- PN-ABC-726

STEP

System for Tracking Educational Progress

User's Guide

Developed by
Research Triangle Institute
with
Harvard University

B • R • I • D • G • E • S

Basic Research and Implementation in Developing Educational Systems

sponsored by
Office of Education
Bureau of Science and Technology
U.S. Agency for International Development

A USER'S GUIDE TO THE SYSTEM FOR TRACKING EDUCATIONAL PROGRESS

(STEP)

• AUGUST 1988 •

Developed by:
Luis Crouch and Scott Moreland
Research Triangle Institute

with:
Ernesto Cuadra and Noel McGinn
Harvard University

BRIDGES Project

Sponsored by the Office of Education, Bureau for Science and Technology,
and
Office of Development Resources, Bureau for Latin America and the Caribbean
United States Agency for International Development

The Basic Research and Implementation in DevelopinG Education Systems (BRIDGES) Project is directed by Harvard Institute for International Development and the Harvard Institute for International Development and the Harvard Graduate School of Education under Cooperative Agreement No. DPE-5824-A-5076 with the Office of Education, USAID.

TABLE OF CONTENTS

CHAPTER 1. INTRODUCTION TO THE STEP SYSTEM 1

Introduction to the STEP System 1

STEP Modules 1

Using STEP on the Microcomputer 1
Hardware Environment 1
Software Environment 2

Organization of this Manual 2

CHAPTER 2. USING STEP 3

Interactive vs. Batch Modes 3

File Management Under STEP 4

STEP Modules: POP 5
Description of the POP Module. 5

Use of the POP module 6
Specifying rates and initial values 6
Specifying mortality 7
Specifying International and Internal Migration 8
Base Year Population 8
Sex Ratio 9
Index of Females 9

STEP Modules: ESTIMATE 9
Description of the ESTIMATE Module 9
Use of the ESTIMATE Module 11

STEP Modules: PROJRATE 13
Description of the PROJRATE Module 13
Use of the PROJRATE Module 15

STEP Modules: EDPROJ 16
Description of the EDPROJ Module 16
Use of the EDPROJ Module 17

STEP Modules: COST 18
Description of the COST module 18
Use of the COST Module 19

STEP Modules: COHORT 20
Description of the COHORT Module 20
Use of the COHORT Module 20

Running the STEP System in Batch Mode 22

CHAPTER 3. CONFIGURING STEP TO A NEW APPLICATION 25

Step 1: Setting the Current Maximum Levels of Dimensions 25
Dimensions Used by STEP Variables in the STEP Data Base..... 26
Dimensions Used by POP Variables in the POP Data Base..... 26

Step 2: Defining and Editing the Label Library 27

Step 3: Entering New Data for the New Application 27

Step 4: Defining New Tables and Graphs 27

Variables in the STEP Data Files 28

Variables in the POP Data Files 29

Appendix to Chapter 3 31
Label Format Definitions 31

CHAPTER 4. TECHNICAL DESCRIPTION OF THE MODEL METHODOLOGY 39

The Grade-transition Enrollment Projection Method 39
Projections with Fixed Parameters 40
Projections with Variable Parameters 40
Linear Trend 40
Logistic Curve Fitting 40
Target Setting Method 41

Efficiency Indicators Based on Cohort Analysis 41
Basic Concept 41
Efficiency and Other Educational Indicators 42
Completion or Graduation Rate 42
"On Time" Graduates 42
Student/Years per Graduate 43
Input-output Ratio 43
Years Spent in School 43

Indicators Based on Enrollment Data 43
Gross Enrollment Ratio 43
Net Enrollment Ratio 43

Estimation of Transition Rates 44
Standard Method 44
Schiefelbein Method - Description 45
Schiefelbein Method - Example of Calculations 49

Projections of Resource Requirements and Costs 50

Projecting Population 52

Appendix to Chapter 4 55
Glossary of Variables and Symbols 55

CHAPTER 1. INTRODUCTION TO THE STEP SYSTEM

Introduction to the STEP System

The System for Tracking Educational Progress (STEP)[1] is designed to allow education planners and policy makers to keep track of and to project a set of basic educational indicators of an education system with an emphasis on the primary level. With it, based on a time series of enrollment data, and population data, the user can:

- estimate the transition rates within a primary school system;
- project those rates into the future using various approaches;
- project annual population by single-year age groups;
- project student enrollment into the future;
- project resource requirements, teacher demand and supply, and recurrent and capital costs; and
- calculate various educational efficiency measures.

The projection methodologies used in STEP have been used for several years in developing country settings. Those methodologies are described in more detail under the Technical Description section of this document. In this section we describe the implementation of the system on the microcomputer and how to use it.

STEP Modules

STEP uses a modular approach. This gives the user the flexibility to run modules independently and/or to concentrate on those modules that are of particular interest. Each of the six modules of STEP corresponds to a set of related tasks. The module ESTIMATE estimates the historical education transition rates for entry into the first grade of primary school, the repetition rates and the drop out rates; the module PROJRATE projects those rates forward for a specified time period; the module EDPROJ projects school enrollment; the module COHORT estimates various educational efficiency measures; the module POP allows the user to project annual population by sex and single-year age groups; and the module COST projects teacher supply and demand by level of experience and qualification, learning resource requirements, construction costs, teacher salary costs, and other recurrent and capital costs.

Using STEP on the Microcomputer

Hardware Environment

Because of the portability, access and ease of use offered by today's microcomputers, and because IBM or compatible equipment has become the de

---

1. Formerly the Primary Education Tracking System (PETS)

facto standard for social science analysis and simulation modeling, STEP has been implemented for the IBM/PC, IBM/XT or IBM/AT microcomputer and compatibles. Because of the large amount of data used by STEP and because of the software under which it is implemented the following configuration of equipment is necessary:

- IBM microcomputer or compatible;
- 256K of RAM;
- 3 MB of hard disk space;
- numeric coprocessor;
- printer (optional);
- color monitor (optional);
- graphics display adapter (optional).

Software Environment

STEP uses and produces a large amount of information and data. Two different software approaches were tried during the initial stages of development of STEP. The first one used the spreadsheet program Lotus 1-2-3. This has the advantage of being familiar to many microcomputer users and is fairly easy to use. It has the disadvantage of becoming awkward and even difficult to use when the application becomes complex, large, modular, or menu-driven.

The second approach, and the one that was chosen, was development of a set of Turbo PASCAL programs that runs under a software "shell" called **Host**. **Host** was specifically developed for modeling applications such as STEP. It was developed by the Research Triangle Institute with funds provided by USAID through the Integrated Population Development Planning (INPLAN) project and the BRIDGES project. It is as easy to use as Lotus 1-2-3, can manipulate large amounts of data, and can project faster than Lotus 1-2-3. It also has several features that make properly implemented models extremely flexible and adaptable to a wide variety of settings. An introduction to the facilities of **Host**, that it is necessary to know to use STEP, is given in a separate manual called "STEP by Steps".

Organization of this Manual

The rest of this manual is organized in three sections. Chapter 2 describes the use of the STEP modules. In Chapter 3, how STEP can be applied to a new application is described. Chapter 4 presents a complete technical description of the methodology underlying STEP.

CHAPTER 2. USING STEP

In running STEP there are three fundamental tasks that the user performs with the help of **Host**: data entry and editing, running modules, and viewing results. Each of these tasks can be done by different facilities available under **Host**. Thus the **Host** EDIT facility allows the user to enter data values into the data base and/or to change existing values. The **Host** RUN facility allows the user to execute a STEP module for a specified number of periods. The **Host** TABULATE facility allows one to view variables in the system in predefined tabular formats and to display those tables on the screen, the printer or to a disk file. Alternatively, the **Host** GRAPH facility allows one to view the variables in predefined graphs. Each of these facilities is described in more detail in the **Host User's Guide** and in the **STEP by Steps** Manual.

STEP has over thirty variables. Some of these variables or parameters must be supplied by the user, and others may be calculated by a STEP module as an output. Some outputs of one module may be used as inputs by another module. For example, repetition rates (REPEATRT) are calculated in ESTIMATE, projected forward in PROJRATE, and used as input to project enrollment in modules EDPROJ and COHORT. The population is projected in POP and used as an input in ESTIMATE and in EDPROJ. Enrollment is projected in EDPROJ and used as input in COST.

Interactive vs. Batch Modes

The STEP system can be run either interactively or in "batch" mode. In the interactive mode, the computer queries the user as to choices to be made in estimating key parameters, methods to be used for projecting them, etc. For example, the user might be queried as to whether he or she wants to use a linear or a logistic curve to project future repetition rates. This interactive mode is very useful the first few times one runs the STEP system. After one gets some experience, however, it becomes tedious to have to answer all these questions every time one runs the system. STEP will remember the answers given the previous time the system was used, and will give them back as default answers to each question during the current run.

In batch mode, the system runs automatically and reads the answers to the questions from the data file. They will have been stored the previous time the system was run in the variable called "BATCHOPT", and can be modified manually at any time using the editor. In batch mode no questions are asked, so the user should be sure that the settings contained in variable BATCHOPT are correct. The interactive mode is also useful if one wants to change a few of the answers. In this case, one simply hits ENTER in order to give the default (remembered) answer for all questions whose answers do not change, and enters a new value for the ones to be changed. Default answers are presented

in square brackets []. The use of Batch mode is described in more detail toward the end of this chapter in the section entitled, "Running the STEP System in Batch Mode".

In addition to the variable BATCHOPT, the variables POPFILE and EDFILE contain information regarding the transfer of population and enrollment data respectively from on STEP data file to another. These variables function similarly to BATCHOPT in interactive and batch mode.

File Management Under STEP

The STEP system uses two separate Host data bases. These are linked together and information is passed between them. A single data base would have been too large for most systems and would therefore be impractical.

While the Host software provides a facility for backing up and deleting data files, the STEP user might find it helpful to use DOS commands for file management. It is likely that the user may accumulate a number of data files and that he/she may wish to store these on floppy disks rather than on the hard disk. In this case the user should use the DOS "COPY" command to copy the files to a floppy disk and then after they are safely copied, use the DOS commands "DEL" or "ERASE" to delete the files from the hard disk.

Host data files are actually six inter-linked DOS files with the same name but different extensions. Therefore the use of the DOS "wild-card" character is recommended. In the following example a set of Host data files with the name FILE1, corresponding to a projection database called FILE1, is transferred to a diskette and then erased. We assume that the user is in the data subdirectory and that the target diskette is in drive A. To copy the files to the floppy disk, use

```
C>copy file1.* a:[Enter]
```

or for added safety, use this command with the "verify" option:

```
C>copy file1.* a: /v[Enter]
```

To erase from the hard disk, use

```
C>del file1.*[Enter]
```

Make sure that you have indeed copied to the floppy disk before you erase from the hard disk.

Eventually, you may wish to copy the files back from the floppy disk to the hard disk. Assuming that you are in the sub-directory which contains the data files, and that the diskette containing the archived files is in drive A, use

```
C>copy a:file1.*[Enter]
```

There are two other sets of files that should be backed up occasionally, but not necessarily as frequently as the data files. These are the LABEL files, the TABLE files and the GRAPH files. To manipulate these, follow the

same sets of commands as above, substituting "table", "label" and "graph" for "file1". Note that data files such as FILE1, FILE2, etc., share the same LABEL, TABLE and GRAPH files, so if you intend to continue working with the system, but with new data files (such as FILE2), do not delete your TABLE, LABEL and GRAPH files.

All STEP variables (input and output) are stored on disk in a file, called a projection file under the **Host** system. One of the most powerful features of the **Host** implemented STEP system is that it is possible to save a set of assumptions in a file and the results of a projection based on those assumptions. For example, the user could project the various education transition rates using a logistic curve in PROJRATE and then project school enrollment using EDPROJ. These rates and the resulting enrollment projections might be stored in a file called, for example, FILE1. Then the user might wish to project the rates using a linear trend in module PROJRATE and subsequently re-run module EDPROJ. In this case the user could store the second set of rates and the results in a second file called, for example, FILE2. The assumptions and results that were stored in FILE1 would still exist and it is even possible to produce tables which compare the two files.

Because POP uses so much data, it has its own database which is separate from the primary STEP database which is used by ESTIMATE, PROJRATE, EDPROJ, COHORT and COST. When ESTIMATE or EDPROJ, which use population data, are run, the user is asked (if in interactive mode), whether he or she wants to use the population data from a POP database, or population data already in the STEP database. If the user chooses to use data from a POP database, the computer will ask for the name of the POP database. So it is important that when the user intends to take population from a POP database, he or she knows the name of this database in advance. If these modules are run in the batch mode, the answers to these questions must be put into the batch options variable. (See the last section of this chapter, entitled "Running the STEP System in Batch Mode".)

