INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

AN INVESTIGATION INTO THE IMPACT OF HIV ON POPULATION DYNAMICS IN AFRICA

Samuel J. Clark

A DISSERTATION

in

Demography

Presented to the Faculties of the University of Pennsylvania in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

2001

Approved by: Samuel H Preston

Dissertation Supervisor: Samuel H. Preston

Chair, Graduate Group in Demography: Herbert L. Smith

To

Clarissa, Sophia & Paula

&

The Gwembe Tonga residents of:

Mazulu Village
Musulumba Village
Siameja Village
Sinafala Village

in Gwembe District, Southern Zambia

ACKNOWLEDGMENTS

I am greatly indebted to the 15,000 Gwembe Tonga who patiently contributed their time and personal information toward the creation of the data set that is the basis of this dissertation.

Making it all possible from the beginning to the end has been the constant support and encouragement of my wife Clarissa. I wish to thank both her and my daughters Sophia and Paula for prevailing with me!

I am grateful to Elizabeth Colson and Thayer Scudder who provided the data and expertise pertaining to the Gwembe Tonga. I am thankful to Jane Menken and Tukufu Zuberi who have contributed to the thrust and quality of this work. James Lee introduced me to Demography and encouraged me to pursue this dissertation, and I have valued his support on many levels.

This work has been made possible through the generous support of the National Science Foundation of the United States, the William Penn Foundation, the Fulbright International Scholars Program of the United States, the Mellon Foundation, and the National Institute on Aging of the United States (grant number R37 AG10168). James Lee, Thayer Scudder, Elizabeth Colson and my parents Andrew and Barbara Clark have also provided material support that has been invaluable.

Tom and Thea Savory played a crucial role in facilitating my work by providing a warm, friendly and reliable base of operations in Zambia. More important; however, I greatly appreciate the friendship they have offered my family and me.

Douglas Ewbank provided the support necessary for me to attend a summer workshop on dynamical systems and population biology. My experience there is directly responsible for generating the ideas that led me to build the population simulator that forms a major part of this dissertation.

I have pursued this work primarily under my own direction and on a time scale that has been unusually long for a student in the Demography Group. I am certain that there were times when my primary advisor Samuel Preston wondered where it was going and when it would be complete. I am grateful to him for allowing me the freedom and time to complete what I feel is a satisfying dissertation. Even after becoming Dean of the School of Arts and Sciences, he continued to make time to see me and provide insightful commentary on my work. Without his generous support on many levels this dissertation would not have been possible. Thank you.

ABSTRACT

AN INVESTIGATION INTO THE IMPACT OF
HIV ON POPULATION DYNAMICS IN
AFRICA

Samuel J. Clark

Supervisor: Professor Samuel H. Preston

Data from eighteen sites in Africa are used to identify new mortality patterns for Africa; two of which may result from excess mortality caused by HIV/AIDS. To examine the underlying population processes that produce those patterns, and to understand how HIV/AIDS affects a population *as a whole*, an individual-level, stochastic, computer simulation of an African population infected with HIV is constructed to study the dynamics of an HIV epidemic in Africa. The simulator is tested and demonstrated by constructing uninfected, untreated-infected, and treated-infected virtual populations.

The fundamental demographic parameters used to run the model are calculated from 40 year's worth of demographic data collected from the Gwembe Tonga of Southern Zambia. Data on nuptiality, fertility and mortality are used to simulate the population dynamics of a rapidly growing polygynous population. Added to that is a fully parametric HIV module that governs the transmission and progression of HIV/AIDS within individuals in the population.

Complementing the design and implementation of the computer simulation of a population with HIV, this work also discusses the theoretical basis for an information management system designed to manage the collection, manipulation and retrieval of longitudinal data – the Structured Population Event History Register (SPEHR).

TABLE OF CONTENTS

PART 1 1
***INDEPTH* MORTALITY PATTENS FOR AFRICA** 1
MORTALITY MODELS AND AFRICA 1
Mortality Data 3
PRINCIPAL COMPONENTS ANALYSIS 6
Data Summary 6
Component Model of Mortality 6
PRINCIPAL COMPONENTS OF *INDEPTH* MORTALITY DATA 11
Male 12
Female 14
Male and Female Principal Components Contrasted 16
***INDEPTH* MORTALITY PATTERNS** 21
Component Clustering Method 21
Clusters 23
Mortality Patterns 25
Pattern 1 31
Pattern 2 32
Pattern 3 32
Pattern 4 33
Pattern 5 33
Pattern 6 34
Pattern 7 34
Compared to the Coale and Demeny and United Nations Model Life Tables 35
DEMONSTRATION OF HIV/AIDS MODEL LIFE TABLE SYSTEM 39
Model Life Table Construction 39
Model Life Tables 44
CONCLUSION 51
PART 2 52
THE DEMOGRAPHY OF THE GWEMBE TONGA 52
THE GWEMBE TONGA 53
The Gwembe Valley 53
Social Organization 56
Recent History 59
THE GWEMBE TONGA RESEARCH PROJECT - GTRP 62
History 62
Organization of the GTRP 66
THE STUDY POPULATION 69
Sample 69
Data 70
People Table 84
Unions Table 85

Links Table 85
Size and Composition 85
Population Pyramid 86
Population Size, Growth Rates, and Sex Ratio by Year 86
Age Composition of Male Population by Year 91
Age Composition of Female Population by Year 93
Sex Ratio by Age by Year 95
METHOD OF ANALYSIS 96
Discrete Time Event History Analysis 96
Analysis Files 98
Person-Year File 99
Union-Year File 100
Unmarried Union-Year File 101
DEMOGRAPHIC CHARACTERISTICS 105
Nuptiality 105
Union Formation 106
Levels 106
Trends 118
Union Dissolution 122
Levels 122
Trends 131
Fertility 133
Levels 134
Trends 147
Evidence of HIV 154
Mortality 170
Levels 170
Trends 175
Evidence of HIV 192
PART 3 199
A TWO-SEX STOCHASTIC MICROSIMULATION OF A POPULATION WITH HIV 199
BACKGROUND 199
Models of Human Population 199
Classification of Population Models 202
Macro and Micro Models 203
Macro Deterministic Models 203
Examples of Macro Models Addressing HIV/AIDS 205
Heuveline 206
IWG – The United States Interagency Working Group on AIDS Models and Methods 207
Anderson, Garnett & Gregson 210
John 213
Palloni 214
Bulatao 216
Bongaarts 218
Brouard 218

Lessons from Macro Models 219
Micro Stochastic Models 220
Examples of Micro Models Addressing HIV/AIDS 223
Wachter & Hammel – SOCSIM 223
Auvert – SimulAIDS 226
van der Ploeg – STDSIM 228
Chick, Adams & Koopman 231
HIV Models – Macro or Micro? 232
AIM – WHAT DOES THE SIMULATOR DO? **235**
ARCHITECTURE **240**
Relational Database Implementation 241
Data Representation 242
Record of History 246
Individual-Level 247
Population-Level 251
Results 251
Modular Construction & Extensability 252
Essential Modules 252
Adding a Module 253
Accounting for Parameters and Time 253
Granularity of Time-Step 254
Time Varying Parameters 255
Odds & Odds Ratios 255
Simulation of Intercourse Events 256
Reproduction 256
Transmission of HIV 257
Complete Linkage of Reproduction and Transmission of HIV 257
MORTALITY MODULE **258**
Logic 258
Empirical Mortality Hazards 260
Parameter Structure 263
Desirable additions 264
INTERCOURSE MODULE **265**
logic 265
Empirical Intercourse Probabilities 265
Method 266
Calculation of fecundability: f 267
Calculation of the average number of intercourse events per month: n 269
Parameter Structure 274
FERTILITY MODULE **275**
Logic 275
Fecundability and Conception 276
Married & Unmarried Fertility 278
Reproductive Histories 278
Contraception 279
Fetal Wastage 279

Vertical Transmission of HIV 281
Parameter Structure 282
Desirable Additions 283
HIV MODULE 285
Logic 287
Incubation & Duration Dependence 288
Transmission 289
Empirical Transmission Probabilities 290
Intercourse Events & Transmission 291
Attenuation of Transmission with Condoms 292
Anti Retroviral Treatment 293
Random Infection 293
Parameter Structure 294
Desirable Additions 295
NUPTIALITY MODULE 297
Formation of Unions 298
Logic 298
Unit of Analysis and Union Formation Probabilities 299
Polygynous Marriage 303
Empirical Hazards of Union Formation 304
Parameter Structure 306
Dissolution of Unions 307
Logic 307
Sample, Unit of Analysis, and Hazard of Separation 307
Empirical Union Dissolution Probabilities 309
Parameter Structure 312
AFFAIRS MODULE 313
Theoretical Dynamics of Affairs 313
Sexual Activity Indices 315
Sexual Activity Associativity 315
Age Associativity 315
Formation of Affairs 316
Logic 316
Sample Structure and Unit of Analysis 316
Affair Formation Probabilities 317
Sexual Intercourse within Affairs 322
Dissolution of Affairs 323
Summary of Affairs 324
MODULE INTERACTIONS & ASSOCIATIONS 326
Implicit Interactions 326
Explicit Interactions 327
HIV 327
Nuptiality 327
Affairs 330
Mortality & Morbidity 330
Fertility 331

Union Formation & Dissolution 332
Intercourse 332
Fecundity 333
Fetal Wastage 334
Vertical Transmission 334
Parameter Values Used for HIV Interactions 336
SIMULATION OUTPUT 337
Calibration – Stable Population 338
Stable Population Size, Growth and Composition 339
Stable Population Mortality Rates 342
Stable Population Fertility Rates 344
HIV Simulations 346
Population Growth 349
Mortality 350
Fertility 351
AIDS 352
Incidence 353
Prevalence 354
New Cases 356
Summary of Simulation Output 357
PART 4 359
THE STRUCTURED POPULATION EVENT HISTORY REGISTER – S.P.E.H.R. 359
SPEHR CONCEPT 359
DESIGN OBJECTIVES 362
Standards and Comparability 362
Flexibility and Extensibility 363
Self-Documentation 365
Low Maintenance 366
Security 367
Validity 369
Analysis-Friendly 370
Genealogies 372
SPEHR DATA MODEL 373
Events 376
Item-Episodes 376
Experiences 376
Attributes 379
Shared Experiences 380
Multi-event Processes 380
Meta Data 381
SPEHR Demonstration 381
APPENDIX A 383
THE RELATIONAL MODEL OF DATA 383
RELATIONS 384
OPERATORS 385
CONSTRAINTS 387

APPENDIX B **388**
STRUCTURED QUERY LANGUAGE - SQL **388**
DDL - DATA DEFINITION LANGUAGE **389**
Create Table Example 389
Modify Table Example 390
Drop Table Example 390
DML - DATA MANIPULATION LANGUAGE **391**
Insert Example 391
Update Example 392
Delete Example 393
SELECT and UNION Examples 393
TRANSACTIONS **397**
SQL IN PRACTICAL USE **398**
BIBLIOGRAPHY **399**
INDEX **404**

LIST OF TABLES

Number *Page*

Table 1: *INDEPTH* Mortality Data 4
Table 2: Site-Periods 5
Table 3: First Fifteen Principal Components 13
Table 4: First Fifteen Principal Components 15
Table 5: R^2 Values from Linear Regressions of Existing Model Mortality Patterns on the *INDEPTH* Components 20
Table 6: *INDEPTH* Mortality Age-Pattern Clusters 23
Table 7: *INDEPTH* Mortality Patterns 26
Table 8: Sum of Squared Differences Comparing *INDEPTH* and Existing Mortality Patterns 36
Table 9: Coefficient Values Estimated in Fit of 43
Table 10: Model Life Tables: *INDEPTH* Pattern 1 47
Table 11: Model Life Tables: *INDEPTH* Pattern 1 48
Table 12: Model Life Tables: *INDEPTH* Pattern 1 49
Table 13: Model Life Tables: *INDEPTH* Pattern 1 50
Table 14: Sample of Raw Data Set: 71
Table 15: Records in the People Table 76
Table 16: Records in the Unions Table 78
Table 17: Records in the Links Table 78
Table 18: Sample of Query on Analysis Data Set: 81
Table 19: Description of the People Table 84
Table 20: Description of the Unions Table 85
Table 21: Description of the Links Table 85
Table 22: Size and Sex Composition of Analysis Sample 88
Table 23: Size and Growth of Analysis Sample: Male 89
Table 24: Size and Growth of Analysis Sample: Female 90
Table 25: Male Sample Population by Age and Year 91
Table 26: Male Sample Proportion by Age and Year (percent) 92
Table 27: Female Sample Population by Age and Year 93
Table 28: Description of the Person-Year Table 100
Table 29: Description of the Union-Year Table 101
Table 30: Description of the Unmarried Union-Year Table 104
Table 31: Sample Proportion Males Married by Year and Age 107
Table 32: Sample Proportion Females Married by Year and Age 108
Table 33: Annual Hazard of Union Formation per 10,000 113
Table 34: Trend in Probability of Union Formation 120
Table 35: Separation Probabilities per 1,000 124
Table 36: Probability of Separation by Duration of Marriage 130
Table 37: Trend in Hazard of Separation 131

Table 38: Male Age-Specific Fertility Rates by Period 141
Table 39: Female Age-Specific Fertility Rates by Period 142
Table 40: Significance Levels of Male Trend 156
Table 41: Male Trend in Annual Hazard of Birth by Age 157
Table 42: Significance Levels of Female Trend 164
Table 43: Female Trend in Annual Hazard of Birth by Age 165
Table 44: Male Life Table 1957-1995 173
Table 45: Female Life Table 1957-1995 173
Table 46: Life Expectancy at Birth 175
Table 47: Life Table Probability of Dying ${}_nq_x$ by Age and Period 177
Table 48: Male Life Tables in Five-Year Periods 178
Table 49: Female Life Tables in Five-Year Periods 181
Table 50: Trend in Annual Hazard of Death by Age 186
Table 51: Transformations Substituted 193
Table 52: P-values for the Year Effects Resulting from Estimation of Equation 18 196
Table 53: Population Simulator Base Tables 242
Table 54: Description of the tblALives Table 243
Table 55: Description of the tblAUnions Table 243
Table 56: Description of the tblALnkLivesUnions Table 243
Table 57: Description of the tblAFertilityEvents Table 244
Table 58: Description of the tblAPregnancies Table 244
Table 59: Underlying 262
Table 60: Associativity Parameters 319
Table 61: HIV Influence 336
Table 62: Stable Population Proportional Growth Rates 339
Table 63: Stable Population Period Expectation of Life at Birth 344
Table 64: Stable Population Female 345
Table 65: Example of "Experiences" 378
Table 66: Records in Item-Episodes Table 378
Table 67: Records in Experiences Table 379
Table 68: Records Events Table 379

LIST OF FIGURES

Number *Page*

Figure 1: First Five Principal Components of *INDEPTH* Male Mortality. 14
Figure 2: First Five Principal Components of *INDEPTH* Female Mortality 16
Figure 3: First Principal Component of *INDEPTH* Mortality 16
Figure 4: Second Principal Component of *INDEPTH* Mortality 17
Figure 5: Third Principal Component of *INDEPTH* Mortality 17
Figure 6: Fourth Principal Component of *INDEPTH* Mortality 18
Figure 7: *INDEPTH* Male Mortality Patterns 27
Figure 8: *INDEPTH* Female Mortality Patterns 27
Figure 9: *INDEPTH* Mortality Pattern 1 28
Figure 10: *INDEPTH* Mortality Pattern 2 28
Figure 11: *INDEPTH* Mortality Pattern 3 29
Figure 12: *INDEPTH* Mortality Pattern 4 29
Figure 13: *INDEPTH* Mortality Pattern 5 30
Figure 14: *INDEPTH* Mortality Pattern 6 30
Figure 15: *INDEPTH* Mortality Pattern 7 31
Figure 16: Male Age-Specific Deviations from Existing Models [logit($_nq_x$)] 38
Figure 17: Female Age-Specific Deviations from Existing Models [logit($_nq_x$)] 38
Figure 18: *INDEPTH* Male Mortality Pattern 5 With and Without HIV/AIDS Mortality 42
Figure 19: *INDEPTH* Female Mortality Pattern 5 With and Without HIV/AIDS Mortality 42
Figure 20: *INDEPTH* Male Mortality Pattern 1 With and Without HIV/AIDS Child Mortality 43
Figure 21: Male Life Table Probability of Dying: $_nq_x$ 44
Figure 22: Female Life Table Probability of Dying: $_nq_x$ 44
Figure 23: Male Life Table Probability of Surviving P_x 45
Figure 24: Female Life Table Probability of Surviving P_x 45
Figure 25: Male Expectation of Life e_x 46
Figure 26: Female Expectation of Life e_x 46
Figure 27: Zambia and the GTRP Study Area within Africa 54
Figure 28: Gwembe Valley and Lake Kariba within Zambia 55
Figure 34: Gwembe Study Population Pyramid 1957-1995: Person Years Lived in Each Age Group as a Percent of Total Person Years Lived (Male and Female) 86
Figure 35: Predicted Annual Hazard of Union Formation by Age 1957-1995 110
Figure 36: Annual Hazard of Union Formation: All Marriage Orders 115
Figure 37: Annual Hazard of Union Formation: Male Unmarried 116
Figure 38: Annual Hazard of Union Formation: Male with One Wife 117
Figure 39: Annual Hazard of Union Formation: Male with Two or More Wives 118
Figure 40: Trend in Probability of Union Formation: Male Unmarried 121
Figure 41: Trend in Polygyny 121

Figure 42: Probabilities of Separation by Male and Female Age: All Fertility Orders 126
Figure 43: Probabilities of Separation by Male and Female Age: No Children 127
Figure 44: Probabilities of Separation by Male and Female Age: One to Two Children 128
Figure 45: Probabilities of Separation by Male and Female Age: Three or More Children 129
Figure 46: Annual Hazard of Separation by Duration of Union 130
Figure 47: Trend in the Probability of Separation 132
Figure 48: Male Age-Specific Fertility Rates 1957-1995 139
Figure 49: Female Age-Specific Fertility Rates 1957-1995 140
Figure 50: Female Age-Specific Fertility Rates by Period 140
Figure 51: Annual Hazard of Birth for Pairs: All Duration of Union 144
Figure 52: Odds Ratio of Birth Probability by Duration of Union 145
Figure 53: Sex-Specific TFR by Age of Opposite Sex 146
Figure 54: Trend in Male Total Fertility Rates 151
Figure 55: Male Married Age-Specific Fertility Rates by Period 152
Figure 56: Trend in Female Total Fertility Rates 152
Figure 57: Female Married Age-Specific Fertility Rates by Period 153
Figure 58: Male Annual Hazard of Birth Ages 15-19 159
Figure 59: Male Annual Hazard of Birth Ages 20-24 159
Figure 60: Male Annual Hazard of Birth Ages 25-29 160
Figure 61: Male Annual Hazard of Birth Ages 30-34 160
Figure 62: Male Annual Hazard of Birth Ages 35-39 161
Figure 63: Male Annual Hazard of Birth Ages 40-44 161
Figure 64: Male Annual Hazard of Birth Ages 45-49 162
Figure 65: Male Annual Hazard of Birth Ages 50-54 162
Figure 66: Male Annual Hazard of Birth Ages 55-59 163
Figure 67: Male Annual Hazard of Birth Ages 60-64 163
Figure 68: Female Annual Hazard of Birth Ages 15-19 166
Figure 69: Female Annual Hazard of Birth Ages 20-24 166
Figure 70: Female Annual Hazard of Birth Ages 25-29 167
Figure 71: Female Annual Hazard of Birth Ages 30-34 167
Figure 72: Female Annual Hazard of Birth Ages 35-39 168
Figure 73: Female Annual Hazard of Birth Ages 40-44 168
Figure 74: Female Annual Hazard of Birth Ages 45-49 169
Figure 75: Life Table Probability of Dying 1957-1995 174
Figure 76: Life Table Survivors 1957-1995 174
Figure 77: Trend in Life Expectancy at Birth 175
Figure 78: Annual Hazard of Death 1957-1995 Age 0 187
Figure 79: Annual Hazard of Death 1957-1995 Ages 1-4 187
Figure 80: Annual Hazard of Death 1957-1995 Ages 5-9 188
Figure 81: Annual Hazard of Death 1957-1995 Ages 10-19 188
Figure 82: Annual Hazard of Death 1957-1995 Ages 20-29 189
Figure 83: Annual Hazard of Death 1957-1995 Ages 30-39 189
Figure 84: Annual Hazard of Death 1957-1995 Ages 40-49 190
Figure 85: Annual Hazard of Death 1957-1995 Ages 50-59 190
Figure 86: Annual Hazard of Death 1957-1995 Ages 60-69 191
Figure 87: Annual Hazard of Death 1957-1995 Ages 70-79 191

Figure 88: Predicted Trend in Annual Hazard of Death Ages 0-4 193
Figure 89: Predicted Trend in Annual Hazard of Death Ages 20-29 194
Figure 90: Predicted Trend in Annual Hazard of Death Ages 30-39 194
Figure 91: Predicted Trend in Annual Hazard of Death Ages 30-39 195
Figure 92: Simulator Data Model 246
Figure 93: Hazard of Death and Life Table ${}_nq_x$ 263
Figure 94: Daily Hazard of Intercourse 273
Figure 95: Vertical Transmission Probability 281
Figure 96: Viral Load by Time Since Infection 288
Figure 97: Probability of Infection 292
Figure 98: Simulator Per Intercourse Probability of Transmission 292
Figure 99: Monthly Hazard of Union Formation for Couples with an Unmarried Male 304
Figure 100: Monthly Hazard of Union Formation for Couples with Male with One Wife 305
Figure 101: Monthly Hazard of Union Formation for Couples with Male with Two or More Wives 306
Figure 102: Age/Age-Dependent Monthly Hazard of Separation for Couples with No Children 309
Figure 103: Age/Age-Dependent Monthly Hazard of Separation for Couples with One to Two Children 310
Figure 104: Age/Age-Dependent Monthly Hazard of Separation for Couples with Three or More Children 311
Figure 105: Union Duration Dependence of Monthly Hazard of Separation 312
Figure 106: Male Age / Female Age Associativity Index 318
Figure 107: Monthly Hazard of Extramarital Union Formation 321
Figure 108: Daily Hazard of Intercourse within an Affair 323
Figure 109: Total Stable Population by Age by Period – Male, No AIDS 340
Figure 110: Stable Population Age Structure by Period – Male, No AIDS 340
Figure 111: Total Stable Population by Age by Period – Female, No AIDS 341
Figure 112: Stable Population Age Structure by Period – Male, No AIDS 341
Figure 113: Stable Population Age-Specific Sex Ratio 342
Figure 114: Male Age-Period-Specific Mortality Rates 343
Figure 115: Female Age-Period-Specific Mortality Rates 343
Figure 116: Stable Population Female Age-Specific Fertility Rates 345
Figure 117: Untreated and Treated Relative Viral Load for AIDS Simulations 347
Figure 118: Total Population Annual Proportional Growth Rate for AIDS Simulations 349
Figure 119: Male Age-Specific Mortality Rates for AIDS Simulations 350
Figure 120: Female Age-Specific Mortality Rates for AIDS Simulations 351
Figure 121: Female Period Total Fertility Rate for AIDS Simulations 352
Figure 122: Male Age-Specific HIV Incidence Rate 353
Figure 123: Female Age-Specific HIV Incidence Rate 353
Figure 124: Male Age-Specific HIV Prevalence Rate 355
Figure 125: Female Age-Specific HIV Prevalence Rate 355
Figure 126: Male Age-Specific Number of New HIV Cases 356
Figure 127: Female Age-Specific Number of New HIV Cases 356
Figure 128: Structured Population Event History Register DATA MODEL 374

Preface

The four parts of this document describe largely independent pieces of work that are held together by a common focus on the Demography of Africa. Two themes are explored in detail, the first relating to the impact of HIV/AIDS on demographic processes in Africa, and the second on developing methodologies to increase the quantity and reliability of demographic data collected in Africa.

Part One is based on data collected by a number of small intensively run field sites in Africa which are members of the *INDEPTH* network (*INDEPTH* 2000). The sites generously contributed their mortality data to an *INDEPTH* effort to publish a single volume describing the survival of study participants in *INDEPTH* sites. I had the privilege of being asked to contribute to the analysis of those data for the *INDEPTH* publication (*INDEPTH* Forthcoming 2001). The end result of that effort is contained first in the *INDEPTH* volume and also here as Part One of this document. Underlying regularities in the age pattern of mortality are identified from the combined data provided by nineteen field sites, and it is shown that these new age patterns of mortality are substantially different from the existing model age patterns of mortality (Coale and Demeny 1966; United Nations 1982). A component model of mortality is used to identify the characteristic age patterns of mortality, and a variation of that model is developed to facilitate the production of model life tables based on these new age patterns of mortality.

Part Two conducts a thorough analysis of the demographic data collected over the past 40 years by the Gwembe Tonga Research Project from the Gwembe Tonga living in Southern

Zambia. The project and the data it has collected are described in detail. The data are then used to examine both levels and trends in mortality, fertility and marriage patterns over the past 40 years. Although the results of these analyses are interesting with respect to the small, isolated population to which they relate, they are of substantially more interest in light of the fact that they describe a complete set of parameters governing all critical aspects of the population dynamic of a rapidly growing polygynous African population. The data describing the Gwembe Tonga are among a small number of data sets that are capable of reliably describing the mortality, fertility *and* marriage and divorce dynamics of a complex, rural African population over a long, multigenerational period of time, extending far enough back in time to describe the pre-AIDS era. As such they are uniquely useful in constructing holistic population models, and it is primarily for this purpose that they are analyzed.

Part Three describes the design, implementation and initial testing of an individual-level stochastic simulator whose aim is to model the complete demography of a polygynous African population with HIV. The goal is to produce a tool that adequately captures the interactions between HIV disease at the individual level and demographic processes that operate at both individual and group levels. It is envisioned primarily as a tool to facilitate our understanding of the complex interactions between various population processes affected by the HIV, and to provide a cheap, fast mechanism through which to test the impact of various treatment and intervention strategies – taking into account their effects at both the individual and population level, and the potentially complex interplay between effects at those levels. A *proof of concept prototype* is constructed that confirms that the basic ideas and technologies developed to

support the simulator are valid and useful, and that such a tool is both tractable, affordable and potentially useful.

Part Four briefly describes the Structured Population Event History Register (SPEHR). SPEHR is a relational data model designed to store and manipulate longitudinal descriptions of complex time-evolving populations in a relational database. The data model is unique in providing a completely standardized method for storing and manipulating structured longitudinal information. The benefits of this method include unlimited flexibility and extensibility, built-in meta data describing the database and the data it contains, and the ability to define and structure temporal as well as standard relational integrity constraints. A sister data model still under design extends the SPEHR model to include a questionnaire instrument model that when complete will integrate fully with SPEHR to provide a powerful, general and very flexible data management method that controls the flow and integrity of *longitudinal* data from the point of collection through to analysis.

Although they appear different, parts Three and Four are closely related. My interest in relational databases was sparked by the need to manage the complex data describing the Gwembe population. In the process of designing an adequate data management solution for those data, I realized that the same machinery could be easily utilized to manage a complex *simulated* population, and that moreover, it would not be too difficult to extend the data management machinery to actually *create* the simulated population as well. So out of the housekeeping task of managing and manipulating the Gwembe data grew the core concepts and technology upon which the simulator is built. SPEHR represents a substantial refinement

over the original data models used to manage the Gwembe data, and it will form the foundation for the next generation of the simulator.

Part 1

INDEPTH MORTALITY PATTENS FOR AFRICA[1]

MORTALITY MODELS AND AFRICA

An individual's probability of dying depends primarily on sex, age, health, genetic endowment and the environment; all of which determine the risk of falling ill or victim to an accident. The primary determinants of mortality interact in complex ways and depend in turn on a large and variable set of complex social determinants. As a result, it has not been possible to formulate a general, theory-driven model of individual risk of death. In lieu of a good general model, there are two widely used sets of "model life tables"; created by Coale and Demeny (CD) (Coale and Demeny 1966) and later by the UN (United Nations 1982). In both cases, a large set of empirical mortality rates were summarized to yield a small number of characteristic age-patterns of mortality. CD identified four patterns that they term *North*, *South*, *East* and *West* to reflect the fact that each pattern is representative of the mortality pattern in a particular region of Europe. For a similar reason, the UN's patterns also bear regional names: *Latin America*, *Chile*, *Far East*, *South Asia* and *General*. The UN *General* pattern is, as its name suggests, a general pattern that is not specific to a single location.

[1] This Part of this document was prepared by me for the *INDEPTH* network as part of the Network's forthcoming monograph titled "Population and Health in Developing Countries" (*INDEPTH*. Forthcoming 2001. Population and Health in Developing Countries, vol. 1, Population, Health and Survival at Indepth Sties. Ottawa, Canada: International Development Research Centre). All of the analysis and writing presented here are mine, while all of the data were provided by members of the Network and remain the property of the individual Network members.

Each of the eight existing model mortality patterns (excluding the UN *General* pattern) results from the characteristic epidemiological profile of the region it represents. For example, the UN *South Asia* pattern describes an age pattern of mortality with "very high rates under age fifteen and very high rates again at the oldest ages, with correspondingly lower mortality for the prime age-groups". This pattern is ascribed to a "high incidences of infectious, parasitic and diarrheal diseases at the youngest ages and high mortality from diarrheal and respiratory diseases at the oldest ages" (United Nations 1982).

For large areas of the Developing World, accurate information describing the mortality of the population is not available because vital registration systems are incomplete and inaccurate. Where that is true, model mortality patterns are used to substitute for real information. Two important examples are population projections and estimates of child mortality. All population projections must include information describing existing mortality conditions and educated estimates to describe mortality regimes that will prevail in the future. The Brass estimators of child mortality (United Nations 1983), widely used in areas where accurate data on child mortality are not available, rely on estimates of the age pattern of child mortality, and in most cases a model mortality pattern is used for that purpose. Moreover, model mortality patterns are used to evaluate data, to produce smoothed or corrected versions of faulty data and to extend or fill in the age range of incomplete data. Demographers working in regions where mortality data are inaccurate or incomplete depend heavily on model mortality patterns to allow them to evaluate the data they have and to make reasonable estimates and predictions.

None of the data used to create either of the widely used collections of model mortality patterns came from Sub-Saharan Africa. Consequently, it is not evident that the existing

model mortality patterns adequately describe the age-patterns of mortality in Africa, and it is only because there is nothing else that they are applied to African populations at all. Furthermore, the emergence of the HIV/AIDS pandemic in Africa has radically altered the age pattern of mortality in large areas of the continent. Because the existing model mortality patterns do not contain an "AIDS" pattern of mortality, they are no longer appropriate under any circumstance where AIDS is a significant cause of death, or where AIDS is anticipated as a significant cause of death in the near future. This is an even more significant problem than it might first appear because of the crucial role that model mortality patterns play in routine demographic work relating to Africa – precisely because there is a substantial lack of comprehensive, accurate data describing mortality in Africa.

