



UNESCO Institute for Statistics (UIS)

Adult Education and
Literacy Statistics Programme

**Global Age-specific Literacy Projections Model (GALP):
Rationale, Methodology and Software**

A paper and software commissioned by the UIS and developed by
Dr. Wolfgang Lutz and Dr. Sergei Scherbov

July 2006

Montréal (Québec), Canada
UNESCO Institute for Statistics (UIS)

Abstract

Based on discussions with the Institute's Advisory Board on Literacy Evaluation (ABLE) in 2004, the UIS decided to develop a new approach to projecting the dichotomous variable of self-defined literacy which is fully based on reported empirical data for the starting conditions. Moreover, it does not impose overly rigid structural assumptions about future trends.

The new model thus takes the patterns of age- and sex-specific literacy rates as reported by countries as the starting point of the projections that are carried out by single year of age and single year of time. This allows users to have different jump-off years for the projections (depending on the year for which the most recent national data are available) as well as common target years (such as 2015) for all countries. The literacy projections are directly based on the medium projections of national populations by age and sex (in single years of age and time) as given in the most recent assessments of the United Nations Population Division (UNPD). The model then superimposes onto these population projections assumed future patterns of age-specific proportions of literate men and women starting at the age of 15.

The software, written in Excel, is very flexible in accommodating different assumptions on these future age- and sex-profiles of literacy. As a default option, it is assumed that all transitions from the illiterate to the literate state happen before the age of 15 and that once a person defines himself/herself as literate this status is maintained throughout life. Under this assumption, the user only has to define one scenario variable: the proportion of those literate at the age of 15 in each consecutive year. Here one can choose between a pre-defined trend scenario (an extrapolation of the logits of the values observed for the 15 youngest cohorts), a "constant transition to literacy" scenario or any specific user-defined scenario. If there is reason to assume that the previous assumption does not hold, the user can also make specific assumptions about how cohorts may further gain or even lose literacy after the age of 15. Based on these scenarios the results provide proportions literate and absolute numbers for single years of time and re-aggregated to five-year age groups for men and women up to 2030 and for the broader age groups 15 to 24 or 15 and over. The software allows users to produce results in different formats.

The software was developed and this report was prepared by Dr. Wolfgang Lutz, head of the World Population Programme at the International Institute for Applied Systems Analysis (IIASA) and Director of the Vienna Institute of Demography, Austrian Academy of Sciences and Dr. Sergei Scherbov, who is Research Group Leader at the Vienna Institute of Demography, Austrian Academy of Sciences.

Table of contents

	Page
1. General approach	
1.1 Why project literacy?	4
1.2 Projecting the dichotomous literacy variable	5
1.3 Variation of literacy over age, time and cohort	6
1.4 Demographic model of cohort component projection and the consideration of sub-populations.....	7
1.5 The GALP model	8
1.6 Structure of the software	10
2. Initializing the model for the jump-off year	
2.1 Empirical input required	11
2.2 Transformation into single-year age groups	12
2.3 Combination with UNPD population projections	13
3. Scenario setting	
3.1 Assumption of invariant proportions literate after age 15	14
3.2 “Constant transition to literacy” scenario	15
3.3 Trend scenario	17
3.4 Specific scenarios	19
3.5 How to incorporate other information and assume additional transitions after the age of 15	20
4. Presentation of results	
4.1 Numerical results	21
4.2 Graphical output	22
4.3 Comparison of scenario results	24
References	27

1. General Approach

1.1 Why project literacy?

Literacy is widely considered to be at the heart of human development. For this reason, literacy is also at the core of the Education for All (EFA) program. As stated by UNESCO Director-General Koïchiro Matsuura in the foreword to the 2006 EFA Global Monitoring Report: "... a good quality basic education equips pupils with literacy skills for life and further learning: literate parents are more likely to send their children to school: literate people are better able to access continuing education opportunities; and literate societies are better geared to meet pressing development challenges" (UNESCO, 2005, page 4).

Because of literacy's central role in education and development, a series of global education conferences have also defined quantitative literacy goals. Early formulations of literacy goals by the international community focused on a reduction of adult illiteracy. Paragraph 8 of the 1990 Jomtien Framework of Action suggested that countries set targets for the reduction of the adult illiteracy rate (the appropriate age group to be defined in each country) to half of its 1990 level by the year 2000. In 2000, however, this goal was unfortunately modified in the Dakar Framework of Action, which refers to "achieving a 50 per cent improvement in levels of adult literacy by 2015 ...". Clearly this is not possible for countries with literacy rates exceeding 67 percent because literacy cannot surpass a 100 percent maximum. For this reason, the recent EFA monitoring report returns to the original measurement definition of the goal to reduce adult illiteracy (UNESCO, 2005, p.66).

Measuring progress of countries towards meeting such goals is no trivial task because policies aimed at improving literacy do not linearly translate into improvements for the adult population. National policies typically focus on increasing literacy through expanding school enrolment among children. And even if countries manage to make great progress in terms of youth literacy, it will take decades for such policies to significantly affect adult literacy because of the time it takes for the better educated younger persons to gradually replace those less educated and older. This age-structural momentum is only reduced in the rare cases of strong campaigns aiming at educating adults of adult literacy.

Hence a realistic assessment of progress requires a projection model that captures the full non-linear dynamics of changing literacy among the adult population. It should also show the expected future gains in literacy under alternative scenario assumptions. The concept of a "constant transition to literacy" for instance, can demonstrate the degree to which the literacy of the adult population will still improve under the hypothetical case that school systems do not improve in the future and that the proportion of each cohort that will become literate remains unchanged. However; since in most countries there have been improvements in the past, resulting in the fact that today young cohorts are better educated than older ones, even such a constancy in transition to literacy is likely to result in substantial improvements of adult literacy over time since the older less literate cohorts will dye off and the younger better educated cohorts move up the age pyramid. Hence such a "constant transition to literacy" scenario measures the purely compositional affect. At the other extreme, such a projection model can demonstrate the maximum improvement that might be achieved under the assumption of instant universal

primary education (without provisions for adult education). Even this unrealistically optimistic case will not necessarily lead to the target of halving adult illiteracy by 2015 because of the considerable amount of time required to replace older illiterates by younger literates. Only literacy projection models based on age-specific information can appropriately describe such processes. In the specific case discussed above, the projection may show that universal primary education may not be sufficient to reach the target of halving adult illiteracy by 2015. Schooling policies may need to be complemented by massive adult education programmes.

Looking beyond these two extreme scenarios, a projection model must also be able to calculate more realistic scenarios based on recent trends of literacy improvements among younger cohorts. The model should also be flexible enough to demonstrate the future change in adult literacy (the stock in human capital) as implied by any specific national policies which improve school enrolments (transitions to literacy).

Literacy is essential for development and, therefore, projections must go beyond the assessment of goals. Effective development planning requires a realistic outlook of what are likely to be the proportions of literate men and women at different ages. This may affect the allocation of national or foreign direct investments in which the skills of the labor force often play a key role. But beyond economics, literacy also is a key determinant of health and empowerment in a broader sense. For these reasons, literacy projections deserve a central place in any broader development outlook.

1.2 Projecting the dichotomous literacy variable

There is no standard international definition that covers all dimensions of literacy. Most definitions are concerned with functional literacy which enables a person to engage in all the activities in which literacy is required. In this broader sense literacy refers to a “context-bound continuum of reading, writing and numeracy skills, acquired and developed through processes of learning and application, in schools and in other settings appropriate to youth and adults” (UNESCO, 2005, p. 30).

This broader understanding of literacy has, in recent years, given rise to efforts to empirically assess different dimensions of literacy skills and, accordingly, group persons according to their performance. This has already been done in a series of international adult literacy surveys and lies at the heart of the new UIS Literacy Assessment and Monitoring Programme (LAMP). These assessments provide a more accurate measurement than traditional literacy statistics, which mostly refer to a dichotomous variable (literate/illiterate) and are based largely on self-reported information (or information given by the head of household). However, these assessments currently exist for just a handful of countries and there are hardly any consistent time series available.