The six modules that make up STEP are briefly described below and information is provided on their use. A description of the methodologies underlying these modules can be found in the Technical Description section. Their use is described in this section.

STEP Modules: POP

Description of the POP Module.

POP is a population projection module which provides the user, by default, with single-year age group population data on an annual basis. It can be used as a stand-alone population projection package, because it strikes a balance between ease of use, completeness of options, and completeness of capabilities. In the context of the STEP system, it is designed to provide the modules ESTIMATE and EDPROJ with the school age population.

The INPUT data for POP are the following:

INPUT data for POP module	
ASFR	Age specific fertility rates. Need to be specified only if User-Supplied ASFR option is chosen with the variable FERTOP.
EXTMIG	External migration, in absolute numbers.
FERTOPT	Fertility options.
FEMINDEX	Index for women.
LEXP	Life expectancy.
LT	Life table L values. Need to be specified only if user-supplied option is chosen in MORTOPT.
MIGRATE	Internal migration rates.
MODFERT	Model fertility schedules. Need to be specified only if user-supplied pattern is chosen with FERTOPT.
MORTOPT	Mortality options.
POP	Base-year single-age group population. Can be input directly or, if single-year data are not available, five-year age group data can be input into POP5, and then the program INT5T01 can be run to produce the variable POP.
SEXRATIO	Ratio of boys to girls at birth.
TFR	Total fertility rate. Needs to be specified only if User-Supplied ASFR was not chose with the variable FERTOPT.

The only output variable produced by POP which is used by other STEP modules is:

OUTPUT data from POP module	
POP	annual, single-year age group population data

However, other standard demographic indicators such as the crude birth and death rates, and the population growth rates are also reported.

Use of the POP module

Specifying rates and initial values

Fertility. There are two basic ways to specify fertility:

- enter projected age-specific fertility rates if available;
- enter projected total fertility rates and choose an age-pattern of fertility.

If the user chooses to enter total fertility rates, then an age pattern of fertility must be chosen. Currently there are three model patterns available, for Sub-Saharan Africa, Asia, and Arab countries, and there is room for a user-supplied pattern.

The fertility options are chosen by editing the variable FERTOPT. The choices presented are:

User ASFR;
Sub-Saharan Mod;
Arab Model;
Asian Model;
User Pattern.

The user chooses one of these options by entering a 1 by it and leaving zeroes elsewhere. If the user chooses one of the four patterns (Sub-Saharan Model, Arab Model, Asian Model, or User Pattern) POP assumes that the user has input a total fertility rate. In this case, the program uses the total fertility rate and the chosen pattern to produce the age-specific-rates as an output. If the user chooses the first option (User ASFR) then the program reads the age-specific rates as input and produces the total fertility rate as output.

After having told the program which option he/she has chosen, the user must input the actual rates, either by editing the variable ASFR (if he chose the User ASFR option of FERTOPT), or the variable TFR (if he chose one of the four pattern options in FERTOPT). These data must be input for every period the user wants to project.

Finally, if the user has chosen to use his/her own fertility pattern, this pattern must be entered into the variable MODFERT. It is assumed by POP that the pattern will be expressed as a series of percentages (adding to 100) of births occurring in each fertile age group (from 15-19 to 45-49), each corresponding to a total fertility rate. A maximum of six series of proportions is allowed, to correspond to six total fertility rates.

Specifying mortality

Mortality is specified by entering the projected life expectancies in the variable LEXP. They must be entered for every projection year. The user can choose which life table to use, however. The choice is made through the variable MORTOPT. Editing this variable will present the user with the choices of:

C-D North;
C-D South;
C-D West;
C-D East;
UN General;
UN Lat. America;
UN Far East;
UN So. Asia;
User-Supplied.

The user must specify one of these options by entering a 1 next to it, and zeroes elsewhere. The first four options represent the Coale-Demeny tables, the second four represent the UN tables, and the last option is for a user supplied set of L columns from a life table for a particular country, if available.

The standard life table data are not included as part of the **Host** database for POP. Instead, they are provided as binary DOS files. If one of the first eight options is chosen, the user should make sure that the relevant file is included in the directory where POP assumes the data files are.

If the user chooses to use his/her own life table L values, these must be entered by editing the variable LT for both males and females. (Currently these values must be entered by hand. A supporting program is being written to import these data from a Lotus 1-2-3 print file or any ASCII file corresponding to a certain format.) Note that the first row of the variable LT must contain the life expectancy at age zero to which the column of L values corresponds. Also note that the first age group is 0-4, not 0. The 0-4 group data for L is just the sum of the rows labeled 0 and 1 in most life tables.

Specifying International and Internal Migration

International migration data are entered into the variable EXTMIG. These data are entered as absolute numbers, negative if there is net out-migration, positive otherwise. The data must be entered for every projection period.

Internal migration is handled automatically when the current level of the dimension REGION is set equal to two. Model migration functions are used to calculate the internal migration rate (IMR) and the number of migrants by region and by age (INTMIG).

Base Year Population

The user must enter the base year population. There are two ways to do this. If single-year age group data are available and considered reliable, these can be entered into the first year of the variable POP, for both sexes. If the single-year data are unavailable or unreliable, then five-year group data can be entered into the variable POP5. In order to provide POP with a base-year single-year group population data, the support module INT5T01 must be run, for one period only.

INT5T01 is a general purpose age decomposition routine which uses "multipliers" to decompose variables given in five year age groups into single year age groups. The inputs for INT5T01 are given in the table below. Since most users of STEP will be using INT5T01 to decompose the population, the values of the control variables INT1NAME and INT5NAME should be set as follows:

INT1NAME : POP
INT5NAME : POP5

Also, the values of variable INT5OPT should be left at 1.

INPUT Data for INT5TO1	
POP5	population data in five year age groups
INT5OPT	options for population interpolation
INT1NAME	name of output variable for INT5TO1
INT5NAME	name of INPUT variable for INT5TO1

OUTPUT Data for INT5TO1	
POP	base year population in single year age groups

The base year for the population projection need not correspond to that of the base year in the STEP data base. That is, if the STEP data base has 1980 as the first year for which historic data are entered, the base year for the population data base may be 1980 or earlier. If a period later than 1980 is specified for the population data base Host/STEP will issue a fatal error message and the program will stop.

Sex Ratio

The sex ratio at birth must be specified by editing the variable SEXRATIO. Populate assumes this will not change over the projection period, and so this need be entered only once.

Index of Females

POP needs to know, for its internal calculations, whether all the variables will be arrayed in such a way that men will be listed first, and then women, or vice-versa. If men will always be listed first, the user needs to set the variable FEMINDEX to 2, and to 1 in the opposite case.

STEP Modules: ESTIMATE

Description of the ESTIMATE Module

The ESTIMATE module estimates repetition, dropout, and enrollment rates, using either the standard method based on reported repeater data or the Schiefelbein method. For this module the INPUT data are:

INPUT data for ESTIMATE module	
Transferred from POP or input by the user:	
POP	population by single-year age groups for the historic period (and for as many periods in the future as the user may wish to project)
Input by the user:	
HISTORIC	indicates whether a given year is historic (represented by an "h") or projected (represented by a "p")
ENTAGE	age of entrance to school system
The following information is necessary only when using the "Schiefelbein" method, or if age-specific enrollment rates for the historic period are desired:	
AGGENROL	enrollment in the aggregate (last) age group by grade
ENROLAGE	enrollment by grade and single-year age groups
The following information is necessary only when using the "standard" method:	
NREPEAT	number of repeaters by grade
ENROL	enrollment by grade - can be aggregated automatically from the variable ENROLAGE if data have already been entered there

The use of the variable ENROLAGE deserves special attention. It is important for the Schiefelbein method that the lowest age group in the first grade contains a number of students that is much lower than the next higher age group. Thus, for example, if the lowest age group is "6 and under" and contains 60,000 students, and the next is "7" and contains 70,000, as would be typical of many school systems, then the Schiefelbein method will not work properly. (The main consequence will be that the estimated rates of entrance into the school system from the population at large will be too low.) There are two ways to deal with this situation. First, the data in the "6 and under" category could be disaggregated into two or three categories: "6" and "5 and under", or "6", "5" and "4 and under" until one gets a category with a very small number in it (100 or 1000 times less than the next higher one). This first alternative is relatively difficult because it is not clear where to obtain this disaggregation. If the school system had the data presumably they could provide it. The second, easier alternative is to simply put in a category lower than the lowest one for which one has data, and to put only zeroes in it. This is the alternative we have used in the sample data in this document.

When the user chooses the Schiefelbein method, some additional information is required about the repetition rate in the last (usually 6th) grade, the number of dropouts, and the age distribution of enrollment. See below.

The primary OUTPUTS of this module are:

OUTPUT data from ESTIMATE module	
REPEATRT	repetition rate by grade
DROPRT	dropout rate by grade
ENTRT	enrollment rate of new students in first grade
ASER	age-specific enrollment rates
ENROL	enrollment by grade and sex
PROMRT	promotion rate by grade

Note that the outputs of this module are calculated only for the historic period. When using the Schiefelbein method, the estimated historical repetition rate for the sixth grade is simply the one assumed by the reader and entered into the computer at run-time. Both with the Schiefelbein and the standard methods, the estimated sixth grade dropout rate will be zero by assumption. Note that with the standard method, the estimated dropout rates will be meaningless for any year followed by a year with no enrollment data.

Use of the ESTIMATE Module

While executing ESTIMATE from the RUN facility of **Host**, the user must specify, as with any module, the number of periods to execute the module. In ESTIMATE, the user should specify a number that is equal to the number of historic periods. If you cannot recall the length of the historic period, specify one that is AT LEAST as large as the number of historic periods. That is, if there are five historic periods, the user should specify at least FIVE. Since no actual "projections" take place in this module, any periods beyond the historic period are ignored.

The last year of historic enrollment data should be the last year for which the indicator variable HISTORIC is set to "h", both with the Schiefelbein and the standard method.

When the ESTIMATE module is executed, as with all modules, the user is provided with an opening screen and with a screen that briefly describes the module.

Then he/she is asked whether the module is to be run interactively or in batch mode. If the module is to be run in batch mode, the user must have already entered his/her choices in the BATCHOPT and POPFILE variables, or else the values already there from the previous run will be used. (See section entitled "Running the STEP System in Batch Mode" at the end of this chapter.)

If the user chooses interactive mode a series of questions must be answered. Next, the user is asked whether he/she wishes to use "Standard method" which uses official data, or whether he/she wishes to use the method developed by Schiefelbein.

Since this module requires population data, the user will be asked whether he/she wishes to use the data already in the STEP database, or to bring in data generated by the POP module. If the user wishes to bring in data from the POP module, he/she will need to enter the name of the database containing this data, so it is important to recall this name before running the module. Also, when importing data from a POP database, the first period specified for the ESTIMATE module may not precede the first period in the POP database.

For the standard method, the user must have entered data on the number of repeaters through the Host EDIT facility. Such data are entered into the variable NREPEAT.

For the Schiefelbein method, other information must be entered:

First, the user is prompted to enter an **estimate of the 6th grade repetition rate.** Here, the user would enter his/her best guess of this rate. All other repetition and dropout rates are calculated based on that rate.

Second, the user is prompted for an **assumption about the percentage of non-promoted students who are dropouts.** In other words, if 90 out of 100 students in grade 1 pass to grade 2, the user must supply a percentage of the 10 students who do not pass who dropout. If the user puts that percentage at 50% then, in this example, 5 students would drop out and the other 5 would repeat first grade. In this case the user would enter 50.

Third, some additional information is required about the **disaggregation of enrollment data.** In many countries it is possible that there are some students, particularly in the later grades, who are "over-age". For example it is not unusual to have a number of 17 or 18 year olds in the sixth grade. enrollment data by age and grade that are published in official sources often report the number of such over-age students, but the breakdown of their ages is often not sufficiently detailed for the Schiefelbein method to work. For example, the last reported age group may be 14, rather than 18. **The aggregated data for the last age group of the ENROLAGE matrix must be entered initially into the variable AGGENROL.** When these aggregated data are present the user must tell the computer to disaggregate the data when prompted.

If disaggregation is necessary, the user is prompted to **specify the last age group.** It is important to enter the correct value here, otherwise the disaggregation will be incorrect, as will all the estimated rates.