This part (of this dissertation) presents seven age-patterns of mortality derived almost exclusively from data collected in Africa, including two patterns resulting from excess mortality caused by AIDS. A fifteen-factor model is constructed to summarize the data, and that model is used to isolate the AIDS-related components of mortality in the AIDS pattern. Last, the AIDS components are superimposed in various amounts on one of the patterns to generate a course set of model life tables that illustrates effects the of AIDS mortality.

MORTALITY DATA

To allow maximum flexibility in analysis, individual *INDEPTH* sites provided counts of deaths and person years observed in standard 0 to 85+ age groups by sex for single years of observation for as many years of observation as possible. The majority of sites were able to provide data in this format although one or two provided time-aggregated data. Table 1 summarizes the data on which this work is based.

TABLE 1: *INDEPTH* MORTALITY DATA

Site	Country	Data Begins	Data Ends	Total Years of Data	Aggregated Years	Total Person Years Observed
Agincourt	South Africa	1992	1999	8	-	405,311.46
Dar es Salaam	Tanzania	1992	1999	8	-	485,446.30
Hai	Tanzania	1992	1999	8	-	1,045,152.69
Morogoro	Tanzania	1992	1999	8	-	741,412.41
Bandafassi	Senegal	1980	1999	14	-	144,475.61
Bandim	Guinea Bissau	1990	1997	8	-	193,832.91
Butajira	Ethiopia	1987	1996	10	-	336,075.71
Oubritenga	Bukina Faso	1994	1998	5	-	482,100.40
Farefenni	Senegal	1990	1999	10	-	98,073.70
Gwembe Tonga Research Project	Zambia	1956	1995	39	-	187,034.00
Ifakara	Tanzania	1997	1999	3	-	159,639.00
Manhiça	Mozambique	1998	1999	2	-	67,344.00
Matlab, Comparison Area	Bangladesh	-NA-	-NA-	2	1988, 1998	203,744.00
Matlab, Treatment Area	Bangladesh	-NA-	-NA-	2	1988, 1998	211,770.00
Mlomp	Senegal	1985	1999	14	-	106,593.48
Navrongo	Ghana	1993	1999	7	-	930,187.50
Niakhar	Senegal	1985	1998	14	1985-9, 1990-4, 1995-8	372,880.00
Nouna	Burkina Faso	1993	1998	6	-	174,689.62
Rufiji	Tanzania	1999	1999	1	-	67,842.50
Total				**169**		**6,413,605.29**

The overall aim of this work is to identify age-patterns of mortality for Africa and Asia using longitudinal data from *INDEPTH* field sites. To adequately capture the variation in mortality over time, the data from each site is grouped into three-year intervals, or as close to three-year intervals as possible and practical, to yield 70 site-periods. The annual data in each of those periods is aggregated to yield 70 site-period data sets for each sex: 140 site-period data sets in all. Table 2 describes the periods that were chosen for each site.

TABLE 2: SITE-PERIODS

Site	Year																				
	50-79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99
Agincourt																2				4	
Bandafassi								2						4						6	
Bandim														2				4			
Butajira												2				4					
Dar es Salaam															2		4		6		
Farefenni																		2			
Gwembe				2						4					6						
Hai															2		4		6		
Ifakara																					
Manhica																					
Matlab, Comparison																				2	
Matlab, Treatment																				2	
Mlomp										2						4					
Morogoro															2		4		6		
Navrongo																		2			
Niakhar														2							
Nouna																		2			
Outbritenga																		2			
Rufiji																					

NB: Numerals label the periods chosen within each site. The AMMP sites in Tanzania Dar es Salaam, Hai and Morogoro start and stop observation at midyear, so their data are reported in midyear to midyear intervals instead of the calendar year intervals. In each of those cases, seven one-year periods are reported.

PRINCIPAL COMPONENTS ANALYSIS

DATA SUMMARY

The goal is to identify a compact representation of the information contained in a large set of observations on similar items. Principal components analysis transforms the observations to produce an equal number of components that can reproduce all of the original observations when combined in the appropriate proportions. The components differ from the original observations in that they capture as much variation as possible in as few components as possible. The first component accounts for the maximum variation that can be accounted for with one component. After removing the variation associated with the first component, the second component accounts for as much of the remaining variation as can be accounted for with one component. This process continues until all the variation in the original data set has been accounted for and the number of components equals the number of original observations. The important consequence is that the majority of the variation in the data set is accounted for in the first few components.

In this way a large set of observations may be summarized by a small number of components. After deciding how much of the original variation must be retained, the analyst may choose to discard the higher order components that account for the residual variation.

COMPONENT MODEL OF MORTALITY

The component model of mortality constructed here makes no substantive assumptions regarding the underlying form of the age-specific mortality schedule. The model makes the general assumption that an arbitrary age-specific mortality schedule can be decomposed into a small number of individual components and a negligible residual term. Additionally, it is

assumed that there are a small number of components that together form a universal set of age-specific mortality components, and that when combined in the appropriate proportions they are able to reproduce any age-specific mortality schedule. For the purposes of this work, these assumptions encompass only the complete set of mortality data examined here; however, it is demonstrated that the "universal" mortality components generated from the *INDEPTH* data are capable of reproducing all of the CD and UN model life table mortality schedules to within a very small tolerance.

Assume there are n separate components of the age-specific mortality schedule and g age groups. Let $\boldsymbol{m}$ represent the g x 1 vector of age-specific logit($_nq_x$) values, and let $\boldsymbol{C}$ represent the g x n matrix whose i^{th} column is the g x 1 vector containing the i^{th} component of mortality. Let $\boldsymbol{a}$ be a n x 1 vector of coefficients that determine how much of each component is used to generate the mortality schedule, and let $\boldsymbol{r}$ be a g x 1 vector of residuals, one for each age. Then Equation 1 is a compact representation of the component model of mortality.

$$\mathbf{m} = \mathbf{Ca} + \mathbf{r}$$

expanding this around the row for the 20 to 24 age group reveals

$$\begin{bmatrix} \vdots \\ \text{logit}({}_5q_{20}) \\ \vdots \end{bmatrix} = \begin{bmatrix} \vdots \\ {}_5c^1_{20} \cdot a_1 \\ \vdots \end{bmatrix} + \ldots + \begin{bmatrix} \vdots \\ {}_5c^n_{20} \cdot a_n \\ \vdots \end{bmatrix} + \begin{bmatrix} \vdots \\ r_{20} \\ \vdots \end{bmatrix}$$

where ${}_5c^i_{20}$ is the value of i^{th} component for the 20 to 24 age group,

a_i is the value of the coefficient on the i^{th} component, and

r_{20} is the value of the residual for the 20 to 24 age group.

Each of the column vectors contains g elements, one for each age group.

Equation 1: Full Component Model of Mortality

Once the matrix $\boldsymbol{C}$ has been identified through principal components analysis (described below), the model may be used in many ways. First, it is informative to examine the shape of the components themselves. The primary component (accounting for the bulk of the variation in the data) represents the common underlying shape of the schedule as a function of age. The second and higher order components define age-specific variations on the basic shape. Moreover, it may be possible to associate certain substantive interpretations with the components; for example, one may appear to affect the balance between child and adult mortality, and one may appear to contribute or remove from a particular age group that is affected by a specific condition such as maternal or AIDS-related mortality.

Once the components $\boldsymbol{C}$ are identified, estimates of the coefficients $\boldsymbol{a}$ that transform the components into a given mortality schedule may be obtained through an ordinary linear least squares regression of the mortality schedule against the components $\boldsymbol{C}$. The residual identified in the regression is equivalent to $\boldsymbol{r}$, and the regression coefficients are the elements of the vector $\boldsymbol{a}$ with the addition of an extra element to store the constant estimated in the regression. Let $\boldsymbol{a}'$ be the $(n+1)$ x 1 coefficient vector with the additional element to store the constant generated in the regression model, and let $\boldsymbol{C}'$ be the g x $(n+1)$ matrix of components with one additional column containing all ones to accommodate the constant in $\boldsymbol{a}'$. The constant is interpreted as a measure of the overall *level* of the mortality schedule while the coefficients indicate how much of each age-pattern (component) is necessary to reproduce the overall age-pattern in the original data. Interpreted in this way, the regression controls for level and provides an estimate of how much of each component is contained within the data, or how important each individual age-pattern is in generating the age-pattern observed in the data.

$$\mathbf{m} = \mathbf{C'a'}$$

expanding this around the row for the 20 to 24 age group reveals

$$\begin{bmatrix} \vdots \\ \text{logit}({}_5q_{20}) \\ \vdots \end{bmatrix} = \begin{bmatrix} \vdots \\ {}_5c^1_{20} \cdot a_1 \\ \vdots \end{bmatrix} + \ldots + \begin{bmatrix} \vdots \\ {}_5c^n_{20} \cdot a_n \\ \vdots \end{bmatrix} + \begin{bmatrix} \vdots \\ 1 \cdot a_c \\ \vdots \end{bmatrix}$$

where ${}_5c^i_{20}$ is the value of the i^{th} component for the 20 to 24 age group,

a_i is the value of the coefficient estimated on the i^{th} component, and

a_c is the the constant term estimated in the regression

which takes the same value for all age groups.

Each of the column vectors contains g elements, one for each age group.

Equation 2: Regression Component Model of Mortality

Ignoring the residual and postmultiplying $\boldsymbol{C'}$ by $\boldsymbol{a'}$ (Equation 2) yields the original mortality schedule purged of the residual r. Together with $\boldsymbol{C'}$, the $(n+1)$ x 1 vector $\boldsymbol{a'}$ contains all the information needed to reproduce the original mortality schedule to within r. In most cases the number of components, $(n+1)$, necessary to adequately encode the mortality schedule is much less than g, the number of age groups. As a result, $\boldsymbol{a'}$ is a compact representation of the mortality schedule that encodes the fundamental shape of the schedule without the "noise" associated with the high order components and the residual term. Additionally, by adjusting the constant term contained in the last element of $\boldsymbol{a'}$, it is possible to arbitrarily set the level of the mortality schedule without affecting its age-pattern.

The individual coefficient vectors associated with each mortality schedule represent the most important dimensions of the mortality schedules and can be compared and grouped without worrying about the high-order noise associated with the individual schedules. Moreover, by comparing only the coefficients corresponding to the components and ignoring the constant, it

is possible to compare individual mortality schedules based only on their individual age-patterns and *not on differences in their level.* Correspondingly, by comparing only the constants associated with two mortality schedules, the influence of the age-pattern is effectively removed (controlled for), and it is possible to compare the mortality schedules based only on their level.

PRINCIPAL COMPONENTS OF *INDEPTH* MORTALITY DATA

For each sex, logit($_nq_x$) values are calculated for the eighteen standard 0-85[2] age groups in each of the site-periods, according to Equation 3 and Equation 4[3]. This yields a 70 x 18 data set consisting of one column for each site-period and one row for each age group, with each cell containing a value of logit($_nq_x$) corresponding to the specified site-period and age group.

$$ {}_nq_x = \frac{n\,{}_nM_x}{1+n(1-{}_na_x)\,{}_nM_x} $$

Where $_nq_x$ is the life table probability of death between ages x and $x+n$ for those who survive to age x, $_nM_x$ is the observed mortality rate (the ratio of deaths to person years lived) for those aged x and $x+n$, and $_na_x$ is the average proportion of years between ages x and $x+n$ lived by those who die in that age interval[4].

Equation 3: nqx as a Function of nMx

$$ \text{logit}({}_nq_x) = \frac{1}{2}\ln\left(\frac{{}_nq_x}{1-{}_nq_x}\right) $$

Equation 4: logit($_nq_x$)

---

[2] 0, 1-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79, 80-84.

[3] The raw age-specific logit($_nq_x$) values were mildly smoothed using a weighted moving average over age; the smoothing kernel being [0.29, 0.43, 0.29].

[4] Without substantially more data tabulated by single year of age it is not possible to directly calculate or estimate the values of $_na_x$. Moreover, except for the youngest ages, the value of $_na_x$ is always near 0.5. At the youngest ages, the values are much closer to 0.25. Additionally, the life table is not highly sensitive to the exact values chosen as long as they are close to 0.25 for ages less than five years and close to 0.5 for ages older than five years. In this work, the value of $_na_x$ used for ages older than five years is 0.5. For ages younger than five years, the values for $_na_x$ are for males 0.33 for ages 0-1 and 0.25 for ages 1-4, and for females 0.35 for ages 0-1 and 0.25 for ages 1-4. These are loosely adapted from the CD West model life table system (Coale, Ansley J. and Paul Demeny. 1966. Regional Model Life Tables and Stable Populations. Princeton, New Jersey: Princeton University Press).

The *factor*[5] and *score* routines provided with the STATA® statistical software package release 5.0 (STATA Corporation 1997) are used to calculate the principal components of the 70 x 18 covariance matrix[6] associated with the data set described above. Each age group (row) in the data set is given a weight equal to the total number of person years of observation for the age group summed across all site-periods. Fifteen of the resulting 70 principal components are retained, and for both males and females those fifteen components account for greater than 99.99 percent of the variation in the data.

MALE

The first fifteen principal components calculated using the *INDEPTH* male data are contained in Table 3 and Figure 1. The primary (first) component obviously represents the underlying age-pattern of mortality, and together components two through four modify the age-pattern in a way that is consistent with mortality caused by AIDS. Component two in particular has the shape necessary to account for increased mortality between the ages of 20 and 64. Components three and four allow modifications between 20 and 49 and during childhood.

The primary component crosses the x-axis between ages five and nine and again between ages 30 and 34 with the result that as the coefficient on the primary component increases, child and

[5] The *factor* routine is used with the options [pc] to request principal components analysis, [covariance] to specify that the covariance matrix is analyzed and [weight] to specify the weighting.

[6] The covariance matrix is used so that the observations are not standardized before the calculation. The resulting principal components refer to the unstandardized observations and can be directly recombined to produce age-specific mortality schedules that need no further transformation, except for the inverse logit, to produce values of $_4q_x$.

TABLE 3: FIRST FIFTEEN PRINCIPAL COMPONENTS OF *INDEPTH* MALE MORTALITY UNITS OF LOGIT($_nq_x$)

Age	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15
0	3.5206	-0.9092	-0.5861	0.7908	-0.9517	-0.1341	-0.0818	0.0137	-0.0715	-0.0257	-0.0624	0.0275	-0.0939	-0.0773	-0.0056
1-4	2.1060	-1.3263	-0.4162	0.2302	-0.0323	0.1009	-0.0104	-0.0263	0.0459	-0.0008	0.0371	-0.0160	0.0626	0.0489	0.0025
5-9	-0.6237	-1.3939	-0.2097	-0.2409	0.3600	0.1113	0.0420	0.0000	-0.0094	-0.0129	-0.0139	0.0089	-0.0443	-0.0346	0.0001
10-14	-3.8913	-0.6672	0.2893	-0.2781	-0.2373	-0.2707	-0.0414	0.0790	-0.1137	0.0415	-0.0198	0.0100	0.0234	0.0162	0.0013
15-19	-4.2300	0.2142	0.5140	0.2019	-0.1094	0.0091	0.1223	-0.1077	0.2174	-0.0260	0.0616	-0.0360	-0.0171	-0.0103	-0.0055
20-24	-2.9701	1.0178	0.2872	0.5311	0.2605	0.2447	-0.0343	0.0151	-0.0975	-0.0369	-0.1389	0.0622	0.0185	0.0144	0.0003
25-29	-1.3633	1.5292	-0.2994	0.2827	0.2560	-0.0229	-0.2078	0.1704	-0.1585	0.0290	0.1827	-0.0389	-0.0237	-0.0082	0.0135
30-34	-0.0101	1.6798	-0.8204	-0.1323	0.1790	-0.3111	-0.1093	-0.1431	0.1191	0.1215	-0.1161	-0.0862	0.0321	-0.0231	-0.0333
35-39	1.1271	1.5442	-0.5701	-0.3670	-0.0745	-0.2137	0.0288	-0.0588	0.1346	-0.1331	0.0150	0.1368	-0.0566	0.0521	0.0674
40-44	2.0278	1.4749	0.2600	-0.3663	-0.2405	0.1735	0.3823	0.1487	-0.0608	-0.1717	0.0395	0.0461	0.0852	-0.0358	-0.0863
45-49	2.9895	1.1639	0.0904	-0.3435	-0.3401	0.4071	0.3410	0.1576	-0.0829	0.0946	-0.0489	-0.1475	-0.0459	0.0031	0.0489
50-54	3.9242	0.7764	0.4165	-0.2790	-0.2578	0.4380	-0.0074	-0.2064	-0.0613	0.2585	-0.0078	0.0370	-0.0202	0.0447	0.0390
55-59	5.0339	0.4222	0.6651	-0.3319	-0.1043	0.1733	-0.3776	-0.3370	-0.0928	0.0997	0.1143	0.1160	-0.0040	-0.0381	-0.0896
60-64	6.4716	0.1705	0.9022	-0.2772	-0.0053	-0.0324	-0.4303	-0.1194	-0.0338	-0.3066	-0.0430	-0.1138	0.1070	-0.0853	0.1084
65-69	8.2046	-0.0280	0.8817	-0.1134	0.1607	-0.1749	-0.3034	0.1980	0.0741	-0.2004	-0.0884	-0.1207	-0.1348	0.1284	-0.1121
70-74	9.9266	-0.0834	0.9652	0.1994	0.3252	-0.2939	-0.0499	0.5487	0.3829	0.1610	-0.0376	0.0641	-0.0381	0.0069	-0.0222
75-79	11.7097	-0.0984	1.0259	0.5481	0.5681	-0.6360	0.4503	0.3546	0.2699	0.3511	0.0297	0.1869	0.1476	-0.1304	0.0800
80-84	13.7690	-0.1067	1.0401	1.0822	1.0750	-1.5415	1.4284	-0.8888	-0.9189	-0.1067	0.0850	-0.1085	-0.0877	0.0748	0.0021
% Var.*	0.8712	0.0889	0.0153	0.0077	0.0062	0.0037	0.0024	0.0014	0.0012	0.0007	0.0004	0.0003	0.0002	0.0001	0.0001
Cumm. % Var.*	0.8712	0.9601	0.9755	0.9832	0.9894	0.9931	0.9955	0.9969	0.9981	0.9988	0.9992	0.9995	0.9997	0.9998	0.9999

* Percent Variance Explained; * Cumulative Percent Variance Explained

adult mortality increase while the mortality of teenagers and young adults decreases. Consequently, the first coefficient determines the ratio of child and adult mortality to teenage and young adult mortality. This is likely due to the fact that mortality of the very young and elderly is more sensitive to adverse (or advantageous) conditions than the mortality of the generally healthy and robust teenagers and young adults[7]. Naturally then, this balance accounts for a great deal of the variation in the data and is therefore encoded in the first component. Remember, that the overall level of mortality is governed by the value of the constant term in Equation 2, so the coefficient on the first component is really only responsible for the age-balance, *not* for the absolute level of mortality at any age.

[7] It is also worth noting that the impact of the first component is not constant with age, where the value of the first component is close to zero, the absolute impact is much smaller than when the value of the first component is more different from zero. Examining the curve reveals that the absolute effect of the first component increases significantly with age past age 39.

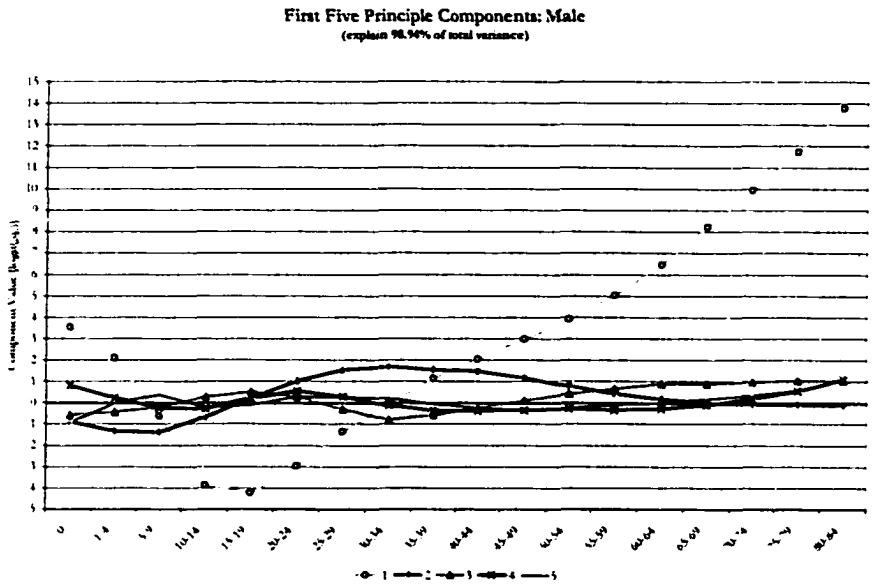

Figure 1: First Five Principal Components of *INDEPTH* Male Mortality.

FEMALE

The first fifteen principal components of *INDEPTH* female mortality are contained in Table 4 and Figure 2. In broad terms they are very similar to the male components. However, the primary component contains a significant positive bulge between ages 20 and 44 which is absent on the male primary component (see Figure 3). The most likely explanation for this is that it accounts for the maternal mortality experienced by the female population. Additionally, the second component describes a somewhat narrower, younger pattern of deviation that at its peak is of slightly greater magnitude than the males (see Figure 4). This likely results from the general fact that the effect of AIDS on female mortality is younger and more focused in age than its effect on male mortality. The third and fourth components are virtually identical for males and females except at older ages. Because the data at older ages are more likely to be

inaccurate, and since the differences are large only at the oldest ages, they will not be interpreted.

TABLE 4: FIRST FIFTEEN PRINCIPAL COMPONENTS OF *INDEPTH* FEMALE MORTALITY UNITS OF LOGIT($_nq_x$)

Age	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15
0	3.5313	-0.4696	-0.2823	0.9837	0.6458	-0.4557	-0.5026	-0.2507	0.0375	-0.0815	-0.0516	-0.0245	0.0183	-0.0096	-0.1023
1-4	2.3801	-1.1570	-0.3886	0.3453	-0.0051	-0.1870	-0.0543	0.0899	-0.0276	0.0054	0.0376	0.0227	-0.0123	0.0075	0.0693
5-9	-0.3553	-1.5460	-0.4143	-0.1775	-0.3376	0.0598	0.0976	0.0563	-0.0250	0.0135	-0.0422	-0.0158	0.0003	-0.0028	-0.0467
10-14	-3.9597	-1.0972	0.1003	-0.2918	0.2469	0.2186	-0.0422	-0.2020	0.0855	-0.0112	0.0366	0.0076	0.0187	0.0038	0.0187
15-19	-3.8199	0.0125	0.6053	0.2285	0.3222	-0.0854	0.1449	0.1722	-0.1569	-0.0147	-0.0341	0.0008	-0.0486	-0.0082	-0.0065
20-24	-2.2312	1.0535	0.2788	0.5568	-0.3124	-0.1589	0.1621	0.0696	0.2258	-0.0418	0.0319	-0.0306	0.0716	-0.0062	-0.0014
25-29	-0.8902	1.6924	-0.1814	0.5157	-0.2909	0.3444	-0.0841	-0.1418	-0.0648	0.1300	0.0070	0.0948	-0.0527	0.0294	-0.0112
30-34	-0.1141	1.8996	-0.6129	-0.0698	0.0584	0.3515	-0.1415	0.0030	-0.0994	-0.0446	-0.0620	-0.1350	0.0061	-0.0375	0.0314
35-39	0.3159	1.7325	-0.6989	-0.7004	0.3716	-0.0558	-0.0309	0.2363	-0.0152	-0.0941	0.0901	0.0743	0.0619	0.0360	-0.0306
40-44	0.7540	1.2876	-0.2148	-0.7254	0.0605	-0.5754	0.1212	-0.0973	0.2371	0.1248	-0.0473	0.0139	-0.1273	-0.0655	0.0098
45-49	1.4083	0.8007	0.4921	-0.5216	-0.2276	-0.5815	0.1042	-0.2427	-0.2023	0.0497	-0.0439	-0.0532	0.0521	0.1314	0.0018
50-54	2.4523	0.2582	0.7844	-0.3784	-0.4306	-0.1563	-0.2427	-0.1569	0.2432	-0.0833	0.0309	0.0400	0.0773	-0.1118	0.0475
55-59	3.8123	-0.0959	1.0571	-0.4104	-0.4334	0.2157	-0.4918	0.1446	0.0872	-0.2270	0.1135	-0.0077	-0.1201	0.0071	-0.0628
60-64	5.4730	-0.2709	0.9026	-0.2793	0.2074	0.4606	-0.3685	0.2549	0.2940	0.1110	-0.2140	-0.0096	0.0205	0.0886	0.0637
65-69	7.4297	-0.1483	0.6562	-0.1330	0.3840	0.3041	0.2144	0.0973	-0.0260	0.3456	-0.0498	0.0983	0.1588	-0.0981	-0.0650
70-74	9.2290	-0.1971	0.4828	0.1217	0.4232	0.3758	0.5258	-0.0544	-0.2405	0.2968	0.2371	-0.1387	-0.0093	-0.0193	-0.0319
75-79	11.2081	0.0265	0.2309	0.2767	0.3649	0.4929	0.8747	-0.2562	0.0128	-0.1368	0.2313	-0.1680	-0.1436	0.0685	-0.0035
80-84	13.4986	0.8028	-0.0993	0.3120	0.2427	0.6651	1.5584	-0.5054	0.0902	-0.9873	-0.3832	0.2796	-0.0119	-0.0200	0.0403
% Var.*	0.8249	0.1176	0.0191	0.015	0.0074	0.0063	0.004	0.0019	0.0014	0.001	0.0004	0.0003	0.0003	0.0002	0.0001
Cumm. % Var.°	0.8249	0.9425	0.9616	0.9766	0.9840	0.9903	0.9944	0.9963	0.9977	0.9987	0.9991	0.9994	0.9996	0.9998	0.9999

* Percent Variance Explained; ° Cumulative Percent Variance Explained

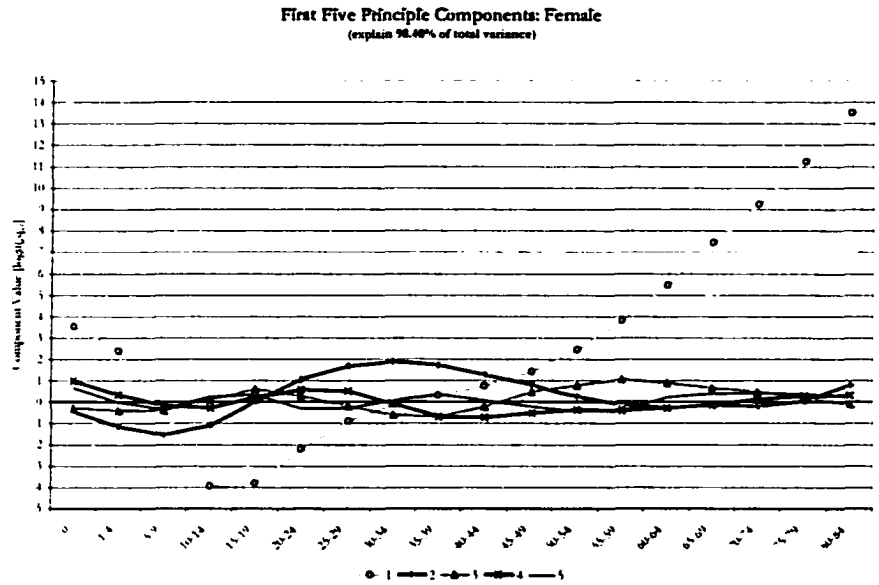

Figure 2: First Five Principal Components of *INDEPTH* Female Mortality

MALE AND FEMALE PRINCIPAL COMPONENTS CONTRASTED

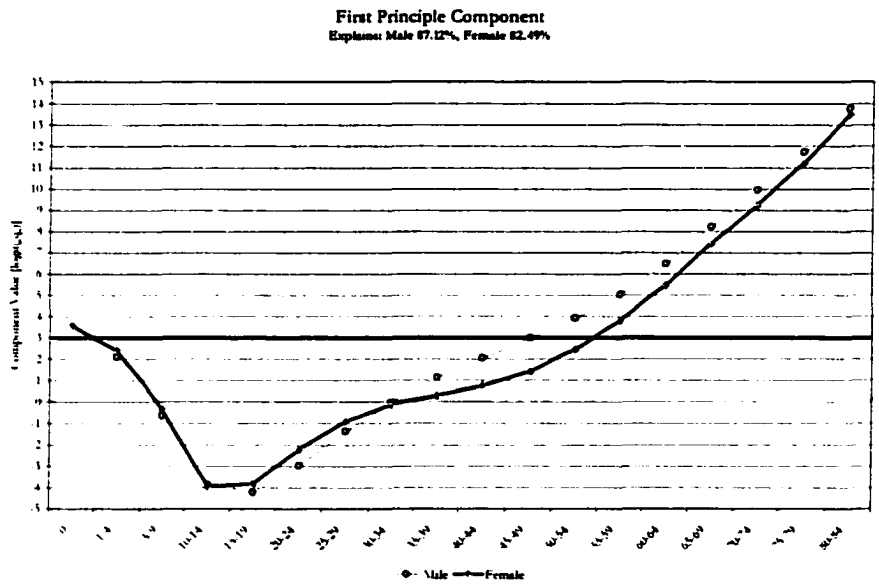

Figure 3: First Principal Component of *INDEPTH* Mortality

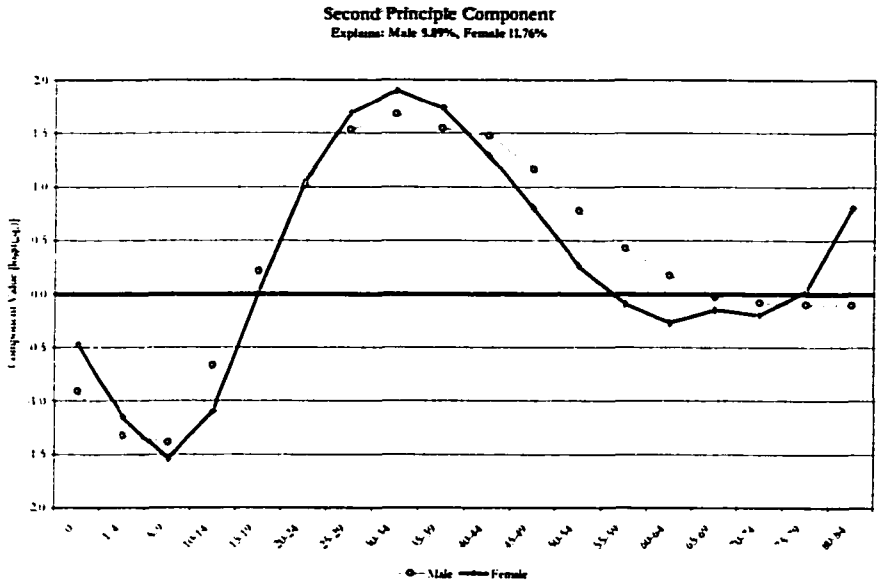

Figure 4: Second Principal Component of *INDEPTH* Mortality

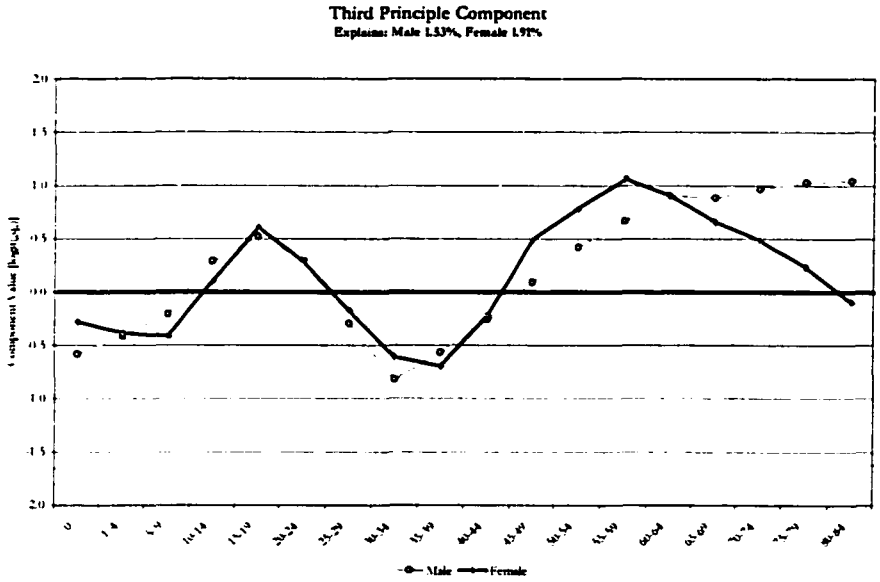

Figure 5: Third Principal Component of *INDEPTH* Mortality

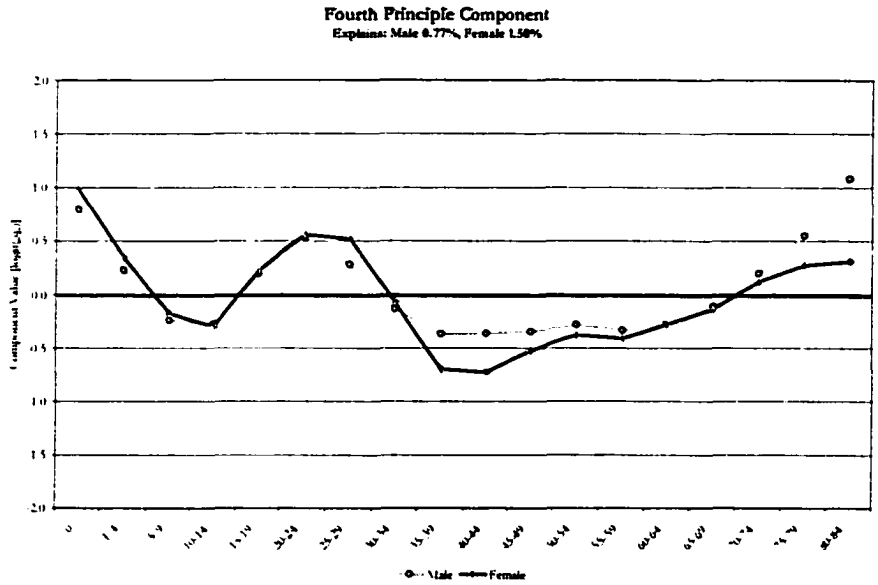

Figure 6: Fourth Principal Component of *INDEPTH* Mortality

Figure 3, Figure 4, Figure 5 and Figure 6 plot the first four principal components of *INDEPTH* mortality for both the males and females together in order to clearly demonstrate the differences between the male and female components. These differences are discussed briefly above.