For this reason any longer term global level analysis, including efforts to project literacy for most countries in the world, must still refer to the old dichotomous literacy variable as it has been used in censuses and surveys in many countries for decades. For this reason, the projection model described here also refers to this dichotomous variable which has been defined by UNESCO as “the ability to read and write, with understanding, a simple short statement on everyday life”

(UNESCO, 2005, p.64). Depending on the country and the data collection instruments (census or surveys), this information—which is typically recorded by age and sex—is either based on the respondents subjective report about his/her literacy abilities or a report of the head of household about his/her assessment of the literacy skills of each household member. The projection model uses these data as they are reported by governments to the UIS.

The basic approach of the projection model described here is in principle also applicable to multi-category literacy variables and those based on test outcomes rather than self-declaration. Only the assumption of literacy stability over the life course will be more problematic in such a context and requires further analysis along the lines described below. But for the time being, the literacy projection model only applies to the dichotomous variable by age and sex as reported by a large number of countries. A future adaptation to the multi-category case is possible but would necessitate further development of the current software.

1.3 Variation of literacy over age, time and cohort

Literacy, like any other characteristic observed with individuals can vary by age and over time. In addition, one is often interested in variations by sex and other categorizations such as urban/rural place of residence, nationality etc. The analysis of such variation helps us to gain important insights into the determinants of literacy and also provides the basis for projections into the future.

When attempting to forecast literacy, it is important to be aware of a significant interaction between the age pattern of literacy and its change over time. For example, consider a typical age profile in which, at one point in time, there are higher literate proportions of young adults than older adults. In theory, this could be due to the (unlikely) fact that people lose literacy skills as they become older. But a more plausible explanation is that greater access to education has benefited the most recent cohorts compared to those who are older. This second effect is called a cohort effect (or a generation effect) in social sciences. A cohort (generation) is defined here as a group of men and women born in the same year and who therefore get older with the passage of time. In almost every country of the world, older cohorts are less educated than younger cohorts because education is concentrated in the younger age groups and most education systems have expanded over time.

These typical cohort patterns also explain the above mentioned case that adult literacy continues to improve with the passage of time because of cohort replacement. However, this does not account for all changes in adult literacy over time. At least in theory, it is also possible that the proportion of a cohort that is literate changes over time, i.e. as the cohort ages. Such changes can be in both directions. Cohorts may experience further improvements in literacy later in their life course through adult education programs or simply by being exposed to a literate environment. On the other hand literacy skills of some members of the cohort can deteriorate and the literate proportion could possibly decline as a cohort ages. In addition, the literate proportion of the adult population can also change through education differentials in mortality or migration.

How can literacy forecasting take into account possible changes due to age, time and cohorts? Virtually all analyses of trends in the traditional dichotomous variable of self-defined literacy

assume that once a cohort has reached its maximum literacy in its youth (around the age of 15), the proportion of literates will stay constant along cohort lines. This assumption is partly based on empirical analysis (UNESCO 1995) and partly on plausible reasoning. We will further discuss this assumption and the possibility to modify it in the context of this model in Section 3.1. Here it shall only be noted that there are formal statistical models (so-called Age-Period-Cohort or APC Models) that can test for the magnitude and significance of these three separate effects in a multi-variant setting. For instance, in the field of changes in political opinions and party preference, such models tend to show that cohort effects are much stronger than age effects. In the more conceptually stable variable of basic literacy, the cohort effect can be expected to be pervasive, although we do not know of any formal APC analysis on literacy trends. If there are two or more data points of literacy by age and sex for a number of countries, such APC analysis is easy to perform if the projection assumptions are tested.

1.4 Demographic model of cohort component projection and the consideration of sub-populations

Since the task is to project the numbers of men and women that will be literate and illiterate at future points in time the literacy projections are closely interwoven with population projections. For example, in order to project the number of 30-year-old women who will be literate in 2015, two different pieces of information must be combined: the total number of women that will be alive at age 30 in 2015 and the proportion of those women that will be able to read and write. The first piece is derived from a population projection following the standard cohort component projection method. The second part involves expanding this model to consider the dynamics of sub-population.

The standard demographic projection model starts with an empirically given population distribution by age and sex. It is performed in either one- or five-year steps using correspondingly one- or five-year age groups. The population projections used here is produced by the UNPD. They were originally performed in five-year steps but were then interpolated to single years of age and time. Since we will be using this single year data set, we will only refer to the single year case in the following description.

This demographic projection model is referred to as a cohort component model because it projects along cohort lines and considers all three components of change, namely fertility, mortality and migration. This is done in the following way: every year, every man and woman is moved up by one age group according to assumed patterns of age- and sex- specific mortality rates. In simpler words, not everybody is made one year older every year, because some people drop out of the model because they are assumed to die over the course of that year. Similarly, assumed age- and sex-specific migration schedules are applied to all men and women, because people may leave or enter the country at each age. Finally, an age-specific fertility pattern is applied to all women in reproductive age and the resulting children (after adjusting for infant mortality) are added to the bottom of the age pyramid to make up the new age group of 0 to one-year-olds in the next year. Since this is the standard model as described in any introductory demographic textbook, its core was only summarized verbally. Formally, it is represented through the so-called Leslie Matrix (Keyfitz 1985).

This simple cohort component model has also been generalized to the multi-dimensional case in which different sub-populations are projected while they interact. This so-called multi-state model, as developed by Rogers (1975) and Keyfitz (1985), was originally applied to simultaneously project the population of different regions where people are allowed to move from one region to another. But the states can also be defined to represent other sub-populations, such as those defined by different levels of educational attainment. Following this approach, Lutz and Goujon (2001) have published the first global level projections of the population by age, sex and four levels of educational attainment. When performing such multi-state projections, we can also consider the fact that people in different states are often subject to different sets of fertility, mortality and migration rates. If the states are defined by educational attainment, women with higher education almost universally tend to have lower fertility than those in the lowest educational categories. Lutz and Goujon (2001) clearly show in their projections how changes in the educational composition of the population lead to changes in the aggregate fertility levels, even if fertility levels within each educational group do not change. In other words, an improvement in the educational attainment of reproductive age women typically leads to declining fertility rates, an effect that comes on top of possible fertility declines within educational groups.

1.5 The GALP Model

The multi-state model described above served as the starting point for designing this specific model for literacy projections. However the method finally chosen does not allow for different vital rates by literacy status because of the potential confusion arising from inevitable discrepancies between the literacy projections and the UN population projections. This does not imply that there is anything wrong with the UN population projections. The discrepancy is simply due to the fact that the dynamics of a system that is assumed to be homogeneous (no differentials by literacy) over time becomes different from that which explicitly considers certain sources of heterogeneity (e.g. different fertility rates for literate and illiterate women).

The solution chosen here to assure perfect conformity with the UN population projections is to take those projections as externally given (and considering them as implicitly including the possible affects of changes in the literacy composition on vital rates). Once these projections by age and sex are given, the only element to be projected is the changing pattern of age- and sex-specific proportions literate over time. In other words, the specific contribution of this literacy projection model is to project the future proportions literate for men and women for each age group and each year and then apply them to the UN population projections:

Let

$N(s,a,t)$ be the number of people of sex s , and age a , at time t

$\text{PropLit}(s,a,t)$ the proportion of the people of sex s , age a , at time t that is literate

then

$L(s,a,t)$ the literate population of sex s , age a , at time t is simply

$$L(s,a,t) = N(s,a,t) * PropLit(s,a,t)$$

and

$I(s,a,t)$ the illiterate population of sex s , age a , at time t is

$$I(s,a,t) = N(s,a,t) - L(s,a,t)$$

Since the full set of $N(s,a,t)$ is given by the UN projections the remaining task is to project the pattern of $PropLit(s,a,t)$ into the future.

With a time horizon of around 30 years (the model runs to 2030 but the jump-off years for countries vary somewhat) and 85 age groups (from age 15 to 99 years), $PropLit(s,a,t)$ theoretically consists of two matrices (male and female), each with up to 2,550 (30*85) data points. It would be nearly impossible to make a meaningful projection of all these data points without imposing any simplifying structure and making certain assumptions to reduce complexity.

For this reason, the following section provides step by step discussions of the plausibility of assumptions intended to facilitate this task. We will also address ways to specifically modify such assumptions if there is reason to believe that the assumption is violated in a specific case.