For the standard method, the computer requests information on whether the enrollment data were entered in disaggregated form (through the variable ENROLAGE) or directly in aggregated form (through the variable ENROL). If the data were entered in disaggregated form, the computer asks whether the last age group is actually an aggregated one, as in the Schiefelbein method.

Finally, after the estimation of the transition rates is completed, the computer asks the user whether age-specific enrollment rates are desired. These rates have meaning only if age-specific enrollment data were entered. Note that there is no need to re-do these calculations unless new data on population or enrollment by grade and age have been entered.

STEP Modules: PROJRATE

Description of the PROJRATE Module

Future dropout, repetition, and first grade enrollment rates are necessary inputs for the calculation of enrollment projections. The user has the choice of entering his/her own scenario about the evolution of these rates, or he/she can let the computer do so by running PROJRATE. The module PROJRATE projects repetition, dropout, and enrollment rates for a specified number of periods, using one of five approaches that can be chosen by the user:

- future rates fixed at the level of the last historic value
- future rates fixed at the level of the historic mean
- varying future rates that follow a logistic curve
- varying future rates that follow a linear trend
- varying future rates that attain target levels at a target date, where the targets and the target year are given by the user

The INPUT data for this module are the historic values of:

INPUT data for PROJRATE module	
REPEATRT	repeater rates by grade
DROPRT	dropout rates by grade
ENTRT	entrance rate for the first grade
TARREP	target repetition rate - necessary only for target setting projections
TARDROP	target dropout rate - necessary only for target setting projections

Before running PROJRATE, the historic values of these rates must have already been estimated using the module ESTIMATE, or they could have been entered manually from some other source.

The module also requires as INPUT information about which periods to use as the basis for the projections:

	Indicator variables for PROJRATE
HISTORIC	signals which periods are historic, which are to be projected
INCLUDE	signals which of the historic periods to be included as the basis for the projections - setting INCLUDE equal to "n" (no) will result in the data for that year being ignored in estimating projection trends

The variable INCLUDE is useful because it allows the user to ignore periods which may be considered atypical and so may not want to be included in the calculation of future trends. Note that any period designated as not included will be subsequently ignored in the module COHORT.

The variable HISTORIC should remain set as it was for the ESTIMATE module. If the Schiefelbein method was used for estimating historical rates, the variable INCLUDE should be set so that the last year for which there are enrollment data IS NOT included. If the standard method was used, then the variable INCLUDE should be set so that the last year with data on enrollment and numbers of repeaters IS NOT included. This is due to the fact that, with the Schiefelbein method, to estimate the rates for any given year, enrollment data for that year and the next are needed. With the standard method, enrollment data for the given year and enrollment and repeater data for the next are needed. For example, if the last year with enrollment data is 1984, the estimated rates for 1984 itself are meaningless. Therefore, 1984 must not be included in estimating the trends - in fact, the rates for 1984 must themselves be projected. NOTE: this applies also to any gaps in the data. Thus, for example, if there are enrollment data for 1974-1976, and 1981-1984, then even though 1974-1984 are designated "h" with the HISTORIC variable, both 1976 and 1984 must be designated "n" (for "no") with the INCLUDE variable: since there were no enrollment data in 1980 or 1985, the estimated rates for 1979 and 1984 are meaningless. Of course, 1977-1980 and 1985 and onward must also be designated "n" with the INCLUDE variable. Here is the rule: if you have n historic years WITH DATA, but with m gaps in the data (regardless of their length), then the number of included years can be AT MOST n-m-1, making sure that the last historic year is not included. All this should be checked carefully. Study the following table to make sure you understand it.

Example of use of HISTORIC and INCLUDE			
Year	Have historic data?	HISTORIC	INCLUDE
1974	Yes.	h	y
1975	yes.	h	y
1976	yes.	h	n
1977	No.	h	n
1978	No.	h	n
1979	No.	h	n
1980	No.	h	n
1981	Yes.	h	y
1982	Yes.	h	y
1983	Yes.	h	y
1984	Yes.	h	n
1985	No.	p	n
1986	No.	p	n
1987	No.	p	n
1988	No.	p	n
1989	No.	p	n
1990	No.	p	n

Note: there are 7 historic years with data, there is 1 gap in the data, therefore there are 7-1-1=5 years with INCLUDE set to "y".

Of course, all of the above applies also to years in which the estimated rates are considered to be "atypical". In these cases, the year with the "atypical" rate should not be included. If it is the enrollment data that are considered atypical, then the year with atypical enrollment data AND THE PREVIOUS ONE should not be included.

The OUTPUTS of this module are the projected values of:

OUTPUT data from PROJRATE module	
REPEATRT	repeater rates by grade
DROPRT	dropout rates by grade
ENTRT	enrollment rate for the first grade
PROMRT	promotion rate by grade
TREP	t-statistic for the slope of the repetition regression
TDROP	t-statistic for the slope of the dropout regression

Use of the PROJRATE Module

Again the user will execute PROJRATE from the RUN facility of **Host** and, after selecting a "projection" or data file, will then be prompted to specify the number of periods to project. In PROJRATE, as with EDPROJ, the number of

periods that are given here include the number of historic periods. Thus if the user has an historic period of data corresponding to 1979 - 1984, and wants to project until the year 1994, he/she would specify a projection period of 15 (1994-1979=15). If the user attempts to project for too short a period the PROJRATE module will detect this and will so inform the user.

Then he/she is asked whether the module is to be run interactively or in batch mode. If the module is to be run in batch mode, the user must have already entered his/her choices in the BATCHOPT variable, or else the values already there from the previous run will be used. (See section entitled "Running the STEP System in Batch Mode" at the end of this chapter.)

When in interactive mode, once the projection period has been specified and the user executes the module, he/she will subsequently be asked to choose the method wanted for projecting the rates. An initial choice is made between fixed and varying rates. Then a choice is made among either the two fixed methods or the three available varying methods, depending on whether the user chose fixed or varying rates.

After projecting the rates, the user should check the t-statistics for the trend estimates. Projections based on weak trends are very risky, and checking this statistic is the minimum that should be done by the careful user.

For all choices except the target method, there are no further decisions required of the user and the program proceeds.

If the user has chosen to project using target rates, he/she is reminded that these targets, that are in the variables TARREP and TARDROP, must be up to date. Targets can be updated through the **Host** EDIT facility. If the target variables are up to date, the user then specifies the date at which the targets are to be achieved. Obviously, this date must be within the projection period.

STEP Modules: EDPROJ

Description of the EDPROJ Module

The module EDPROJ projects enrollment in primary school for a specified number of periods using the standard grade transition method of projecting education enrollment. For this module the INPUT data are:

INPUT data for EDPROJ module	
Transferred from POP or input by the user:	
POP	population by age
REPEATRT	repeater rates by grade
DROPRT	dropout rates by grade
ENTRT	entrance rate for the first grade
ENDAGE	oldest "legal" school age
ENROL	base year enrollment by grade

The rates must have already been supplied by running the other modules in the system, i.e. ESTIMATE and PROJRATES, or they may have been entered manually.

The OUTPUTS of this module are:

	OUTPUT data from EDPROJ module
ENROL	enrollment by grade
GER	gross enrollment rate
TOTENROL	total (aggregated across grades) enrollment in primary school
SCHLPOP	total (aggregated across ages) population of school age

Use of the EDPROJ Module

As with other modules, the user of EDPROJ will have to tell **Host** the number of periods to project. Here, there must be input data for each of the projection years. Hence if the user has an historic period of 1979 - 1984 and wants to project until 1994 he would specify a run of 15 periods, 10 of which would be projections. Here, the module PROJRATE would have had to have been run for 15 periods. In addition, population data for the projection period as well as the historic period would have to be present.

Then he/she is asked whether the module is to be run interactively or in batch mode. If the module is to be run in batch mode, the user must have already entered his/her choices in the BATCHOPT and POPFILE variables, or else the values already there from the previous run will be used. (See section entitled "Running the STEP System in Batch Mode" at the end of this chapter.)

Since this module requires population data, the user will be asked whether he/she wishes to use the data already in the STEP database, or to bring in data generated by the POP module. If the user wishes to bring in data from the POP module, he/she will need to enter the name of the database containing this data, so it is important to recall this name before running the module. As with the ESTIMATE module, if data are being imported from a POP database, the first period specified when running EDPROJ must not precede the first period in the POP database.

Once this module is executed, no further input is required on the part of the user.

STEP Modules: COST

Description of the COST module

The COST module allows the user to forecast

- teacher demand by salary grade
- demand for administrators by salary grade
- other recurrent costs
- capital costs such as classroom construction, land acquisition, etc., by type of resource

As currently configured, the COST module is sufficient only for projection of fairly general cost categories. Detailed cost and resource requirement projections for school systems is complex enough to require specialized software. This software is being developed under the BRIDGES project, but is not yet available.

The input data for the COST module include:

INPUT data for COST module	
TOTENROL	Total enrollment (summed across grades and ages) by sex. Can be input manually or result from EDPROJ module.
SYGRAD2	Student/years per graduate, accounting for dropouts. This variable is an output of the COHORT module. See section on Efficiency and Other Educational Indicators of the Technical Description chapter.
UNITCOST	Cost per unit of each resource, by cost or salary grade level. Needs to be entered for BASE year only.
COSTINCR	Rate of increase for UNITCOST variable.
COSTTYPE	Matix showing the uses and descriptions of cost variables classified by cost or salary grade and by type of cost category. This serves only as an aid in using the variables UNITCOST, QUALDIST, and RES_COEF.
QUALDIST	Distribution of resources by cost or salary grade level. For the base year, the current distribution must be entered. For the target year, the desired distribution must be entered.

INPUT data for COST module (continued)	
RES_COEF	A matrix detailing resource use coefficients and various other indicators. The rows correspond to different types of resources, such as teachers, classrooms, etc. The first two columns show the base year and target year number of students per unit of the resource, such as students per teacher, etc. If the concept is not relevant to the resource, it must be set to 1. The second two columns show the base year and target year number of shifts per unit of the resource, such as number of shifts per teacher, etc. Again, if the concept is irrelevant, these must be set to 1. The fifth column contains the minimum level of resource use. This is relevant for central resources such as administrators, etc., for which there is a minimum level that does not depend on the number of students in the system. If the concept is irrelevant for a particular resource, it must be set to 0. The sixth column indicates whether the resource is a capital good. If the resource is a capital good, a 1 must be entered, otherwise a 0 must be entered. Finally, the seventh column indicates what proportion of total costs the "Other Costs" category represents. For example, if "Other Costs" are 5% of Total Costs, a 0.05 needs to be entered here. Note that this must be entered as a proportion, not as a percentage (e.g., 0.05, not 5). For all resources except "Other", the "Prop." column must have 0s. For the "Other" resources category, both the students-per-unit-of-resource column and the "Shifts" column must have only 0s.

OUTPUT Data for COST Module	
RES_COST	Resource cost by type
RES_USE	Resource use in physical units by type
COST_CAP	Per student costs
COST_GRA	Cost per graduate

Use of the COST Module

The COST module may be run independently of the other modules, as long as the data that are normally output by other modules (the variables TOTENROL and SYGRAD2) are entered manually. If it is run as part of STEP, it should be run after EDPROJ and COHORT have been run, so that the outputs from those modules can be used. Evidently, EDPROJ and COHORT need to be re-run prior to running the COST module only if something has been altered that would change the

projections of enrollment or of internal efficiency indicators. If the changes made affect only categories such as unit cost, or per student resource requirements, then COST can be re-run without re-running EDPROJ or COHORT.

STEP Modules: COHORT

Description of the COHORT Module

The module COHORT projects the history of a cohort of 1000 entrants in primary school according to given repeater and drop out rates. The module allows the user to use two alternative methods:

- fixed drop out and repeater rates;
- varying or contemporaneous dropout and repeater rates.

The required INPUTS of the COHORT module are:

INPUT data for COHORT module	
REPEATRT	repeater rates by grade
DROPRT	drop out rates by grade

As with the EDPROJ module, these rates must have already been estimated by running other modules in the STEP system, or may have been entered manually.

The OUTPUTS of the COHORT module are:

OUTPUT data from COHORT module	
IORATE	the input/output ratio
GRADRATE	number who graduate out of 1000 entrants
YRSSCHL	average years spent in school
SYGRAD	student/years per graduate not counting student/years "wasted" by drop-outs
SYGRAD2	student/years per graduate, counting student/years "wasted" by drop-outs
ONTIME	number of graduates who graduate "on time", i.e. have not repeated

Use of the COHORT Module

Use of the COHORT module is similar to that of module EDPROJ. An appropriate number of periods must be specified and inputs present for the length of the projection period.