In order to examine the generality of the *INDEPTH* components of mortality, the existing CD and UN model mortality patterns (at levels corresponding to an expectation of life at birth of 55.00 years) were regressed against the *INDEPTH* components of mortality in a simple linear ordinary least squares regression. The regressions were run against all fifteen of the *INDEPTH* components, the first ten, and finally the first five. In each case, the fit statistics were examined and the predicted mortality patterns were calculated and visually compared to the patterns that were fit. Table 5 displays the R^2 fit statistic for those regressions. Using all fifteen components produces near-perfect fits that are able to faithfully reproduce the existing

patterns in all respects. Reducing the number of components used has the expected effect of reducing the quality of the overall fit and failing to correctly model the higher frequency variation in the model patterns. Using ten components still produces a reasonable fit, and using five or six components is acceptable in most circumstances; however with a small number of components, there is a substantial "smoothing" due to the lack of high frequency components. This is actually useful if the aim is to capture the fundamental shape of the mortality curve, or if the data are dirty and it is necessary to fit the basic shape and ignore the smaller bumps and dips that may be meaningless.

TABLE 5: R^2 VALUES FROM LINEAR REGRESSIONS OF EXISTING MODEL MORTALITY PATTERNS ON THE *INDEPTH* COMPONENTS

Model	Male	Female
	Fit with First 15 Components	
North	0.99999977	0.99995679
South	0.99942947	0.99904130
East	0.99999192	0.99999787
West	0.99993568	0.99987275
Latin America	0.99971166	0.99908125
Chile	0.99999361	0.99986967
South Asia	0.99984172	0.99933633
Far East	0.99997782	0.99999868
General	0.99995098	0.99977579
	Fit with First 10 Components	
North	0.99985585	0.99897605
South	0.99643755	0.99382783
East	0.99956920	0.99957740
West	0.99955650	0.99792047
Latin America	0.99888354	0.99595977
Chile	0.99949656	0.99909239
South Asia	0.99815676	0.99675165
Far East	0.99965167	0.99910922
General	0.99965960	0.99824604
	Fit with First 5 Components	
North	0.99886669	0.99633704
South	0.99382770	0.98825846
East	0.99656374	0.99403000
West	0.99678475	0.99452655
Latin America	0.99480725	0.98818531
Chile	0.99448695	0.98020460
South Asia	0.99272203	0.98342185
Far East	0.99698749	0.99560908
General	0.99637700	0.99412156

INDEPTH MORTALITY PATTERNS

The overall aim of this work is to identify common age-patterns of mortality in the *INDEPTH* data. The resulting patterns provide a distilled representation of the important mortality conditions experienced by the populations from which the data come. Moreover, some understanding of the age-patterns of mortality in Africa, based on empirical data from Africa, is invaluable to demographers and planners of all kinds who must account for present and future mortality in much of their work.

COMPONENT CLUSTERING METHOD

The most critical task in identifying the common underlying mortality patterns is to identify clusters of similar mortality patterns; in this case clusters of site-periods with similar age-patterns of mortality. This is a particularly difficult exercise that necessarily involves some subjective input from the analyst.

A given ***age***-pattern of mortality can be observed at various *levels* resulting from the fact that there may be causes of mortality that affect all ages in roughly the same way and consequently do not produce an age-pattern. Given that, there are two separate dimensions on which mortality schedules may be clustered: *age-pattern* and *level*. The age-pattern of a mortality schedule contains a lot of information regarding the epidemiological profile of the population from which the data are collected and is consequently of primary interest here.

One of the substantial advantages of the component model of mortality is the distilled, parsimonious representation of a mortality pattern that results from regressing it on the components. The vector of regression coefficients contains *independent* information on the age-

pattern and level of the mortality schedule. That fact allows the creation of clusters of age-patterns without respect to level.

To create the age-pattern clusters, all 70 of the *INDEPTH* mortality schedules for both males and females are regressed against the appropriate sex-specific components of *INDEPTH* mortality. The coefficients corresponding to the first four principal components are retained and the other eleven plus the constant are discarded. The first four principal components account for 98.32 percent of the variation in the male data and 97.66 percent of the variation in the female data, making them sufficient to capture all but the finest nuances in the age-pattern of mortality. These form a collection of 70 4 X 1 coefficient vectors for each sex.

The agglomerative hierarchical clustering algorithm provided with the S-Plus 2000 Professional® statistical software package release 3 is used to identify clusters of similar coefficient vectors for each sex.[8] The "Ward" method used here is described by the provider of the software as follows:

> The basic hierarchical agglomerative algorithm starts with each object in a separate group. At each iteration it merges two groups to form a new group; the merger chosen is the one that leads to the smallest increase in the sum of within-group sums of squares. The number of iterations is equal to the number of objects minus one, and at the end all the objects are together in a single group. (Mathsoft Inc. 1999b)

For a detailed discussion of clustering techniques and this particular algorithm please see Mathsoft Inc. (1999a), Struyf and Hubert (1997), and Kaufman and Rousseeuw (1990)[9]. This

[8] S-PLUS's "agnes" routine was used with options: metric = euclidean, standardize = true, and linkage type = ward.

[9] A number of clustering techniques were applied to both the raw and transformed data and to the coefficient vectors, including agglomerative hierarchical clustering, partitioning around K-means, partitioning around K-medoids, fuzzy

routine was applied separately to the male and female data sets, each consisting of four columns, one for each coefficient described above, and 70 rows, one for each site-period.

CLUSTERS

The method described above identified five robust clusters in the male data and seven robust clusters in the female data, presented in Table 6. Because females are subject to the additional risk of maternal mortality that is not experienced by the males, the female age-patterns are always more complex and so it is not surprising that two more clusters were identified in the female data. Categorizing the male data into the seven female clusters produces seven male

TABLE 6: *INDEPTH* MORTALITY AGE-PATTERN CLUSTERS

MALE			FEMALE		
ID	Site-Period	Cluster / Pattern	ID	Site-Period	Cluster / Pattern
26	Bandafassi: 80-84	1	26	Bandafassi: 80-84	1
27	Bandafassi: 85-87	1	27	Bandafassi: 85-87	1
28	Bandafassi: 88-90	1	28	Bandafassi: 88-90	1
29	Bandafassi: 91-93	1	29	Bandafassi: 91-93	1
30	Bandafassi: 94-96	1	30	Bandafassi: 94-96	1
31	Bandafassi: 97-99	1	31	Bandafassi: 97-99	1
36	Butajira: 87-89	1	32	Bandim: 90-91	1
37	Butajira: 90-91	1	40	Oubritenga: 94-95	1
38	Butajira: 92-93	1	41	Oubritenga: 96-98	1
39	Butajira: 94-96	1	43	Farefenni: 96-97	1
40	Oubritenga: 94-95	1	44	Farefenni: 98-99	1
47	Gwembe Tonga Research Project: 84-86	1	45	Gwembe Tonga Research Project: 50-80	1
65	Niakhar: 85-89	1	49	Gwembe Tonga Research Project: 90-92	1
66	Niakhar: 90-94	1	50	Gwembe Tonga Research Project: 93-95	1
67	Niakhar: 95-98	1	51	Ifakara: 97-99	1
69	Nouna: 96-98	1	52	Manhiça: 98-99	1
54	Matlab, Comparison Area: 98	2	65	Niakhar: 85-89	1
55	Matlab, Treatment Area: 88	2	66	Niakhar: 90-94	1
56	Matlab, Treatment Area: 98	2	67	Niakhar: 95-98	1
59	Mlomp: 91-93	2	70	Rufiji: 99	1
60	Mlomp: 94-96	2	53	Matlab, Comparison Area: 88	2
1	Agincourt: 92-93	3	54	Matlab, Comparison Area: 98	2
2	Agincourt: 94-95	3	55	Matlab, Treatment Area: 88	2
3	Agincourt: 96-97	3	56	Matlab, Treatment Area: 98	2

partitioning, divisive hierarchical clustering. Three different statistical software packages, STATA®, S-PLUS® and MVSP®, were used, and in each case all of their clustering routines were tried. All of the methods produced essentially the same clusters but differed in the clarity of their output and in how they managed ambiguous cases. The agglomerative hierarchical algorithm provided with S-PLUS was eventually chosen based on its clear and robust theoretical underpinnings and the fact that its output is easily understood and interpreted. Moreover, it appeared to provide the most robust clusters and the most efficient means of categorizing ambiguous cases.

TABLE 6: *INDEPTH* MORTALITY AGE-PATTERN CLUSTERS

MALE			FEMALE		
ID	Site-Period	Cluster / Pattern	ID	Site-Period	Cluster / Pattern
11	Dar es Salaam: 1998-99	3	57	Mlomp: 85-87	2
35	Bandim: 96-97	3	61	Mlomp: 97-99	2
5	Dar es Salaam: 1992-93	4	2	Agincourt: 94-95	3
32	Bandim: 90-91	4	3	Agincourt: 96-97	3
33	Bandim: 92-93	4	4	Agincourt: 98-99	3
34	Bandim: 94-95	4	7	Dar es Salaam: 1994-95	3
42	Farefenni: 94-95	4	8	Dar es Salaam: 1995-96	3
43	Farefenni: 96-97	4	9	Dar es Salaam: 1996-97	3
44	Farefenni: 98-99	4	10	Dar es Salaam: 1997-98	3
45	Gwembe Tonga Research Project: 50-80	4	11	Dar es Salaam: 1998-99	3
46	Gwembe Tonga Research Project: 81-83	4	35	Bandim: 96-97	3
49	Gwembe Tonga Research Project: 90-92	4	1	Agincourt: 92-93	4
53	Matlab, Comparison Area: 88	4	33	Bandim: 92-93	4
57	Mlomp: 85-87	4	34	Bandim: 94-95	4
58	Mlomp: 88-90	4	42	Farefenni: 94-95	4
61	Mlomp: 97-99	4	62	Navrongo: 93-95	4
62	Navrongo: 93-95	4	63	Navrongo: 96-97	4
63	Navrongo: 96-97	4	64	Navrongo: 98-99	4
64	Navrongo: 98-99	4	68	Nouna: 93-95	4
68	Nouna: 93-95	4	5	Dar es Salaam: 1992-93	5
70	Rufiji: 99	4	6	Dar es Salaam: 1993-94	5
4	Agincourt: 98-99	5	12	Hai: 1992-93	5
6	Dar es Salaam: 1993-94	5	13	Hai: 1993-94	5
7	Dar es Salaam: 1994-95	5	14	Hai: 1994-95	5
8	Dar es Salaam: 1995-96	5	15	Hai: 1995-96	5
9	Dar es Salaam: 1996-97	5	16	Hai: 1996-97	5
10	Dar es Salaam: 1997-98	5	17	Hai: 1997-98	5
12	Hai: 1992-93	5	18	Hai: 1998-99	5
13	Hai: 1993-94	5	19	Morogoro: 1992-93	5
14	Hai: 1994-95	5	20	Morogoro: 1993-94	5
15	Hai: 1995-96	5	21	Morogoro: 1994-95	5
16	Hai: 1996-97	5	22	Morogoro: 1995-96	5
17	Hai: 1997-98	5	23	Morogoro: 1996-97	5
18	Hai: 1998-99	5	24	Morogoro: 1997-98	5
19	Morogoro: 1992-93	5	25	Morogoro: 1998-99	5
20	Morogoro: 1993-94	5	36	Butajira: 87-89	6
21	Morogoro: 1994-95	5	37	Butajira: 90-91	6
22	Morogoro: 1995-96	5	38	Butajira: 92-93	6
23	Morogoro: 1996-97	5	39	Butajira: 94-96	6
24	Morogoro: 1997-98	5	58	Mlomp: 88-90	6
25	Morogoro: 1998-99	5	69	Nouna: 96-98	6
41	Oubritenga: 96-98	5	46	Gwembe Tonga Research Project: 81-83	7
48	Gwembe Tonga Research Project: 87-89	5	47	Gwembe Tonga Research Project: 84-86	7
50	Gwembe Tonga Research Project: 93-95	5	48	Gwembe Tonga Research Project: 87-89	7
51	Ifakara: 97-99	5	59	Mlomp: 91-93	7
52	Manhiça: 98-99	5	60	Mlomp: 94-96	7

In many cases periods from the same site are grouped in the same cluster reassuring us that the clustering algorithm is identifying and grouping mortality schedules that are fundamentally similar. Where periods from the same site are assigned to different clusters, mortality has been

changing significantly over time, and the mortality schedules from two different periods in time are substantially different.

MORTALITY PATTERNS

After the clusters are identified, a characteristic age-pattern of mortality is identified for each cluster. In keeping with the use of the component model of mortality, for each of the fifteen coefficients derived from the regression of the individual site-period mortality schedules on the fifteen components of *INDEPTH* mortality, the weighted average across the site-periods in each cluster is calculated. The weights used are the person years of observation in each site-period. This yields the average amount of each of the fifteen components and the constant that is needed by each of the mortality schedules in a given cluster. When these average values are combined with the components through Equation 2, the result is the weighted average mortality schedule for each cluster. By varying the constant, the mortality schedules can be adjusted to an arbitrary level, and for convenience's sake, the seven *INDEPTH* mortality patterns presented in Table 7 are adjusted to a level that yields a life expectancy at birth of 55.00 years. Table 7 presents the male and female patterns organized into the seven female-derived clusters. This is done to facilitate comparison of the male and female patterns. The five male-derived patterns are retained when the male data are organized into the female-derived patterns; this simply creates two sets of two slightly redundant male patterns. The author verified this by producing the male patterns based on both the male and female-derived clusters.

TABLE 7: *INDEPTH* MORTALITY PATTERNS

LOGIT($_nq_x$)

Age	Pattern 1	2	3	4	5	6	7
				Male			
0	-1.1821	-1.0939	-1.6252	-1.3192	-1.3260	-1.3778	-1.2170
1-4	-1.3230	-1.4728	-1.7509	-1.4661	-1.5931	-1.3428	-1.3911
5-9	-1.6722	-1.9849	-2.0255	-1.7771	-1.9413	-1.5184	-1.7003
10-14	-2.1807	-2.3702	-2.3544	-2.1811	-2.2056	-1.8187	-2.0821
15-19	-2.2586	-2.5108	-2.2378	-2.2402	-2.1341	-1.8875	-2.1865
20-24	-2.1049	-2.4333	-1.9393	-2.1120	-1.8661	-1.8463	-2.1345
25-29	-1.9047	-2.2779	-1.6891	-1.9157	-1.6286	-1.8062	-2.0705
30-34	-1.7481	-2.1099	-1.5053	-1.7563	-1.4667	-1.7737	-2.0261
35-39	-1.6588	-1.9003	-1.3908	-1.5743	-1.3647	-1.7097	-1.8115
40-44	-1.5905	-1.7467	-1.2490	-1.4380	-1.2778	-1.5853	-1.6832
45-49	-1.4908	-1.5228	-1.1515	-1.3033	-1.2277	-1.4725	-1.4792
50-54	-1.3599	-1.2380	-1.0762	-1.1844	-1.2131	-1.3443	-1.3307
55-59	-1.2138	-0.9758	-0.9546	-1.0316	-1.1841	-1.2052	-1.1678
60-64	-1.0475	-0.7508	-0.7807	-0.8254	-1.0605	-1.0625	-0.8985
65-69	-0.8344	-0.5340	-0.5862	-0.6689	-0.8813	-0.8767	-0.6242
70-74	-0.6132	-0.3143	-0.3531	-0.5276	-0.6934	-0.6775	-0.3279
75-79	-0.3790	-0.0674	-0.1027	-0.3782	-0.4948	-0.4865	-0.1158
80-84	-0.1107	0.2082	0.1747	-0.2005	-0.2477	-0.3257	-0.0226
				Female			
0	-1.1678	-1.0304	-1.4926	-1.2429	-1.2667	-1.4005	-1.1935
1-4	-1.2698	-1.3893	-1.6489	-1.4084	-1.5306	-1.3479	-1.2674
5-9	-1.6070	-1.9119	-1.9691	-1.7526	-1.8930	-1.5252	-1.5658
10-14	-2.1126	-2.3759	-2.3076	-2.1760	-2.1958	-1.8319	-2.1678
15-19	-2.0958	-2.3195	-2.1232	-2.2106	-2.0281	-1.8767	-2.5014
20-24	-1.9525	-2.1988	-1.8469	-2.0725	-1.6854	-1.8322	-2.3502
25-29	-1.8484	-2.1152	-1.6241	-1.9094	-1.4610	-1.7935	-2.1065
30-34	-1.8019	-2.1711	-1.4641	-1.8040	-1.3720	-1.7781	-1.7919
35-39	-1.7623	-2.1811	-1.3715	-1.7224	-1.3793	-1.7215	-1.5495
40-44	-1.7020	-1.9609	-1.3386	-1.6330	-1.4161	-1.6174	-1.5311
45-49	-1.6005	-1.6935	-1.2734	-1.4865	-1.4478	-1.4856	-1.6743
50-54	-1.4831	-1.4249	-1.2305	-1.3010	-1.4333	-1.2875	-1.5927
55-59	-1.3321	-1.1522	-1.0773	-1.0693	-1.3500	-1.1067	-1.4082
60-64	-1.1252	-0.8883	-0.9092	-0.7946	-1.1827	-0.9697	-1.0982
65-69	-0.8707	-0.6080	-0.6508	-0.6352	-0.9797	-0.8405	-0.7822
70-74	-0.6243	-0.3002	-0.4577	-0.4904	-0.7919	-0.7177	-0.5037
75-79	-0.3983	-0.0193	-0.2001	-0.3331	-0.5537	-0.6067	-0.1865
80-84	-0.2084	0.2012	0.1935	-0.1574	-0.2269	-0.4946	0.1573

Figure 7 through Figure 15 plot the seven *INDEPTH* age-patterns of mortality for both males and females. The patterns are arbitrarily named 1-7[10], and a discussion of the patterns follows the plots.

[10] This is done to avoid the potential stigmatization that could result from more descriptive names.