Before going into more, we wish to stress again the most important of such simplifying assumptions: literacy, once achieved, will remain stable along cohort lines. The cohorts refer to the diagonals of the above mentioned matrices by age and time. In the above notation a female ($s=f$) cohort that enters the model at age 15 is represented by the series

$$N(f,15,t), N(f,16,t+1), N(f,17,t+2) \dots$$

The above mentioned simplifying assumption would then be

$$PropLit(f,15,t) = PropLit(f,16,t+1) = PropLit(f,17,t+2) = \dots$$

This assumption of constancy along cohort lines includes all female cohorts that are already above age 15 in the jump-off (= latest observed) year t , i.e. also

$$\begin{aligned} PropLit(f,16,t) &= PropLit(f,17,t+1) = PropLit(f,18,t+2) = \dots \\ PropLit(f,17,t) &= PropLit(f,18,t+1) = PropLit(f,19,t+2) = \dots \end{aligned}$$

As well as cohorts that will only enter the model in future years, i.e.

$$\begin{aligned} PropLit(f,15,t+1) &= PropLit(f,16,t+2) = PropLit(f,17,t+3) = \dots \\ PropLit(f,15,t+2) &= PropLit(f,16,t+3) = PropLit(f,17,t+4) = \dots \end{aligned}$$

The assumption hence applies to around 115 female cohorts and 115 male cohorts, 85 that enter the model in the jump-off year at ages 15-99 and another around 30 (depending on the exact difference between the jump-off year and 2030) that will enter the model over the coming years. Hence this assumption alone reduces the number of parameters to be specified from more than 5000 to around 230. Of these only 60 (30 for women and 30 for men) have to be assumed because they concern still unobserved cohorts. The Poplins are provided empirically for all cohorts aged 15 and above in the jump-off year.

Note that this is the key underlying assumption being used in the current methodology, but as will be discussed below, the model will also be able to handle certain deviations from this basic simplifying assumption.

1.6 Structure of the software

The software is written in Excel and has a shell that makes it easy to use for people without advanced knowledge of the programme. In addition, Excel and similar spreadsheet programmes are essentially self documenting. All of the formulas behind the model can be identified by the users by looking at the appropriate sheets. For this reason, in the following description we will focus on the substantive choices that the user will have to make rather than a detailed description of the algorithms.

The model projects one country after another and produces the standard output as described below. The results are also suitable for database import where all kinds of regional aggregations and comparative analysis can be derived with only minimal effort. This sequential nature of the projection model results from the need to critically examine the assumptions made for each individual country. There are no uniform scenarios (except for the implausible “constant transition to literacy” scenario) that can meaningfully be run across all countries in an identical manner. As will be discussed extensively below, even the default trend scenario requires a critical look for each country to assess whether a continuation of the past trend is always the most plausible and likely future trend.

As shown in Figure 1, the main screen brings together all of the different components of the programme:

- On the left hand side, users can choose the country for which the projection should be performed.
- The button “Define Initial Files” helps the user to identify the names of the data input files for each country and also change the input files if other input data should be chosen for a specific country.
- When clicking the “Initialize” button, the data for this country will be loaded. This will only work if the input data are in the standard format. These data will then be matched with demographic projections from the UNPD, graduated into single years of age and adjusted as described in Section 2.
- Once the country is initialized, users can go to the scenario setting module which allows for the choice of a predefined trend scenario, “constant transition to literacy” scenario or any other country specific scenario as extensively discussed in Section 3.

- The results can be presented in either graphical or tabular form. This is discussed in more detail in Section 4.

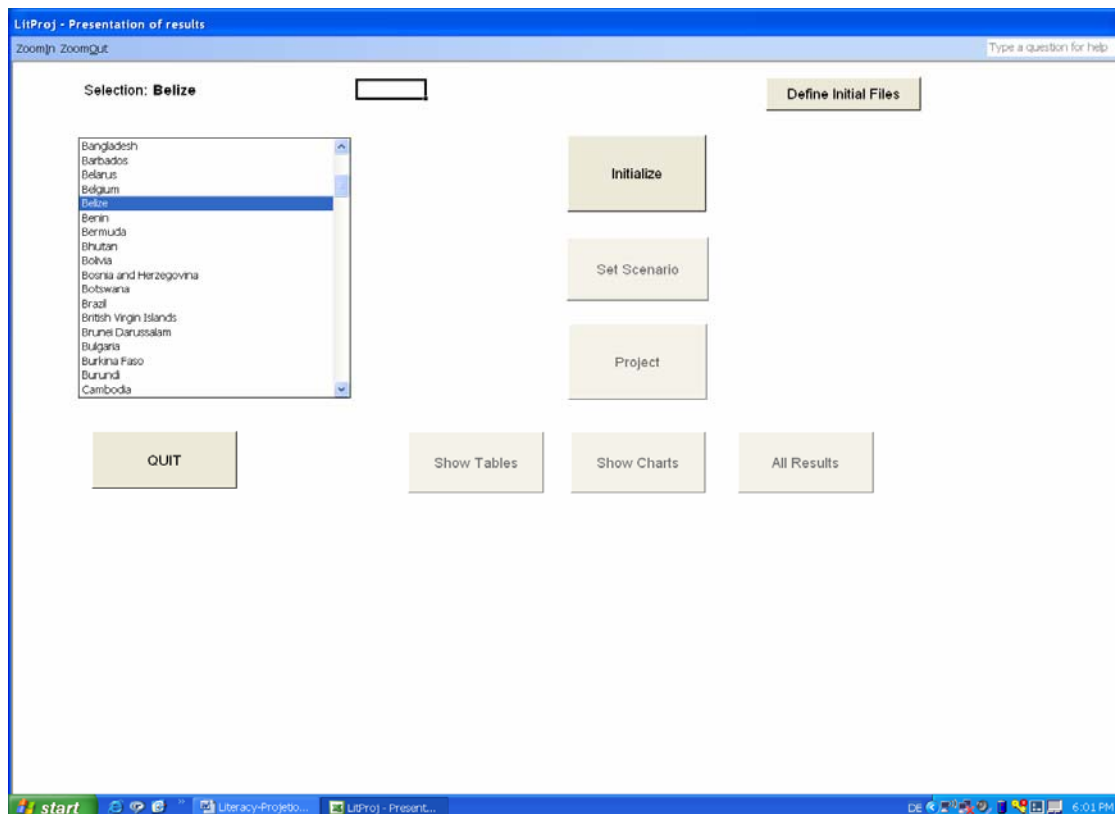


Figure 1: Screen shot from the main menu of the literacy projection program.

2. Initializing the model for the jump-off year

2.1 Empirical input required

The literacy projection model starts with the most recent information provided by countries. As requested by the UIS this information should include the population by literacy status, sex and age in five-year age groups starting with the age group 10-14. This information is then converted by UIS into standardized tables for the year of reporting, which includes both the age and sex specific proportions literate and the absolute numbers of literate men and women. These tables of literate and illiterate men and women, in standardized format, are used as the empirical input for the jump-off year of the projections.

The model is rather flexible in dealing with the following irregularities in the input tables:

- (a) For some countries, the age group of 10 to 14 years in the standard UIS tables has data and in other cases it is empty (gives “n” or no number). This is a function of what

countries reported. If this information is available, the data for this age group will be considered in the calculation that converts the five-year data into single year age groups (see below).

- (b) The model can deal with different open ended categories, such as the highest age group (as long as this category is marked with a “+” the model can recognize it whether it is “50+” or “95+” or anything in between).

Other irregularities in the standard input tables cannot be accommodated and the tables have to be corrected (manually) before entered into the model.

2.2 Transformation into single-year age groups

As discussed above, the model operates in single years of time in order to accommodate for the fact that the empirical input data give different jump-off years for different countries, while the target years of the projections should be uniform. This implies projections of different length for different countries, which requires single-year time steps in the projections and, as a consequence, also single years of age because the age and time intervals need to correspond.

This computational necessity of single-year steps in the projection mildly conflicts with the statistical principle of staying as close as possible to the empirically provided data because it implies the need for a transformation of the given five-year age groups into single-year age groups. Hence we chose a two-step graduation procedure that in the second step makes sure that the data re-aggregated to five-year age groups are identical to those given in the original empirical five-year age groups.