Then he/she is asked whether the module is to be run interactively or in batch mode. If the module is to be run in batch mode, the user must have already entered his/her choices in the BATCHOPT variable, or else the values already there from the previous run will be used. (See section entitled "Running the STEP System in Batch Mode" at the end of this chapter.)

Once the model is run the user is asked to choose between two alternative methods.

In the first method, all transition rates are assumed constant and equal to those during the entry year. For this method there must be as many periods of data for the transition rates as the current number of periods for which COHORT is being executed. This method is applied to the historic period for periods which are "included" by the INCLUDE variable being set to "y". It is also applied to the whole projection period.

In the second method, the transition rates that are applied change according to the actual calendar year in the future school history of the cohort of entrants. Experience with STEP has shown that it may take as long as 15 years for a cohort of entrants to finally graduate or dropout. In this case, a ten year projection would require AT LEAST twenty five years of future rates. For this reason, the program reminds the user who has chosen method 2 that a sufficient amount of data for future repeater and dropout rates must be in the data base. The user has the option of exiting the program and re-running PROJRATE if he/she determines that there are not enough periods of transition rate data.

Because this second method breaks down if there are gaps in the data, and since gaps in the data can only occur in the historic period, it cannot be applied in any periods designated "h" with the HISTORIC variable. The experienced user can trick the system by designating even the historic periods as "p", but this is fairly dangerous, since one might forget that this was done, and run into trouble later, when re-estimating or re-projecting the transition rates. Also, this should be done only if one is absolutely sure of the validity of the historical data, and if there are no gaps in them.

Note that the module COHORT can be used for analyzing the impact of programs oriented to the first few grades. For example, one can calculate the "fourth grade completion rate" by forcing the computer program to think there are only 4 grades right before running the module COHORT. This requires EXTREME caution, because it is easy to forget one has done this, and it will bias any subsequent exercises with the ESTIMATE, PROJRATE, and EDPROJ modules. The way to do this is as follows. Use the D)efine function in the Main Menu, then the D)imension function in the Define Menu, and then the E)dit function in the Dimension Menu. Enter the name of the GRADE dimension. Then use the cursor to go to the cell for the Current No. of Levels, and input the number 4, and press RETURN. Then press [s] for Save, [Enter] to save the dimension under its own name, and press [ESC] several times until you get back to the main menu. After running the module COHORT, go back and change the current number of levels of the grade dimension to what it was.

Running the STEP System in Batch Mode

In order to run the system in Batch mode, the user needs to be satisfied that all the options that were used last time (e.g., linear vs. logistic projection of rates, Schiefelbein vs. standard estimation method) are still valid. If not, the user can edit the variable BATCHOPT. This variable contains a record of the choices that were made the last time the system was used. Any of these choices can be changed by editing this variable. The list of choices appears as follows:

BATCHOPT Variable
ESTIMATE Method (1=Std, 2=Schief)
Default Grade 6 Rep. Rte MF
Default Grade 6 Rep. Rte Male
Default Grade 6 Rep. Rte Female
Drop out proportion Male, Female
Drop out proportion Male
Drop out proportion Female
Last age disagg.? (1=no, 2=yes)
Aggregated Age
Disagg. by Age (1=disagg., 2=agg.)
Age Spec. Enrol. Rte? (1=yes, 2=no)
PROJRATE Meth. (1=Varying, 2=Fix)
Varying Meth. (1=log, 2=lin, 3=tar)
Target year for Target Rates
Fixed Rates val. (1=mean, 2=last)
COHORT Meth (1=fixed, 2=varying)
Target Year for COST Module

The meaning of each choice is as follows:

Estimate Method (1=Std, 2=Schief)

Whether to use the Schiefelbein or standard method in estimating transition rates. 1=standard, 2=Schiefelbein.

Default Grade 6 Rep. Rte MF

Assumed 6th grade repetition rate to be used if data are not available and are not disaggregated by sex. Must be entered as a %, i.e., 4, not 0.04.

Default Grade 6 Rep. Rte Male

Assumed 6th grade repetition rate to be used for males if data are not available and are disaggregated by sex. Must be entered as a %, i.e., 4, not 0.04.

Default Grade 6 Rep. Rte Female

Assumed 6th grade repetition rate to be used for males if data are not available and are disaggregated by sex. Must be entered as a %, i.e., 4, not 0.04.

Drop out proportion Male, Female

Proportion of non-promoted who will be assumed to be dropouts when data are not disaggregated by sex. NOT a dropout rate. Must be entered as a %, e.g., 4, not .04.

Drop out proportion Male

Proportion of male non-promoted who will be assumed to be dropouts when data are disaggregated by sex. NOT a dropout rate. Must be entered as a %, e.g., 4, not .04.

Drop out proportion Female

Proportion of female non-promoted who will be assumed to be dropouts when data are disaggregated by sex. NOT a dropout rate. Must be entered as a %, e.g., 4, not .04.

Last age disagg.? (1=no, 2=yes)

Whether the last age group in the age-grade enrollment data was an open-ended (e.g., 14 and older) or closed (e.g., just 14) age group.

Aggregated Age

If the variable above was open-ended, the open-ended age group needs to be entered here (e.g., 14).

Disagg. by Age (1=disagg., 2=agg.)

If the standard method is chosen for estimating the transition rates, the computer must be told whether the enrollment data were entered disaggregated by age (in the ENROLAGE variable) or aggregated across ages (in the ENROL variable). 1=disaggregated, 2=aggregated.

Age Spec. Enrol. Rte? (1=yes, 2=no)

After the ESTIMATE module is finished estimating all the transition rates, it can go on to calculate age-specific enrollment rates. These can be useful in calculating the historical net enrollment ratios. But it does take some time. 1=yes, calculate; 2=not, don't calculate.

PROJRATE Meth. (1=Varying, 2=Fix)

In the PROJRATE module, the rates can be projected at a fixed level, or according to a trend. 1=vary according to a trend, 2=fixed.

Varying Meth (1=log, 2=lin, 3=tar)

If varying future rates are chosen, the trend can be either logistic, linear, or according to a target. 1=logistic, 2=linear, 3=according to a target. Remember that if the target method is chosen, the target repetition and dropout rates must be entered in the variables TARREP and TARDROP.

Target year for Target Rates

If the trend for the transition rates is projected according to a target, the target year (e.g., 2000) must be entered here.

Fixed Rates val. (1=mean, 2=last)

If the future rates are to projected at a fixed level, this level can be fixed at the historical mean, or at the value of the last historical year. 1=historical mean, 2=value of last historical year.

COHORT Meth (1=fixed, 2=varying)

When using the cohort method, either varying or fixed transition rates can be used. 1=fixed, 2=varying/contemporaneous. (This is an important but somewhat slippery concept. Refer to the sections entitled "Use of the COHORT Module" in Chapter 2, and "Efficiency Indicators Based on Cohort Analysis in Chapter 4.)

Target Year for COST Module

Target year for cost module.

CHAPTER 3. CONFIGURING STEP TO A NEW APPLICATION

As indicated earlier, one of the advantages of the **Host** modeling system under which STEP runs is that it can be adapted from one situation to another without the need to alter the original computer programs of the system. **Host** and STEP have been set up in such a way that the user can change a few settings in the STEP data base, enter new data and the system is ready to use. For more detailed knowledge about **Host** related fuctions, the user should consult the **Host** User's Guide. Some of the material described below is also covered in the "STEP by Step" tutorial.

In applying any **Host** model to a new situation there are several steps involved:

Step 1: Configure the data base to conform to the structure of the application by setting the dimension levels.

Step 2: Edit existing labels and define new labels which correspond to the new dimensions.

Step 3: Enter new data into the data base.

Step 4: Define new tables and graphs if desired.

In configuring the data files for the new application, existing **Host** STEP and POP files should be used. There are three possibilities. First, you can use the two "empty" files distributed with the STEP system. These **Host** files are called STEPO and POPO and contain variable definitions but no data. A second strategy is to use data bases with existing data, such as was also distributed (e.g. GUAT and POPGUAT). Of course with this approach you will need to replace existing values with new ones as appropriate. Yet a third possibility is to start with an exicting data file, "clear" the values of selected variables using **Host**'s Define/Variable/Clear facility, and then enter new variable values.

Whichever approach you use, be sure to use the **Host** data file you want when changing dimensions, etc. The easiest way to do this is to specify that file as the primary data file in **Host**'s Configure/Files facility.

Step 1: Setting the Current Maximum Levels of Dimensions

To configure the STEP and POP data bases, the first step is to set the levels of the dimensions of the variables to be equal to the correct and appropriate current levels for the application. For example, the dimension GRADE, which is used by many STEP varibles, (see Table below), has a maximum potential level for any application of twelve. For most applications, however, this would be a lower number, for example 6. This means that under the current application there would be six grades (i.e. a current maximum of six). Similar remarks apply to the other dimensions in both the STEP and POP data bases.

The tables immediately below show the dimensions for the STEP and POP data bases. To change the current maximum levels, use the Define/Dimension/Edit command of **Host**. For more information regarding dimensions, consult the **Host** User's Guide.

Dimensions Used by STEP Variables in the STEP Data Base

Name	Maximum Levels	Current Levels	Starting Value	Increment Between Levels	Maximum Effective Level	Default Label Format
AGE	86	76	0	1	85	AGEGP
ANSWER	40	17	1	1	40	BATCHQUE
COEF_TYP	10	7	1	1	10	COEF_TYP
COST_TYP	7	5	1	1	7	COST_TYP
ENAGE	19	19	5	1	23	ENAGE
GRADE	12	6	1	1	12	GRADE
PERIOD	1024	27	1974	1	2997	NONE
QUAL_TYP	14	10	1	1	14	QUAL_TYP
R	6	2	1	1	6	R
RES_TYPE	11	8	1	1	11	RES_TYPE
SEX	2	1	1	1	2	SEX
YEAR	40	27	1974	1	2013	NONE

Dimensions Used by POP Variables in the POP Data Base

Name	Maximum Levels	Current Levels	Starting Value	Increment Between Levels	Maximum Effective Level	Default Label Format
AGE	86	76	0	1	85	NONE
AGE5	18	16	0	1	17	AGGPOP
ASFR	7	7	1	1	7	ASFR
BEERSCOL	5	5	1	1	5	NONE
BEERSROW	25	25	1	1	25	NONE
FERTAGES	2	2	1	1	2	FERTAGES
FERTOPT	5	5	0	1	4	FERTOPT
INT5OPT	2	2	1	1	2	INT5OPT
LTCOLS	24	6	1	1	24	NONE
LTFILE	8	8	1	1	8	LTFILE
LTROWS	20	19	1	1	20	LTROWS
MFCOLS	8	8	1	1	[illegible]	MFCOLS
MFROWS	6	1	1	1	6	MFROWS
MODFERT	4	4	1	1	4	MODFERT
MORTOPT	9	9	1	1	9	LTFILE
PERIOD	1024	32	1970	1	2993	NONE
REGION	2	1	1	1	2	REGION
SEX	2	2	1	1	2	SEX

Step 2: Defining and Editing the Label Library

After the current maximum levels of the dimensions are set, the next step is to make sure that the labels which correspond to each level of a dimension are correct. For complete information about labels, once again the user should consult the **Host** User's Guide. In the Appendix to this chapter the labels from the STEP and POP data bases which are distributed with the system are listed. To avoid confusion, there is only one set of label libraries although there are two data bases.

In the table describing the dimensions, you will see that many of these are "attached" to dimensions. Others may be used in tables and graphs.

To alter old labels or define new ones use the Define/Labels/Edit or Define/Labels/Add commands of **Host**. Altered existing labels can be saved under new names so that the originals can be left intact.

Step 3: Entering New Data for the New Application

Once the structure of the data base is ready it is time to enter the data for the new application. If you have been working with the "empty" **Host** files, then you will need to fill in the necessary input data for each module as required. If you are using the example **Host** files which contain data you can either overwrite the existing values with new ones or, perhaps better, you can use **Host**'s Define/Variable/Clear facility to clear data from the appropriate variables.

Of course, once the new base period values for the variables have been entered for the new application you will run the various modules. This will also generate new data for the new application and test to see if any non-sensical inputs have been made.

Be sure to reset the starting value of the dimension PERIOD by using Define/Dimension/Edit. Also be careful to have appropriate values in the control variables INCLUDE and HISTORIC.