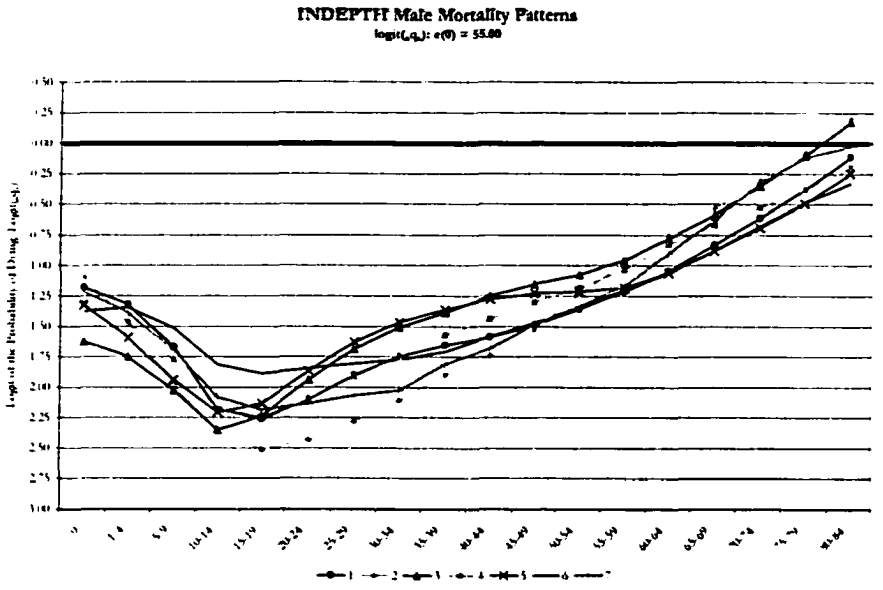

Figure 7: *INDEPTH* Male Mortality Patterns

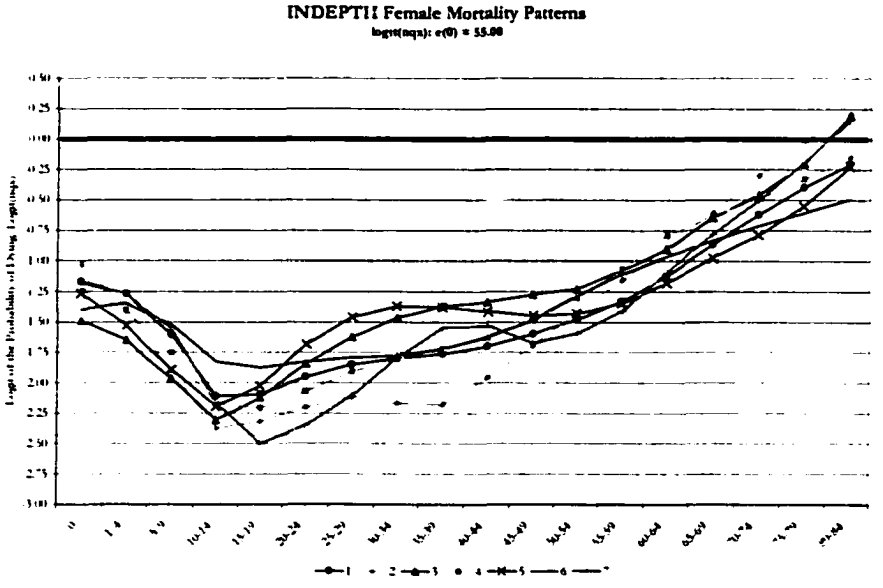

Figure 8: *INDEPTH* Female Mortality Patterns

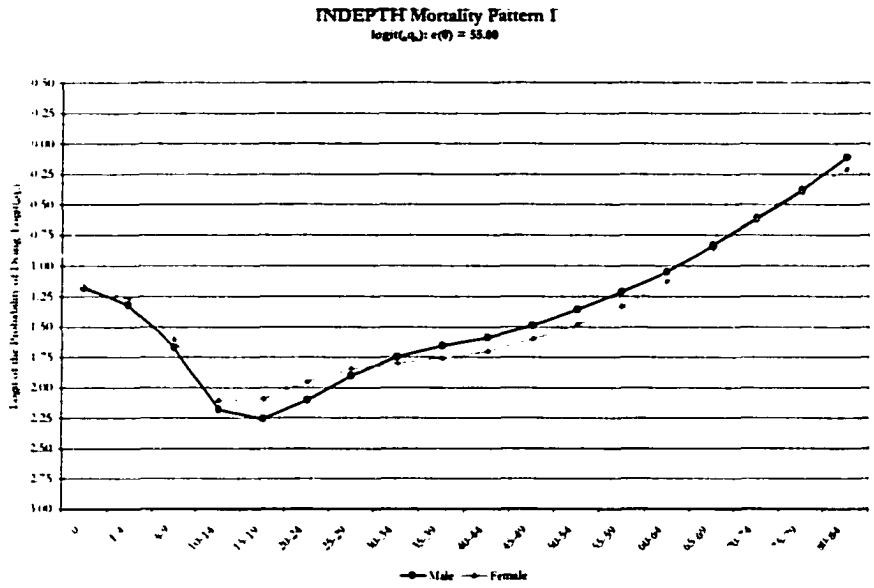

Figure 9: *INDEPTH* Mortality Pattern 1

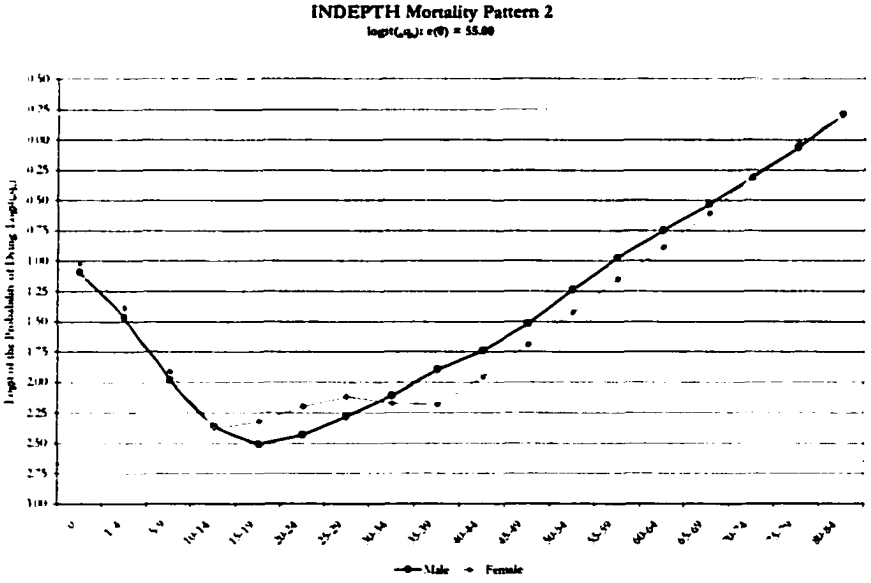

Figure 10: *INDEPTH* Mortality Pattern 2

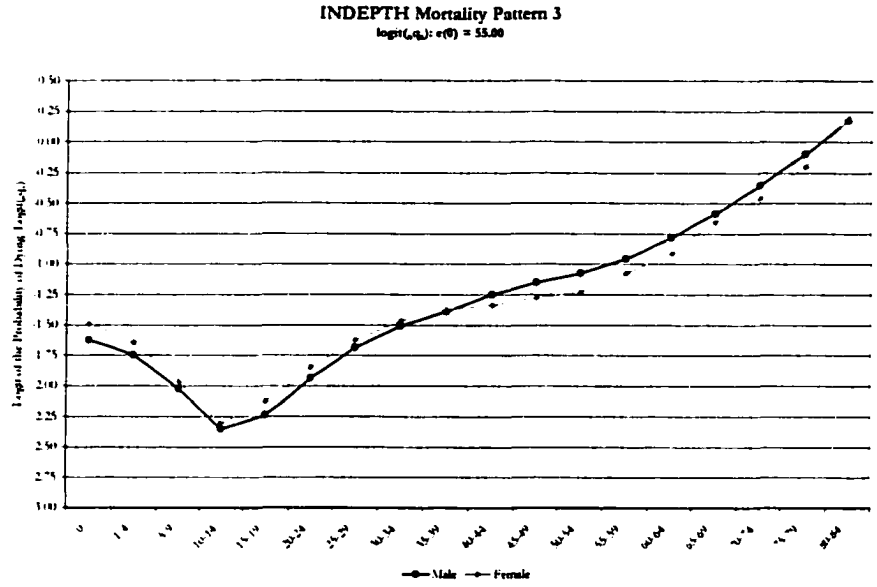

Figure 11: *INDEPTH* Mortality Pattern 3

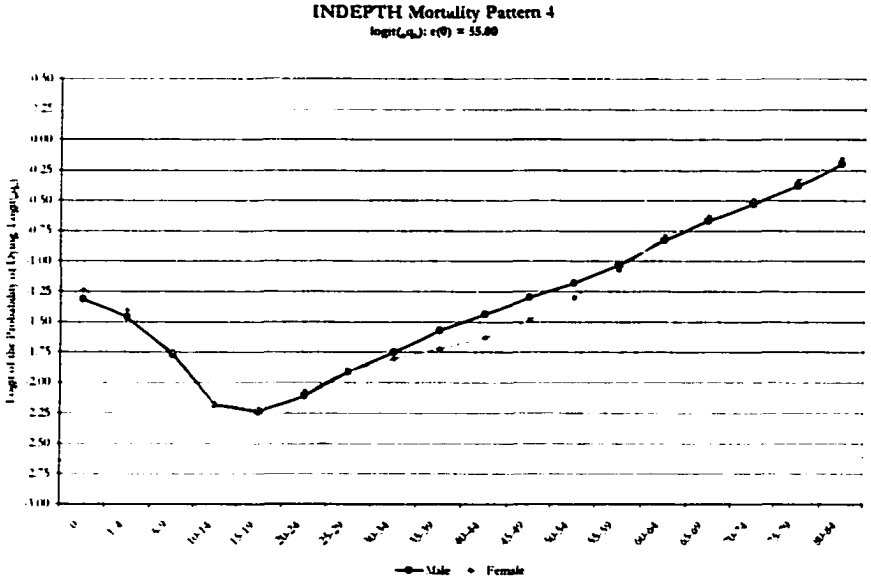

Figure 12: *INDEPTH* Mortality Pattern 4

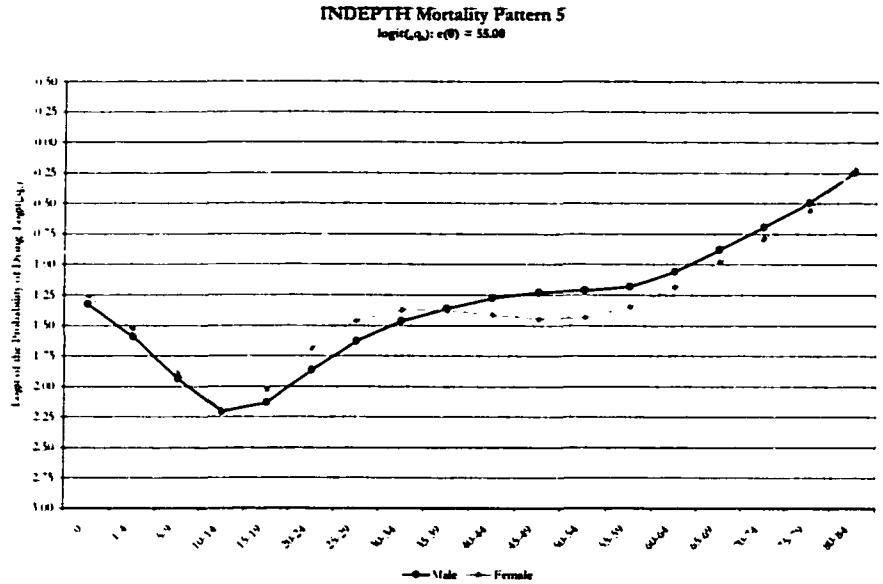

Figure 13: *INDEPTH* Mortality Pattern 5

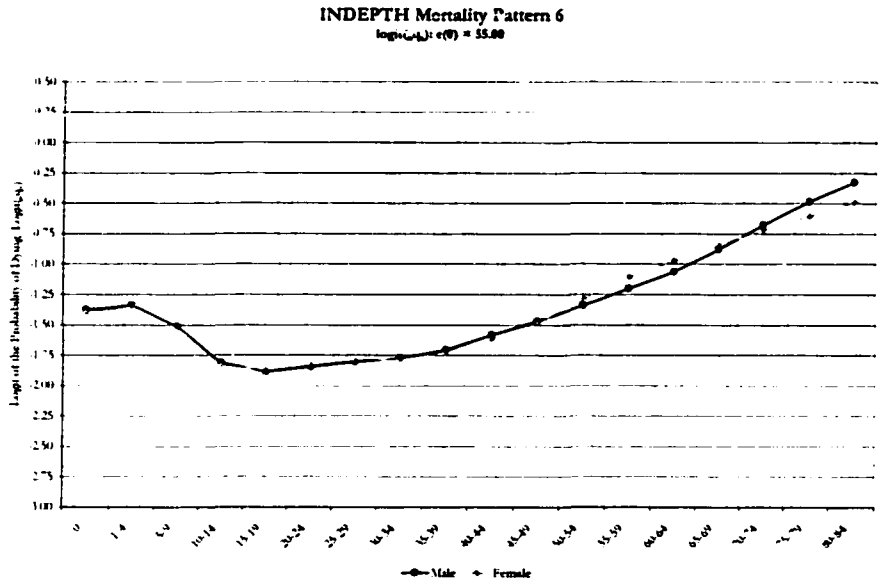

Figure 14: *INDEPTH* Mortality Pattern 6

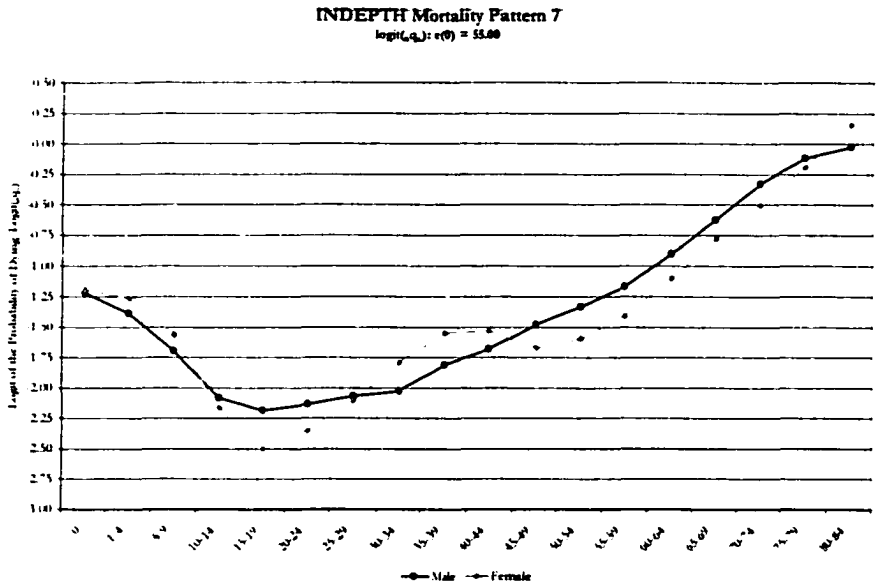

Figure 15: *INDEPTH* Mortality Pattern 7

PATTERN 1

The first pattern is similar to the CD North and UN Latin American model life table age-patterns of mortality, see Compared to the Coale and Demeny and United Nations Model Life Tables, below. There is no indication that HIV/AIDS affects pattern 1, and the male and female age-patterns are similar with the exception of a bulge in the female pattern during the reproductive years, presumably caused by maternal mortality. Pattern 1 is primarily derived from sites in West Africa over the entire period covered by the *INDEPTH* data set. HIV/AIDS has not yet become as significant a problem in West Africa as it is in Central and Southern Africa, so it is not expected to see a large impact of AIDS in the data from West Africa. It is worth noting that child mortality between the ages of one and nine is significant and substantially elevated above the most similar existing models, below. This is in keeping with the fact that malaria prevalence is high in West Africa, and it has a large impact on those ages.

PATTERN 2

Pattern 2 is the only pattern to contain significant contributions from Asia, and it is in fact dominated by data from the Matlab project in Bangladesh. The only other site to contribute data to this pattern is the Mlomp site located in Senegal. Again, the male and female patterns are similar with the exception of maternal mortality. However, pattern 2 is strikingly different from all of the others in that the mortality of children, teenagers, and young adults is comparatively very low, and correspondingly the mortality of older adults is comparatively high. In keeping with the fact that the data contributing to this pattern come from Bangladesh and Senegal, it is not surprising that there is no evidence at all of an HIV/AIDS impact. Pattern 2 is very similar to the UN South Asia pattern, as it should be coming largely from South Asia, see below.

PATTERN 3

The sites contributing to pattern 3 are almost exclusively located in South and East Africa; South Africa and Tanzania in particular. This pattern obviously contains some influence of HIV/AIDS, but not nearly to the degree observed in pattern 5. The South African data come from the Agincourt site, where mortality is extraordinarily low compared to the other *INDEPTH* sites in Africa, and where HIV/AIDS is recognized but not yet impacting the population in the catastrophic sense that it is in other parts of Southern and East Africa. The remainder of the data come from the Dar es Salaam site where there appears to be a greater impact of HIV/AIDS. This pattern is most similar to the UN Far East pattern of mortality corresponding to the fact that infant and child mortality are very low compared to mortality at older ages. A noteworthy feature of this pattern is the fact that infant and child mortality does

not appear to be substantially elevated as might be expected when HIV/AIDS is an important contributor to mortality.

PATTERN 4

Pattern 4 is a variation on pattern 1 with the important difference manifested in the 35-69 age range. At all other ages, patterns 1 and 4 are negligibly different except that infant and child mortality in pattern 4 is consistently slightly lower than pattern 1. But between ages 35 and roughly 69, pattern 4 reveals significantly higher mortality than pattern 1. This pattern is most similar to the UN General pattern for females and UN Latin America for males. As was the case with pattern 1, most of the data contributing to pattern 4 comes from West Africa.

PATTERN 5

The HIV/AIDS pattern of mortality is most clearly visible in pattern 5. The data contributing to pattern 5 are derived from the three Tanzanian sites run by the Adult Morbidity and Mortality Project (AMMP) in Dar es Salaam, Hai, and Morogoro. There is a striking bulge in the mortality of males between the ages of 20 and 54 and for females between the ages of 15 and 49. Additionally, the female bulge is significantly narrower and more pronounced corresponding to the fact that the female population is infected earlier and within a tighter age range than the male population. This pattern is not particularly similar to any of the existing model patterns, but it is most closely matched (below) with the UN General (female) and Latin American (male) model patterns. Pattern 5 differs from pattern 3 mainly in the shape of the HIV/AIDS impact. The effect is more diffuse with age in pattern 3; meaning that mortality is elevated through a broader age range, the magnitude of the elevation is more consistent, and the differences between the sexes are less apparent. Pattern 3 is derived largely from the Dar

es Salaam data, and this may reflect the fact that the epidemic is more mature in Dar es Salaam and has consequently had enough time to infect a wider age range of people of both sexes. As with pattern 3, it is worth noting that infant and child mortality do not appear to be substantially affected in a manner comparable to adult mortality, and this is in contradiction to what is known about HIV prevalence and vertical transmission. Further investigation is necessary to determine why this effect is not prominently measured in these data.

PATTERN 6

Pattern 6 is one of the two additional patterns identified in the female data. It is an interesting pattern that reveals high mortality of children and teenagers together with comparatively low mortality of infants and adults of all ages. This pattern is exhibited by sites in Northeast and West Africa with most of the data coming from Ethiopia. Without additional information, it is not possible to speculate on what may be producing this unique pattern. The male pattern is most similar to the CD North model pattern, and the female pattern is closest to the CD West model - both of which embody high mortality in the same age ranges. They deviate from those patterns in that infant mortality is substantially less than would be found in the either model pattern, and child and adolescent mortality is significantly higher: this might be called the "Super North" pattern.

PATTERN 7

Pattern 7 is the other additional pattern identified in the female data, and it is also interesting. It is derived from two sites located in Central and West Africa. The reason why it was identified in the female data are obvious; there is a substantial bulge in the female age-pattern between ages 25 and 44. This most likely results from serious maternal mortality, the risk of

which increases with age. The site in Zambia is a rural site without easy access to modern medical facilities, and this may contribute to an unusual risk of maternal mortality. The corresponding male pattern is similar to pattern 6, and both are similar to the CD North model pattern. The North model pattern contains relatively high child and teenage mortality coupled with comparatively low mortality at older ages. This is consistent the with fact that malaria is an important contributor to mortality in both sites.

COMPARED TO THE COALE AND DEMENY AND UNITED NATIONS MODEL LIFE TABLES

The *INDEPTH* mortality patterns are explicitly compared to the existing CD and UN models to ensure that they are indeed new patterns, and to demonstrate exactly how they differ from the existing model mortality patterns.

The method used is a simple minimum sum of squared differences. Each *INDEPTH* mortality pattern is compared with all of the existing CD and UN model mortality patterns: CD patterns North, South, East, West; and UN patterns Latin America, Chile, South Asia, Far East and General. For each pair of patterns, the difference between the $\text{logit}({}_nq_x)$ of the two patterns is calculated for each age group, and those differences are squared and summed to yield the sum (over all ages) of the squared differences (SSD) between the two patterns. For each *INDEPTH* pattern, the SDDs derived from the seven comparisons to the model mortality patterns are ranked, and the members of the pair with the smallest SDD are considered to be most similar. All of the mortality patterns used in the comparisons are

adjusted to a level corresponding to an expectation of life at birth of 55.00[11]. The SDDs are presented in Table 8 where both the minimum and next greater SDD for each comparison are identified.

TABLE 8: SUM OF SQUARED DIFFERENCES COMPARING *INDEPTH* AND EXISTING MORTALITY PATTERNS

Pattern	North	South	East	West	LA	CH	SA	FE	GL
					Male				
1	***0.2676***	0.5550	0.7599	0.6035	**0.3111**	1.1682	0.8744	1.6724	0.6458
2	1.3819	0.6787	**0.6538**	0.9190	0.8760	1.0439	***0.2265***	1.1394	0.6918
3	1.1060	1.6448	1.2875	0.8313	1.0938	0.9075	2.3094	***0.5066***	**0.7774**
4	**0.4041**	0.7742	0.7219	0.4561	***0.3646***	0.8273	1.0159	0.9664	0.4344
5	***0.6766***	1.4443	1.3789	0.8996	**0.7961**	1.1767	2.3346	1.6279	1.0265
6	***0.5118***	1.4451	1.6459	1.2315	**0.9998**	2.2118	2.1365	2.5333	1.4486
7	***0.4017***	0.4548	0.5344	0.4985	**0.4233**	1.1231	0.5451	1.1824	0.4866
					Female				
1	**0.1763**	0.4823	0.4573	0.3666	***0.1727***	0.6523	0.6724	1.0096	0.3428
2	1.4695	1.0966	**0.8080**	1.2731	1.0356	1.0373	***0.4703***	1.4209	0.9744
3	1.4447	2.2012	1.5312	1.0859	1.3253	1.1886	2.4018	***0.4426***	**0.9283**
4	0.4823	1.0188	0.7003	0.4749	**0.3982**	0.6471	1.0570	0.6098	***0.3752***
5	0.7861	1.5676	1.2496	***0.7118***	0.8045	0.9274	2.1636	0.7916	**0.7728**
6	***0.3966***	1.1897	1.2256	**0.7723**	0.7730	1.4854	1.7386	1.6242	0.9320
7	***0.3837***	0.5397	0.4040	0.4859	**0.3905**	0.7709	0.4704	1.0079	0.4050

LA-Latin America: CH-Chile: SA-South Asia: FE- Far East: GL- General: ***Minimum***: **Next Best**

For each *INDEPTH* pattern, the age-specific deviations from the closest fit existing model pattern are calculated and presented in Figure 16 and Figure 17. Those figures clearly reveal that *all* of the *INDEPTH* patterns are systematically different from the existing model mortality patterns. Both figures reveal clear peaks in the deviations during the childhood (1-14) and young to middle-aged adult (25-49) years. Interestingly, infant and child mortality between one and four is generally lower than the corresponding pattern. The peak in the

[11] The level of the *INDEPTH* patterns is set by adjusting the constant term in the component model of mortality, and the CD and UN model mortality patterns used in the comparisons are generated by the United Nation's computer program for the analysis of mortality data, MortPak (United Nations, Department of International Economic and Social Affairs. 1988. Computer Program: *Mortpak-Lite* 3.0 (IBM PC Compatible): New York: United Nations) at a level corresponding to an expectation of life at birth of 55.00.

deviations during childhood *may* be due to malaria and other diseases that have a large impact on children in Africa but not elsewhere in the world, and it is clear that continued investigation is necessary to identify all of the factors contributing to the childhood deviations. The peak during the adult years is most pronounced for patterns three and five, which are the two patterns that are affected by HIV/AIDS, and it is reasonable to assume that this peak is primarily due to the impact of HIV/AIDS. It is curious to note that infant and child mortality in patterns three and five does not appear to be elevated in a manner corresponding to the increase in adult mortality. This suggests that the HIV/AIDS epidemic does not have an enormous impact on infant and child mortality, or that all of the data used to generate patterns 3 and 5 are defective with regard to infants and children. It seems unlikely that all the data would be defective, and defective to the same degree, pointing to the fact that considerable investigation into the impact of HIV/AIDS on infant and child mortality needs to be conducted.

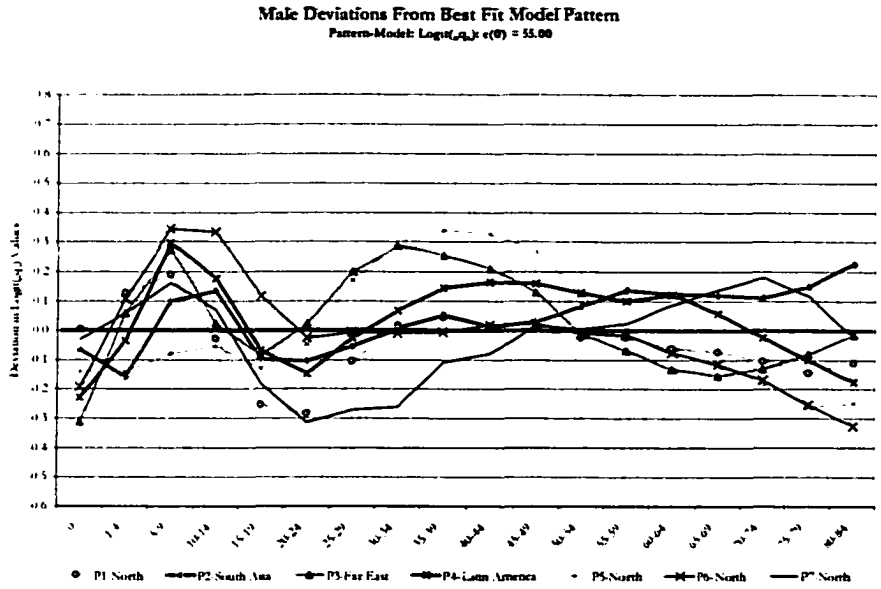

Figure 16: Male Age-Specific Deviations from Existing Models [logit(q_x)]

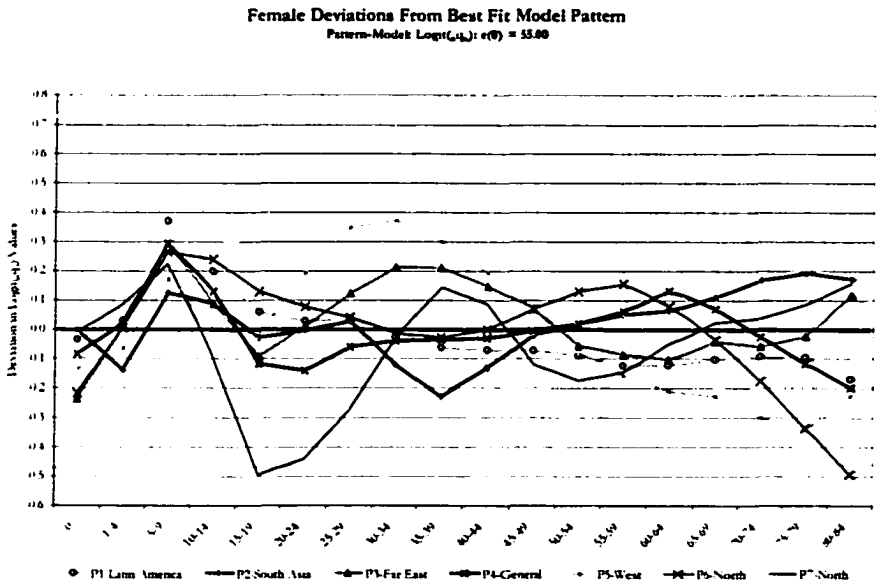

Figure 17: Female Age-Specific Deviations from Existing Models [logit(q_x)]

DEMONSTRATION OF HIV/AIDS MODEL LIFE TABLE SYSTEM

MODEL LIFE TABLE CONSTRUCTION

The component model of mortality is capable of generating (and fitting) a very wide range of mortality patterns. This makes it particularly well-suited for the creation of model life tables. In order to demonstrate how the component model can be used to create a set of model life tables, the *INDEPTH* mortality components are used to isolate (in a set of coefficient deviations) the general age-pattern of the impact of HIV/AIDS, and then to add that impact in increasing quantities to the *INDEPTH* pattern 1 mortality schedule to create a set of life tables with decreasing expectations of life at birth corresponding to an increasing impact of HIV. The result is a set of life tables with the underlying age pattern defined by *INDEPTH* pattern 1 set so that expectation of life at birth is 55.00, but with various levels of HIV/AIDS morality added to that.

Figure 18 and Figure 19 display the male and female *INDEPTH* pattern 5 mortality schedules with and without what is presumed to be the increase in mortality due to HIV/AIDS. Figure 20 presents the male *INDEPTH* pattern 1 mortality with and without an increase in mortality over the infant and childhood ages[12]. In each case, the difference between the two curves is fitted against the first fifteen components of mortality (for the appropriate sex) to yield the coefficients presented in Table 9.

The model life tables are constructed to produce a family of life tables with the underlying mortality of *INDEPTH* pattern 1 mortality. The HIV/AIDS pattern of mortality is added to

[12] There is no empirical pattern used to create the increase in infant and child mortality. It is simply created so that it could be included in the model life tables.

each of the members of the family in amounts sufficient to reduce the expectation of life at birth in five year increments. Equation 5 is a simple extension of the component model of mortality that describes the relationship used to accomplish this. In this case the, the $(n+1)$ x 1 vector $\boldsymbol{d}'$ of HIV/AIDS coefficient deviations is multiplied by α and added to the $(n+1)$ x 1 vector of coefficients, $\boldsymbol{a}$[13]. The scaling factor α determines how much of the HIV/AIDS pattern to add to the basic pattern of mortality represented by the vector of coefficients, $\boldsymbol{a}'$. Once that addition has been accomplished, the resulting vector is premultiplied by the matrix of components C' to yield the logit transformed probabilities of dying, $\text{logit}({}_nq_x)$.

$$\mathbf{m} = \mathbf{C}'(\mathbf{a}'+\alpha\mathbf{d}')$$

expanding this around the row for the 20 to 24 age group reveals

$$\begin{bmatrix} \vdots \\ \text{logit}({}_5q_{20}) \\ \vdots \end{bmatrix} = \begin{bmatrix} \vdots \\ {}_5c^1_{20}\cdot(a_1+\alpha\cdot d_1) \\ \vdots \end{bmatrix} + \ldots + \begin{bmatrix} \vdots \\ {}_5c^n_{20}\cdot(a_n+\alpha\cdot d_n) \\ \vdots \end{bmatrix} + \begin{bmatrix} \vdots \\ 1\cdot(a_c+\alpha\cdot d_c) \\ \vdots \end{bmatrix}$$

where ${}_5c^i_{20}$ is the value of the i^{th} component for the 20 to 24 age group,

a_i is the value of the coefficient on the i^{th} component,

α is the scale factor applied to the vector of coefficient deviations,

d_i is the coefficient deviation for the i^{th} component,

a_c is the the constant term which takes the same value for all age groups, and

d_c is the deviation for the constant term .

Each of the column vectors contains g elements, one for each age group.

Equation 5: Relationship Governing HIV-Augmented Model Life Table

Once the $\text{logit}({}_nq_x)$ values have been calculated, the inverse logit produces values for ${}_nq_x$ that are substituted into a life table and used to calculate the other columns of the life table, including

[13] Remember that the prime (') indicates that the matrices and vectors include the column and row necessary to store the constant and its coefficient. Also, n is the number of components used, and g is the number of age groups.

expectation of life. The model life tables are calculated through an iterative, target-seeking process that varies α until the desired value for the expectation of life is attained.

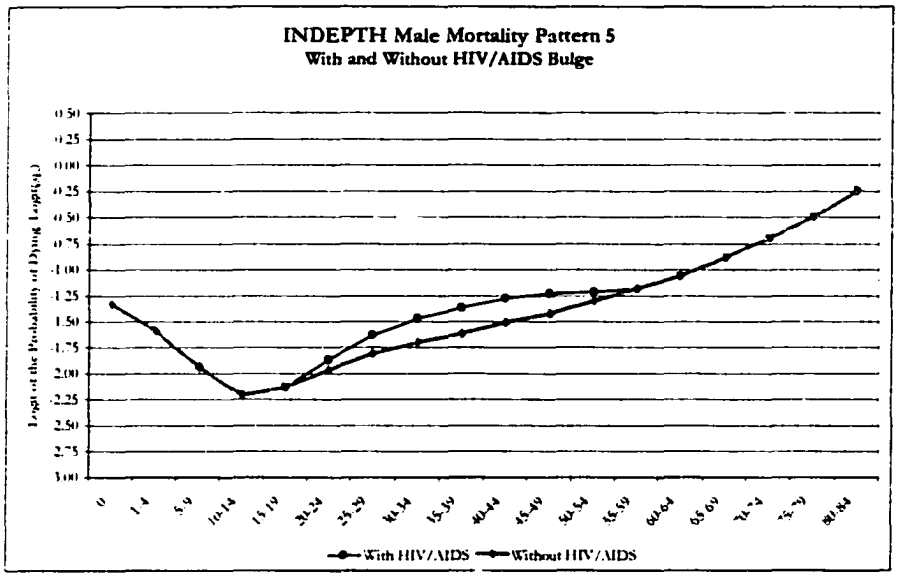

Figure 18: *INDEPTH* Male Mortality Pattern 5 With and Without HIV/AIDS Mortality

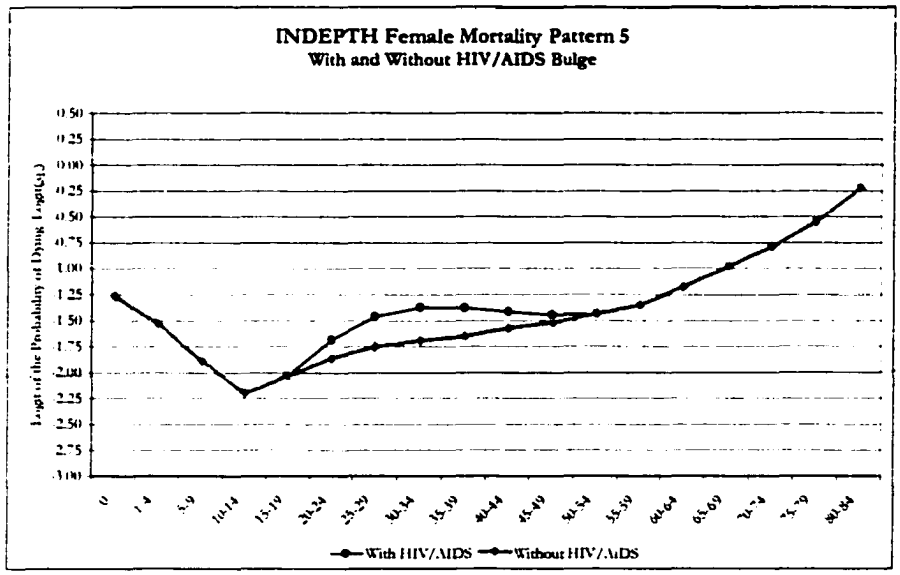

Figure 19: *INDEPTH* Female Mortality Pattern 5 With and Without HIV/AIDS Mortality

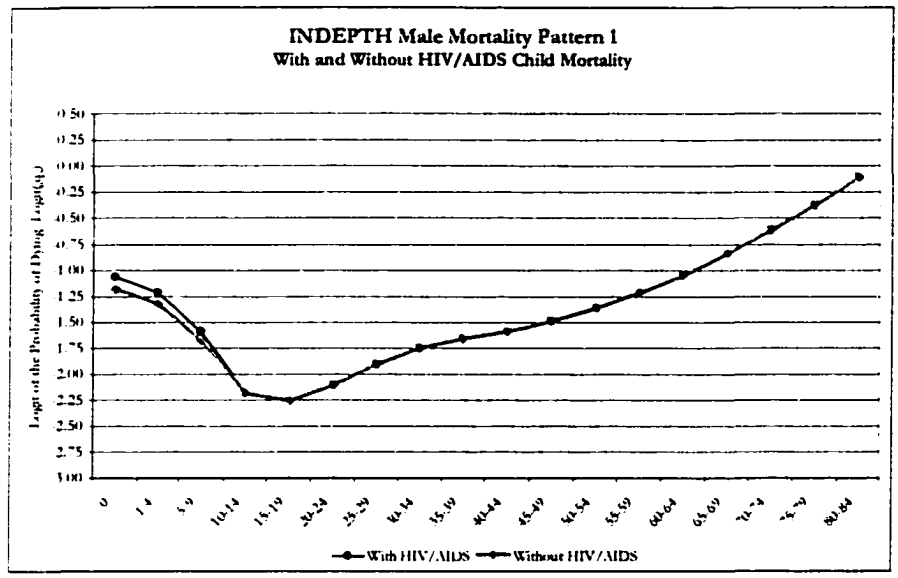

Figure 20: *INDEPTH* Male Mortality Pattern 1 With and Without HIV/AIDS Child Mortality

TABLE 9: COEFFICIENT VALUES ESTIMATED IN FIT OF HIV-DERIVED DEVIATIONS IN LOGIT($_nq_x$) ON THE MORTALITY COMPONENTS

	Fit of Adult Deviations		Fit of Child Deviations	
Factor	Male	Female	Male	Female
1	0.001794	-0.004217	-0.002822	-0.003926
2	0.069515	0.086812	0.030939	0.024063
3	-0.087825	-0.093468	0.048722	0.046546
4	-0.046538	0.007340	-0.030034	-0.033062
5	0.014998	-0.053602	-0.002600	0.017291
6	0.007024	0.071480	-0.042015	0.044176
7	0.057843	-0.026918	-0.001601	0.029769
8	0.067342	0.011817	0.015266	-0.036031
9	-0.035387	0.055790	-0.012263	0.029199
10	-0.030752	0.070519	0.028062	0.006903
11	-0.048241	0.037762	-0.013452	0.043736
12	0.040329	-0.028917	-0.001339	-0.004031
13	0.003209	0.082885	0.032003	0.025621
14	0.091293	0.089362	0.050373	-0.013287
15	0.126678	-0.048030	-0.008452	0.030330
Constant	0.062364	0.079854	-0.030420	-0.028344