Figure 2 below illustrates this process in the example of female literacy in Guatemala for data reported to the UIS for 2002: The step-wise green line shows the empirically given (observed) data in terms of age-specific proportions literate for five-year age groups. A linear graduation (linear interpolation between mid-points of five-year intervals) of these data into single year steps then yields the smooth black line (diamonds called “graduated”). For the youngest ages 15 and 16, this procedure also considers the observed information for the age group 10 to 14, which makes a difference if there have been strong recent improvements in literacy. In the case of the few countries for which there is no empirical information on the 10 to 14 age group, the values for ages 15 and 16 are calculated by assuming constant proportions in the first step (for lack of other information), which are then adjusting as described below.

It is important to ensure perfect consistency of the constructed single-year age structure with the empirically based information for the five-year groups. If the graduated line is not straight, then a further iteration is necessary in order to be able to re-aggregate the five-year age groups to fully reproduce the original pattern. This will not be the case when the curve bends slightly upwards or downwards as a function of the literacy level in the adjacent five-year age group. In order to create a set of single-year data that makes it possible to move between the original five-year data and the new single-year age data, we further adjust the graduated data by re-aggregating them into five-year age groups. These are then compared to the original five-year data and a correction factor is calculated for each five-year age group that is then applied to all single-year ages within

the corresponding five-year age group. This results in the adjusted curve which is shown in red (squares) in Figure 1. This adjusted graduated curve, which in most cases tends to be rather close to the original graduated curve, then serves as the input to the projection model.

Observed, graduated and adjusted proportion of literate Females, Guatemala

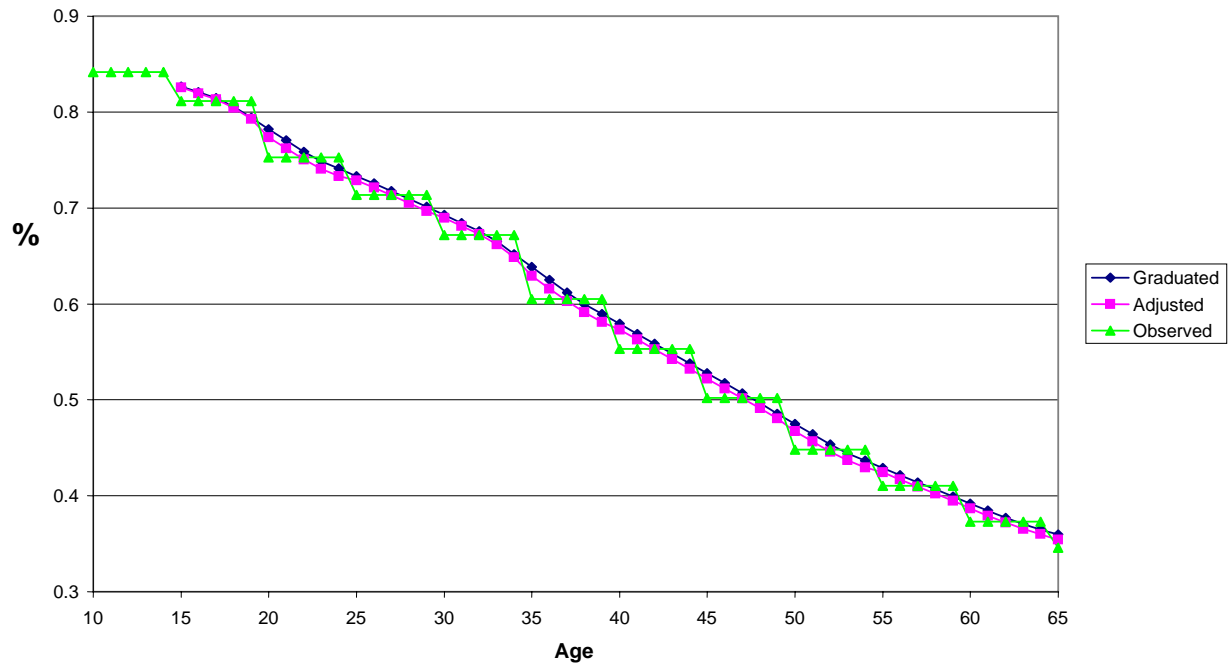


Figure 2: Illustration of graduation of five-year age groups into single-year age groups and adjustment to fit with original five-year age groups. Guatemala 2002.

2.3 Combination with UNPD population projections

To ensure the full compatibility of the literacy projections, it is essential to incorporate the UNPD population projections by age and sex.

In its current version, the model includes a population data base (stemming from the UNPD) of men and women by single years of age and for single years of time, from 1950 to 2050, for all countries of the world. It is based on the medium variant of the most recent (2004) assessment. When future assessments become available, the database can be updated accordingly.

It is important to be aware of the fact, however, that these data as provided by UNPD result from original projections carried out by five-year age groups and in five-year time steps through linearly interpolating over age and time. Hence the data do not cover sharp inter-cohort differences or other single-year irregularities as may be included in national data (e.g. of children by single years). While this may present an issue in the calculations of school enrollment rates it

is of little or no practical concern in the projection of adult literacy, which is the focus of this model.

At any stage of processing, all data referring to absolute numbers of men and women are generated by applying the proportions literate that are projected in this model to the single year population data as given by the UNPD. This assures that the aggregates (sum of literate and illiterate) fully correspond to the data as provided by the UNPD.

3 Scenario Setting

3.1 The assumption of invariant proportions literate after age 15

In principle, the model provides the user with a wide range of choices concerning the future proportions literate by age and sex. Yet in practical terms, it is useful to make assumptions that will reduce the number of these choices.

As discussed in Section 1.5, the projection for each country has to come up with more than 5,000 different proportions literate (around 30 years for 85 age groups and two sexes). Theoretically, the forecaster will have to come up with assumptions for all these values which are still unknown because they lie in the future and cannot yet be observed. To independently choose values for some 5,000 cells would represent a daunting and even silly task because it would assume full independence of these values. Fortunately these values are not all independent and some plausible assumptions can greatly reduce the number of parameter values. In this section, we discuss the assumption of having an entire cohort represented by one value. Sections 3.3 and 3.4 will examine further assumptions reducing the possible variation across cohorts.

There are important substantive reasons as well as overwhelming empirical evidence that literacy does not randomly fluctuate over the individual life cycle but shows great persistence. Substantively, literacy is a skill that once attained will continue to stay with a person for many years unless the person becomes mentally or physically disabled or forgets this skill over time for lack of practice. While the lack of practice leading to loss of literacy may be an issue for literacy skills when measured by actual assessments and multiple categories (people may have different test results at different points of their life), this is much less likely to be the case when projecting the dichotomous variable of self-defined literacy. Once a person has declared that he/she can read and write a simple sentence, it seems very unlikely that he/she would later claim to be illiterate again. Of course in actual measurements, there can be all kinds of misreporting and random fluctuations. However these fluctuations can go in both directions and hence have a tendency to cancel at the aggregate cohort level.

So in the typical case, self reported literacy of a whole cohort (a group of men or women born in the same year) tends to behave similar to the average height of a cohort. While at younger ages the average height increases substantially with age, at a certain point a maximum height is reached, which can then be expected to be maintained for many years (until possibly at very high ages people start to shrink again a bit due to curvature of the spine). Hence when charged to project the average height of the adult population in the future, it would be a rather safe choice to reduce the task to projecting for each cohort the height at the age when the maximum is reached

and thereafter assuming invariant height along cohort lines (leaving out, for the time being, the issue of shrinking at higher ages).

In the case of literacy, what would be the age at which the maximum literacy in a cohort is reached? Typically, literacy skills are acquired through formal education or other forms of learning in childhood and youth. There are of course also examples of adult literacy programmes particularly in socialist countries during the 1960s and 70s and rarely thereafter (UNESCO 2005). However, such programmes were more the exception than the rule in the past and there is no reason to assume that it will be much different in the future. Yet it also implies that the model should provide for the possibility of additional transitions to literacy at certain adult ages in exceptional cases where there is reason to assume such programmes.