Step 4: Defining New Tables and Graphs

Once the new data have been entered and the dimensions and the labels have been updated for the new application you may wish to add new graphs and tables or change the presentation of the existing ones. This can be done by using the Define/Tables/Edit, Define/Tables/Add or Define/Graph/Edit of Define/Graph/Add facilities of **Host**.

It should be noted that most of the tables and graphs in the distribution graph and table libraries use the default labels of the dimensions. This means that any changes to the dimensions and their corresponding labels will automatically be used by the graphs and tables.

Variables in the STEP Data Files

Name	Size	Type	Periodic	Number of Dimensions	Dimension Names	Maximum Levels
AGGENROL	192	real	Yes	2	SEX	2
					GRADE	12
ASER	1376	real	Yes	2	AGE	86
					SEX	2
BATCHOPT	80	integer	No	1	ANSWER	40
COSTINCR	616	real	No	2	RES_TYPE	11
					COST_TYP	7
COSTTYPE	19712	string	No	2	RES_TYPE	11
					COST_TYP	7
COST_CAP	16	real	Yes	1	SEX	2
COST_GRA	16	real	Yes	1	SEX	2
DROPRT	192	real	Yes	2	SEX	2
					GRADE	12
ENDAGE	2	integer	No	0	NONE	NONE
ENROL	192	real	Yes	2	SEX	2
					GRADE	12
ENROLAGE	3648	real	Yes	3	SEX	2
					GRADE	12
					ENAGE	19
ENTAGE	2	integer	No	0	NONE	NONE
ENTRT	16	real	Yes	1	SEX	2
GER	16	real	Yes	1	SEX	2
GRADRATE	16	real	Yes	1	SEX	2
HISTORIC	40	char	No	1	YEAR	40
INCLUDE	40	char	No	1	YEAR	40
IORATE	16	real	Yes	1	SEX	2
NREPEAT	192	real	Yes	2	SEX	2
					GRADE	12
ONTIME	16	real	Yes	1	SEX	2
POP	1376	real	Yes	2	AGE	86
					SEX	2
POPFILE	1536	string	No	1	R	6
PROMRT	192	real	Yes	2	SEX	2
					GRADE	12
QUALDIST	2464	real	No	3	RES_TYPE	11
					QUAL_TYP	14
					SEX	2
REPEATRT	192	real	Yes	2	SEX	2
					GRADE	12
RES_COEF	1760	real	No	3	RES_TYPE	11
					COEF_TYP	10
					SEX	2
RES_COST	176	real	Yes	2	RES_TYPE	11
					SEX	2

Variables in the STEP Data Files (continued)

Name	Size	Type	Periodic	Dimensions	Names	Levels
RES_USE	176	real	Yes	2	RES_TYPE	11
					SEX	2
SCHLPOP	16	real	Yes	1	SEX	2
SYGRAD	16	real	Yes	1	SEX	2
SYGRAD2	16	real	Yes	1	SEX	2
TARDROP	192	real	No	2	SEX	2
					GRADE	12
TARREP	192	real	No	2	SEX	2
					GRADE	12
TDROP	192	real	No	2	SEX	2
					GRADE	12
TOTENROL	16	real	Yes	1	SEX	2
TREP	192	real	No	2	SEX	2
					GRADE	12
UNITCOST	1232	real	Yes	3	RES_TYPE	11
					COST_TYP	7
					SEX	2
YRSSCHL	16	real	Yes	1	SEX	2

Variables in the POP Data Files

Name	Size	Type	Periodic	Number of Dimensions	Dimension Names	Maximum Levels
ASFR	112	real	Yes	2	ASFR	7
					REGION	2
BEERS	1000	real	No	2	BEERSROW	25
					BEERSCOL	5
BIRTHS	32	real	Yes	2	SEX	2
					REGION	2
CBR	16	real	Yes	1	REGION	2
CDR	16	real	Yes	1	REGION	2
EMR	16	real	Yes	1	REGION	2
EXTMIG	8	real	Yes	0	NONE	NONE
FEMINDEX	1	byte	No	0	NONE	NONE
FERTAGES	2	byte	No	1	FERTAGES	2
FERTOPT	5	byte	No	1	FERTOPT	5
IMR	16	real	Yes	1	REGION	2
INT1NAME	256	string	No	0	NONE	NONE
INT5NAME	256	string	No	0	NONE	NONE
INT5OPT	2	byte	No	1	INT5OPT	2
INTMIG	2752	real	Yes	3	AGE	86
					SEX	2
					REGION	2
LEXP	32	real	Yes	2	SEX	2
					REGION	2
LT	7680	real	No	3	SEX	2
					LTROWS	20
					LTCOLS	24

Variables in the POP Data Files (continued)

Name	Size	Type	Periodic	Dimensions	Names	Levels
LTFILES	2048	string	No	1	LTFILE	8
MIGRATE	16	real	Yes	1	REGION	2
MODFERT	1536	real	No	3	MODFERT	4
					MFROWS	6
					MFCOLS	8
MORTOPT	9	byte	No	1	MORTOPT	9
NATLCBR	8	real	Yes	0	NONE	NONE
NATLCDR	8	real	Yes	0	NONE	NONE
NATLEMR	8	real	Yes	0	NONE	NONE
NATLIMR	8	real	Yes	0	NONE	NONE
NATLNG	8	real	Yes	0	NONE	NONE
NATLOG	8	real	Yes	0	NONE	NONE
NG	16	real	Yes	1	REGION	2
OG	16	real	Yes	1	REGION	2
POP	2752	real	Yes	3	AGE	86
					SEX	2
					REGION	2
POP5	576	real	Yes	3	AGE5	18
					SEX	2
					REGION	2
SEXRATIO	8	real	No	0	NONE	NONE
TFR	16	real	Yes	1	REGION	2
TOTALPOP	8	real	Yes	0	NONE	NONE
TOTPOP	16	real	Yes	1	REGION	2

Appendix to Chapter 3

Label Format Definitions

Format: AGE5

Description: Labels for five-year age groups

	Lower Bound	Upper Bound	Size of Interval (Inclusive)	or	Interval Label
1	0	84	5		
2	85	100			85+

Format: AGE611

Description: Labels for pop. 6 to 11 yrs.

	Lower Bound	Upper Bound	Size of Interval (Inclusive)	or	Interval Label
1	0	5			0 - 5
2	6	11			6 - 11
3	12	100			12+

Format: AGEGP

Description: Age group

	Lower Bound	Upper Bound	Size of Interval (Inclusive)	or	Interval Label
1	0	6			0-6
2	7	7			7
3	8	8			8
4	9	9			9
5	10	10			10
6	11	11			11
7	12	12			12
8	13	13			13
9	14	14			14
10	15	15			15
11	16	16			16
12	17	17			17
13	18	100			18+

Format: AGGPOP

Description: Labels for pop. aggreg. by 5-yr. groups

	Lower Bound	Upper Bound	Size of Interval (Inclusive)	or	Interval Label
1	0	0			0 - 4
2	1	1			5 - 9
3	2	2			10 - 14
4	3	3			15 - 19
5	4	4			20 - 24
6	5	5			25 - 29
7	6	6			30 - 34
8	7	7			35 - 39
9	8	8			40 - 44
10	9	9			45 - 49
11	10	10			50 - 54
12	11	11			55 - 59
13	12	12			60 - 64
14	13	13			65 - 69
15	14	14			70 - 74
16	15	15			75 +

Format: ASFR

Description: Label of age-specific fertility rates

	Lower Bound	Upper Bound	Size of Interval (Inclusive)	or	Interval Label
1	1	1			15-19
2	2	2			20-24
3	3	3			25-29
4	4	4			30-34
5	5	5			35-39
6	6	6			40-44
7	7	7			45-49

Format: BATCHQUE

Description: Questions for Batch run.

	Lower Bound	Upper Bound	Size of Interval (Inclusive)	or	Interval Label
1	1	1			ESTIMATE Method (1=Std, 2=Schief)
2	2	2			Default Grade 6 Rep. Rte MF
3	3	3			Default Grade 6 Rep. Rte Male
4	4	4			Default Grade 6 Rep. Rte Female

Format: BATCHQUE (continued)

5	5	5	Drop out proportion Male, Female
6	6	6	Drop out proportion Male
7	7	7	Drop out proportion Female
8	8	8	Last age disagg.? (1=no, 2=yes)
9	9	9	Aggregated Age
10	10	10	Disagg. by Age (1=disagg.,2=agg,)
11	11	11	Age Spec.Enrol.Rte? (1=yes, 2=no)
12	12	12	PROJRATE Meth. (1=Varying, 2=Fix)
13	13	13	Varying Meth(1=log, 2=lin, 3=tar)
14	14	14	Target year for Target Rates
15	15	15	Fixed Rates val. (1=mean, 2=last)
16	16	16	COHORT Meth (1=fixed, 2=varying)
17	17	17	Target Year for COST Module

Format: COEF_TYP

Description: Label for columns of res. coef. table

	Lower Bound	Upper Bound	Size of Interval (Inclusive)	or	Interval Label
1	1	1			Stud. per, base
2	2	2			Stud. per, trgt
3	3	3			Shifts, base
4	4	4			Shifts, target
5	5	5			Minimum level
6	6	6			Recurr. or cap.
7	7	7			Prop.

Format: COST_TYP

Description: Label for cost grades

	Lower Bound	Upper Bound	Size of Interval (Inclusive)	or	Interval Label
1	1	1			Type 1
2	2	2			Type 2
3	3	3			Type 3
4	4	4			Type 4
5	5	5			Type 5

Format: ENAGE

Description: Labels for enrolment age groups

	Lower Bound	Upper Bound	Size of Interval (Inclusive)	or	Interval Label
1	5	5			5 yr. olds
2	6	6			6 yr. olds
3	7	7			7 yr. olds
4	8	8			8 yr. olds
5	9	9			9 yr. olds
6	10	10			10 yr. olds
7	11	11			11 yr. olds
8	12	12			12 yr. olds
9	13	13			13 yr. olds
10	14	14			14 yr. olds
11	15	15			15 yr. olds
12	16	16			16 yr. olds
13	17	17			17 yr. olds
14	18	18			18 yr. olds
15	19	19			19 yr. olds
16	20	20			20 yr. olds
17	21	21			21 yr. olds
18	22	22			22 yr. olds
19	23	23			23 yr. olds

Format: FERTAGES

Description: Labels for range of fertile ages

	Lower Bound	Upper Bound	Size of Interval (Inclusive)	or	Interval Label
1	1	1			Min. fert. age
2	2	2			Max. fert. age

Format: FERTOPT

Description: Labels for fertility table options

	Lower Bound	Upper Bound	Size of Interval (Inclusive)	or	Interval Label
1	0	0			User ASFR
2	1	1			Sub-Saharan
3	2	2			Arab
4	3	3			Asian
5	4	4			User Pattern

Format: GRADE

Description: Labels for grade names

	Lower Bound	Upper Bound	Size of Interval (Inclusive)	or	Interval Label
1	1	1			First
2	2	2			Second
3	3	3			Third
4	4	4			Fourth
5	5	5			Fifth
6	6	6			Sixth
7	7	7			Seventh
8	8	8			Eighth
9	9	9			Ninth
10	10	10			10
11	11	11			11
12	12	12			12

Format: INT5OPT

Description: 1=INT5VAR structure, 2=INTVAR structure

	Lower Bound	Upper Bound	Size of Interval (Inclusive)	or	Interval Label
1	1	1			INT5VAR levels
2	2	2			INTVAR levels

Format: LTFILE

Description: Labels for life table files

	Lower Bound	Upper Bound	Size of Interval (Inclusive)	or	Interval Label
1	1	1			C-D North
2	2	2			C-D South
3	3	3			C-D West
4	4	4			C-D East
5	5	5			UN General
6	6	6			UN Lat. America
7	7	7			UN Far East
8	8	8			UN So. Asia
9	9	9			User-supplied

Format: LTROWS

Description: Labels for life-table rows

	Lower Bound	Upper Bound	Size of Interval (Inclusive)	or	Interval Label
1	1	1			Life Expectancy
2	2	2			0-4
3	3	3			5-9
4	4	4			10-14
5	5	5			15-19
6	6	6			20-24
7	7	7			25-29
8	8	8			30-34
9	9	9			35-39
10	10	10			40-44
11	11	11			45-49
12	12	12			50-54
13	13	13			55-59
14	14	14			60-64
15	15	15			65-69
16	16	16			70-74
17	17	17			75-79
18	18	18			80-84
19	19	19			85+