MODEL LIFE TABLES

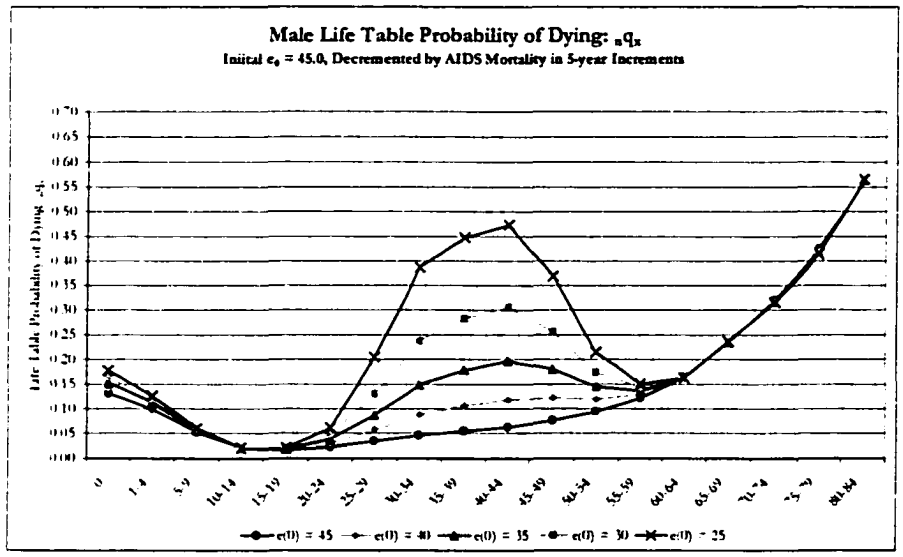

Figure 21: Male Life Table Probability of Dying: $_nq_x$

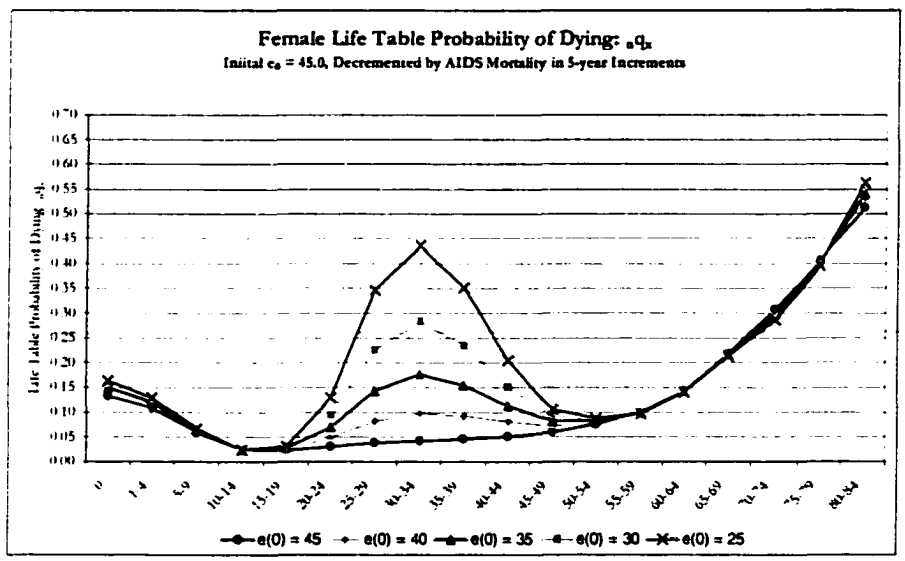

Figure 22: Female Life Table Probability of Dying: $_nq_x$

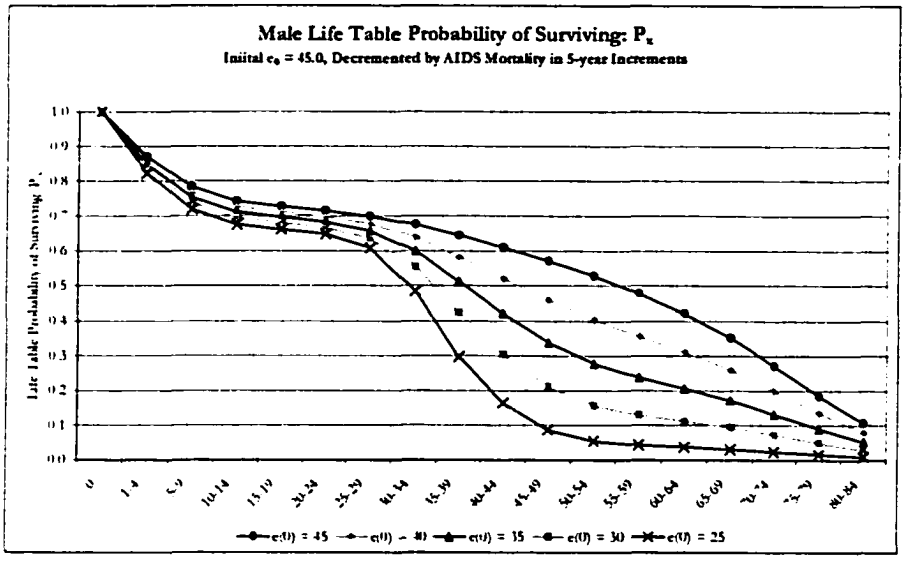

Figure 23: Male Life Table Probability of Surviving P_x

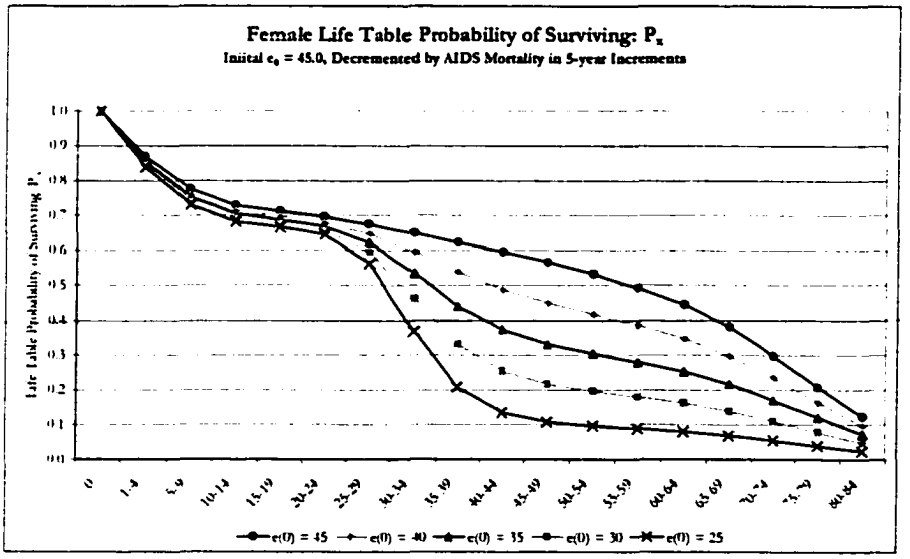

Figure 24: Female Life Table Probability of Surviving P_x

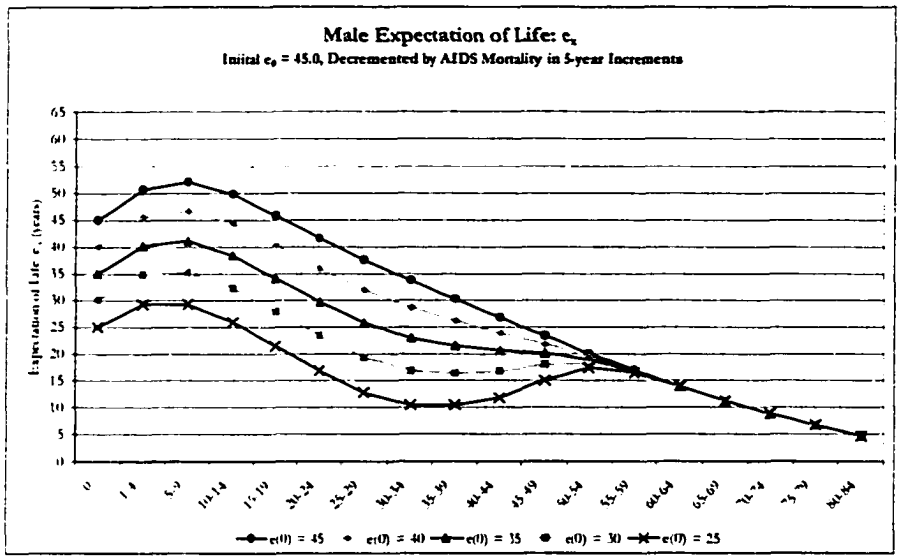

Figure 25: Male Expectation of Life e_x

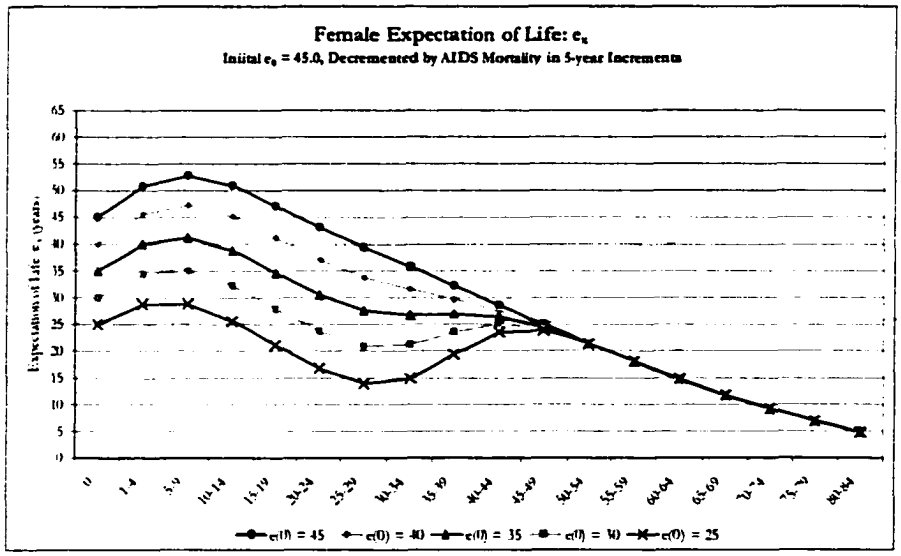

Figure 26: Female Expectation of Life e_x

TABLE 10: MODEL LIFE TABLES: *INDEPTH* PATTERN 1

EXPECTATION OF LIFE OF 60.0

DECREMENTED BY HIV/AIDS MORTALITY

	Male					Female				
	Reduction in e_0					Reduction in e_0				
Age	0.0	5.0	10.0	15.0	20.0	0.0	5.0	10.0	15.0	20.0
					$_nq_x$					
0	0.068354	0.076013	0.081963	0.087230	0.092413	0.070648	0.076841	0.081025	0.084427	0.087546
1-4	0.051414	0.055626	0.058835	0.061634	0.064355	0.057153	0.061250	0.063991	0.066204	0.068222
5-9	0.026085	0.027408	0.028394	0.029240	0.030051	0.030118	0.031523	0.032448	0.033187	0.033854
10-14	0.010159	0.010162	0.010164	0.010165	0.010167	0.011862	0.011849	0.011840	0.011834	0.011828
15-19	0.008511	0.009057	0.009469	0.009825	0.010169	0.011452	0.012871	0.013861	0.014683	0.015450
20-24	0.011362	0.015591	0.019531	0.023537	0.027986	0.015359	0.027156	0.038818	0.051094	0.064944
25-29	0.016697	0.030949	0.047772	0.068061	0.093905	0.018947	0.048985	0.087589	0.134870	0.193185
30-34	0.022747	0.049920	0.085953	0.132372	0.193248	0.021038	0.059561	0.111710	0.176657	0.256096
35-39	0.027520	0.061020	0.105336	0.161762	0.234230	0.022694	0.054981	0.094380	0.140965	0.196943
40-44	0.031406	0.068631	0.117010	0.177538	0.253830	0.025472	0.045196	0.064552	0.084710	0.107149
45-49	0.038409	0.069160	0.103772	0.143428	0.191043	0.030371	0.038020	0.043802	0.048876	0.053829
50-54	0.048610	0.064737	0.079184	0.093381	0.108628	0.038946	0.041139	0.042592	0.043757	0.044813
55-59	0.063546	0.068230	0.071772	0.074846	0.077819	0.051958	0.051479	0.051178	0.050944	0.050738
60-64	0.087337	0.087156	0.087026	0.086919	0.086818	0.076859	0.076105	0.075630	0.075262	0.074937
65-69	0.129704	0.130249	0.130641	0.130967	0.131271	0.121951	0.120671	0.119865	0.119240	0.118690
70-74	0.186536	0.185707	0.185117	0.184626	0.184170	0.180792	0.175174	0.171681	0.168998	0.166653
75-79	0.265402	0.262574	0.260564	0.258900	0.257356	0.254473	0.251341	0.249368	0.247837	0.246489
80-84	0.384085	0.385585	0.386658	0.387551	0.388385	0.343509	0.361047	0.372376	0.381299	0.389263
					P_x					
0	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
1-4	0.931646	0.923987	0.918037	0.912770	0.907587	0.929352	0.923159	0.918975	0.915573	0.912454
5-9	0.883747	0.872589	0.864024	0.856513	0.849180	0.876236	0.866616	0.860169	0.854958	0.850205
10-14	0.860695	0.848673	0.839491	0.831468	0.823661	0.849846	0.839297	0.832258	0.826585	0.821122
15-19	0.851951	0.840049	0.830959	0.823016	0.815287	0.839765	0.829352	0.822403	0.816803	0.811707
20-24	0.844700	0.832441	0.823091	0.814930	0.806996	0.830148	0.818677	0.811004	0.804810	0.799166
25-29	0.835103	0.819462	0.807015	0.795749	0.784412	0.817397	0.796445	0.779523	0.763690	0.747265
30-34	0.821160	0.794100	0.768462	0.741590	0.710751	0.801910	0.757431	0.711246	0.660691	0.602905
35-39	0.802481	0.754459	0.702410	0.643424	0.573400	0.785039	0.712318	0.631793	0.543975	0.448503
40-44	0.780396	0.708421	0.628421	0.539342	0.439092	0.767223	0.673154	0.572164	0.467294	0.360173
45-49	0.755887	0.659802	0.554890	0.443589	0.327637	0.747681	0.642730	0.535230	0.427709	0.321581
50-54	0.726854	0.614170	0.497308	0.379966	0.265045	0.724973	0.618293	0.511786	0.406804	0.304271
55-59	0.691522	0.574411	0.457929	0.344484	0.236253	0.696739	0.592857	0.489988	0.389004	0.290636
60-64	0.647578	0.535218	0.425062	0.318701	0.217868	0.660537	0.562337	0.464912	0.369186	0.275889
65-69	0.591020	0.488571	0.388071	0.291000	0.198953	0.609769	0.519541	0.429751	0.341401	0.255215
70-74	0.514363	0.424935	0.337373	0.252889	0.172837	0.535408	0.456847	0.378239	0.300692	0.224924
75-79	0.418415	0.346021	0.274920	0.206199	0.141005	0.438610	0.376819	0.313302	0.249876	0.187440
80-84	0.307367	0.255165	0.203285	0.152814	0.104717	0.326996	0.282109	0.235175	0.187947	0.141238
					e_x					
0	60.00	55.00	50.00	45.00	40.00	60.00	55.00	50.00	45.00	40.00
1-4	63.37	58.48	53.42	48.25	43.02	63.52	58.54	53.36	48.10	42.79
5-9	62.69	57.81	52.63	47.29	41.84	63.25	58.23	52.88	47.37	41.78
10-14	59.30	54.37	49.10	43.64	38.06	60.14	55.04	49.57	43.91	38.15
15-19	54.89	49.90	44.58	39.06	33.43	55.83	50.67	45.13	39.41	33.58
20-24	50.34	45.33	39.98	34.43	28.75	51.45	46.30	40.73	34.96	29.07
25-29	45.89	41.01	35.73	30.19	24.50	47.21	42.52	37.27	31.71	25.91
30-34	41.62	37.24	32.39	27.22	21.78	43.08	39.58	35.61	31.26	26.52
35-39	37.53	34.07	30.20	25.99	21.40	38.95	36.93	34.77	32.43	29.79
40-44	33.52	31.12	28.47	25.52	22.18	34.79	33.93	33.14	32.34	31.48
45-49	29.53	28.23	26.91	25.49	23.88	30.64	30.42	30.25	30.10	29.96
50-54	25.61	25.14	24.73	24.34	23.92	26.52	26.52	26.52	26.52	26.52
55-59	21.79	21.71	21.64	21.59	21.54	22.49	22.55	22.59	22.62	22.64
60-64	18.10	18.11	18.13	18.13	18.14	18.59	18.64	18.68	18.70	18.72
65-69	14.59	14.61	14.61	14.62	14.63	14.93	14.97	15.00	15.02	15.04
70-74	11.40	11.42	11.44	11.45	11.46	11.65	11.68	11.70	11.71	11.72
75-79	8.44	8.45	8.46	8.47	8.48	8.67	8.64	8.61	8.59	8.57
80-84	5.58	5.57	5.57	5.56	5.56	5.78	5.69	5.64	5.59	5.55

TABLE 11: MODEL LIFE TABLES: *INDEPTH* PATTERN 1

EXPECTATION OF LIFE OF 55.0

DECREMENTED BY HIV/AIDS MORTALITY

	Male					Female				
	Reduction in e0					Reduction in e0				
Age	0.0	5.0	10.0	15.0	20.0	0.0	5.0	10.0	15.0	20.0
					$_nq_x$					
0	0.086457	0.095211	0.102372	0.109000	0.115902	0.088792	0.095816	0.100835	0.105077	0.109142
1-4	0.065345	0.070215	0.074132	0.077708	0.081390	0.072101	0.076786	0.080106	0.082896	0.085555
5-9	0.033394	0.034950	0.036176	0.037278	0.038398	0.038281	0.039916	0.041057	0.042006	0.042903
10-14	0.013065	0.013069	0.013071	0.013074	0.013076	0.015155	0.015139	0.015129	0.015120	0.015112
15-19	0.010951	0.011599	0.012116	0.012585	0.013065	0.014633	0.016296	0.017522	0.018584	0.019622
20-24	0.014608	0.019565	0.024393	0.029535	0.035628	0.019603	0.033111	0.046933	0.062021	0.079914
25-29	0.021433	0.037844	0.057782	0.082836	0.116851	0.024158	0.057774	0.101617	0.156753	0.227435
30-34	0.029149	0.060017	0.101670	0.157114	0.233660	0.026809	0.069618	0.128095	0.202532	0.296503
35-39	0.035218	0.073105	0.123957	0.190613	0.280180	0.028906	0.065058	0.109890	0.164281	0.232142
40-44	0.040146	0.082148	0.137483	0.208616	0.302143	0.032418	0.054837	0.077539	0.101974	0.130432
45-49	0.048998	0.083893	0.124091	0.171658	0.231688	0.038600	0.047420	0.054425	0.060814	0.067345
50-54	0.061832	0.080307	0.097550	0.115199	0.135262	0.049381	0.051917	0.053699	0.055186	0.056597
55-59	0.080487	0.085865	0.090160	0.094061	0.098056	0.065641	0.065090	0.064720	0.064421	0.064146
60-64	0.109876	0.109669	0.109513	0.109377	0.109242	0.096432	0.095576	0.095001	0.094538	0.094110
65-69	0.161244	0.161849	0.162310	0.162712	0.163110	0.151128	0.149709	0.148758	0.147989	0.147281
70-74	0.228271	0.227379	0.226704	0.226116	0.225538	0.220511	0.214469	0.210459	0.207246	0.204304
75-79	0.317886	0.314969	0.312765	0.310851	0.308969	0.304365	0.301131	0.298958	0.297202	0.295583
80-84	0.445796	0.447247	0.448348	0.449309	0.450256	0.401458	0.418649	0.430372	0.439940	0.448835
					P_x					
0	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
1-4	0.913543	0.904789	0.897628	0.891000	0.884098	0.911208	0.904184	0.899165	0.894923	0.890858
5-9	0.853847	0.841259	0.831085	0.821762	0.812141	0.845509	0.834755	0.827136	0.820738	0.814640
10-14	0.825334	0.811858	0.801020	0.791128	0.780957	0.813142	0.801435	0.793176	0.786262	0.779690
15-19	0.814550	0.801248	0.790549	0.780785	0.770745	0.800819	0.789302	0.781177	0.774374	0.767907
20-24	0.805630	0.791954	0.780971	0.770959	0.760675	0.789101	0.776440	0.767489	0.759983	0.752839
25-29	0.793862	0.776459	0.761920	0.748188	0.733573	0.773632	0.750731	0.731468	0.712848	0.692677
30-34	0.776846	0.747075	0.717895	0.686211	0.647854	0.754942	0.707358	0.657138	0.601107	0.535138
35-39	0.754202	0.702238	0.644907	0.578398	0.496476	0.734703	0.658114	0.572962	0.479363	0.376468
40-44	0.727641	0.650901	0.564966	0.468148	0.357374	0.713466	0.615298	0.510000	0.400613	0.289074
45-49	0.698429	0.597430	0.487293	0.370485	0.249396	0.690337	0.581556	0.470455	0.359761	0.251370
50-54	0.664207	0.547310	0.426824	0.306888	0.191614	0.663690	0.553979	0.444850	0.337882	0.234441
55-59	0.623138	0.503357	0.385187	0.271535	0.165696	0.630916	0.525218	0.420962	0.319236	0.221172
60-64	0.572984	0.460137	0.350459	0.245994	0.149448	0.589502	0.491032	0.393718	0.298670	0.206985
65-69	0.510027	0.409674	0.312079	0.219088	0.133122	0.532655	0.444101	0.356314	0.270435	0.187506
70-74	0.427788	0.343368	0.261426	0.183440	0.111409	0.452156	0.377615	0.303310	0.230413	0.159890
75-79	0.330137	0.265294	0.202159	0.141961	0.086282	0.352451	0.296628	0.239475	0.182661	0.127223
80-84	0.225191	0.181734	0.138931	0.097832	0.059623	0.245177	0.207304	0.167882	0.128374	0.089618
					e_x					
0	55.00	50.00	45.00	40.00	35.00	55.00	50.00	45.00	40.00	35.00
1-4	59.16	54.21	49.08	43.83	38.52	59.31	54.25	48.99	43.64	38.23
5-9	59.15	54.15	48.84	43.36	37.76	59.76	54.59	49.08	43.40	37.62
10-14	56.11	51.02	45.58	39.94	34.17	57.04	51.76	46.08	40.19	34.19
15-19	51.82	46.66	41.15	35.43	29.59	52.88	47.51	41.75	35.77	29.68
20-24	47.37	42.18	36.63	30.85	24.94	48.63	43.26	37.45	31.40	25.22
25-29	43.03	37.97	32.48	26.72	20.77	44.55	39.66	34.17	28.31	22.19
30-34	38.92	34.37	29.32	23.90	18.19	40.59	36.93	32.75	28.11	22.99
35-39	35.01	31.41	27.36	22.89	17.98	36.64	34.51	32.19	29.62	26.63
40-44	31.20	28.68	25.87	22.70	19.00	32.66	31.74	30.86	29.95	28.92
45-49	27.40	26.03	24.60	23.02	21.14	28.67	28.43	28.24	28.07	27.89
50-54	23.68	23.18	22.73	22.27	21.77	24.72	24.72	24.72	24.72	24.72
55-59	20.08	19.99	19.92	19.85	19.78	20.88	20.94	20.99	21.02	21.05
60-64	16.62	16.63	16.64	16.65	16.66	17.17	17.23	17.26	17.30	17.32
65-69	13.36	13.37	13.38	13.39	13.40	13.73	13.78	13.81	13.84	13.86
70-74	10.45	10.47	10.49	10.50	10.52	10.73	10.77	10.79	10.81	10.83
75-79	7.80	7.82	7.83	7.84	7.85	8.06	8.03	8.00	7.98	7.96
80-84	5.27	5.26	5.26	5.25	5.25	5.49	5.41	5.35	5.30	5.26

TABLE 12: MODEL LIFE TABLES: *INDEPTH* PATTERN 1

EXPECTATION OF LIFE OF 50.0

DECREMENTED BY HIV/AIDS MORTALITY

	Male					Female				
	Reduction in e_0					Reduction in e_0				
Age	0.0	5.0	10.0	15.0	20.0	0.0	5.0	10.0	15.0	20.0
					$_nq_x$					
0	0.107006	0.117042	0.125659	0.134058	0.143537	0.109195	0.117155	0.123135	0.128417	0.133799
1-4	0.081323	0.086976	0.091759	0.096363	0.101503	0.089043	0.094403	0.098400	0.101912	0.105474
5-9	0.041910	0.043748	0.045276	0.046726	0.048322	0.047685	0.049590	0.050992	0.052212	0.053439
10-14	0.016485	0.016490	0.016493	0.016496	0.016499	0.018990	0.018971	0.018958	0.018947	0.018936
15-19	0.013826	0.014601	0.015253	0.015877	0.016571	0.018338	0.020296	0.021816	0.023193	0.024630
20-24	0.018424	0.024301	0.030294	0.037051	0.045852	0.024536	0.040164	0.056745	0.075678	0.099802
25-29	0.026984	0.046177	0.070323	0.102350	0.150137	0.030202	0.068421	0.119279	0.185467	0.275133
30-34	0.036623	0.072313	0.121688	0.190360	0.292481	0.033492	0.081882	0.148955	0.236782	0.352657
35-39	0.044177	0.087770	0.147553	0.229008	0.345758	0.036092	0.077161	0.129137	0.194322	0.280152
40-44	0.050293	0.098477	0.163232	0.249528	0.370014	0.040442	0.066165	0.093074	0.123237	0.160738
45-49	0.061241	0.101344	0.148810	0.207374	0.286862	0.048078	0.058291	0.066783	0.074881	0.083717
50-54	0.077021	0.098342	0.119063	0.141348	0.168860	0.061337	0.064274	0.066449	0.068348	0.070266
55-59	0.099771	0.105966	0.111170	0.116151	0.121678	0.081198	0.080563	0.080113	0.079733	0.079361
60-64	0.135167	0.134932	0.134744	0.134572	0.134388	0.118362	0.117392	0.116703	0.116122	0.115552
65-69	0.195759	0.196429	0.196968	0.197464	0.197994	0.182977	0.181413	0.180303	0.179365	0.178447
70-74	0.272473	0.271517	0.270753	0.270052	0.269305	0.262462	0.256013	0.251477	0.247668	0.243960
75-79	0.371096	0.368105	0.365717	0.363526	0.361195	0.355004	0.351692	0.349338	0.347346	0.345392
80-84	0.504579	0.505975	0.507094	0.508122	0.509220	0.457623	0.474353	0.486322	0.496497	0.506508
					P_x					
0	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
1-4	0.892994	0.882958	0.874341	0.865942	0.856463	0.890805	0.882845	0.876865	0.871583	0.866201
5-9	0.820373	0.806162	0.794112	0.782497	0.769529	0.811486	0.799502	0.790581	0.782758	0.774839
10-14	0.785991	0.770893	0.758158	0.745934	0.732344	0.772790	0.759855	0.750268	0.741889	0.733432
15-19	0.773034	0.758182	0.745654	0.733629	0.720261	0.758115	0.745440	0.736044	0.727832	0.719544
20-24	0.762346	0.747112	0.734280	0.721981	0.708326	0.744212	0.730310	0.719987	0.710952	0.701822
25-29	0.748301	0.728956	0.712036	0.695231	0.675847	0.725952	0.700978	0.679131	0.657148	0.631778
30-34	0.728108	0.695295	0.661963	0.624074	0.574378	0.704027	0.653016	0.598125	0.535269	0.457955
35-39	0.701443	0.645016	0.581411	0.505275	0.406383	0.680448	0.599546	0.509031	0.408527	0.296454
40-44	0.670455	0.588403	0.495622	0.389564	0.265873	0.655889	0.553284	0.443297	0.329142	0.213402
45-49	0.636736	0.530459	0.414720	0.292356	0.167496	0.629364	0.516676	0.402038	0.288579	0.179100
50-54	0.597741	0.476700	0.353006	0.231729	0.119448	0.599105	0.486559	0.375188	0.266970	0.164106
55-59	0.551703	0.429820	0.310976	0.198975	0.099278	0.562358	0.455286	0.350257	0.248723	0.152575
60-64	0.496659	0.384274	0.276405	0.175864	0.087198	0.516695	0.418507	0.322197	0.228892	0.140467
65-69	0.429527	0.332423	0.239161	0.152197	0.075480	0.455538	0.369466	0.284596	0.202312	0.124235
70-74	0.345443	0.267125	0.192054	0.122144	0.060535	0.372185	0.302440	0.233282	0.166025	0.102066
75-79	0.251319	0.194596	0.140055	0.089159	0.044233	0.274500	0.225011	0.174617	0.124906	0.077166
80-84	0.158056	0.122964	0.088834	0.056747	0.028256	0.177052	0.145877	0.113617	0.081520	0.050513
					e_x					
0	50.00	45.00	40.00	35.00	30.00	50.00	45.00	40.00	35.00	30.00
1-4	54.93	49.90	44.68	39.34	33.94	55.07	49.91	44.55	39.08	33.56
5-9	55.62	50.46	44.99	39.32	33.55	56.25	50.90	45.19	39.29	33.28
10-14	52.94	47.66	42.00	36.13	30.13	53.95	48.42	42.48	36.32	30.02
15-19	48.79	43.41	37.67	31.69	25.59	49.94	44.31	38.26	31.97	25.55
20-24	44.43	39.02	33.21	27.16	20.98	45.83	40.18	34.05	27.67	21.13
25-29	40.22	34.93	29.17	23.11	16.87	41.92	36.76	30.95	24.73	18.19
30-34	36.27	31.50	26.19	20.46	14.41	38.15	34.27	29.81	24.79	19.15
35-39	32.55	28.76	24.47	19.69	14.33	34.38	32.10	29.59	26.71	23.22
40-44	28.94	26.29	23.27	19.79	15.59	30.58	29.58	28.60	27.55	26.29
45-49	25.34	23.88	22.32	20.54	18.27	26.76	26.50	26.28	26.07	25.84
50-54	21.83	21.30	20.79	20.26	19.61	22.98	22.98	22.98	22.98	22.97
55-59	18.44	18.35	18.26	18.18	18.09	19.32	19.39	19.44	19.48	19.52
60-64	15.21	15.22	15.23	15.24	15.25	15.81	15.87	15.92	15.95	15.99
65-69	12.20	12.21	12.22	12.22	12.23	12.60	12.65	12.69	12.72	12.75
70-74	9.56	9.58	9.60	9.62	9.63	9.86	9.90	9.93	9.95	9.98
75-79	7.20	7.22	7.23	7.25	7.26	7.47	7.45	7.42	7.41	7.39
80-84	4.98	4.97	4.96	4.96	4.95	5.21	5.13	5.07	5.02	4.97

