULANJE	MISHONI
25075	2096	7	0.000719	0.002043	MULANJE	MKANDA
25076	2497	7	0.000292	0.001569	MULANJE	MKHUMBA
25084	1087	7	0.000373	0.001523	MULANJE	MULOMBA
25089	738	7	0.000253	0.001712	MULANJE	NACHIWALA
25091	1044	7	0.000358	0.002038	MULANJE	NAKHWIKWI
25094	1771	7	0.000607	0.002177	MULANJE	MBUNUMBU
25095	1502	7	0.000515	0.003403	MULANJE	NALUSO
25099	1482	7	0.000508	0.002902	MULANJE	NAMBITI
25101	1144	7	0.000392	0.001908	MULANJE	NAMINDORA
25104	1362	7	0.000467	0.001999	MULANJE	NAMULENGA GIRLS
25108	868	7	0.000298	0.002677	MULANJE	NASIYAYA
25110	1159	7	0.000397	0.001566	MULANJE	NDANGA
25113	2689	7	0.000922	0.002407	MULANJE	NJEDZA
25117	1359	6	0.000466	0.001756	MULANJE	NTATA
25133	2434	7	0.000834	0.002731	MULANJE	SATEMWA

25136	1378	7	0.000472	0.002236	MULANJE	SUPUNI
25171	1223	7	0.000419	0.001897	MULANJE	MTEPUWA
25182		7	0.000292	0.001727	MULANJE	NDINDI
26013	1080	7	0.000370	0.002313	MWANZA	KALANGA
26015	1178	7	0.000404	0.001731	MWANZA	KHOLOMBIDZO
26016	536	7	0.000184	0.002328	MWANZA	MUWALE
26022	318	7	0.000109	0.001506	MWANZA	MATANDANI
26034	256	7	0.000088	0.001388	MWANZA	MWANZA
26040	635	6	0.000218	0.001419	MWANZA	MPANDADZI
26045	1108	7	0.000380	0.001963	MWANZA	TSUPE
26048		7	0.000292	0.001947	MWANZA	CHITHUMBWI
26054	157	7	0.000054	0.001574	MWANZA	NSENJELE
27003	1321	7	0.000453	0.002098	CHIKWAWA	CHANGA
27009	1582	6	0.000542	0.001924	CHIKWAWA	CHIKWAWA
27017	1059	7	0.000363	0.002942	CHIKWAWA	GOMA
27021	440	7	0.000151	0.001819	CHIKWAWA	KAMPOMO
27033	3377	7	0.001158	0.003043	CHIKWAWA	MAKANDE
27035	469	7	0.000161	0.001560	CHIKWAWA	MANDRADE
27038	524	7	0.000180	0.001543	CHIKWAWA	MASANDUKO
27042	1718	6	0.000589	0.003595	CHIKWAWA	MFERA
27045	2160	7	0.000740	0.004214	CHIKWAWA	MITONDO
27053	1510	7	0.000518	0.003138	CHIKWAWA	MWANZA
27063	690	7	0.000237	0.001902	CHIKWAWA	NDAKWERA
27068	1926	6	0.000660	0.001973	CHIKWAWA	NKHWANGWA
27069	921	7	0.000316	0.002029	CHIKWAWA	MKUMANIZA
27077	1933	7	0.000663	0.002146	CHIKWAWA	PHANDA
27078	1556	7	0.000533	0.002646	CHIKWAWA	PHIMBI
27085	1231	7	0.000292	0.002110	CHIKWAWA	THABWANI
27086	721	6	0.000247	0.001686	CHIKWAWA	THENDO
27091	787	7	0.000270	0.001883	CHIKWAWA	OLE-OLE
28014	1164	7	0.000399	0.001509	NSANJE	KHULUBVI
28023	736	7	0.000252	0.002097	NSANJE	MANKHOKWE
28033	281	7	0.000096	0.003231	NSANJE	MPISAMANJA
28038	906	7	0.000311	0.001756	NSANJE	NAMIYALA
28048	1137	7	0.000390	0.002039	NSANJE	NYAMITHUTHU