Format: MFCOLS

Description: Labels for model fertility age groups

	Lower Bound	Upper Bound	Size of Interval (Inclusive)	or	Interval Label
1	1	1			TFR
2	2	2			15-19
3	3	3			20-24
4	4	4			25-29
5	5	5			30-34
6	6	6			35-39
7	7	7			40-44
8	8	8			45-49

Format: MFROWS

Description: Labels for fert. levels in fert. tables

	Lower Bound	Upper Bound	Size of Interval (Inclusive)	or	Interval Label
1	1	1			TFR 1
2	2	2			TFR 2
3	3	3			TFR 3
4	4	4			TFR 4
5	5	5			TFR 5
6	6	6			TFR 6

Format: MODFERT

Description: Labels for fertility table regions

	Lower Bound	Upper Bound	Size of Interval (Inclusive)	or	Interval Label
1	1	1			Sub-saharan
2	2	2			Arab
3	3	3			Asian
4	4	4			User-supplied

Format: QUAL_TYP

Description: Label for cost grade distribution

	Lower Bound	Upper Bound	Size of Interval (Inclusive)	or	Interval Label
1	1	1			Type 1, base
2	2	2			Type 1, target
3	3	3			Type 2, base
4	4	4			Type 2, target
5	5	5			Type 3, base
6	6	6			Type 3, target
7	7	7			Type 4, base
8	8	8			Type 4, target
9	9	9			TYpe 5, base
10	10	10			Type 5, target

Format: R

Description: label for popfile

	Lower Bound	Upper Bound	Size of Interval (Inclusive)	or	Interval Label
1	1	1			ext (1) int (2)
2	2	2			filename

Format: REGION

Description: Labels for regions

	Lower Bound	Upper Bound	Size of Interval (Inclusive)	or	Interval Label
1	1	1			Total
2	2	2			Rural

Format: RES_TYPE

Description: Label for type of resource

	Lower Bound	Upper Bound	Size of Interval (Inclusive)	or	Interval Label
1	1	1			Teachers
2	2	2			Administrator
3	3	3			Other admin.
4	4	4			Learning res.
5	5	5			Classrooms
6	6	6			Laboratories
7	7	7			Land
8	8	8			Furniture

Format: SEX

Description: Labels for sexes

	Lower Bound	Upper Bound	Size of Interval (Inclusive)	or	Interval Label
1	1	1			Both sexes

CHAPTER 4. TECHNICAL DESCRIPTION OF THE MODEL METHODOLOGY

The Grade-transition Enrollment Projection Method

Of the many available methods for projecting enrollment, the grade-transition method, while perhaps not the most sophisticated, is a good compromise for developing countries since it does not require as much data as the more difficult methods, yet is useful for policy analysis.

The method, at the elementary school level, is straightforward, and consists of two basic equations. The first equation applies to grade 1 only:

$$E_{1,t+1} = r_{1,t}E_{1,t}+n_{t+1}POP_{entage,t+1}$$

Here $E_{1,t}$ is enrollment in grade 1 at time t, $r_{1,t}$ is the repetition rate in grade 1 at time t, n_{t+1} is the proportion of the population of age entage ("entrance age") that enrolls in grade 1 for the first time at time t+1, and $POP_{entage,t+1}$ is the population of age entage at time t+1. All these rates from now on will be collectively referred to as transition rates or paramaters. All these variables can be disaggregated by sex within the STEP system, but the sex subscript is left out of the technical description to simplify the notation.

This equation means that the number of students in grade 1, in any given year t+1, comes from two sources: a proportion $r_{1,t}$ of last year's enrollment in grade 1, $E_{1,t}$, which repeats the grade, and a proportion n_{t+1} of the population $POP_{entage,t+1}$ that enters school for the first time at age entage, in year t+1.

For grades 2-6, the basic equation is:

$$E_{g,t+1} = r_{g,t}E_{g,t} + (1-r_{g-1,t} - d_{g-1,t})E_{g-1,t}$$

where the symbols are analogous to those above, and $d_{g,t}$ is the dropout rate from grade g at time t.

That is, enrollment in year t+1 in any grade g comes from two sources: a proportion $r_{g,t}$ of last year's enrollment $E_{g,t}$, which repeats the same course, and a proportion $(1-r_{g-1,t} - d_{g-1,t})$ of last year's enrollment in the previous grade, $E_{g-t,t}$, that is promoted forward. Note that $(1-r_{g-1,t} - d_{g-1,t})$ is the same thing as the promotion rate.

This methodology can be used with either fixed or varying transition parameters. In either case, it is necessary to project the parameters themselves **before projecting enrollment.**

Projections with Fixed Parameters

Projecting with fixed parameters simply means that $r_{g,t} = r_{g,t-1}$ and $d_{g,t} = d_{g,t-1}$ for all grades g, and for all periods t. When projecting with fixed parameters, two intuitively appealing ways to fix the parameters are:

- at the historical mean, and
- at the last historical level

The computer program allows the user to choose between these. Simple as it may seem, projecting with fixed parameters is often the more correct option (see below).

Projections with Variable Parameters

In projecting with variable parameters, there are several common options:

- projecting with a linear trend
- projecting with a logistic trend
- projecting according to optimistic, medium, and pessimistic targets

Linear Trend. If one uses a linear trend to project the transition parameters, the equation is (using the repetition rate for any grade g as an example, but the method applies to any rate and any grade):

$$r_{g,t} = \begin{cases} a+bt & \text{if} \quad 0 \leq a+bt \leq 1 \\ 0 & \text{if} \quad a+bt < 0 \\ 1 & \text{if} \quad a+bt > 1 \end{cases}$$

where t is the time period, and the parameters a and b can be estimated using simple linear regression. Of course, if the statistical significance of these parameters, especially b, is questionable (one determines this by at least looking at its t-statistic - ideally the Durbin-Watson statistic should also be checked if available) it would be preferable to use fixed rates.

Logistic Curve Fitting. The problem with linear projections is a familiar one: there is nothing inherent in the method itself to prevent the projected rate from being greater than 1 or less than 0. Thus, it is necessary to impose those (or any other) bounds on it, as we have done above. Since this seems somewhat artificial (specially if the b coefficient is large and the rate is approaching either 1 or 0 fast), fitting a logistic curve, which automatically keeps the rate between 0 and 1, without restraining it artificially as it approaches those bounds, is usually a preferred method. Logistic curves are commonly used to model phenomena that are naturally bound from above and below, like the spread of disease or technology: the proportion of a population with a certain disease or using a certain technology cannot ever be below 0 or above 1. Furthermore, a logistic curve is somewhat "S"-shaped ("sigmoid"): at first, as it departs from the lower (or upper, if negatively sloped) bound, the slope is gentle, then it gets steeper, and as it approaches its upper (or lower, if negatively sloped) bound, the slope gets

gentle again. This seems like a good approximation to the time trend of a variable like a repetition rate: at first, it may be hard to reduce it because of lack of experience and ingrained habits. But with success comes learning and more success, so it gets easier. At the end, it gets difficult again, because the few remaining repeaters are "hard-core". The projecting equation would be:

$$r_{g,t} = \frac{1}{1+e^{a+bt}}$$

where e refers to the mathematical constant e. This implies that in estimating the parameters a and b, simple linear regression can be used, but only after first transforming the rate $r_{g,t}$ as follows:

$$\tilde{r}_{g,t} = \ln\left(\frac{1-r_{g,t}}{r_{g,t}}\right)$$

It is this rate which is then regressed on t to obtain a and b. The same cautions about the statistical significance of a, and specially b, apply here.

Target Setting Method. Since these two methods are somewhat mechanical, in that they simply extrapolate current trends, a method based on target-setting is often used. In this case, and assuming a linear approach to the target, the projection equation is

$$r_{g,t} = r_{g,t-1} + \frac{\bar{r}_{g,tt} - r_{g,lht}}{tt-lht}$$

where the bar in $r_{g,tt}$ indicates that this is a rate fixed by the user of the user method, tt refers to the target period, and lht refers to the last historical period. With this method, the user has to determine both the target rate and the target year. If the statistical projection equations discussed above are untrustworthy, because of the low significance of the slope coefficients b, then the target-setting method might be the more forthright. The method can be used to set optimistic, pessimistic, and medium targets.

Efficiency Indicators Based on Cohort Analysis

Basic Concept

The internal efficiency measures of a school system which are described in the next section are derived from calculating the school history of an artificial cohort of 1000 entrants to the school system and tracking them as they progress (or drop out, or remain) in the school system. The method consists of starting with a matrix of enrollment by grade and sex with 1000

students in grade 1 and nobody in grades 2 - 6. Here we assume a six year primary cycle for exposition purposes. The grade transition model is then repeatedly applied to this enrollment matrix until there is nobody left in the school system, i.e. the initial 1000 children have either dropped out or have graduated. As the grade transition model is iteratively applied, the number of students remaining in the system is tracked as well as the number of iterations. The number of iterations is conceptually equivalent to the amount of time the students remaining in the system have spent in school. Hence, if 100 of the 1000 entrants graduate "on time" after six years (i.e. they did not repeat), then they would be counted as graduates after six iterations of the method. Those members of the cohort who repeated a grade will obviously remain in the system for longer than six years and it will require more than six iterations for them to graduate, if at all.

In doing these calculations the number of student/years spent in the system for various categories of students is also computed. Thus the 100 "on time" graduates spent 600 student/years in the system. If 100 more students took seven years to graduate and no one else graduates, then the second group spent 700 years in the system and the total of 200 graduates spent 1300 student/years or an average of 6.5 years in the system per graduate. Similar calculations can be applied to students who leave the system without graduating.

Note that the method depends only on the drop out and repeater rates. For this reason the assumptions underlying the future transition rates during the future school history of the cohort are important. There are two basic methods which can be used here.

In the first method, all transition rates are assumed constant and equal to those during the entry year. In other words, the method projects the future history of a cohort entering school in 1986 according to the transition rates in 1986, and those rates are assumed to remain at 1986 levels.

In the second method, the transition rates change according to the actual "year" in the future projection of the cohort of entrants. In other words a cohort of 1000 entrants in 1986 would be subject to the transition rates in 1986 during the first iteration or "year", to the 1987 rates during the second iteration, to the 1988 rates during the third iteration, etc. If drop out rates are expected to fall over time, the second method would produce completion rates which are consistently higher and growing faster than those produced by the first approach.

Efficiency and Other Educational Indicators

Completion or Graduation Rate. Perhaps the most used, and certainly the simplest measure, this is the number of graduates per 1000 entrants.

"On Time" Graduates. This is the number of graduates per 1000 entrants who complete the prescribed cycle within the prescribed number of years.

Student/Years per Graduate. This is average amount of time spent in a school system per graduate. There are two methods for calculating this. In the first method, the number of student/years spent in the system by all graduates is divided by the number of graduates. In the second method, the number of student/years spent both by graduates and by drop-outs is divided by the number of graduates. In a six year system with repetition and drop out rates equal to zero these would be equal to six.

Input-output Ratio. A problem with the above measure is that is does not take into account the length of a cycle. A six year school system will require ideally six student/years per graduate which, from an efficiency standpoint should be no less efficient than an eight year system requiring ideally eight years. The input-output ratio standardizes for the length of the sysle by dividing the calculation for student/years per graduate by the length of the cycle. The ratio is ideally equal to one in any system.

Years Spent in School. This is the average number of years spent in school by all of the 1000 entrants before leaving the system. It is the average of the student/years per graduate and student/years per drop out, weighted respectively by the number of graduates and the number of drop outs. Note that the number of drop outs and the number of graduates must sum to 1000.

Indicators Based on Enrollment Data

Gross Enrollment Ratio This is the ratio of enrollment in primary school, in all age groups, divided by the population in the "legal schooling" age groups, usually 6-14 or 7-15, or thereabouts. Since the ratio is of actual or projected enrollment to actual or projected population, it has to be derived either from historical data or from an enrollment projection. It is not based on the analysis of an idealized cohort.

Net Enrollment Ratio This is a similar concept to the one above, except that the numerator is not all students, but only those in the same age groups as the population. Thus, when compared with the Gross enrollment Ratio, it is an indicator of the proportion of "overage" students in the system. It is important to note that this "overage" might be only in a legal sense: if a perfectly efficient system, with little repetition or dropouts, was unable to absorb the population right at the legal starting age, then the difference between the Net enrollment Ratio and the Gross enrollment Ratio would simply be a reflection of this fact, rather than a measure of inefficiency or wastage. In this sense, it differs from measures of "overage" derived from cohort analysis, such as the number of graduates not on time.