TABLE 13: MODEL LIFE TABLES: *INDEPTH* PATTERN 1

EXPECTATION OF LIFE OF 45.0

DECREMENTED BY HIV/AIDS MORTALITY

	Male					Female				
	Reduction in e_0					Reduction in e_0				
Age	0.0	5.0	10.0	15.0	20.0	0.0	5.0	10.0	15.0	20.0
					$_nq_x$					
0	0.130408	0.142000	0.152464	0.163365	0.177364	0.132228	0.141286	0.148429	0.155094	0.162575
1-4	0.099735	0.106357	0.112256	0.118335	0.126060	0.108342	0.114505	0.119334	0.123819	0.128833
5-9	0.051903	0.054101	0.056029	0.057987	0.060440	0.058597	0.060833	0.062566	0.064160	0.065927
10-14	0.020546	0.020551	0.020555	0.020560	0.020565	0.023498	0.023475	0.023459	0.023444	0.023427
15-19	0.017243	0.018183	0.019017	0.019874	0.020961	0.022695	0.025023	0.026924	0.028750	0.030861
20-24	0.022951	0.030035	0.037641	0.046910	0.061008	0.030319	0.048669	0.068959	0.093586	0.128831
25-29	0.033543	0.056429	0.086545	0.129663	0.204725	0.037270	0.081539	0.142109	0.224795	0.346988
30-34	0.045415	0.087531	0.147904	0.237255	0.386298	0.041297	0.097061	0.176087	0.283708	0.435645
35-39	0.054680	0.105837	0.178214	0.282343	0.446914	0.044476	0.091937	0.153658	0.234709	0.350954
40-44	0.062155	0.118483	0.196399	0.305597	0.472510	0.049783	0.079693	0.112158	0.150603	0.203719
45-49	0.075480	0.122298	0.179636	0.254615	0.368924	0.059074	0.071008	0.081403	0.091897	0.104601
50-54	0.094559	0.119443	0.144724	0.173778	0.214795	0.075127	0.078551	0.081218	0.083682	0.086423
55-59	0.121807	0.128996	0.135355	0.141867	0.150082	0.098982	0.098248	0.097699	0.097209	0.096682
60-64	0.163599	0.163330	0.163104	0.162881	0.162612	0.143019	0.141916	0.141091	0.140355	0.139561
65-69	0.233495	0.234239	0.234870	0.235492	0.236244	0.217770	0.216044	0.214754	0.213601	0.212358
70-74	0.319129	0.318106	0.317242	0.316393	0.315369	0.306695	0.299831	0.294730	0.290194	0.285335
75-79	0.424780	0.421720	0.419136	0.416599	0.413540	0.406242	0.402864	0.400332	0.398063	0.395614
80-84	0.560367	0.561711	0.562846	0.563963	0.565311	0.511918	0.528150	0.540312	0.551190	0.562900
					P_x					
0	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
1-4	0.869592	0.858000	0.847536	0.836635	0.822636	0.867772	0.858714	0.851571	0.844906	0.837425
5-9	0.782863	0.766745	0.752396	0.737631	0.718934	0.773756	0.760388	0.749949	0.740291	0.729537
10-14	0.742231	0.725264	0.710240	0.694858	0.675482	0.728416	0.714131	0.703028	0.692793	0.681441
15-19	0.726981	0.710359	0.695641	0.680572	0.661591	0.711300	0.697367	0.686536	0.676552	0.665477
20-24	0.714445	0.697442	0.682412	0.667046	0.647724	0.695157	0.679916	0.668051	0.657101	0.644940
25-29	0.698048	0.676495	0.656725	0.635755	0.608207	0.674080	0.646825	0.621983	0.595605	0.561851
30-34	0.674634	0.638321	0.599888	0.553322	0.483692	0.648958	0.594084	0.533594	0.461716	0.366896
35-39	0.643995	0.582448	0.511163	0.422043	0.296843	0.622157	0.536422	0.439635	0.330724	0.207060
40-44	0.608781	0.520803	0.420067	0.302882	0.164180	0.594486	0.487104	0.372081	0.253100	0.134391
45-49	0.570942	0.459097	0.337566	0.210322	0.086603	0.564891	0.448285	0.330350	0.214982	0.107013
50-54	0.527848	0.402951	0.276927	0.156771	0.054653	0.531520	0.416154	0.303458	0.195226	0.095819
55-59	0.477935	0.354821	0.236849	0.129528	0.042914	0.491589	0.383741	0.278812	0.178889	0.087538
60-64	0.419719	0.309050	0.204790	0.111152	0.036473	0.442931	0.346039	0.251572	0.161500	0.079075
65-69	0.351053	0.258573	0.171388	0.093047	0.030542	0.379583	0.296931	0.216078	0.138832	0.068039
70-74	0.269084	0.198005	0.131134	0.071136	0.023327	0.296921	0.232780	0.169674	0.109178	0.053591
75-79	0.183212	0.135019	0.089533	0.048629	0.015970	0.205857	0.162986	0.119666	0.077495	0.038299
80-84	0.105387	0.078079	0.052006	0.028370	0.009366	0.122229	0.097325	0.071760	0.046647	0.023148
					e_x					
0	45.00	40.00	35.00	30.00	25.00	45.00	40.00	35.00	30.00	25.00
1-4	50.67	45.54	40.21	34.76	29.28	50.78	45.50	40.01	34.42	28.76
5-9	52.07	46.72	41.04	35.16	29.22	52.71	47.12	41.16	35.00	28.71
10-14	49.78	44.25	38.32	32.17	25.94	50.83	45.01	38.74	32.22	25.56
15-19	45.77	40.12	34.08	27.79	21.43	47.00	41.04	34.62	27.94	21.12
20-24	41.53	35.82	29.69	23.30	16.83	43.03	37.03	30.50	23.69	16.71
25-29	37.45	31.85	25.75	19.33	12.76	39.30	33.79	27.58	20.88	13.81
30-34	33.66	28.61	22.96	16.83	10.41	35.72	31.57	26.73	21.21	14.82
35-39	30.14	26.11	21.51	16.29	10.38	32.15	29.69	26.91	23.62	19.33
40-44	26.74	23.91	20.63	16.72	11.76	28.53	27.45	26.34	25.09	23.43
45-49	23.35	21.78	20.06	17.98	15.05	24.90	24.61	24.36	24.10	23.79
50-54	20.05	19.47	18.90	18.26	17.38	21.30	21.30	21.29	21.29	21.28
55-59	16.88	16.77	16.68	16.58	16.45	17.83	17.90	17.95	18.00	18.05
60-64	13.88	13.89	13.90	13.91	13.92	14.51	14.58	14.63	14.67	14.72
65-69	11.10	11.11	11.12	11.12	11.13	11.52	11.58	11.62	11.66	11.70
70-74	8.72	8.75	8.76	8.78	8.80	9.03	9.08	9.11	9.14	9.18
75-79	6.64	6.66	6.67	6.69	6.71	6.92	6.89	6.88	6.86	6.84
80-84	4.70	4.69	4.69	4.68	4.67	4.94	4.86	4.80	4.74	4.69

CONCLUSION

Data describing mortality in nineteen different sites in Africa and Asia is used to identify seven new age-patterns of mortality, six of which originate solely from Africa. A component model of mortality is constructed from the raw data and used to identify clusters of similar age-patterns of mortality. These patterns are compared to the existing CD and UN model life table age-patterns of mortality and demonstrated to be systematically and individually different from the existing models. This finding supports the notion that there are unique age-patterns of mortality found in Africa, and that routine demographic and epidemiological estimations calculated from African data must take this into account. In order to make these data useful to the practicing demographers and epidemiologists, a set of model life tables based on these patterns must be constructed. *INDEPTH* is pursuing the construction of a set of *INDEPTH Model Life Tables for Africa* using the component model of mortality and based on the age-patterns of mortality presented here.

Part 2

THE DEMOGRAPHY OF THE GWEMBE TONGA

The Tonga inhabit the Gwembe Valley in Gwembe District, Southern Province, Zambia. Over the past 44 years they have been the subjects of a long-term study in Anthropology conducted by Elizabeth Colson and Thayer Scudder. The first section of this part introduces the Gwembe Tonga, their home area and the project that has studied them. I then discuss the comprehensive demographic data collected from four villages by the Gwembe Tonga Research Project. Finally, the part concludes with a detailed summary of the demographic indicators calculated from that data between 1957 and 1995. The values of those indicators form the basis for the fundamental parameters in the simulation described in Part 3.

THE GWEMBE TONGA

THE GWEMBE VALLEY

Defined by the course of the Zambezi River, the Gwembe Valley extends from the Southwest to the Northeast through an area bounded by longitudes 26° and 29° East and latitudes 16° and 18° South. Before the construction of the Kariba Dam in 1956-1958, the banks of the Zambezi River provided a rich bed of alluvial soil watered by the annual floods of the Zambezi and the rains between November and March. Following the construction of the Dam, a large portion of the Gwembe Valley floor was covered by Lake Kariba, at that time the largest man-made lake in the world – stretching for nearly 300 kilometres over the former course of the Zambezi River and occupying in excess of 5,000 square kilometres (Scudder and Colson 1977). The Zambezi River demarcates the boundary between Zambia and Zimbabwe thereby situating the Gwembe Tonga near the border with Zimbabwe in the extreme South of Zambia.

Compared to the adjacent plateaus at roughly 1,200 metres elevation, the Gwembe Valley is a low-lying area with an average elevation of about 400 metres, surrounded by imposing escarpments that mark the meeting of the Valley and its surrounding plateaus. The climate is semi-arid with an annual rainfall of 60 ± 40 cm (Scudder 1962) falling between November and March-April during the hot season. However, the pattern of rainfall in the Valley varies considerably from place to place and year to year. The temperature varies between a few degrees above 0° C at night during the cold season between May and August to over 40° C during the hot season between August and March-April.

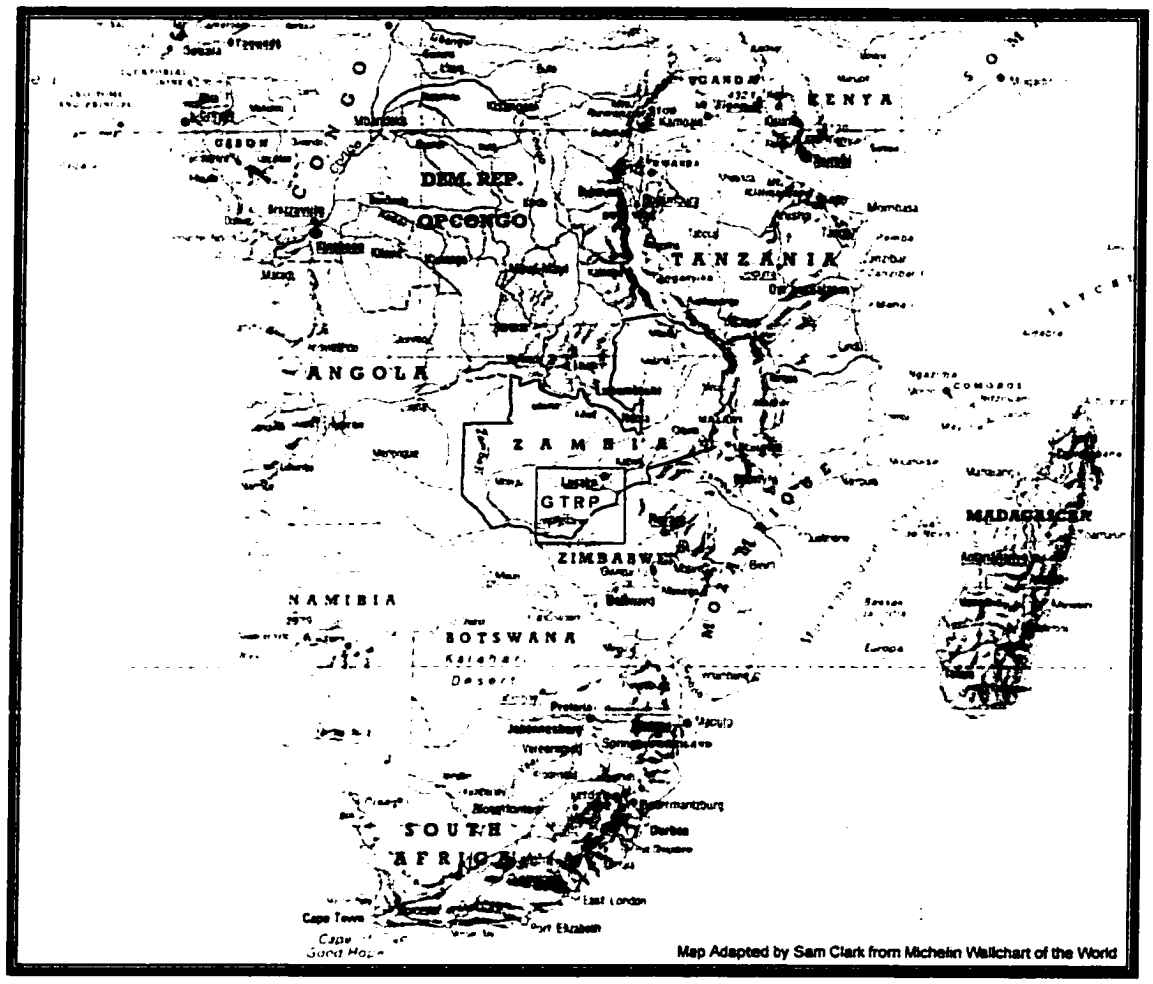

Figure 27: Zambia and the GTRP Study Area within Africa

The predominant vegetation complex is mopane woodland consisting of mopane trees (*Colophospermum mpane*) growing over various types of grass. The soils come in various types, none of which are particularly fertile. Scudder has written an exhaustive description of the

Figure 28: Gwembe Valley and Lake Kariba within Zambia

geological and ecological features of the Gwembe Valley in The Ecology of the Gwembe Tonga (Scudder 1962), and I refer the reader to that book for more information regarding the geography of the Gwembe Valley.

SOCIAL ORGANIZATION

The social organization of the Gwembe Tonga is complex and constantly evolving. The following provides the reader with the roughest kind of overview, just enough to form a crude picture in one's mind and to orient the demographic analysis that follows. The reader is directed to Colson's The Social Organization of the Gwembe Tonga (Colson 1960a) and The Social Consequences of Resettlement (Colson 1971a) for a detailed and complete treatment of this topic.

Figure 29: Nine Chakasala, Headman of Chakasala Village, 1992

The Gwembe Tonga are a matrilineal, virilocal group who practice polygynous marriage. With some variation, there are two social aggregates around which they organize themselves: the clan and the matrileage (Colson 1971a). The clan is a geographically dispersed, named body consisting of matrilineages. The matrilineage is a group of people related through the female side of their genealogy. Through the matrilineage passes the inheritance of *shades* and land-use rights. A shade or *muzimu* is the perpetual manifestation of a dead ancestor and must be inherited by one of the remaining members of the lineage upon someone's death. The shade carries with it the status, responsibilities and privileges possessed by the deceased and confers a portion of those on the inheritor along with some portion of the deceased's estate. Land is cleared and claimed by men, and when they are no longer capable of utilizing it or they die, the land is passed to an appropriately chosen member of the matrilineage.

Figure 30: Woman Pounding Millet in Sinafala Village, 1992

A visitor to the Gwembe Valley is immediately aware of the *homestead* which is the basic unit of residence for the Tonga. The physical homestead consists of a closely grouped collection of mud-brick, thatch-roofed huts, granaries and corrals. The inhabitants usually include a group of men related through various kinship ties and their dependents. Because women move to the homes of their husband at the time of marriage, the women living in a homestead may be drawn from unrelated lineages and a variety of locations (Colson 1971a). Several dozen adjacent homesteads comprise a village, and a group of half a dozen or so nearby villages forms a neighborhood or *cisi*.

Although it is becoming less common, the Tonga form polygynous families. Older men often marry more than one wife, and capable young men are also able to take additional wives. Around each wife grows a distinct household that when combined form a homestead or one of the core families of a multi-family homestead. Although all wives usually have access to the husband's stores of grain and other property, the first or senior wife, often occupies a special position in the family. To acquire a wife, a man must pay bridewealth in an amount agreed between his family and that of the prospective wife. The first payment or *lubolo* needs to be made at the time of the union or soon after. Following that, additional payments are made over a variable length of time until the bridewealth is fully paid. Unions may be broken over a

wide range of complaints originating from both the man and the woman, and from their families. It is often the case that the woman's family is displeased with the amount or pace of the marriage payments and subsequently "recalls" their daughter until acceptable payment has been made. I refer the reader to chapter four of The Social Organization of the Gwembe Tonga (Colson 1971a) for a comprehensive discussion of the Gwembe Tonga family.

The Tonga primarily occupy themselves with subsistence and small-scale cash crop agriculture. Many young people migrate to urban areas to seek wage labor, and there is some circular local migration to seek seasonal wage labor available at large commercial farming enterprises. Almost all people inhabiting the study area have access to primary education, and if they can afford it, some have access to secondary education as well. Few are able to complete secondary education, and few move on to tertiary education.

Access to the four study villages varies: Mazulu and Musulumba in the Lusitu are readily accessible via tarred and short dirt roads, but Siameja and Sinafala are much less accessible; only after several hours of driving on poor quality dirt roads. Water is provided by shallow wells, the Zambezi River, Lake Kariba, and in rare instances bore holes. The urban water supply comes mainly from bore holes and the Kafue River. The vast majority of the water is of poor quality. None of the rural villages are electrified, although there are some electrified structures in the rural areas.

All of the villages have a clinic in or near the village, and all have access, although sometimes difficult, to reasonable district hospitals. The village clinics are stocked and staffed sporadically so that the service varies considerably over time. The hospitals are more reliable, but

depending on the season, it is difficult for a villager to get to them. Additionally, the cost of medical treatment at the hospitals is a burden to most villagers, especially since IMF structural adjustment in 1992. Immunization programs have been fairly reliable and reasonably successful although I cannot quote exact figures. There are weekly or bi-weekly peri-natal clinics that operate some of the time in most of the villages. The villagers themselves would probably rate hunger as their primary health problem followed by malaria, dysentery and HIV.

RECENT HISTORY

Beginning with their forced relocation in 1957-1958, the population of the Gwembe Valley has faced a challenging series of circumstances. Relocation destroyed their old way of life and forced them to adapt to new geographical, physical, social and economic circumstances. Trends in the demographic indicators indicate that relocation had a significant negative effect on the population, but only for a brief period – about three years between 1957 and 1959.

Following relocation during the 1960s, the situation improved dramatically. The local environment was hospitable and productive with Lake Kariba producing an abundant supply of fish (Scudder 1967). 1964 brought independence for Zambia and an opening up and concurrent development of rural areas that lead to a wider incorporation of the Gwembe Tonga into the new nation state. During this period the economy functioned well and the Gwembe Valley became home to a growing number of small retailers and merchants. Villagers had access to both healthy incomes and to a range of basic consumer goods on which to spend them.

That all changed during the 1970s. Zambia does not have any natural oil reserves and it is a land-locked nation. Consequently, the world oil problems during the 1970s were particularly severe in Zambia, and when combined with the plummeting price of copper, on which Zambia was and is most dependent for national income and foreign exchange, they were sufficient to propel the national economy into a permanent decline that has not yet reversed. The cumulative result has been devastating in the Gwembe Valley. Many of the services that were provided by the colonial and immediately post-independent Zambia government have slowly decayed and disappeared - medical programs, infrastructure maintenance, education programs, agricultural outreach programs, water supply programs and many more.

Fast on the heels of the economic downturn came the War for Zimbabwean Independence. Zambia acted as a safe haven for many of the freedom fighters who waged war on the white Rhodesian government. As a result, the frontier between Zambia and Zimbabwe - running through the middle of the Gwembe Valley - became a conflict zone. Hundreds of land mines were laid, bridges and roads were destroyed and occasionally civilians were shot. Several members of the study population lost their lives in this fashion. The more lasting consequence is that the Gwembe Valley was further cut off from the rest of the country and spent a number of years in the late 1970s and early 1980s being largely neglected on the grounds that it was too dangerous and in fact pointless to do very much in the Valley.

Unfortunately, the history of the Valley did not improve after that. During the mid 1980s, a large portion of Eastern and Southern Africa suffered punishing droughts, and the Gwembe was significantly affected. The result was a prolonged shortage of food and water.

Starting at about that time, the nation began a series of structural adjustment programs. The World Bank organized the most effective set of programs beginning in 1992. Those have resulted in an accelerated decline in the economic circumstances of the poorest people in Zambia. Already an area low on the national priority list, the Gwembe Valley suffers considerably. The frail health and agricultural service provision system is in almost total collapse, and at the same time the price of basic food staples has doubled or tripled while the income a villager can obtain from selling agricultural produce has declined. The Gwembe Tonga are caught firmly in the vice grip of the tough love policy of international organizations, and at this time they have virtually no recourse but to accept what is handed down to them.

Compounding this truly bleak outlook is the HIV/AIDS crisis which appears to be having a major impact in the Gwembe. The downward trend in mortality at all ages except 10-14 has reversed as a result of the general hardship endured by the Gwembe Tonga over the past decade or so, but most strikingly, the mortality in the age groups most likely to suffer from HIV has increased much more rapidly beginning in the early 1990s. It is this dramatic reversal in fortunes that has prompted me to undertake the work presented here.

THE GWEMBE TONGA RESEARCH PROJECT - GTRP

HISTORY

In 1955, the government of the Central African Federation (Zambia and Zimbabwe today) decided to build the Kariba Dam on the Zambezi River. Henry Fosbrooke, then the director of the Rhodes Livingstone Institute (now the Institute for Economic and Social Research at the University of Zambia), realized that there would be an enormous impact on the Tonga population of the Gwembe Valley through which the Zambezi flows. He obtained funding to conduct a two-part study of the forced relocation of the Gwembe Tonga who lived in the basin of the future reservoir – what was to become Lake Kariba. His intent was to study the socio-ecological processes of change and adaptation forced upon the Tonga by their relocation and subsequent settlement into a new environment (Scudder and Colson 1977). He was also concerned that the relocation of the Tonga be based on some reasonable understanding of their existing situation: their social organization and their use of and adaptation to the environment in which they lived (Colson 1971a). To accomplish this task he recruited Social Anthropologist Elizabeth Colson who had spent several years studying the Tonga living on the plateau adjacent to the Gwembe Valley on the north and who spoke a dialect of chi-Tonga. She in turn recruited Anthropologist Thayer

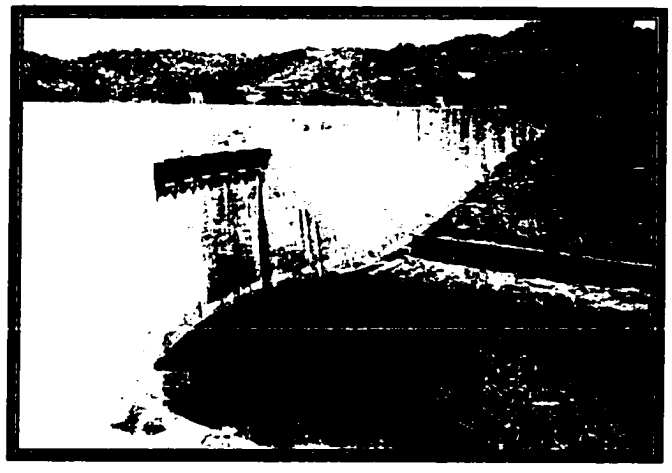

Figure 31: Kariba Dam from the South (Zimbabwe), Downstream Perspective, 1992

Scudder who had an interest in geography and ecology. Together they have been very successful in investigating the range of issues envisioned by Fosbrooke.

Colson and Scudder spent a year in 1956 conducting the first round of the study. Their plan was to make intensive investigation into seven villages, some of whom were to be moved, and some of whom were to receive relocatees. In this way they hoped to study both the relocated and the receiving communities. Unfortunately, their plans were altered when the height of the dam wall was raised significantly at a late stage in the project, with the result that the reservoir would be much larger than originally predicted. This required that many villages previously unaffected by the filling reservoir were to be flooded, and those included Colson and Scudder's proposed "receiving" villages (Colson 1971a).

Figure 32: Boys Fetching Water in the Morning, Sinafala Village, 1992

During the first round of the study, Colson and Scudder enumerated everyone in the seven villages[14] that they had chosen for study and recorded in great detail the Tonga's social organization and the relationship that they shared with their riverine environment. As part of the enumeration, Colson and Scudder recorded genealogies and administered a census form. The census form elicited information on: "place of birth, parents, marital history,

[14] Because of its large size and for other logistical reasons, half of the population of Siameja village, the largest and one of the four to remain in the study until today, was enumerated during the initial study in 1956.

children and siblings, current residence, migration histories, ownership of fields and stock, occupation, schooling, bridewealth payments, religious affiliation, participation in cults, observation of food taboos and other status attributes" (Scudder and Colson 1977). These two basic instruments would form the basis for all of the demographic and economic data collected from the study population over the next 40 years.

During 1957 and 1958, roughly 57,000 Gwembe Valley Tonga were relocated out of their riverine homelands, mainly to areas up the tributaries of the Zambezi. Two of the permanent study villages, Sinafala and Siameja, fall into this category. Siameja was moved up the Mweemba River into an area with poor soil, and Sinafala was moved to an area up the Chezia River. However, because of the lack of suitable receiving areas and the large number of people involved, some Tonga were relocated to areas along the Zambezi River below the new Dam. The remaining permanent study villages, Mazulu and Musulumba, fall into this category.

They were both moved to an area called Lusitu next to the Zambezi River about 50 kilometres downstream from Kariba Dam. The Lusitu was already inhabited, and with the addition of the immigrants, the area became (and remains today) severely overpopulated. During the two years following relocation, there were an unusually large number of deaths to women and

Figure 33: Musulumba Village, Lusitu, 1992

children, and the Lusitu immediately acquired a reputation for disaster. It was suggested that the cause of the excess deaths was severe protein malnutrition accompanied by the ingestion of poisonous plants (Colson 1971a). The new inhabitants were unfamiliar with the fauna in their new home and to alleviate the hunger they experienced as a result of losing their most recent crop to the relocation effort, they experimented with the unfamiliar plants, and some died as a result.

Colson visited the Gwembe briefly in 1960, and then both Colson and Scudder returned for one year during 1962-1963 to conduct the follow-up study as proposed in the original plan. At that time, they updated the genealogies and census in the four study villages and collected a wide range of additional information. During and after their period of fieldwork in 1962-1963, they decided to stop conducting comprehensive updates of their information pertaining to the remaining three villages that had been part of the original enumeration. It was clear by that time that the effort involved was too great to keep all seven villages.

After 1963, the study continued to examine a range of interesting issues unfolding in the Gwembe. Some of those include: the Kariba Lake fishery (Scudder 1967), patterns of labor migration, a study of shopkeepers, a large study of early secondary school leavers from the Gwembe (Scudder and Colson 1980), a study of urban migrants, a investigation into the role of locally brewed beer in society (Colson and Scudder 1987) and a number of other specific studies. From 1963 onward, the census and the genealogies have been updated on a regular basis in the four permanent study villages: Mazulu (Lusitu), Musulumba (Lusitu), Siameja[15]

[15] Updates in Siameja have been conducted on the roughly half of the village that was originally enumerated by Colson in 1956.

(Siameja) and Sinafala (Chezia). Additionally beginning in the 1970s, local informants in each village[16] have kept detailed diaries of village life, including such things as transcripts of village court cases, records of droughts and hunger periods and descriptions of conflicts between village members and how they were resolved.

Ever since the second round in 1963, Colson and Scudder have made a concerted effort to track and locate emigrants from the villages. This has resulted in an expanding study area which now includes the major urban centres of Lusaka (Zambia's capital), Ndola and Livingstone, and the smaller urban Centres of Choma, Monze, Mazabuka, Chirundu and Siavonga. Additionally, a substantial number of the study participants have emigrated to frontier rural areas to the northwest of the Gwembe Valley in an area known as Chikanta. Moreover, there are smaller numbers of study participants in other rural areas scattered throughout Zambia and even outside of Zambia. All of these emigrants are tracked to the best of our ability, and they provide the potential for a range of interesting studies.

ORGANIZATION OF THE GTRP

The GTRP has always been organized around a close association between the *Rhodes-Livingstone Institute* come *Institute for African Studies* come *Institute for Economic and Social Research* at the University of Zambia and the home institutions of Colson and Scudder, the University of California at Berkeley and Caltech respectively for most of the duration of the study. In addition to the local research assistants, a number of Zambians have worked on the project at

[16] Diaries began to be collected in Siameja in 1995-1996.

various times including Bennett Siamwiza who wrote his dissertation on hunger in the Gwembe (Siamwiza 1993).

Based on a clear understanding that all the data collected is available to both of them, and that they are both free to write on everything and pursue their own interests unhindered by the partnership (Scudder and Colson 1977), Colson and Scudder have maintained a close collaboration between the two of them over the entire duration of the study until now. Recently, they have recruited three younger scientists to carry on the study and grow it in new directions. Those include Social Anthropologist Lisa Cliggett, Physical Anthropologist Rhonda Gillett-Netting, and myself, a Demographer. This second generation is gradually taking over while Colson and Scudder focus more on writing from the data they have collected over the past 40 years.

Between 1963 and the present a number of organizations have supported the GTRP including: the African Committee of the SSRC/ACLS, the National Science Foundation, the John Simon Gugenheim Memorial Foundation, the California Institute of Technology (Caltech), the University of California at Berkeley, the Mellon Foundation, the Fulbright International Scholars program, the William Penn Foundation, the Mellon Foundation, the National Institute on Aging, the University of Pennsylvania, the Summer Undergraduate Fellowship program at Caltech and private funds from Thayer Scudder, Elizabath Colson, Lisa Cliggett, Rhonda Gillett-Netting and myself. Invaluable friendship and logistical support in Zambia has been provided by Tom Savory and his family for many years.

Colson and Scudder have conducted the bulk of the work, but they have been assisted at various times by Jonathan Habarard, Lisa Cliggett, Rhonda Gillett-Netting and myself. Without the consistent help and insight provided by the following local research assistants none of the work would have been possible: for Mazulu; Senete Adam Sikagoma, Bernard Siakanomba and Emmy Musanja; for Musulumba; Christopher Kiwani, Bernard Simalabali and Stanard; for Siameja; Ward; for Sinafala; Paul Siamwinga, Bunyika Chibilika, Jelena Chasomba, Willy Chikuni, Chester, and Shadrack; and for Lusaka; Jailos Mazambani.

THE STUDY POPULATION

SAMPLE

The sample consists of the original inhabitants of the four study villages in 1956 plus all of their progeny, those who have married original inhabitants or their progeny, and close relatives of or illegitimate parents of original inhabitants or their progeny who have taken up residence with sample members. Where possible all emigrating diaspora of the original inhabitants of the villages have been followed to their new home(s). One can be admitted to the sample in one of three ways[17]: 1) by being born to a sample member, 2) by marrying a sample member, or 3) by either giving birth to the child of a sample member or by siring a child with a sample member. Those born into the sample can exit the sample in one of two ways: 1) through death, 2) by emigrating from the village and becoming completely lost from the knowledge of anyone remaining in the sample. Those who enter the sample through marriage or parentage can exit in three ways: 1) through death, 2) by breaking or disowning the relationship that brought them into the sample, or 3) by emigrating from the village and becoming completely lost from the knowledge of anyone remaining in the sample. The distinction between exit conditions for those who are born into the sample and those who marry or parent their way into the sample is that we will make a vigorous effort to follow those who are born into the sample wherever they move, while we will make less effort to follow those who marry or parent their way into the sample once they disown their associating relationship. There are cases of villagers who have emigrated to foreign nations such as Zimbabwe, England, the

[17] There are few exceptions to these rules, but when they do occur they are evaluated on a case by case basis and usually involve someone who has moved within the physical confines of the village and has successfully integrated themselves (or their whole family) into the village group.

U.S.A., and Japan. These individuals remain in the sample as long as we have access to information on them through their relatives.