Aside from the above mentioned possibility, it seems safe to assume that the typical case of literacy being acquired during childhood leads to a peak in cohort literacy before the age of 15, (e.g. typically in the 10 to 14 age group). Strong empirical support for this assertion is provided by the empirical patterns of age-specific literacy rates as reported by countries in five-year age groups. Here many countries show a pattern similar to that of Guatemala given in Figure 1 where young men and women in the 10 to 14 age group tend to even have higher literacy rates than in the 15 to 19 age group.

Hence in the literacy projection model the age of 15 is chosen as the age at which—unless specified otherwise—a cohort has reached its maximum literacy level, which is then kept constant throughout the life of that cohort. This generally plausible assumption leads to a very significant reduction in the number of parameters for which scenario assumptions have to be made. As explained in Section 1.5, it reduces the more than 5000 parameters to about 60 for which values have to be chosen characterizing 30 male and 30 female cohorts that reach the age of 15 during the projection period from around 2000 to 2030.

3.2 The “constant transition to literacy” scenario

However choosing 60 parameters characterizing the individual male and female cohorts may still be a difficult task. To further reduce the complexity, we can also utilize the fact that literacy does not jump or, in more technical terms, that the serial correlation among literacy rates of subsequent cohorts is very high.

The most extreme reduction of complexity is to simply assume that the proportions literate at age 15 will stay constant at the level of the cohort that was aged 15 in the jump-off year. That is to say,

$$\text{Prop Lit (f, 15, t)} = \text{Prop Lit (f, 15, t+1)} = \text{Prop Lit (f, 15, t+2)} = \dots$$

where t is the jump-off year.

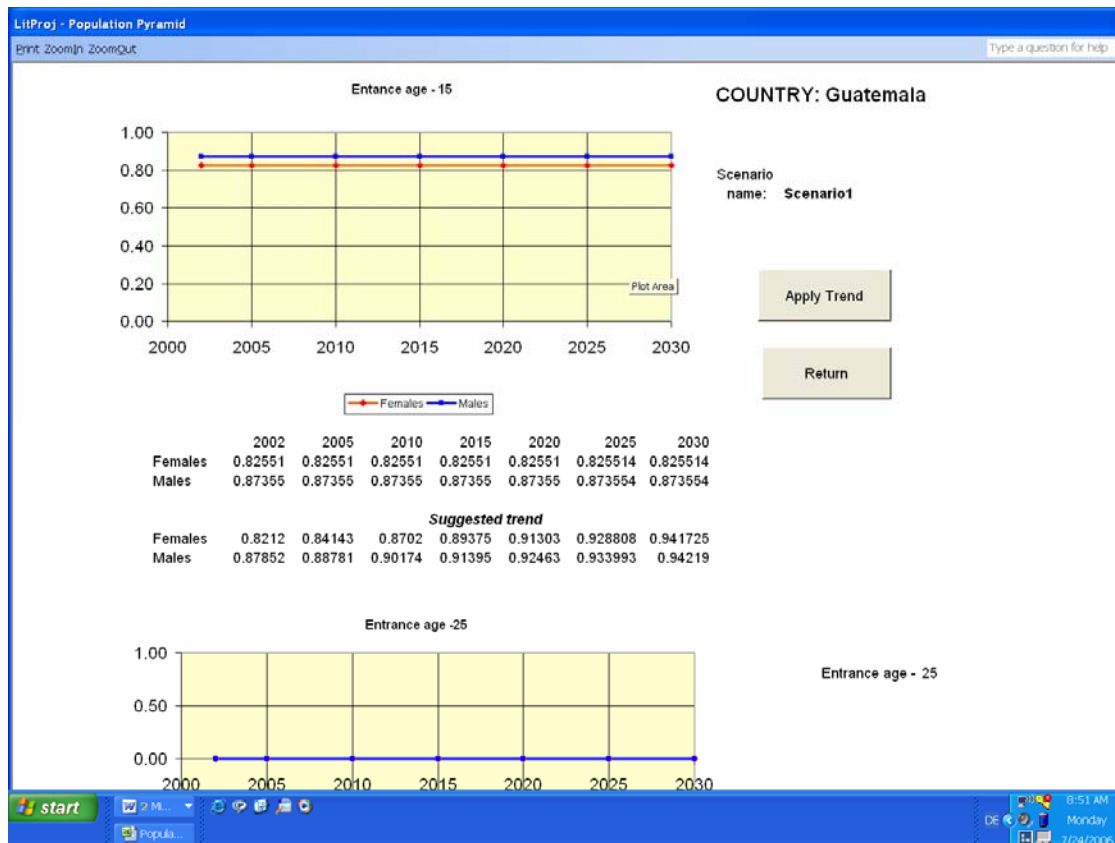


Figure 3: Screen shot from the scenario setting module of the software showing the default constant scenario setting. The values give the proportions literate at age 15 in the year indicated (in this case 2003 is the jump-off year). All values for the subsequent years are set equal to the proportions in the jump-off year. The graph above always displays graphically the numbers set in the corresponding table.

In the case of this strong assumption the user does not have to choose any parameter values for the future because the Prop Lit's at time t are empirically given and everything else is determined by this assumption together with the assumption of constancy along cohort lines. The number of parameters to be chosen hence comes down from over 5000 to zero. Substantively, this choice implies the assumption of "constant transition to literacy" over the coming years, which in most cases (except for countries that already have universal literacy) is an unrealistically pessimistic assumption. However it can serve an interesting analytical purpose.

Comparing the results of a "constant transition to literacy" scenario to the literacy pattern in the jump-off year illustrates the important effect of compositional momentum. It shows by how much the literacy of the adult population would increase, if no further improvements in the transition rates to literacy are being made. In many countries which have experienced increasing proportions of youth literacy over the past decades, this structural momentum (which is simply caused by the fact that illiterate elderly will slowly die off and the more literate younger cohorts will move up in the age pyramid) will be very high. Actually, in most countries much of the

projected gain in adult literacy until 2015 is driven by this structural momentum rather than by the specific assumptions made on future proportions literate of the youngest cohorts. Hence this constant scenario can shed new light on our understanding of the dynamics of improving adult literacy.

In the software, the scenario is predefined with all values set to the same level as in the jump-off year as a default, once the screen is opened. By clicking on the “Apply Trend” button, the other predefined scenario (i.e. the “trend scenario” described in the next section) will be applied (see screen shot in Figure 5). This trend scenario is the prime candidate to be considered as a plausible forecast.

3.3 The trend scenario

The trend scenario reduces complexity not by keeping the proportions literate constant along cohort lines but by assuming that the trend of proportions literate from one cohort to the next follows a smooth pattern that can be described as a linear function of the logit of literacy. Typically, the trend has been improving over time and therefore such a scenario assumes further improvements along the extrapolated curve but with the increments being reduced and 100% literacy never fully reached. If there has been a recent stagnation or even deterioration in youth literacy (as has been the case in some African countries), this trend scenario projects continued decline along the same declining trend. In addition, male and female literacy trends are assessed completely independently, which is clearly important in situations where the one increase while the other declines.

In a more technical sense, the regression line is fitted according to

$$\text{Logit}(y) = ax + b$$

where y is the proportion of literates

x is age (or year in which corresponding cohort was at age 15).

The advantage of this function is that it slowly approaches but never fully reaches 100%. The line is fitted through 16 data points given by the proportions literate of the cohorts aged 15-30 in the jump-off year. Figure 4 illustrates this procedure for the literacy rates at the age of 15 for females in Guatemala.

Let us accept that the previous assumption about the constancy of proportions literate along cohort lines holds. Then the proportions literate of women aged 16-30 in 2003 (the year of observation) directly reflect the proportions literate at age 15 in the (unobserved) years 1988-2002. That is to say:

$$\text{Prop Lit}(f, 16, 2003) = \text{Prop Lit}(f, 15, 2002)$$

$$\text{Prop Lit}(f, 17, 2003) = \text{Prop Lit}(f, 15, 2001)$$

$$\text{Prop Lit}(f, 18, 2003) = \text{Prop Lit}(f, 15, 2000)$$

...

$$\text{Prop Lit}(f, 28, 2003) = \text{Prop Lit}(f, 15, 1990)$$

Prop Lit (f, 29, 2003) = Prop Lit (f, 15, 1989)
 Prop Lit (f, 30, 2003) = Prop Lit (f, 15, 1988).