Estimation of Transition Rates

Standard Method

The simplest standard method of estimating the repetition and dropout rates involves using the official data on numbers of repeaters and enrollment to get the repetition rates, and solving the grade-transition equations for the dropout rates, using historical enrollment data and the previously estimated repetition rates.

The repetition rates are estimated as follows:

$$r_{g,t} = \frac{R_{g,t+1}}{E_{g,t}}$$

where $R_{g,t}$ is the number of repeaters in grade g, period t. Evidently, what we have here is just the ratio of repeaters to enrollment in the same grade the previous year. From this, we can go on to solve the grade-transition equations.

As we saw above, the standard equation of the grade-transition method for grades higher than 1 is:

$$E_{g,t+1} = r_{g,t}E_{1,t} + (1-r_{g-1,t} - d_{g-1,t})E_{g-1,t}$$

Solving $d_{g-1,t}$ for, we get

$$d_{g-1,t} = 1 - \frac{(E_{g,t+1} - r_{g,t}E_{g,t})}{E_{g-1,t}} - r_{g-1,t}$$

which is just 1 minus the promotion rate minus the repetition rate. Using this method we can solve for all the repetition and dropout rates. Note that the method implies that it will be easier to make the calculations if all repetition rates are calculated before the dropout rates, since to calculate the dropout rate for grade g we need to know the repetition rate not only for grade g, but also for grade g+1. Also note that the dropout rate implied in this method is the beginning-of-year to beginning-of-next-year dropout rate, i.e., it includes all those who abandon the school system between grades. Finally, it includes all those who die. This makes it unnecessary to refer to a death rate in making the enrollment projections.

The entrance (or first grade enrollment) rate, n_t, is calculated using the estimate repeater rate and the population and enrollment data:

$$n_{t+1} = \frac{E_{1,t+1} - r_{1,t}\ E_{1,t}}{POP_{entage,t+1}}$$

Schiefelbein Method - Description

The motivation for the Schiefelbein[2] method is the suspicion, confirmed in some cases, that the numbers of repeaters in most developing countries are under-reported. The reasons given for this are many:

- repetition is a social stigma, so children, when asked by a new teacher whether they are repeaters, are not likely to be 100% truthful
- some children who appear as new in a classroom are actually repeaters who have migrated from another region or school
- children may repeat a course even if they have not officially failed the previous one

It is easy to see from the equations for the standard method that, under these circumstances, the repetition rates will tend to be under-estimated, and the dropout rates consequently over-estimated.

The basic idea behind the Schiefelbein method is that there has to be a certain mathematical structure that determines the differences between two consecutive tables of enrollment by age and grade. This underlying mathematical structure must have age-specific repetition and dropout rates as its parameters. From the data on enrollment by age and grade for two consecutive years it may not be possible to uniquely identify a set of age-specific repetition and dropout rates, but, with certain unrestrictive assumptions, one can at least obtain a range within which the overall (i.e., not age-specific) rates MUST lie, given the age-grade enrollment data and the underlying mathematical structure.

The method consists of the following. We assume a primary school system consisting of six grades, an entrance age of 7, and a maximum school age of 23. These assumptions are for ease of exposition, but the method is easily generalizable. Also, the method can be used to calculate rates disaggregated by sex, if the enrollment data are available by sex. In the discussion below we abstract from this to keep things a little simpler.

We start with age-grade enrollment data for two consecutive years. Thus, we have $E_{g,a,t}$ and $E_{g,a,t-1}$, where g stands for grade, a stands for age, and t stands for time.

The calculations then proceed grade by grade, for each pair of consecutive ages. For any age group a, therefore, we first make the calculations for grade 6. We start by calculating the number of repeaters of age a, in grade 6 at time t:

$$R_{6,a,t} = r_6 E_{6,a-1,t-1}$$

where r_6 is an assumed value. From here we can then go on to calculate the following. The total number of students promoted from grade 5 to grade 6 has to be:

---

2. See Schiefelbein, Ernesto, "Statistical Report on Repetition in Latin America," report prepared for UNESCO, Division of Statistics on Education, Office of Statistics, Paris, 1980.

$$P_{6,a,t} = \begin{cases} E_{6,a,t} - R_{6,a,t} & \text{if} \quad E_{5,a-1,t-1} + R_{6,a,t} \geq E_{6,a,t} \\ E_{5,a-1,t-1} & \text{otherwise} \end{cases}$$

In words: if the sum of enrollment in grade 5 last year, plus those repeating grade 6 this year, is greater than enrollment in grade 6 this year, then we have to put a cap on how many students could have been promoted from grade 5 to grade 6. This cap is enrollment in grade 6 this year, minus repeaters in grade 6 this year. But if the sum of enrollment in grade 5 last year, plus those repeating grade 6 this year, is less than enrollment in grade 6 this year, then we have to promote everyone who was in grade 5 last year, unless we are ready to admit that there were a lot of students new to grade 6, which would not be logical.

The number of new students in grade 6[3] will then be:

$$N_{6,a,t} = \begin{cases} E_{6,a,t} - P_{6,a,t} & \text{if} \quad E_{5,a-1,t-1} + R_{6,a,t} \leq E_{6,a,t} \\ 0 & \text{otherwise} \end{cases}$$

which follows from the determination of the number of promoted: if there were not enough people in grade 5 last year, such that the sum of enrollment in grade 5 last year, plus repeaters in grade 6 this year is less than enrollment in grade 6 this year, then there must have been some new students in grade 6 to make up the difference.

Having done the calculations for grade 6, we can then proceed with the earlier grades, still working with the same pair of consecutive age groups. The number of non-promoted students, that is, students who will either repeat any given grade g or drop out, will be:

$$NP_{g,a,t} = E_{g,a-1,t-1} - P_{g+1,a,t}$$

That is, the number of non-promoted from grade g is equal to the enrollment in that grade, minus those who were promoted to a higher grade. (Note that this implies that the calculation has to proceed from the higher grades to the lower ones.)

---

3. Note: for national-level projections the only new students in grade 6 can be international migrants. In regional projections, they may be inter-regional migrants.

The number of repeaters is calculated according to:

$$R_{g,a,t} = \begin{cases} (1\text{-dropprop})NP_{g,a,t} & \text{if} \quad NP_{g,a,t} \leq E_{g,a,t} \\ E_{g,a,t} & \text{otherwise} \end{cases}$$

where the proportion dropprop determines the proportion of non-promoted students who drop out. The number of repeaters in grade g is equal to a proportion of the non-promoted if the pool of non-promoted is less than enrollment, but if there are more non-promoted students from the previous grade than enrollment in grade g, then the maximum number of repeaters is set equal to enrollment. Note that the proportion dropprop plays a role, but only under certain conditions (the first condition in the equation above). The smaller the number of cases where those conditions are met, the less dependent on the particular assumed value of dropprop will be the results.

The number of students promoted to grade g will be:

$$P_{g,a,t} = \begin{cases} E_{g,a,t} - R_{g,a,t} & \text{if} \quad E_{g-1,a-1,t-1} + R_{g,a,t} \geq E_{g,a,t} \\ E_{g-1,a-1,t-1} & \text{otherwise} \end{cases}$$

The number of new students in grade g will be:

$$N_{g,a,t} = \begin{cases} E_{g,a,t} - P_{g,a,t} - R_{g,a,t} & \text{if} \quad E_{g-1,a-1,t-1} + R_{g,a,t} \leq E_{g,a,t} \\ 0 & \text{otherwise} \end{cases}$$

That is, if the pool of enrollment in the previous grade last year, plus repeaters of grade g, is larger than enrollment in grade g this year, then the number promoted has to be limited, and the limit is set to equal enrollment in grade g minus repeaters of grade g. But if the the number of repeaters of grade g plus enrollment in grade g-1 last year is smaller than enrollment in grade g this year, then we set the number promoted to equal the number enroled in g-1 last year.

And, finally, the number of dropouts can be calculated as:

$$D_{g,a,t} = NP_{g,a,t} - R_{g,a,t}$$

or the difference between the total that does not go on to grade g+1 and those who repeat grade g.

The equations for grade 1 are slightly different only because there is no previous grade. We list them for completeness only.

$$NP_{1,a,t} = E_{1,a-1,t-1} - P_{2,a,t}$$

$$R_{1,a,t} = \begin{cases} (1\text{-dropprop})NP_{1,a,t} & \text{if} \quad NP_{1,a,t} \leq E_{1,a,t} \\ E_{1,a,t} & \text{otherwise} \end{cases}$$

$$P_{1,a,t} = 0$$

(since there cannot be anyone promoted to grade 1)

$$N_{1,a,t} = \begin{cases} E_{1,a,t} - R_{1,a,t} & \text{if} \quad R_{1,a,t} \leq E_{1,a,t} \\ 0 & \text{otherwise} \end{cases}$$

$$D_{1,a,t} = NP_{1,a,t} - R_{1,a,t}$$

After having calculated all these flows, we finally get the transition paramaters as follows:

$$r_{g,t} = \frac{\sum_{a=5}^{23} R_{g,a,t+1}}{\sum_{a=5}^{23} E_{g,a-1,t}}$$

$$d_{g,t} = \frac{\sum_{a=5}^{23} D_{g,a,t+1}}{\sum_{a=5}^{23} E_{g,a-1,t}}$$

$$n_t = \frac{\sum_{a=5}^{23} N_{1,a,t}}{POP_{entage,t}}$$

Thus, the grade-specific transition rates are in fact weighted averages of the age-and-grade-specific transition rates. The presumption is that this averaging process will smooth out some of the irregularities inherent to the methodology, which arise because the data are not enough to uniquely identify all the age-specific transition rates.

Schiefelbein Method - Example of Calculations

Since the method is somewhat involved, a numerical example might help to understand the calculations required.

The following example uses data from Guatemala, from the years 1982 and 1983, for the age groups 11 (in 1982) and 12 (in 1983).

Grade	$E_{11,1982}$	P	NP	R	D	N	$E_{12,1983}$
First	17737		6477	6348	130	8299	14647
Second	21605	11260	6202	6078	124	0	17338
Third	24863	15403	5613	5500	112	0	20903
Fourth	23167	19250	2224	2180	44	0	21430
Fifth	19964	20493	457	448	9	0	21391
Sixth	6044	19507		302	0	0	19809

We start with the number of repeaters in grade 6, 302, which is 6044 multiplied by the assumed grade 6 repetition rate (0.05 here). This appears in the R column. Since we have 302 students repeating grade 6, and since 19964+302>19809, we cannot promote everyone that was in grade 5 in 1982. We therefore only promote 19809-302=19507. They appear in the P column. If grade 5 enrollment in 1982 had been, say, 19000, then, since 19000+302<19809, we would have promoted all 19000, and the difference between 19809 and 19302, which is 506, would have had to be new students. By promoting everyone when we have a situation like 19000+302<19809 we minimize the need to suppose there are new students entering grade 6.

Since 19507 of the 19964 students in grade 5 in 1982 were promoted, 457 (=19964-19507) must not have been promoted, so they appear in the NP column. These 457 students are split into repeaters and dropouts according to the assumed ratio dropprop, mentioned above, but only if the number of non-promoted is smaller than enrollment in grade 6 in 1983. Since 457 is indeed smaller than 21391, then the ratio is used in this case. Had we had, say, 23000 non-promoted (from grade 5) students, then only 21391 could have been repeaters, and the rest would have been dropouts. Now we can calculate the number promoted to grade 5. Since the enrollment pool in grade 4 in 1982, plus the number repeating grade 5 in 1983, is sufficient to supply the 22229 in grade 5 in 1983 (23167+448>21391), we need to cap the number that were promoted. Only 20493 (=21391-448) could have been promoted. The number of new students is 0, for grade 5 in 1983, because enrollment in grade 4 in 1982 (23167), plus those repeating grade 5 (448), was sufficient to supply the 21391 that were enroled in grade 5 in 1983 (since 23167+448>21391).

The process continues upward until we complete this table. Similar tables have to be completed for all consecutive age groups. This gives us total numbers of repeaters and dropouts by age. We aggregate these, and aggregate the enrollment as well, divide the repeaters and dropouts by enrollment, and get the repetition and dropout rates.

Projections of Resource Requirements and Costs

The COST module uses common formulas for projecting resource requirements and costs. This makes it possible to add cost categories for specific countries, as long as the methodology used to calculate costs as outlined here is acceptable.