DATA

The demographic data collected by the GTRP was originally coded and input to machine readable format by Colson and Scudder beginning in the 1960s and 1970s. At that time they devised a clever but complicated scheme to code the relationships between people in the sample and all of the other information so that it could be recorded in an ASCII flat file. For the most part, they used the WordStar word processing program to enter data into plain ASCII files, an example of which is presented in Table 14:

Several features of this record keeping system have provided me with considerable challenges in the production of demographic indicators. Although they worked from the same template and agreed carefully on the organization and format they would use when coding the data, Colson and Scudder often deviated from their rules in small ways. They were much more consistent with themselves than they were with each other, but over time even their individual coding schemes varied considerably. Compounding this is the fact that they were representing very complex social organization that has evolved rapidly over the past 40 years. This lead to many unforeseen circumstances that each of them had to handle individually as they were coding the data. For the most part they found solutions that were within the general theme of the coding system, but occasionally, the solution was a one-off. Last, the files that they worked with became very large, and they sometimes placed codes in columns adjacent to the one designated for the code.

TABLE 14: SAMPLE OF RAW DATA SET: AARONNSIPA'S FAMILY, MAZULU VILLAGE

24	1 AARONNSIPA	MAZ	M19254444
24	2 FULAU	MAZW1	F19324444
24	3 BRAIN	MAZW1. 1	M19504444
24	4 ELITA	MAZW1. 1W1	F19564444
24	5 CHIKWEKWE	MAZW1. 1W1. 1	F19744444
24	5aMASOLI	MAZW1. 1W1. 1H1	M19734444
24	5bSUZEN	MAZW1. 1W1. 1H1. 1	F19924444
24	6 LORENCE	MAZW1. 1W1. 2	M19764444
24	7 MALILA	MAZW1. 1W1. 3	F19794444
24	8 MANKALIA	MAZW1. 1W1. 4	M19814444
24	8aKOPO	MAZW1. 1W1. 5	M19844444
24	8bBOY	MAZW1. 1W1. 6	M19861986
24	8cSIAMPIYE	MAZW1. 1W1. 7	F19874444
24	8dLENTI	MAZW1. 1W1. 8	F19904444
24	8eDUTY	MAZW1. 1W1. 9	F19934444
24	9 JENNY	*MAZW1. 1W2	F19604444
24	10 ANDERSON	MAZW1. 2	M19534444
24	11 GEORGINA	MAZW1. 2W1	F19574444
24	12 GIRL	MAZW1. 2W1. 1	F19761976
24	13 LONISI	MAZW1. 2W1. 2	F19774444
24	14 MBOZI	MAZW1. 2W1. 3	F19804444
24	15 VICENT	MAZW1. 2W1. 4	M19844444
24	16 JENIPHER	MAZW1. 2W2	F99994444
24	17 SUNDAY	MAZW1. 2W2. 1	M19794444
24	18 MUTINTA	MAZW1. 2W2. 2	F19814444
24	19 RUTH	MAZW1. 2W2. 3	F19831984
24	20 MISCARRIAGE	MAZW1. 2W2. 4	mS19841984
24	21 IRENE	MAZW1. 2W2. 5	F19864444
24	22 LYDEE	MAZW1. 2W2. 6	F19884444
24	22aMWAKA	MAZW1. 2W2. 7	F19904444
24	22bCHIMUNYA	MAZW1. 2W2. 8	F19924444
24	23 LOTIA	MAZW1. 2W3	F19734444
24	23aMONICA	MAZW1. 2W3. 1	F19924444
24	23bVELLA	MAZW1. 2W3. 2	F19944444

These idiosyncracies ensured that no single logic could fully decipher the data files and transform them into a representation suitable for statistical estimation. Work on the files started in 1991 and continued through the Fall of 1999. I started by going through each file record-by-record and adjusting the columns so that the codes occupied the correct columns in

the raw files. I then read them into dBASE and wrote and ran a number of programs designed to identify gross inconsistencies, and I slowly identified and fixed a number of those. The coding scheme that represented genealogies is complex and took a long time to fully understand. When I finally did understand it, it was evident that many people in the data files were represented with more than one record, in some cases many records. Most individuals had a record to code their position in their natal family and additional records to record some or all of their marriages and affairs. Comparing the raw data describing Aaronnsipa's family in Table 14 with the genealogy generated from the relational equivalent in Table 18, it is evident that many of the people appearing in Table 14 share more relationships than are represented in that section of the raw data. All of those extra relationships were coded elsewhere in the raw data, and for each, there was another record to represent the related person appearing in Table 14. These duplicate records posed a serious challenge. The first step in solving this was to go over all of the data record-by-record with Colson and Scudder and have them ensure that all of the records associated with each individual were correctly identified so that my programs could associate them. This was a tedious job that required hours of checking and rechecking with the paper records. Once that was complete, I began writing programs to reconstruct families from Colson and Scudder's coding scheme and to create my own links between family members. This step in particular was challenging and occupied the better part of one and a half years. The middle group of columns in Table 14 contains the codes that link people to one another. All of the codes in those columns have meaning including the spaces and periods. The result is a complex representation system that when recorded in an ASCII file with no input-time consistency checking resulted in a painfully inconsistent data set. I literally wrote and ran hundreds of programs to decipher those relationships; every new run revealed

another unique case or set of cases that were incorporated into the next version of the program. After some months of iterating through that cycle, I gave up and decided to print out the links I had generated and manually compare them to the raw files. Hence commenced by far the most tedious task associated with this dissertation. I spent most of a year reading each link individually and comparing it to the raw files. I fixed any inconsistencies that I uncovered and also verified that I had recorded all of the links that were represented in the raw files.

Concurrently, I experimented with various relational (see Appendix A and Appendix B) schemes for representing the data, and after several false starts, I settled on a simple but flexible data model that balances representational generality with computational efficiency and ensures that new data elements can be easily incorporated in the future. Part 4 discusses the issues involved with designing a good data model and presents an expanded version of the data model that I chose for the Gwembe demographic data. The fundamental elements of the model responsible for representing relationships within families are displayed in Figure 33, see Part 4 for details on the notation.

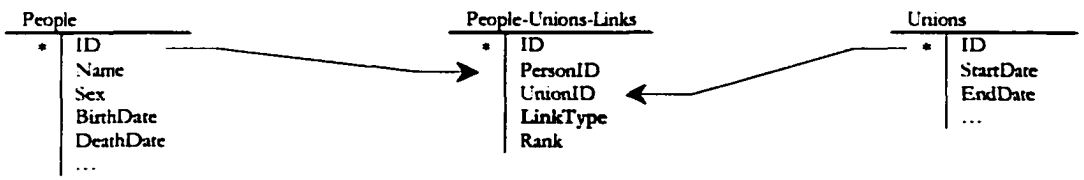

Figure 33: The Basic Relational Model of the Analysis Data Set

The data model (Figure 33) needs to be able to represent all kinds of households: monogynous, polygynous and potentially homosexual (not yet necessary, but possible in the future and certainly worth foreseeing in the basic design of the data model). For that reason, I

had to choose a scheme that relates individuals to unions without reference to their sex, that can relate more than one wife to each man concurrently (or the other way around, although polyandronous unions are not supposed to be possible, they do on occasion exist for brief periods), does not rely on accurate dates to make the associations or represent the chronological order of the unions, does not require complete knowledge of both partners to represent a union, and finally, easily allows children to be associated with the unions that created them. A very brief and incomplete description of the model follows.

People and unions are the basic items that are represented, the two outermost tables in Figure 33. The table in the middle is a relationship that links people with unions in two ways. One, each spouse participating in a union has a record in the relationship table (a link record) containing their ID and the ID of the union that they share. The fact that they are spouses is represented by the "spouse" value in the **LinkType** field in both of their link records, and the chronological order of the unions is represented by the values in the **Rank** field if the set of dates is incomplete. Two, each child produced by a union has a link record containing their ID and the ID of the union that produced them, and the fact that they are children of that union is represented by the "child" value in the **LinkType** field of their link record. If birth dates are not known, their birth order is represented by values in the **Rank** field. Using SQL (Appendix B), it is easy to associate spouses, parents and children, to count spouses or children by any combination of attributes and to make complex associations between large numbers of people, such as genealogy, and example of which is presented in Table 18.

To illustrate the organization of the analysis data set, I have located the records that describe Aaronnsipa's family in the three base tables shown in Figure 33. Table 15 is an extract from the people table containing all of the members of Aaronnsipa's family; Table 16 is the set or records from the unions table that describes the unions formed between members of Aaronnsipa's family; and Table 17 is the list of links from the people-unions-links table that represent the connections between the people and unions in Aaronnsipa's family. As an example, let us take the first record in Aaronnsipa's genealogy, Table 18. That record indicates that with his wife Fulau, Aaronnsipa had a son named Brain who married Elita and together with her had a daughter named Chikwekwe. Chikwekwe subsequently married Masoli and the two of them had a daughter named Suzen. To see how this was constructed from the three base tables, we first examine the people table and identify the people involved: 1292:Aaronnsipa, 1540:Fulau, 17:Brain, 1541:Elita, 1542:Chikwekwe, 1543:Masoli, and 1544:Suzen. The union between Aaronnsipa and Fulau is number 925 beginning in 1950 and ending in 1982. Both Aaronnsipa and Fulau have a link record to that union; numbers 14110 and 16416 respectively. Brain in linked to Aaronnsipa and Fulau's union by link record 18931. The reader can verify for themselves the unions between Brain and Elita and Chikwekwe and Masoli and the fact that Chikwekwe is a product of the first union and Suzen of the second.

TABLE 15: RECORDS IN THE PEOPLE TABLE DESCRIBING MEMBERS OF AARONNSIPA'S FAMILY[18]

ID	Name	Village	Sex	BirthYear	DeathYear
16	Jenny	Mazulu	Female	1960	~~4444~~
17	Brain	Mazulu	Male	1950	~~4444~~
18	Midress	Mazulu	Female	1978	~~4444~~
19	Charles	Mazulu	Male	1973	~~4444~~
20	Siakaligonya	Mazulu	Male	1981	1984
21	Judy	Mazulu	Female	1983	~~4444~~
22	Siamuzambi	Mazulu	Male	1986	1988
23	Mbozi	Mazulu	Female	1988	~~4444~~
24	Mudongo	Mazulu	Male	1990	1993
25	Boy	Mazulu	Male	1992	1992
26	Cheelo	Mazulu	Male	1993	~~4444~~
1291	Ciibwa	Mazulu	Female	1949	~~4444~~
1292	Aaronnsipa	Mazulu	Male	1925	~~4444~~
1296	Gertrude	Mazulu	Female	1975	~~4444~~
1297	Davis	Mazulu	Male	1960	~~4444~~
1298	Sitikila	Mazulu	Male	1969	~~4444~~
1299	Boy	Mazulu	Male	1993	1993
1300	Clever	Mazulu	Male	1969	~~4444~~
1301	Morden	Mazulu	Male	1978	~~4444~~
1501	Bridget	Mazulu	Female	1971	~~4444~~
1502	Siamuzambi	Mazulu	Male	1970	~~4444~~
1503	Siamunkala	Mazulu	Male	1991	~~4444~~
1504	Muntanga	Mazulu	Female	1993	~~4444~~
1540	Fulau	Mazulu	Female	1932	1996
1541	Elita	Mazulu	Female	1956	~~4444~~
1542	Chikwekwe	Mazulu	Female	1974	~~4444~~
1543	Masoli	Mazulu	Male	1973	~~4444~~
1544	Suzen	Mazulu	Female	1992	~~4444~~
1545	Lorence	Mazulu	Male	1976	~~4444~~
1546	Malila	Mazulu	Female	1979	1983
1547	Mankalia	Mazulu	Male	1981	~~4444~~
1548	Kopo	Mazulu	Male	1984	~~4444~~
1549	Boy	Mazulu	Male	1986	1986
1550	Siampiye	Mazulu	Female	1987	~~4444~~
1551	Lenti	Mazulu	Female	1990	~~4444~~
1552	Duty	Mazulu	Female	1993	~~4444~~
1553	Manembwa	Mazulu	Female	1995	~~4444~~
1554	Anderson	Mazulu	Male	1953	~~4444~~
1555	Georgina	Mazulu	Female	1957	~~4444~~
1556	Girl	Mazulu	Female	1976	1976
1557	Lonisi	Mazulu	Female	1977	~~4444~~
1558	Monitor	Mazulu	Male	1975	~~4444~~
1559	Mbozi	Mazulu	Female	1980	~~4444~~
1560	Vicent	Mazulu	Male	1984	~~4444~~
1570	Lotia	Mazulu	Female	1973	~~4444~~

[18] ~~4444~~ represents "forever", so a death year of ~~4444~~ indicates that the person is still alive at the time when their information was last updated.

TABLE 15: RECORDS IN THE PEOPLE TABLE DESCRIBING MEMBERS OF AARONNSIPA'S FAMILY18

ID	Name	Village	Sex	BirthYear	DeathYear
1571	Monica	Mazulu	Female	1992	++++
1572	Vella	Mazulu	Female	1994	++++
1573	Mugandanga	Mazulu	Female	1955	1996
1580	Siachikede	Mazulu	Male	1958	1960
1582	Siakagoma	Mazulu	Male	1960	1967
1583	Chikwekwe	Mazulu	Female	1963	++++
1584	David	Mazulu	Male	1957	++++
1585	Jacklin	Mazulu	Female	1983	++++
1586	Gibson	Mazulu	Male	1986	++++
1587	Tinosi	Mazulu	Male	1989	++++
1588	Chikwekwe	Mazulu	Female	1992	++++
1589	Sondo	Mazulu	Male	1956	++++
1590	Maina	Mazulu	Female	1995	++++
1591	Eunice	Mazulu	Female	1965	++++
1592	Robby	Mazulu	Male	1965	++++
1593	Fatless	Mazulu	Female	1984	++++
1594	Fidence	Mazulu	Female	1988	++++
1595	Lud	Mazulu	Female	1995	++++
1596	Dobadoba	Mazulu	Male	1968	1995
1597	Shelly	Mazulu	Female	1976	++++
1598	Boy	Mazulu	Male	1995	1995
1599	Rontiva	Mazulu	Female	1973	++++
1600	Peter	Mazulu	Male	1972	++++
1601	Mazuba	Mazulu	Female	1994	++++
1611	Maniya	Mazulu	Female	1943	++++
1612	Boy	Mazulu	Male	1962	1965
1613	Girl	Mazulu	Female	1965	1967
1614	Ngulube	Mazulu	Female	1968	1981
1615	George	Mazulu	Male	1970	++++
1619	Molly	Mazulu	Female	1972	++++
1620	Davidson	Mazulu	Male	1962	1993
1621	Boy	Mazulu	Male	1993	1993
1622	Munamwide	Mazulu	Female	1975	1975
1623	Ever	Mazulu	Female	1976	++++
1624	Bigboy	Mazulu	Male	1979	++++
1625	Mujibelo	Mazulu	Female	1983	++++
1626	Julia	Mazulu	Female	1944	++++
1628	Mbozi	Mazulu	Female	1970	1971
1629	Siachibone	Mazulu	Male	1972	++++
1630	Malala	Mazulu	Female	1978	++++
1631	Chikwekwe	Mazulu	Female	1995	++++
1632	Kephas	Mazulu	Male	1975	++++
1633	Maclina	Mazulu	Female	1978	++++
1634	Grantea	Mazulu	Male	1981	1991
1635	Miscarriage	Mazulu	Male	1991	1991

TABLE 16: RECORDS IN THE UNIONS TABLE DESCRIBING AARONNSIPA'S FAMILY

ID	StartYear	StopYear
21	1995	####
778	1993	1993
779	1993	1995
780	1995	####
924	1974	1977
925	1950	1982
930	1962	####
931	1969	####
935	1978	####
936	1973	####
939	1992	####
940	1976	1984
942	1991	1994
945	1984	####
955	1982	1992
956	1993	####
958	1983	####
959	1990	####
960	1994	1996
963	1992	####
981	1994	####
1195	1990	1993

TABLE 17: RECORDS IN THE LINKS TABLE DESCRIBING AARONNSIPA'S FAMILY

ID	lnkPeopleID	lnkUnionsID	RelationshipType
13798	17	935	Spouse
13799	17	936	Spouse
13800	19	21	Spouse
14110	1292	925	Spouse
14113	1292	930	Spouse
14114	1292	931	Spouse
14116	1292	924	Spouse
14120	1297	778	Spouse
14121	1298	779	Spouse
14122	1300	780	Spouse
14146	1502	959	Spouse
14147	1502	960	Spouse
14154	1543	939	Spouse
14155	1554	940	Spouse
14156	1554	942	Spouse
14158	1558	945	Spouse

TABLE 17: RECORDS IN THE LINKS TABLE DESCRIBING AARONNSIPA'S FAMILY

ID	lnkPeopleID	lnkUnionsID	RelationshipType
14159	1584	955	Spouse
14160	1589	956	Spouse
14161	1592	958	Spouse
14162	1600	963	Spouse
14165	1620	1195	Spouse
14166	1629	981	Spouse
16050	16	935	Spouse
16051	18	21	Spouse
16361	1291	924	Spouse
16365	1296	780	Spouse
16366	1296	779	Spouse
16367	1296	778	Spouse
16404	1501	959	Spouse
16416	1540	925	Spouse
16417	1541	936	Spouse
16418	1542	939	Spouse
16419	1555	940	Spouse
16420	1557	945	Spouse
16421	1570	942	Spouse
16424	1583	955	Spouse
16425	1583	956	Spouse
16426	1591	958	Spouse
16427	1597	960	Spouse
16428	1599	963	Spouse
16435	1611	930	Spouse
16438	1619	1195	Spouse
16439	1626	931	Spouse
16440	1630	981	Spouse
18939	26	935	Child
19905	1296	924	Child
19906	1299	778	Child
19907	1301	924	Child
20038	1502	925	Child
18931	17	925	Child
18932	18	935	Child
18933	20	935	Child
18934	21	935	Child
18935	22	935	Child
18936	23	935	Child
18937	24	935	Child
18938	25	935	Child
20039	1503	959	Child
20040	1504	959	Child
20060	1542	936	Child
20061	1544	939	Child
20062	1545	936	Child
20063	1546	936	Child
20064	1547	936	Child
20065	1548	936	Child
20066	1549	936	Child

TABLE 17: RECORDS IN THE LINKS TABLE DESCRIBING AARONNSIPA'S FAMILY

ID	lnkPeopleID	lnkUnionsID	RelationshipType
20067	1550	936	Child
20068	1551	936	Child
20069	1552	936	Child
20070	1553	936	Child
20071	1554	925	Child
20072	1556	940	Child
20073	1557	940	Child
20074	1559	940	Child
20075	1560	940	Child
20085	1571	942	Child
20086	1572	942	Child
20087	1573	925	Child
20094	1580	925	Child
20096	1582	925	Child
20097	1583	925	Child
20098	1585	955	Child
20099	1586	955	Child
20100	1587	955	Child
20101	1588	955	Child
20102	1590	956	Child
20103	1591	925	Child
20104	1593	958	Child
20105	1594	958	Child
20106	1595	958	Child
20107	1596	925	Child
20108	1598	960	Child
20109	1599	925	Child
20110	1601	963	Child
20118	1612	930	Child
20119	1613	930	Child
20120	1614	930	Child
20121	1615	930	Child
20124	1619	930	Child
20126	1621	1195	Child
20127	1622	930	Child
20128	1623	930	Child
20129	1624	930	Child
20130	1625	930	Child
20131	1628	931	Child
20132	1629	931	Child
20133	1631	981	Child
20134	1632	931	Child
20135	1633	931	Child
20136	1634	931	Child
20137	1635	931	Child

TABLE 18: SAMPLE OF QUERY ON ANALYSIS DATA SET: GENEALOGY OF AARONNSIPA'S FAMILY, MAZULU VILLAGE[19]

Generation 0				Generation 1				Generation 2				Generation 3	
ID	Name	Spouse ID	Spouse	ID	Name	Spouse ID	Spouse	ID	Name	Spouse ID	Spouse	ID	Name
1292	Aaronnsipa	1540	Fulau	17	Brain	1541	Elita	1542	Chikwekwe	1543	Masoli	1544	Suzen
1292	Aaronnsipa	1540	Fulau	17	Brain	1541	Elita	1545	Lorence				
1292	Aaronnsipa	1540	Fulau	17	Brain	1541	Elita	1546	Malila				
1292	Aaronnsipa	1540	Fulau	17	Brain	1541	Elita	1547	Mankalia				
1292	Aaronnsipa	1540	Fulau	17	Brain	1541	Elita	1548	Kopo				
1292	Aaronnsipa	1540	Fulau	17	Brain	1541	Elita	1549	Boy				
1292	Aaronnsipa	1540	Fulau	17	Brain	1541	Elita	1550	Siampiye				
1292	Aaronnsipa	1540	Fulau	17	Brain	1541	Elita	1551	Lenti				
1292	Aaronnsipa	1540	Fulau	17	Brain	1541	Elita	1552	Duty				

[19] Using Microsoft Access 2000, the unoptimized SQL used to generate Table 18 from the three base tables is:

```
qryGenealogy: SELECT Generation_1.EgoID AS Gen0ID, Generation_1.EgoName AS Gen0Name,
Generation_1.SpouseID AS Gen0SpouseID, Generation_1.SpouseName AS Gen0Spouse, Generation_1.ChildID AS
Gen1ID, Generation_1.ChildName AS Gen1Name, Generation_2.SpouseID AS Gen1SpouseID,
Generation_2.SpouseName AS Gen1Spouse, Generation_2.ChildID AS Gen2ID, Generation_2.ChildName AS
Gen2Name, Generation_3.SpouseID AS Gen2SpouseID, Generation_3.SpouseName AS Gen2Spouse,
Generation_3.ChildID AS Gen3ID, Generation_3.ChildName AS Gen3Name FROM (qryGeneration AS Generation_1
LEFT JOIN qryGeneration AS Generation_2 ON Generation_1.ChildID = Generation_2.EgoID) LEFT JOIN
qryGeneration AS Generation_3 ON Generation_2.ChildID = Generation_3.EgoID WHERE
(((Generation_1.EgoName)="Aaronnsipa")) ORDER BY Generation_1.EgoID, Generation_1.StartYear,
Generation_1.ChildID, Generation_2.StartYear, Generation_2.ChildID, Generation_3.StartYear,
Generation_3.ChildID;.

qryGeneration: SELECT qryAllPeopleSpouses.EgoID, qryAllPeopleSpouses.EgoName,
qryAllPeopleSpouses.SpouseID, qryAllPeopleSpouses.SpouseName, qryPeopleWithNatalUnions.ChildID,
qryPeopleWithNatalUnions.ChildName, qryAllPeopleSpouses.StartYear FROM qryAllPeopleSpouses LEFT JOIN
qryPeopleWithNatalUnions ON qryAllPeopleSpouses.UnionID = qryPeopleWithNatalUnions.UnionID ORDER BY
qryAllPeopleSpouses.EgoID, qryAllPeopleSpouses.StartYear.

qryPeopleWithNatalUnions: SELECT Children.NumericID AS ChildID, Children.Name AS ChildName, Unions.ID AS
UnionID, Unions.StartYear, Unions.StopYear FROM tblitmPeople AS Children LEFT JOIN (tblepdUnions AS
Unions RIGHT JOIN tbllnkPeopleUnions AS LinksChildrenToUnions ON Unions.ID =
LinksChildrenToUnions.lnkUnionsID) ON Children.NumericID = LinksChildrenToUnions.lnkPeopleID WHERE
(((LinksChildrenToUnions.RelationshipType)=1)).

QryAllPeopleSpouses: SELECT AllPeopleWithUnions.NumericID AS EgoID, AllPeopleWithUnions.Name AS EgoName,
UnionsWithAllPeople.NumericID AS SpouseID, UnionsWithAllPeople.Name AS SpouseName,
AllPeopleWithUnions.UnionID, AllPeopleWithUnions.NumericID, UnionsWithAllPeople.StartYear,
UnionsWithAllPeople.StopYear FROM qryAllPeopleUnions AS AllPeopleWithUnions LEFT JOIN qryAllPeopleUnions
AS UnionsWithAllPeople ON AllPeopleWithUnions.UnionID = UnionsWithAllPeople.UnionID WHERE
(((AllPeopleWithUnions.NumericID)<>[UnionsWithAllPeople].[NumericID])).

QryAllPeopleUnions: SELECT tblitmPeople.NumericID, tblitmPeople.Name, Unions.ID AS UnionID,
Unions.StartYear, Unions.StopYear FROM tblitmPeople LEFT JOIN (tblepdUnions AS Unions RIGHT JOIN
tbllnkPeopleUnions ON Unions.ID = tbllnkPeopleUnions.lnkUnionsID) ON tblitmPeople.NumericID =
tbllnkPeopleUnions.lnkPeopleID WHERE (((tbllnkPeopleUnions.RelationshipType)=0 Or
(tbllnkPeopleUnions.RelationshipType) Is Null)).
```

TABLE 18: SAMPLE OF QUERY ON ANALYSIS DATA SET: GENEALOGY OF AARONNSIPA'S FAMILY, MAZULU VILLAGE[19]

Generation 0				Generation 1				Generation 2				Generation 3	
ID	Name	Spouse ID	Spouse	ID	Name	Spouse ID	Spouse	ID	Name	Spouse ID	Spouse	ID	Name
1292	Aaronnsipa	1540	Fulau	17	Brain	1541	Elita	1553	Manembwa				
1292	Aaronnsipa	1540	Fulau	17	Brain	16	Jenny	18	Midress	19	Charles		
1292	Aaronnsipa	1540	Fulau	17	Brain	16	Jenny	20	Siakaligonya				
1292	Aaronnsipa	1540	Fulau	17	Brain	16	Jenny	21	Judy				
1292	Aaronnsipa	1540	Fulau	17	Brain	16	Jenny	22	Siamuzambi				
1292	Aaronnsipa	1540	Fulau	17	Brain	16	Jenny	23	Mbozi				
1292	Aaronnsipa	1540	Fulau	17	Brain	16	Jenny	24	Mudongo				
1292	Aaronnsipa	1540	Fulau	17	Brain	16	Jenny	25	Boy				
1292	Aaronnsipa	1540	Fulau	17	Brain	16	Jenny	26	Cheelo				
1292	Aaronnsipa	1540	Fulau	1502	Siamuzambi	1501	Bridget	1503	Siamunkala				
1292	Aaronnsipa	1540	Fulau	1502	Siamuzambi	1501	Bridget	1504	Muntanga				
1292	Aaronnsipa	1540	Fulau	1502	Siamuzambi	1597	Shelly	1598	Boy				
1292	Aaronnsipa	1540	Fulau	1554	Anderson	1555	Georgina	1556	Girl				
1292	Aaronnsipa	1540	Fulau	1554	Anderson	1555	Georgina	1557	Lonisi	1558	Monitor		
1292	Aaronnsipa	1540	Fulau	1554	Anderson	1555	Georgina	1559	Mbozi				
1292	Aaronnsipa	1540	Fulau	1554	Anderson	1555	Georgina	1560	Vicent				
1292	Aaronnsipa	1540	Fulau	1554	Anderson	11163	Jenipher	1561	Sunday				
1292	Aaronnsipa	1540	Fulau	1554	Anderson	11163	Jenipher	1562	Mutinta				
1292	Aaronnsipa	1540	Fulau	1554	Anderson	11163	Jenipher	1563	Ruth				
1292	Aaronnsipa	1540	Fulau	1554	Anderson	11163	Jenipher	1564	Miscarriage				
1292	Aaronnsipa	1540	Fulau	1554	Anderson	11163	Jenipher	1565	Irene				
1292	Aaronnsipa	1540	Fulau	1554	Anderson	11163	Jenipher	1566	Lydee				
1292	Aaronnsipa	1540	Fulau	1554	Anderson	11163	Jenipher	1567	Mwaka				
1292	Aaronnsipa	1540	Fulau	1554	Anderson	11163	Jenipher	1568	Chimunya				
1292	Aaronnsipa	1540	Fulau	1554	Anderson	11163	Jenipher	1569	Kisco				
1292	Aaronnsipa	1540	Fulau	1554	Anderson	1570	Lotia	1571	Monica				
1292	Aaronnsipa	1540	Fulau	1554	Anderson	1570	Lotia	1572	Vella				
1292	Aaronnsipa	1540	Fulau	1573	Mugandanga								
1292	Aaronnsipa	1540	Fulau	1580	Siachikede								
1292	Aaronnsipa	1540	Fulau	1582	Siakagoma								
1292	Aaronnsipa	1540	Fulau	1583	Chikwekwe	1584	David	1585	Jacklin				
1292	Aaronnsipa	1540	Fulau	1583	Chikwekwe	1584	David	1586	Gibson				
1292	Aaronnsipa	1540	Fulau	1583	Chikwekwe	1584	David	1587	Tinosi				
1292	Aaronnsipa	1540	Fulau	1583	Chikwekwe	1584	David	1588	Chikwekwe				
1292	Aaronnsipa	1540	Fulau	1583	Chikwekwe	1589	Sondo	1590	Maina				
1292	Aaronnsipa	1540	Fulau	1591	Eunice	1592	Robby	1593	Fatless				
1292	Aaronnsipa	1540	Fulau	1591	Eunice	1592	Robby	1594	Fidence				
1292	Aaronnsipa	1540	Fulau	1591	Eunice	1592	Robby	1595	Lud				
1292	Aaronnsipa	1540	Fulau	1596	Dobadoba								
1292	Aaronnsipa	1540	Fulau	1599	Rontiya	1600	Peter	1601	Mazuba				
1292	Aaronnsipa	1611	Mariya	1612	Boy								
1292	Aaronnsipa	1611	Mariya	1613	Girl								
1292	Aaronnsipa	1611	Mariya	1614	Ngulube								
1292	Aaronnsipa	1611	Mariya	1615	George								
1292	Aaronnsipa	1611	Mariya	1619	Molly	1620	Davidson	1621	Boy				
1292	Aaronnsipa	1611	Mariya	1622	Munamwide								
1292	Aaronnsipa	1611	Mariya	1623	Ever								
1292	Aaronnsipa	1611	Mariya	1624	Bigboy								
1292	Aaronnsipa	1611	Mariya	1625	Mujibelo								

TABLE 18: SAMPLE OF QUERY ON ANALYSIS DATA SET: GENEALOGY OF AARONNSIPA'S FAMILY, MAZULU VILLAGE[19]

Generation 0				Generation 1				Generation 2				Generation 3	
ID	Name	Spouse ID	Spouse	ID	Name	Spouse ID	Spouse	ID	Name	Spouse ID	Spouse	ID	Name
1292	Aaronnsipa	1626	Julia	1628	Mbozi								
1292	Aaronnsipa	1626	Julia	1629	Siachibone	1630	Malala	1631	Chikwekwe				
1292	Aaronnsipa	1626	Julia	1632	Kephas								
1292	Aaronnsipa	1626	Julia	1633	Maclina								
1292	Aaronnsipa	1626	Julia	1634	Grantea								
1292	Aaronnsipa	1626	Julia	1635	Miscarriage								
1292	Aaronnsipa	1291	Ciibwa	1296	Gertrude	1298	Sitikila						
1292	Aaronnsipa	1291	Ciibwa	1296	Gertrude	1297	Davis	1299	Boy				
1292	Aaronnsipa	1291	Ciibwa	1296	Gertrude	1300	Clever						
1292	Aaronnsipa	1291	Ciibwa	1301	Morden								

The final step in preparing the data involved the conversion of the cleaned and fully linked raw files into the simple relational representation shown in Figure 33. I constructed the new database using Microsoft Access, and from that point on, the data were all managed and manipulated using Microsoft Visual Basic and Microsoft Access SQL, both embedded in Microsoft Access. The conversion step was relatively straightforward and set the stage for rapid progress with analysis.