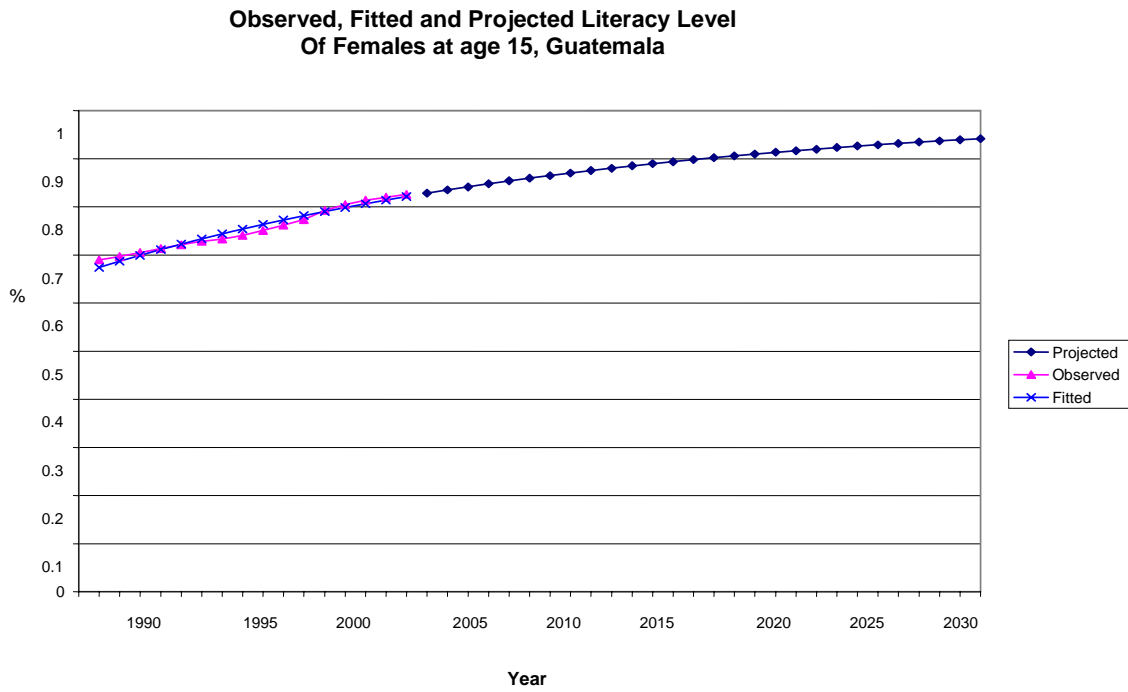


Figure 4: Calculation of “Trend Scenario”: the future values of those aged 15 in the stated year are derived from linear extrapolation along a line fitted to the logits of the literacy rates in the 16 most recent observed data (youngest cohorts in jump-off year).

Hence this assumption of constancy along cohort lines directly allows the translation of cross-sectional observed data into longitudinal data. In effect 16 observed data points for a single year (2003) are used as proxy measurements for 15-year-olds across 16 years (cohorts). As compared to real longitudinal data, this procedure has the advantage of being immune to the frequent problem of changes in measurements over time. Nevertheless it is still useful to study the data of several time points (if available), not least to evaluate the assumption of constancy along cohort lines in specific historical settings (see also Section 3.5).

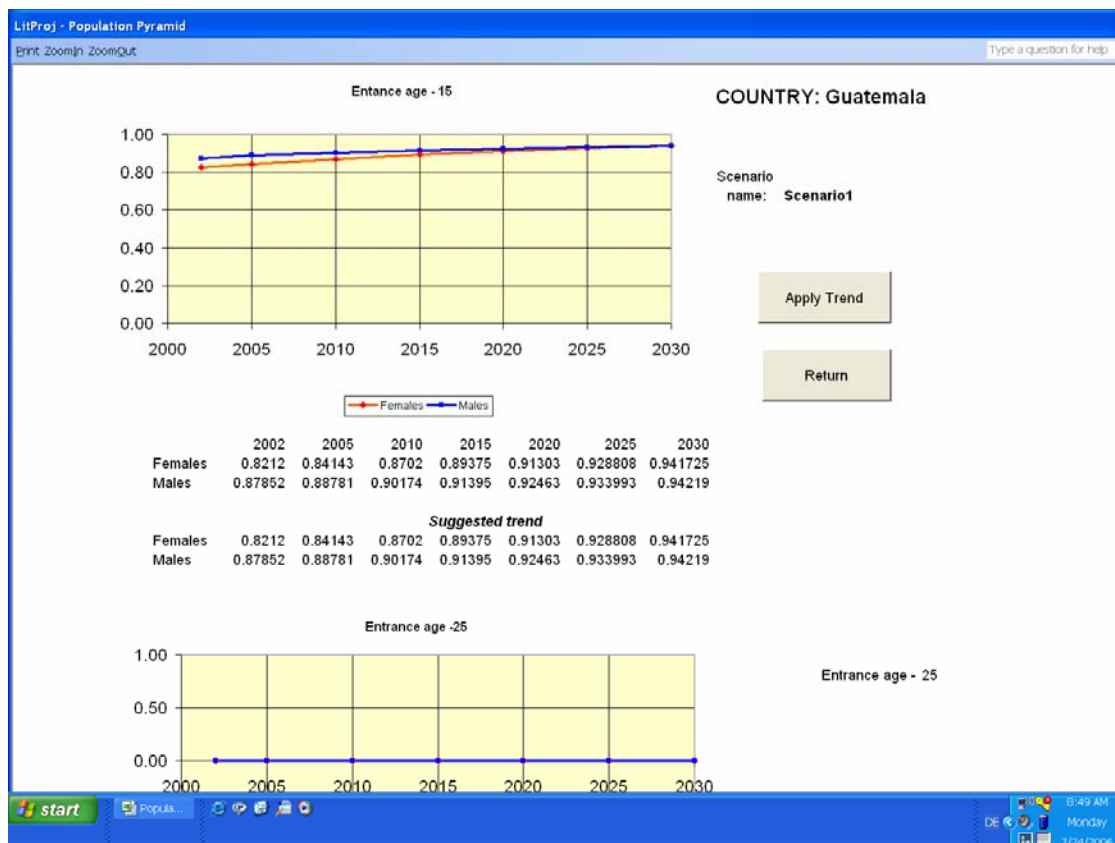


Figure 5: Screen shot from the scenario setting module for the trend scenario. This scenario has been activated by clicking on the “Apply Trend” button.

3.4 Specific scenarios

In addition to the trend scenario and the “constant transition to literacy” scenario described above, the user may still want to choose other specific scenarios. Such scenarios may either be deemed desirable because there are reasons to assume that the trends, as extrapolated by the default trend scenario, are not the most realistic path for the future. Alternatively, the user may want to explore the hypothetical achievement of certain politically defined goals.

There are many cases in which the future trend in proportions literate, as previously described, may not seem appropriate for the best guess scenario. One example is the case depicted in Figure 3, where further improvements in female literacy are expected while male literacy has been stagnant or even declined minimally. This would result in close to a constant scenario assumption for men and a greatly diverging trend between men and women. What should the user of the model do in such a case? One option is to mechanically apply the trend assumptions as produced by the default scenario setter and forget about plausibility for the sake of consistency in applying a set of chosen rules. The other option is to have a closer look at the country concerned and see whether there are problems with the data or reasons why over the past years male transitions to literacy have done so badly in comparison to those of females. It is then

possible to make an informed judgment about whether these trends are likely to continue in the future or whether there is reason to assume that over the coming years young men will start to catch up with women again. But even in cases when the trends are in principle plausible, there the government may have stated plans that will affect the future transitions to literacy in a certain way that should be reflected in the projections. In all these cases, it makes sense to define country specific scenarios that are different from what is suggested by the trend.