First, total resource use by type of resource is calculated as follows:

$$RES_USE_{j,t} = TOTENROL_t \ / \ (RES_COEF_{j,t} * SHIFTS_{j,t})$$

for recurrent costs except the "other" resources category, and where $RES_USE_{j,t}$ is total requirement of resource j at time t, $RES_COEF_{j,t}$ is the number of students per unit of resource j at time t, and $SHIFTS_{j,t}$ is the number of times per day that resource j is used at time t. Physical requirements for the "other" resources category are not calculated since because the concept is irrelevant. As most variables in the cost calculations, all of the variables in this formula can be disaggregated by sex, but the sex subscript is left out to simplify the notation. Both RES_COEF and SHIFTS for every time period between the base year and the target year are interpolated by the module from the base year and target year values given by the user in the Host variable RES_COEF. For every year after the target year these variables remain at the target year level. For capital costs (i.e., investment) the formula is

$$RES_USE_{j,t} = (TOTENROL_{t+1} - TOTENROL_t) \ / \ (RES_COEF_{j,t} * SHIFTS_{j,t})$$

Note that according to this expression investment requirements might become negative if enrollment declines. This can have several interpretations. It is not likely that there would be significant disinvestment in the short run in any real situation. Thus, the simplest interpretation is to take any negative numbers here as a measure of slack in the use of the existing resource base.

Second, physical resource use or requirements need to be converted into monetary costs. This is done in several steps.

Unit costs by type of resource and cost level (or salary grade) are projected as

$$UNITCOST_{j,k,t} = UNITCOST_{j,k,t-1} * (1 + COSTINCR_{j,k})$$

where COSTINCR is the yearly cost increase rate for resource j of cost grade k.

Average unit cost calculations for each resource j, weighted by the distribution across cost categories, must then be made:

$$AVUNITCOST_{j,t} = \sum_{k} QUAL_{j,k,t} * UNITCOST_{j,k,t}$$

where $QUAL_{j,k,t}$ is the distribution of resource j across cost or salary grades k at time t. This is interpolated by the module from the base year and target year distributions given by the user in **Host** variable QUALDIST.

Once the average unit costs are known, the cost of the resources can be calculated according to the equation

$$RES_COST_{j,t} = RES_USE_{j,t} * AVUNITCOST_{j,t}$$

for all resources except the "other" resources category.
For the "other" category, costs are calculated as

$$RES_COST_{other,t} = (PROP/(1-PROP)) \sum_{j \neq \text{"other"}} RES_COST_{j,t}$$

where PROP is the ratio of "other" costs to total costs.

Finally, two per-student cost indicators are calculated: cost per capita (COST_CAP) and cost per graduate (COST_GRA). Cost per student is calculated according to the equation

$$COST_CAP_t = \sum_{j} RES_COST_{j,t} / TOTENROL_t$$

If resource costs for a given category of resources are negative they are excluded from the above sum. Cost per graduate is calculated according to the formula equation

$$COST_GRA_t = \sum_{r=0}^{5+} (COST_CAP_{t+r} * SYGRAD2_{t+r})/6$$

where $SYGRAD2_t$ is the number of student-years it takes to produce one graduate at time t. Thus, if $COST_CAP_t$ is the cost/student-year at time t, and $SYGRAD2_t$ is the student-years/graduate at time t, then $COST_CAP_t*SYGRAD2_t$ is the cost per graduate at time t. But since each graduate starting at time t will be in the system for at least 6 years, we take the average of the costs per graduate for the 6 years starting with year t. Note that this implies that one can only produce data on the cost per graduate for 6 years less than there are data for the cost per student. Note also that in fact the average number of years it takes a graduate to get through the system is a variable greater than 6 and thus one would ideally take the average of a variable number of years. The disadvantage would be that then one would be able to produce even less data on cost per graduate. Using 6 gives a good approximation without losing too much data.

Projecting Population

POP uses the cohort-component method to project the base population forward. This method has been adapted to make annual, single-year age group population projections. The base-year population can be entered either in single-year age groups, if available, or in five-year age groups and then disaggregated using Beers multipliers. The survival ratios used are calculated from the disaggregated (again, using Beers coefficients) L values of a life table.

The basic projection equation is

$$POP_{a,s,r=1,t} = SURVIVORS_{a,s,r=1,t} + INTMIG_{a,s,r=2,t} - INTMIG_{a,s,r=1,t} + EXTMIG_{a,s,r=1,t}$$

for region 1, and

$$POP_{a,s,r=2,t} = SURVIVORS_{a,s,r=2,t} + INTMIG_{a,s,r=1,t} - INTMIG_{a,s,r=2,t} + EXTMIG_{a,s,r=2,t}$$

for region 2.

The components are calculated as follows.

First, the survivors are determined from last year's population. There are three relevant equations. For those newly born,

$$SURVIVORS_{a=0,s,r,t} = SURVIVALRATE_{a,s,r,t} * BIRTHS_{s,r,t},$$

then for the age groups between the first and the last,

$$SURVIVORS_{a,s,r,t} = SURVIVALRATE_{a-1,s,r,t-1} * POP_{a-1,s,r,t-1}$$

and for the last age group,

$$SURVIVORS_{a,s,r,t} = SURVIVALRATE_{a-1,s,r,t-1} * (POP_{a-1,s,r,t-1} + POP_{a,s,r,t-1})$$

The survival rates in all these equations are determined as the ratio between the Lx values for successive age groups in the life tables chosen by the user. But the Lx values needed here are for single-year age groups. Thus, the Lx values from the standard five-year tables (Coale-Demeny and U.N.) are first interpolated, using Beers coefficients, into single-year values. The justification for this is that the Lx values from the standard five-year life tables are conceptually equivalent to a stable population, and thus using Beers coefficients to interpolate them is equivalent to using Beers coefficients to interpolate population data grouped in five-year age groups into population data grouped into single-year age groups. However, whatever

objections can be made against using Beers coefficients for interpolating population values (especially in the early and late age groups) also hold against using them for interpolating the Lx values from the life tables. After interpolating using Beers coefficients to get single-year age group values, the POP module interpolates linearly to get Lx values appropriate for the life expectancies entered by the user under the variable LEXP.

In the first equation for calculating the number of survivors above, the variable BIRTHS is used. The number of births is calculated according to

$$BIRTHS_{male,r,t} = (SEXRATIO/(1 + SEXRATIO)) * \sum_{a=15}^{49} ASFR_{a,r,t} * POP_{a,female,r,t-1}$$

$$BIRTHS_{female,r,t} = (1/(1 + SEXRATIO)) * \sum_{a=15}^{49} ASFR_{a,r,t} * POP_{a,female,r,t-1}$$

where SEXRATIO is the ratio of boys to girls at birth, and ASFR is the age-specific fertility rate, disaggregated in for single-year age groups, disaggregated assuming births occur uniformly within each five-year age group in the fertile ages.

Internal migrants are calculated from a total migration rate given by the user by region only, and then disaggregated by age and sex within the program according to a model function. The equation for the model is

$$INTMIG_{a,s,r,t} = MIGRATE_{r,t}$$

$$* \; m_{a,s}*SURVIVORS_{a,s,r,t}/(\sum_{a} m_{a,s}*SURVIVORS_{a,s,r,t}))$$

$$*\sum_{a} \sum_{s} SURVIVORS_{a,s,r,t}$$

where $m_{a,s}$ is a function of age a and sex s. The function for $m_{a,s}$ is

$$m_{a,s} = a1_s*\exp(-alpha1_s*a)$$

$$+ a2_s*\exp(-alpha2_s*(a-mu2_s) - \exp(-lambda2_s*(a-mu2_s)))$$

$$+ c_s$$

where a is age, and a1, alpha1, a2, alpha2, mu2, lambda2, and c in turn are functions of sex:

	Male	Female
a1	0.01992	0.02209
alpha1	0.10390	0.10883
c	0.00263	0.00350
mu2	19.67385	19.88280
a2	0.06471	0.06935
alpha2	0.10618	0.13434
lambda2	0.37244	0.37870

Finally, for external migrants it is assumed that only total numbers are known to the user. The number of migrants by age, sex and region are calculated simply by assuming that the distribution by age, sex and region is the same as for the population:

$$EXTMIG_{a,s,r,t} = EXTMIG_t$$

$$* \; POP_{a,s,r,t-1} / (\sum_a \sum_s \sum_r POP_{a,s,r,t-1})$$

This is evidently not the best solution, but there is no model function for external migration equivalent to the one used for internal migration.

Appendix to Chapter 4
Glossary of Variables and Symbols

The following table shows, for each variable, its symbol in this technical description, its name in the user's manual or in the computer program, and a brief description of its meaning. The variables are listed in the alphabetical order of their symbols.

Symbol	Name	Meaning
$COSTINCR_{j,k}$	COSTINCR	rate of cost increase for input j, cost or salary grade k
$d_{g,t}$	DROPRT	dropout rate grade g, time t
dropprop	DROPPROP	assumed share of dropouts in non-promoted students
$E_{g,a,t}$	ENROLAGE	enrollment in grade g, age a, at time t
$E_{g,t}$	ENROL	enrollment grade g, period t
endage	ENDAGE	oldest "legal" school age
entage	ENTAGE	age of entry into schooling
n_t	ENTRT	rate of entry from pop. to 1st grade
$POP_{a,t}$	POP	population of age a at time t
$r_{g,tt}$	TARREP	target repetition rate
$R_{g,t}$	NREPEAT	repeaters, grade g, time t
$QUAL_{j,k,t}$	QUALDIST	proportion of input j of quality level k at time t
$r_{g,t}$	REPEATRT	repeater rate grade g, time t
$RES_COEF_{j,t}$	RES_COEF	resource use and other coefficients for resource j at time t
tt	TARYEAR	target time period
$UNITCOST_{j,k,t}$	UNITCOST	cost per unit of input j, cost or salary grade k at time t
None	AGGENROL	enrollment in the aggregated age group
None	ASER	age-specific enrollment rate
None	COSTTYPE	cost categories and descriptions
None	GER	gross enrollment ratio
None	GRADRATE	number who graduate out of 1000 entrants
None	HISTORIC	indicator var.: historic=h, projected=p
None	INCLUDE	indicator var.: if include this year, include=y, if not, include=n.
None	IORATIO	input-output ratio: student years per graduate divided by the nominal school cycle length
None	ONTIME	number of "on-time" graduates,i.e., have not repeated
None	SCHLPOP	total (aggregated across ages) population of school age
None	SYGRAD	student/years per graduate not counting student/years "wasted" by drop-outs

Glossary of Variables and Symbols (continued)

None	SYGRAD2	student/years per graduate, counting student/years "wasted" by drop-outs
None	TARDROP	target dropout rate
None	TDROP	t-statistic for drop. rate regr. slope
None	TOTENROL	total (aggregated across grades) enrollment in primary school
None	TREP	t-statistic for rep. rate regr. slope
None	YRSSCHL	average no. of yrs. spent in school

The following variables are used only for intermediate calculations, and so the user need not be concerned about them. Within the program, some of these variables are simply various columns of the matrix "flowmatrix".

Symbol	Name	Meaning
a	various	regression intercepts
b	various	regression slopes
$D_{g,a,t}$	flowmatrix	dropouts from grade g, of age a, at time t
lht	lasthist	last historical period
$N_{g,a,t}$	flowmatrix	new students in grade g, of age a, at time t
$NP_{g,a,t}$	flowmatrix	students not promoted from grade g, of age a, at time t
$P_{g,a,t}$	flowmatrix	students promoted to grade g, of age a, at time t
$R_{g,a,t}$	flowmatrix	repeaters of grade g, of age a, at time t
t	t	time period

The following variables are used in the population projection database only.

Symbol	Name	Meaning
$ASFR_{a,rt}$	ASFR	age specific fertility rates for age a, region r, at time t
$ESTMIG_{a,s,r,t}$	EXTMIG	external migration in absolute numbers, for age a, sex s, region r, at time t
$MIGRATE_{r,t}$	MIGRATE	internal migration rate for region r at time t
$POP_{a,s,r,t}$	POP	population by single-year age groups, age a, sex s, region r, at time t
SEXRATIO	SEXRATIO	ration of boys to girls at birth
NONE	FEMINDEX	index denoting women in internal calculations
NONE	LEXP	life expectancy
NONE	LT	life table L values
NONE	MODFERT	model fertility schedules
NONE	MORTOPT	mortality options
NONE	TFR	total fertility rate