When trying to assess the effect of national or international literacy goals, it is important to recognize that the scenarios can only be set in terms of the input (proportions literate at age 15 or other specified age) rather than aggregate output measures for the entire adult population. In this sense the goals of universal primary education and closing the gender gap in education can be operationalized in terms of scenarios assuming a convergence of male and female proportions literate at the age of 15 to close to 100% by the target year set. The results then show how the achievement of this goal in terms of “inputs” will subsequently slowly increase the proportion of literates among the total adult population. The goal of halving illiteracy by 2015, however, is defined in terms of output and therefore can only be explored in an iterative manner: one can assume certain rapid increases in the literacy at age 15 and then look at the output and see how this affects the change in the proportion of literates among the total adult population. If this more than halves the illiteracy of the adults, one can try next to specify a milder increase in literacy at age 15 and vice versa. As stated in the introduction, this model can also show that the EFA goal of halving illiteracy is simply unachievable by 2015 through youth education alone. In this case, the model yields a politically important finding, suggesting that either the goal needs to be modified or that youth education should be complemented quickly by adult education. The model can also help to estimate the extent of additional adult education needed. This will be discussed in Section 4.3 in the case of India.

3.5 How to incorporate other information and assume additional transitions after the age of 15

The previous sections discussed several cases in which it is useful to deviate from the default trend scenario. Here we will discuss more such cases and consider various options to relax the otherwise central assumption of constancy of literacy along cohort lines.

For many countries more than one data point in time exists with age- and sex-specific literacy information. This led us to ask whether this information could be meaningfully incorporated into the projection model thus enriching its empirical basis. The answer is “yes” generally but not automatically. First it is important to see that additional data from earlier points in time are crucial for understanding the past trend. However, they are of only limited importance when it comes to making assumptions about the likely future trend. It is for this reason that the current methodology limits itself to projections based on data from a single point in time rather than using two time points from censuses or surveys, which may be spaced a decade apart.

Nevertheless, there are still meaningful uses of this historical information. Most importantly, they may provide a basis for empirically assessing the validity of the constancy along cohort lines assumptions for the past experience of a specific country. This can be done in a descriptive manner (which looks at whether the same cohorts actually show the same proportions literate at

different points in time) or by applying the multivariate Age-Period-Cohort models discussed in Section 1.3. This latter use is particularly powerful in the case of information from a larger number of time points because it also allows for statistical testing of the validity of the constancy assumption. However, few countries have the necessary data.

We shall therefore consider the more common case in which a specific country has two data points (e.g. from a census in 1991 and a survey in 1998). If the data do indeed show nearly constant proportions literate along cohort lines (i.e. comparing people aged x in 1991 to people aged $x+7$ in 1998) then the constancy assumption is confirmed and there is no further need to consider 1991 information in the projections. But if there are some significant deviations in either direction, one must try to find the reason. The reason could be as simple as not comparing identical cohorts because the data is only given in five-year age groups and the difference between the data points is seven years. But even if this problem can be solved the difference may still be due to different data collection instruments (census versus survey), slightly different wordings of the questions asked or slightly differing coverage of two surveys. Even in the rare case when all these problems can be ruled out and a real change in the number of literates of an adult cohort can be established, further work is required to find the reasons before being able to utilize this information from the past for specific assumptions concerning future trends. What are, for example, the future implications if the increased proportion within the cohort between two data points can be attributed to a specific adult education programme? Does this suggest that we should assume equally effective recurring adult education programs at certain intervals in the future? This is unlikely. But what else should we do with the information? The answer will require substantive expert knowledge and will vary from one country to another. There is no automatic response.

The only case where the consideration of additional past data points would have real impact on the specification of the projections would be in the hypothetical case that the analysis along cohort lines shows pervasive and consistent deviations across countries from the constancy assumption, with either literacy increasing or decreasing in a systematic manner along cohort lines. Although there is no empirical evidence in this direction, the model does provide the option of both adding or subtracting literate men and women to the cohorts at any future point in time and any age chosen. This can simply be done by entering the assumed numbers into the lowest panel on the scenario setting page. Even if this feature is unlikely to be often used (given that it is often difficult to translate the impact of specific country policies planned for the future into literacy proportions for any given cohort at a point in time), it adds to the flexibility of the model. It allows the user to produce better and more realistic projections if there is good reason to assume that literacy in one or several countries will increase or decrease along cohort lines.

4. Presentation of results

4.1 Numerical results

The model provides some standardized output for every scenario for each country. As shown in Figure 5, this output gives the tables for any year as specified in the top section of the screen (use blue arrows for choice of year). These years go in single year steps from the jump-off year to 2030. This standard output screen shows four tables. Table 1 gives the proportions literate for all

five-year age groups for men and women as well as for both sexes combined. Table 2 provides the same information in terms of absolute numbers of literate and illiterate men and women.

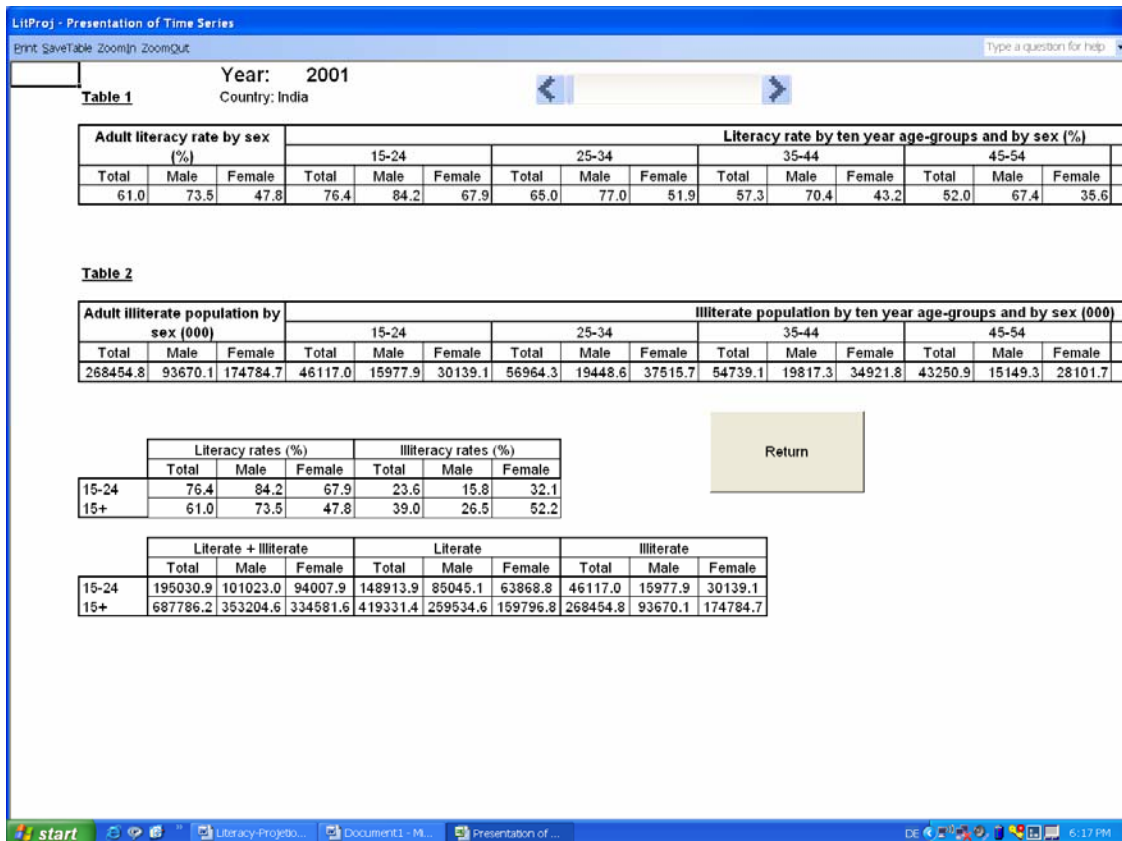


Figure 6: Screen shot from the “Show Tables” presentation of results: the example of India with the jump-off year 2001.

In the lower part of the results table, the same information is given in terms of the two aggregate age groups which are routinely used in UIS and EFA reporting, namely 15-24 for youth literacy and 15+ for adult literacy. For both of these age groups the information is given as proportions and in absolute numbers for men, women and both sexes combined.

4.2 Graphical output

The same information presented in tabular form is also given in graphical form if the user chooses the “Show Charts” option on the main screen. As with the tables, the blue arrows on top can be used to select the specific year for which the output shall be shown. The pie charts in the first line of the screen show results for the two broad age groups 15+ on the left and 15-24 on the right, in both cases for men and women separately. This kind of presentation makes it visually very clear that in India; more than half of all adult women and about one-quarter of all adult men are still illiterate in the jump-off year 2001. The picture improves for youth literacy, with only

about one-third of young women and 16% of young men still illiterate. This clearly shows that literacy has been improving for younger cohorts.



Figure 7: Screen shot from the “Show Graphs” presentation of results for India 2001.

The lower part of the screen then shows the results in bar charts for 10-year age groups of men and women. The graph on the left gives the proportions illiterate for each age group and sex and the right graph gives the proportions literate. This graph is particularly illuminating as one moves from one year to the next, with the bars on the left shrinking and those on the right increasing over time.

Figure 8 gives the same graphical output for the end year of the projection, 2030, based on the assumptions of the trend scenario with additional female education efforts as discussed below. A comparison of the two graphs clearly shows the projected progress by 2030. Female adult illiteracy is now just slightly above one-quarter of the population or half the 2001 level. This also shows that even in this optimistic scenario for India, the EFA goal of halving illiteracy rates by 2015 will only be reached in 2030. This is also true for men. The lower panel with the age structure of literacy helps to illustrate the reason for this slow improvement: namely that it takes a long time for the younger better educated cohorts to replace the older less literate cohorts.



Figure 8: Screen shot from the “Show Graphs” presentation of results for India 2030, based on the trend scenario with extra female adult education (see definition in next section)

4.3 Comparison of scenario results

In this final section, we will demonstrate how the literacy projection model can be used to evaluate the feasibility of literacy goals and to calculate the measures required to try to meet these goals. We will use the case of female literacy in India where the data for the jump-off year and projections to 2030 were discussed above.

The first scenario calculated for India is the “constant transition to literacy” scenario. In substantive terms, this implies that all future cohorts entering the model at age 15 will be literate to the same extent as the cohort of 15-year-old women in the jump-off year 2001. In this cohort, 75% of women aged 15 are literate. This proportion will be applied to all younger cohorts while for the older cohorts the empirically given (lower) proportions will apply (as shown in Figure 5). In the software, the choice of this constant scenario is extremely simple, because it corresponds to the default value given for each year when opening the “Set Scenario” screen. One simply has to hit the “return” button to run the projection.

The second scenario is the trend scenario according to the procedures discussed extensively in Section 3. It assumes that the trend of improvement that observed for the last 15 cohorts will

continue in the future and gradually increase the proportions literate of new cohorts entering the model at age 15. In the software, this is also a predefined scenario: by hitting the “Apply Trend” button, the trend will immediately be applied to the incoming cohorts of both men and women, who will be also shown on the charts depicting the assumed transition to literacy values (see upper chart of Figure 9).

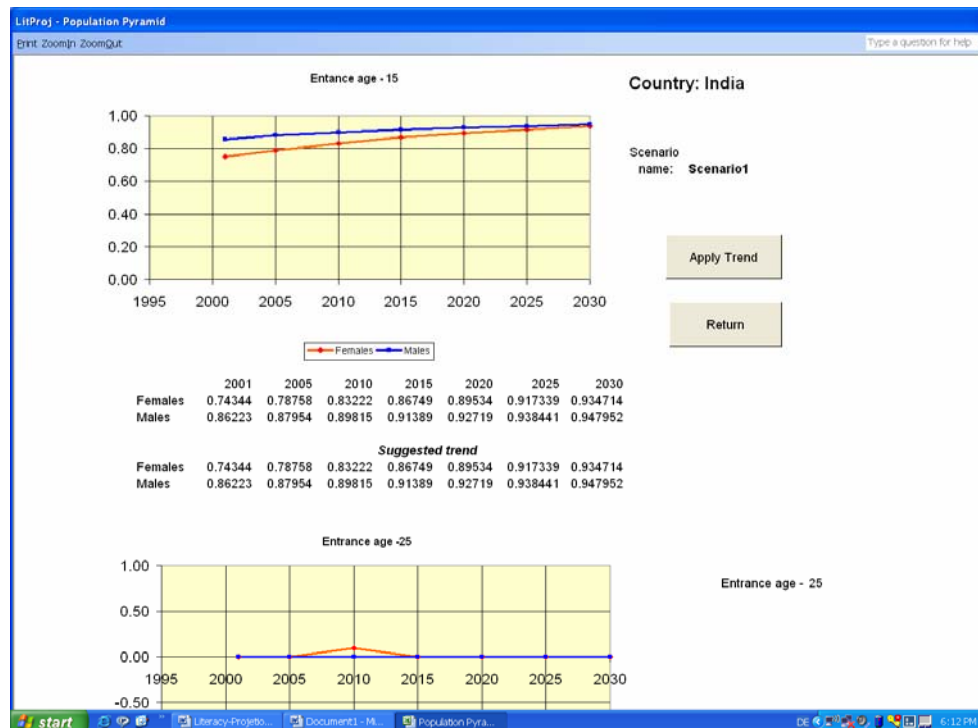


Figure 9: Screen shot of scenario setting page for trends scenario plus adult literacy efforts.

The third scenario is shown in Figure 8. It is based on the trend scenario and, in addition, assumes that in about 2010, India will also make a strong effort to educate another 10% of all women at the age of 25. This is roughly equivalent to halving the proportion of women that will still be illiterate in this age group. It is quite unusual to find adult literacy campaigns launched on top of an ongoing expansion of youth school enrolment. Nevertheless, we defined this scenario after anticipating that the improving trend for youngsters would not be sufficient to reach the goal of halving adult illiteracy by 2015.

The results of these three scenarios are quite sobering: in the jump-off year 2001, India has 174 million illiterate adult women (15+) which imply an adult illiteracy rate of 52%. Under the constant scenario, the illiteracy rate will fall to 41% by 2015 while the absolute number of illiterate adult women will actually increase to 183 million because of population growth. The trend scenario alone will bring the illiteracy rate further down to 39% with the absolute number of illiterate women being 172 million, roughly the same as today. Only the third scenario combining the trend with an adult literacy campaign will bring the absolute number down to 166

million and the illiteracy rate to 37% by 2015. All three scenarios fall far short of reaching the EFA goal of halving illiteracy by 2015. As discussed above, the third scenario will halve the illiteracy rate by 2030, twice the time foreseen in the EFA goals.



Figure 10: Graphical output for 2015 for Scenario 4 (100% transition to literacy for women as of 2005, "constant transition to literacy" for men).

The model can also be used to try to find out whether, at least theoretically, there is any way for India to still meet the EFA goals. For this, we defined another four overly-optimistic scenarios, which assume progressively stronger efforts. In scenario 4 we assume that from 2005 onwards, all women (100%) at the age of 15 will be literate. By 2015, this scenario brings female adult illiteracy down to 33%, still short of the goal that would imply a reduction to 26%. Therefore, the next scenario assumes that the adult literacy campaign results in 100% percent literacy of incoming cohorts, which reduces the rate to 31%. If the adult education campaign is further strengthened so that all remaining illiterate women at age 25 are literate and is carried out continuously over all projection years, then the illiteracy rate comes down to 27% by 2015, still not quite half of the 2001 level. Hence, the EFA goal can only be reached if the full campaign was extended to older women. This exercise dramatically demonstrates the great momentum of improvements in adult literacy and the need to set realistic goals based on clear assessments of what can possibly be achieved.

This type of highly policy-relevant consideration is not possible without the literacy projection model discussed in this paper.

References:

Keyfitz, Nathan. 1985: Applied Mathematical Demography. 2nd edition. New York: Stringer Verlag.

Lutz, Wolfgang and Anne Goujon. 2001: The World's Changing Human Capital Stock: Multi-state Population Projections by Educational Attainment. Population and Development Review, 27 (2):323-339 (June 2001).

Rogers, Andrei. 1975: Introduction to Multi-regional Mathematical Demography. New York. John Wiley.

UNESCO. 1995: Methodology Used in the 1994 Estimation and Projection of Adult Literacy, Statistical Issues, Division of Statistics, STE-18, Paris, June 1995.

UNESCO. 2005: Literacy for life. EFA Global Monitoring Report, 2006. UNESCO, Paris, 2005.

United Nations Population Division. 2005. World Population Prospects: The 2004 Revision. New York