Utilizing the benefits of the relational representation, I was able to write efficient, consistent SQL and Visual Basic code to select appropriate records for analysis and finally create the analysis files. The first step identified individuals in the sample for whom all important dates were known, and for whom connections between spouses and parents and children were also known. The dates I required are: birth date, death date, date of censorship if lost to observation, and all dates describing conjugal unions. These data selection processes removed roughly 3,000 individual records from the original 14,201 unique individual records to result in

a clean and complete individual data set containing 11,162 individuals, the composition of which is described below. Similar selection on the unions table removed roughly 700 unions from the original 4,000 to yield 3,300 well-described unions. The time period covered by this dataset extends from the earliest retrospective data that Colson and Scudder collected on the original sample in 1956 up until 1995. Because of the retrospective data, there are dates stretching back into the late 1800s for the oldest members of the original enumeration.

The tables with which I finally started the analysis had the following attributes:

PEOPLE TABLE

TABLE 19: DESCRIPTION OF THE PEOPLE TABLE

Name: TblitmPeople
Records: 11,162
Fields:

Name	Type	Description
ID	Number	Unique ID for this person
Name	Text	This person's name
Village	Text	This person's village
Sex	Number	This person's sex
Twin	Binary	This person is a member of a pair of twins
SameSexTwin	Binary	This person is a same-sex twin
BirthYear	Number	This person's year of birth
DeathYear	Number	This person's year of death
EntryYear	Number	The year this person entered observation
InEvent	Number	Event that initiated observation of this person
ExitYear	Number	The year this person exited observation
OutEvent	Number	Event that concluded observation of this person

UNIONS TABLE

TABLE 20: DESCRIPTION OF THE UNIONS TABLE

Name: TblepdUnions
Records: 3,353
Fields:

Name	Type	Description
ID	Number	Unique ID for this union
StartType	Number	Type of initiation for this union
StopType	Number	Type of termination for this union
StartYear	Number	Year this union started
StopYear	Number	Year this union ended

LINKS TABLE

TABLE 21: DESCRIPTION OF THE LINKS TABLE

Name: tbllnkPeopleUnions
Records: 13,789
Fields:

Name	Type	Description
ID	Number	Unique ID for this link
LnkPeopleID	Number	Link ID to the person associated with this link
lnkUnionsID	Number	Link ID to the Union associated with this link
LnkType	Number	Type of this link
LnkRank	Number	Rank of this link

You will notice that all of the date fields are labeled **Year**. That reveals one of the major limitations of this dataset. The majority of dates are defined with the precision of one year; hence it is not possible to do any calculations with periods less than one year without making assumptions about the timing of the event of interest over the period of one year.

SIZE AND COMPOSITION

The following sections present the overall size and composition of the well-defined sample of the Gwembe Study population over time. This information is included to familiarize the

reader with the general characteristics of the data set from which all of the demographic indicators were calculated. The notation "1/1/19XX" indicates that the values refer to reference dates at the beginning of a calendar year; for example, a growth rate is calculated as the proportional growth between the first of January on consecutive years.

POPULATION PYRAMID

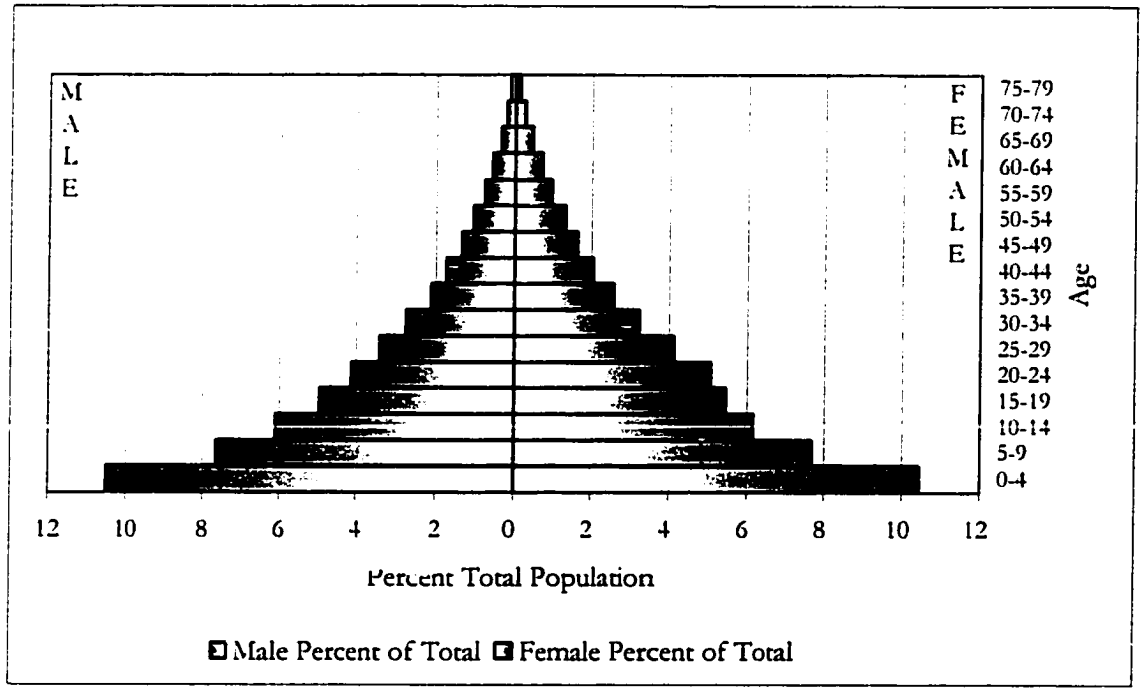

Figure 34: Gwembe Study Population Pyramid 1957-1995: Person Years Lived in Each Age Group as a Percent of Total Person Years Lived (Male and Female)

POPULATION SIZE, GROWTH RATES, AND SEX RATIO BY YEAR

Table 22 through Table 24 display the size of the sample at the beginning of each year of observation with the numbers of births and deaths during each subsequent year. The demographic balancing equation is used to calculate the net additions to the sample that

cannot be accounted for by the births and deaths during the year. Two growth rates are calculated from those numbers; 1) the total growth rate including growth (or decline) from sources other than natural increase, and 2) the growth of the population due to natural increase only.

The weighted average of the annual "net additions" to the sample is 21 indicating that a non-negligible number of people join the sample for various reasons during each year. These are primarily women (and their children) marrying into the sample.

TABLE 22: SIZE AND SEX COMPOSITION OF ANALYSIS SAMPLE OF THE GWEMBE STUDY POPULATION BY YEAR

	Additions and Subtractions			Growth Rates			
Year	Births During Year	Deaths During Year	Net Additions to Sample During Year	Recorded Sample Population (1/1/19XX)	Annual Proportional Growth of Recorded Sample Population (1/1/19XX)	Annual Proportional Growth due to Natural Increase (1/1/19XX)	Sex Ratio Male:Female per 100 (1/1/19XX)
1957	74	40	22	1199	-NA-	-NA-	81.7
1958	113	53	8	1255	4.56%	2.80%	82.1
1959	79	46	19	1323	5.28%	4.67%	85.0
1960	97	33	22	1375	3.86%	2.46%	87.1
1961	71	30	55	1461	6.07%	4.55%	87.1
1962	115	31	38	1557	6.36%	2.77%	86.7
1963	107	33	24	1679	7.54%	5.25%	87.2
1964	112	55	35	1777	5.67%	4.31%	87.6
1965	153	42	22	1869	5.05%	3.16%	86.3
1966	121	35	25	2002	6.87%	5.77%	86.6
1967	116	60	40	2113	5.40%	4.21%	86.3
1968	168	47	25	2209	4.44%	2.62%	87.5
1969	142	40	28	2355	6.40%	5.33%	88.4
1970	181	41	26	2485	5.37%	4.24%	89.0
1971	137	49	57	2651	6.47%	5.48%	90.0
1972	206	38	14	2796	5.33%	3.27%	90.7
1973	179	50	19	2978	6.31%	5.83%	91.9
1974	156	57	32	3126	4.85%	4.24%	92.7
1975	203	48	35	3257	4.11%	3.12%	94.0
1976	211	40	38	3447	5.67%	4.65%	93.4
1977	181	42	29	3656	5.89%	4.84%	92.8
1978	192	79	42	3824	4.49%	3.73%	92.5
1979	201	79	37	3979	3.97%	2.91%	90.7
1980	229	58	57	4138	3.92%	3.02%	90.3
1981	239	60	44	4366	5.36%	4.05%	90.4
1982	209	63	-81	4589	4.98%	4.02%	91.0
1983	218	46	26	4654	1.41%	3.13%	92.0
1984	238	55	50	4852	4.17%	3.63%	92.2
1985	243	74	32	5085	4.69%	3.70%	91.8
1986	277	77	38	5286	3.88%	3.27%	91.0
1987	297	86	-16	5524	4.40%	3.71%	90.9
1988	253	83	-24	5719	3.47%	3.75%	90.1
1989	287	90	29	5865	2.52%	2.93%	90.5
1990	358	83	44	6091	3.78%	3.30%	90.5
1991	370	114	28	6410	5.10%	4.42%	90.4
1992	314	80	-59	6694	4.34%	3.92%	90.4
1993	301	145	36	6869	2.58%	3.44%	90.8
1994	299	102	28	7061	2.76%	2.25%	90.4
1995	332	123	-NA-	7286	3.14%	2.75%	91.3
			21		4.33%	3.68%	90.5

TABLE 23: SIZE AND GROWTH OF ANALYSIS SAMPLE: MALE

	Additions and Subtractions			Growth Rates		
Year	Births During Year	Deaths During Year	Net Additions to Sample During Year	Recorded Sample Population (1/1/19XX)	Annual Proportional Growth of Recorded Sample Population (1/1/19XX)	Annual Proportional Growth due to Natural Increase (1/1/19XX)
1957	45	23	5	539	-NA-	-NA-
1958	64	27	5	566	4.89%	4.00%
1959	40	20	12	608	7.16%	6.33%
1960	55	19	4	640	5.13%	3.24%
1961	37	17	23	680	6.06%	5.47%
1962	56	15	18	723	6.13%	2.90%
1963	54	16	10	782	7.84%	5.52%
1964	49	25	12	830	5.96%	4.74%
1965	72	16	7	866	4.25%	2.85%
1966	62	21	9	929	7.02%	6.27%
1967	60	29	21	979	5.24%	4.32%
1968	92	29	11	1031	5.18%	3.12%
1969	74	21	12	1105	6.93%	5.93%
1970	100	21	7	1170	5.72%	4.68%
1971	68	24	30	1256	7.09%	6.53%
1972	110	20	6	1330	5.72%	3.44%
1973	89	22	11	1426	6.97%	6.55%
1974	88	26	12	1504	5.33%	4.59%
1975	100	27	14	1578	4.80%	4.04%
1976	98	19	16	1665	5.37%	4.52%
1977	87	19	10	1760	5.55%	4.64%
1978	75	39	18	1838	4.34%	3.79%
1979	91	34	15	1892	2.90%	1.94%
1980	117	29	21	1964	3.73%	2.97%
1981	121	29	22	2073	5.40%	4.38%
1982	113	36	-34	2187	5.35%	4.34%
1983	115	27	9	2230	1.95%	3.46%
1984	120	34	21	2327	4.26%	3.87%
1985	118	43	10	2434	4.50%	3.63%
1986	138	43	17	2519	3.43%	3.03%
1987	143	53	-11	2631	4.35%	3.70%
1988	126	41	-9	2710	2.96%	3.36%
1989	145	44	7	2786	2.77%	3.09%
1990	180	45	14	2894	3.80%	3.56%
1991	192	65	8	3043	5.02%	4.56%
1992	148	29	-28	3178	4.34%	4.09%
1993	162	87	9	3269	2.82%	3.68%
1994	167	50	7	3353	2.54%	2.27%
1995	166	74	-NA-	3477	3.63%	3.43%
			7		4.40%	3.89%

TABLE 24: SIZE AND GROWTH OF ANALYSIS SAMPLE: FEMALE

	Additions and Subtractions			Growth Rates		
Year	Births During Year	Deaths During Year	Net Additions to Sample During Year	Recorded Sample Population (1/1/19XX)	Annual Proportional Growth of Recorded Sample Population (1/1/19XX)	Annual Proportional Growth due to Natural Increase (1/1/19XX)
1957	29	17	17	660	-NA-	-NA-
1958	49	26	3	689	4.30%	1.80%
1959	39	26	7	715	3.70%	3.28%
1960	42	14	18	735	2.76%	1.80%
1961	34	13	32	781	6.07%	3.74%
1962	59	16	20	834	6.57%	2.65%
1963	53	17	14	897	7.28%	5.03%
1964	63	30	23	947	5.42%	3.93%
1965	81	26	15	1003	5.75%	3.43%
1966	59	14	16	1073	6.75%	5.34%
1967	56	31	19	1134	5.53%	4.11%
1968	76	18	14	1178	3.81%	2.18%
1969	68	19	16	1250	5.93%	4.81%
1970	81	20	19	1315	5.07%	3.85%
1971	69	25	27	1395	5.91%	4.53%
1972	96	18	8	1466	4.96%	3.11%
1973	90	28	8	1552	5.70%	5.18%
1974	68	31	20	1622	4.41%	3.92%
1975	103	21	21	1679	3.45%	2.26%
1976	113	21	22	1782	5.95%	4.77%
1977	94	23	19	1896	6.20%	5.03%
1978	117	40	24	1986	4.64%	3.68%
1979	110	45	22	2087	4.96%	3.80%
1980	112	29	36	2174	4.08%	3.07%
1981	118	31	22	2293	5.33%	3.75%
1982	96	27	-47	2402	4.64%	3.72%
1983	103	19	17	2424	0.91%	2.83%
1984	118	21	29	2525	4.08%	3.41%
1985	125	31	22	2651	4.87%	3.77%
1986	139	34	21	2767	4.28%	3.48%
1987	154	33	-5	2893	4.45%	3.72%
1988	127	42	-15	3009	3.93%	4.10%
1989	142	46	22	3079	2.30%	2.79%
1990	178	38	30	3197	3.76%	3.07%
1991	178	49	20	3367	5.18%	4.29%
1992	166	51	-31	3516	4.33%	3.76%
1993	139	58	27	3600	2.36%	3.22%
1994	132	52	21	3708	2.96%	2.23%
1995	166	49	-NA-	3809	2.69%	2.13%
			14		4.26%	3.50%

AGE COMPOSITION OF MALE POPULATION BY YEAR

TABLE 25: MALE SAMPLE POPULATION BY AGE AND YEAR

	Age at Beginning of Five-Year Interval																	
Year	0-79	0	1	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75
1957	584	45	103	87	58	62	33	69	34	30	19	14	9	6	7	3	5	0
1958	630	64	112	86	76	49	41	67	38	29	24	10	12	8	5	4	5	0
1959	648	40	138	93	77	58	39	66	38	28	28	8	13	9	4	4	4	1
1960	695	55	143	91	87	55	51	42	63	25	35	10	14	8	5	7	3	1
1961	717	37	149	107	85	59	56	45	69	25	35	12	13	9	4	8	1	3
1962	779	56	150	116	89	60	73	43	70	32	31	17	15	9	6	7	2	3
1963	836	54	155	142	87	81	60	55	72	35	29	24	11	12	7	5	3	4
1964	879	49	168	154	93	85	71	51	75	36	26	30	10	13	7	4	4	3
1965	938	72	160	169	91	95	67	61	51	63	24	38	10	15	9	3	7	3
1966	991	62	194	161	110	92	76	65	54	69	23	38	12	15	9	2	8	1
1967	1,039	60	194	178	117	98	72	80	47	74	31	32	19	16	8	4	7	2
1968	1,123	92	194	176	143	93	93	67	61	76	36	30	25	10	14	5	5	3
1969	1,179	74	228	185	151	98	96	77	59	77	37	29	29	9	15	7	4	4
1970	1,270	100	229	205	170	93	107	72	67	57	63	24	40	9	16	9	3	6
1971	1,324	68	269	222	161	111	103	84	68	61	69	24	40	10	15	10	2	7
1972	1,440	110	285	220	179	121	109	83	85	52	76	34	32	20	16	9	3	6
1973	1,515	89	307	261	171	148	104	102	69	66	75	38	29	24	11	13	4	4
1974	1,592	88	322	282	181	155	112	104	82	60	81	37	30	27	9	14	5	3
1975	1,678	100	311	305	204	171	102	121	75	75	57	64	25	36	10	13	6	3
1976	1,763	98	336	318	225	164	120	117	86	75	60	70	25	38	11	12	7	1
1977	1,847	87	333	367	225	183	126	124	85	90	51	74	33	31	18	13	6	1
1978	1,913	75	329	371	262	176	152	121	105	75	65	74	37	29	23	8	9	2
1979	1,983	91	307	385	280	188	163	130	105	89	58	78	35	29	26	6	10	3
1980	2,081	117	298	383	301	210	182	113	124	78	79	53	62	26	31	8	11	5
1981	2,194	121	315	403	315	232	177	130	128	87	78	59	66	25	33	9	10	6
1982	2,300	113	356	379	361	229	202	131	136	85	95	50	73	33	25	15	11	6
1983	2,345	115	374	347	361	261	197	161	127	102	78	62	70	33	24	17	7	9
1984	2,447	120	396	352	375	273	214	166	135	103	90	57	73	31	27	19	5	11
1985	2,552	118	397	372	371	297	220	208	113	126	80	77	53	53	24	27	6	10
1986	2,657	138	391	387	391	308	240	204	127	126	88	77	57	57	23	28	7	8
1987	2,774	143	414	407	372	353	236	229	127	134	85	90	46	64	31	22	13	8
1988	2,836	126	430	440	343	357	266	218	158	127	99	75	61	64	33	19	14	6
1989	2,931	145	433	455	341	374	277	231	164	135	100	86	54	65	30	21	16	4
1990	3,074	180	463	455	356	370	302	220	210	112	120	76	74	47	48	20	17	4
1991	3,235	192	514	467	375	388	315	237	209	121	121	82	72	52	48	20	18	4
1992	3,326	148	551	499	391	373	356	233	224	119	125	80	83	41	53	27	15	8
1993	3,431	162	558	515	428	342	356	264	216	148	117	93	70	56	53	29	15	9
1994	3,520	167	551	534	455	343	374	276	218	153	119	93	81	49	55	27	17	8
1995	3,643	166	557	574	451	360	369	301	205	197	101	108	74	65	44	38	19	14

TABLE 26: MALE SAMPLE PROPORTION BY AGE AND YEAR (PERCENT)

	Age at Beginning of Five-Year Interval																			
Year	0-14	65-79	10-79	0	1	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75
1957	50.17	1.37	59.76	7.71	17.64	14.90	9.93	10.62	5.65	11.82	5.82	5.14	3.25	2.40	1.54	1.03	1.20	0.51	0.86	0.00
1958	53.65	1.43	58.41	10.16	17.78	13.65	12.06	7.78	6.51	10.63	6.03	4.60	3.81	1.59	1.90	1.27	0.79	0.63	0.79	0.00
1959	53.70	1.39	58.18	6.17	21.30	14.35	11.88	8.95	6.02	10.19	5.86	4.32	4.32	1.23	2.01	1.39	0.62	0.62	0.62	0.15
1960	54.10	1.58	58.42	7.91	20.58	13.09	12.52	7.91	7.34	6.04	9.06	3.60	5.04	1.44	2.01	1.15	0.72	1.01	0.43	0.14
1961	52.72	1.67	59.14	5.16	20.78	14.92	11.85	8.23	7.81	6.28	9.62	3.49	4.88	1.67	1.81	1.26	0.56	1.12	0.14	0.42
1962	52.76	1.54	58.66	7.19	19.26	14.89	11.42	7.70	9.37	5.52	8.99	4.11	3.98	2.18	1.93	1.16	0.77	0.90	0.26	0.39
1963	52.39	1.44	58.01	6.46	18.54	16.99	10.41	9.69	7.18	6.58	8.61	4.19	3.47	2.87	1.32	1.44	0.84	0.60	0.36	0.48
1964	52.79	1.25	57.79	5.57	19.11	17.52	10.58	9.67	8.08	5.80	8.53	4.10	2.96	3.41	1.14	1.48	0.80	0.46	0.46	0.34
1965	52.45	1.39	57.25	7.68	17.06	18.02	9.70	10.13	7.14	6.50	5.44	6.72	2.56	4.05	1.07	1.60	0.96	0.32	0.75	0.32
1966	53.18	1.11	57.92	6.26	19.58	16.25	11.10	9.28	7.67	6.56	5.45	6.96	2.32	3.83	1.21	1.51	0.91	0.20	0.81	0.10
1967	52.84	1.25	58.42	5.77	18.67	17.13	11.26	9.43	6.93	7.70	4.52	7.12	2.98	3.08	1.83	1.54	0.77	0.38	0.67	0.19
1968	53.87	1.16	58.86	8.19	17.28	15.67	12.73	8.28	8.28	5.97	5.43	6.77	3.21	2.67	2.23	0.89	1.25	0.45	0.45	0.27
1969	54.11	1.27	58.69	6.28	19.34	15.69	12.81	8.31	8.14	6.53	5.00	6.53	3.14	2.46	2.46	0.76	1.27	0.59	0.34	0.34
1970	55.43	1.42	57.95	7.87	18.03	16.14	13.39	7.32	8.43	5.67	5.28	4.49	4.96	1.89	3.15	0.71	1.26	0.71	0.24	0.47
1971	54.38	1.44	57.78	5.14	20.32	16.77	12.16	8.38	7.78	6.34	5.14	4.61	5.21	1.81	3.02	0.76	1.13	0.76	0.15	0.53
1972	55.14	1.25	57.29	7.64	19.79	15.28	12.43	8.40	7.57	5.76	5.90	3.61	5.28	2.36	2.22	1.39	1.11	0.63	0.21	0.42
1973	54.65	1.39	56.63	5.87	20.26	17.23	11.29	9.77	6.86	6.73	4.55	4.36	4.95	2.51	1.91	1.58	0.73	0.86	0.26	0.26
1974	54.84	1.38	56.53	5.53	20.23	17.71	11.37	9.74	7.04	6.53	5.15	3.77	5.09	2.32	1.88	1.70	0.57	0.88	0.31	0.19
1975	54.83	1.31	57.33	5.96	18.53	18.18	12.16	10.19	6.08	7.21	4.47	4.47	3.40	3.81	1.49	2.15	0.60	0.77	0.36	0.18
1976	55.42	1.13	57.35	5.56	19.06	18.04	12.76	9.30	6.81	6.64	4.88	4.25	3.40	3.97	1.42	2.16	0.62	0.68	0.40	0.06
1977	54.79	1.08	57.39	4.71	18.03	19.87	12.18	9.91	6.82	6.71	4.60	4.87	2.76	4.01	1.79	1.68	0.97	0.70	0.32	0.05
1978	54.21	0.99	59.49	3.92	17.20	19.39	13.70	9.20	7.95	6.33	5.49	3.92	3.40	3.87	1.93	1.52	1.20	0.42	0.47	0.10
1979	53.61	0.96	60.51	4.59	15.48	19.42	14.12	9.48	8.22	6.56	5.30	4.49	2.92	3.93	1.77	1.46	1.31	0.30	0.50	0.15
1980	52.81	1.15	61.65	5.62	14.32	18.40	14.46	10.09	8.75	5.43	5.96	3.75	3.80	2.55	2.98	1.25	1.49	0.38	0.53	0.24
1981	52.60	1.14	61.76	5.52	14.36	18.37	14.36	10.57	8.07	5.93	5.83	3.97	3.56	2.69	3.01	1.14	1.50	0.41	0.46	0.27
1982	52.57	1.39	63.13	4.91	15.48	16.48	15.70	9.96	8.78	5.70	5.91	3.70	4.13	2.17	3.17	1.43	1.09	0.65	0.48	0.26
1983	51.04	1.41	64.35	4.90	15.95	14.80	15.39	11.13	8.40	6.87	5.42	4.35	3.33	2.64	2.99	1.41	1.02	0.72	0.30	0.38
1984	50.80	1.43	64.53	4.90	16.18	14.38	15.32	11.16	8.75	6.78	5.52	4.21	3.68	2.33	2.98	1.27	1.10	0.78	0.20	0.45
1985	49.29	1.68	65.24	4.62	15.56	14.58	14.54	11.64	8.62	8.15	4.43	4.94	3.13	3.02	2.08	2.08	0.94	1.06	0.24	0.39
1986	49.19	1.62	65.53	5.19	14.72	14.57	14.72	11.59	9.03	7.68	4.78	4.74	3.31	2.90	2.15	2.15	0.87	1.05	0.26	0.30
1987	48.16	1.55	65.25	5.16	14.92	14.67	13.41	12.73	8.51	8.26	4.58	4.83	3.06	3.24	1.66	2.31	1.12	0.79	0.47	0.29
1988	47.21	1.38	64.88	4.44	15.16	15.51	12.09	12.59	9.38	7.69	5.57	4.48	3.49	2.64	2.15	2.26	1.16	0.67	0.49	0.21
1989	46.88	1.40	64.76	4.95	14.77	15.52	11.63	12.76	9.45	7.88	5.60	4.61	3.41	2.93	1.84	2.22	1.02	0.72	0.55	0.14
1990	47.30	1.33	64.28	5.86	15.06	14.80	11.58	12.04	9.82	7.16	6.83	3.64	3.90	2.47	2.41	1.53	1.56	0.65	0.55	0.13
1991	47.85	1.30	63.74	5.94	15.89	14.44	11.59	11.99	9.74	7.33	6.46	3.74	3.74	2.53	2.23	1.61	1.48	0.62	0.56	0.12
1992	47.78	1.50	63.98	4.45	16.57	15.00	11.76	11.21	10.70	7.01	6.73	3.58	3.76	2.41	2.50	1.23	1.59	0.81	0.45	0.24
1993	48.47	1.54	64.00	4.72	16.26	15.01	12.47	9.97	10.38	7.69	6.30	4.31	3.41	2.71	2.04	1.63	1.54	0.85	0.44	0.26
1994	48.49	1.48	64.43	4.74	15.65	15.17	12.93	9.74	10.63	7.84	6.19	4.35	3.38	2.64	2.30	1.39	1.56	0.77	0.48	0.23
1995	47.98	1.95	64.40	4.56	15.29	15.76	12.38	9.88	10.13	8.26	5.63	5.41	2.77	2.96	2.03	1.78	1.21	1.04	0.52	0.38
57-95	50.90	1.39	61.90	5.41	16.66	16.02	12.80	10.40	8.68	7.11	5.75	4.47	3.59	2.79	2.18	1.59	1.15	0.71	0.43	0.25

AGE COMPOSITION OF FEMALE POPULATION BY YEAR

TABLE 27: FEMALE SAMPLE POPULATION BY AGE AND YEAR

	Age at Beginning of Five-Year Interval																	
Year	0-79	0	1	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75
1957	689	29	112	79	70	85	76	68	51	34	22	14	17	16	7	4	4	1
1958	738	49	109	89	69	87	78	69	62	35	26	14	17	15	11	2	5	1
1959	754	39	114	99	68	93	65	85	58	44	23	14	15	18	11	2	5	1
1960	777	42	118	100	74	75	91	66	67	44	31	15	17	14	13	4	5	1
1961	815	34	114	119	81	87	88	72	67	51	29	20	13	16	12	6	3	3
1962	893	59	127	119	90	88	104	76	68	52	33	21	14	16	15	7	3	1
1963	950	53	144	131	99	88	107	82	69	61	34	26	13	17	13	9	2	2
1964	1,010	63	163	132	111	84	119	74	83	59	43	22	15	15	14	9	2	2
1965	1,084	81	169	148	107	101	100	105	67	67	45	29	19	17	14	10	1	4
1966	1,132	59	203	142	122	109	110	98	72	70	50	27	24	13	18	10	2	3
1967	1,190	56	211	167	122	117	102	111	78	70	51	32	22	15	18	12	3	3
1968	1,254	76	208	170	134	118	112	113	84	69	62	34	26	13	19	9	5	2
1969	1,318	68	224	192	138	129	109	123	76	83	60	43	24	15	14	12	6	2
1970	1,396	81	215	222	153	117	130	108	105	67	70	45	31	18	15	9	9	1
1971	1,464	69	242	232	145	139	135	118	100	74	72	51	28	23	13	13	8	2
1972	1,562	96	259	228	170	144	142	108	115	81	71	53	32	20	17	12	11	3
1973	1,642	90	276	253	176	154	140	117	116	88	70	63	32	24	15	14	9	5
1974	1,690	68	292	266	190	157	150	113	123	80	84	59	40	22	17	12	11	6
1975	1,782	103	272	267	219	176	135	135	109	106	69	68	45	27	17	17	9	8
1976	1,895	113	308	280	231	170	164	135	119	99	76	70	50	25	21	14	13	7
1977	1,990	94	325	320	231	198	164	145	106	117	81	70	51	30	19	16	13	10
1978	2,103	117	330	330	257	195	187	146	116	116	89	68	62	30	23	14	15	8
1979	2,197	110	374	324	269	213	198	155	112	121	85	83	58	39	20	16	11	9
1980	2,286	112	370	346	266	245	218	138	136	105	109	70	68	42	26	15	14	6
1981	2,411	118	374	386	282	257	212	172	137	116	100	82	68	47	25	17	11	7
1982	2,498	96	389	388	316	262	241	175	143	104	117	84	70	47	30	17	11	8
1983	2,527	103	370	388	324	272	232	196	145	109	110	88	68	57	26	21	11	7
1984	2,643	118	379	430	315	281	262	201	156	105	115	79	81	54	34	17	10	6
1985	2,776	125	386	438	342	291	279	232	138	131	98	104	68	65	38	23	9	9
1986	2,906	139	397	444	379	309	293	214	172	134	107	94	78	64	44	22	10	6
1987	3,047	154	431	448	376	343	303	244	169	140	95	111	82	65	44	26	9	7
1988	3,136	127	469	452	387	349	324	237	186	143	106	107	87	66	54	23	12	7
1989	3,221	142	468	460	424	338	323	263	184	153	96	111	77	80	51	32	11	6
1990	3,375	178	485	460	436	356	341	275	214	130	122	93	101	64	63	37	18	2
1991	3,545	178	525	486	438	400	348	292	199	164	127	100	90	73	61	44	17	3
1992	3,682	166	538	529	447	404	381	292	225	164	136	86	107	75	64	41	22	5
1993	3,739	139	558	536	441	407	384	310	229	179	141	95	101	79	64	48	21	7
1994	3,840	132	548	549	455	444	389	311	260	178	149	94	106	72	73	44	29	7
1995	3,975	166	505	600	452	461	398	330	264	209	125	120	88	93	58	59	32	15

FEMALE SAMPLE PROPORTION BY AGE AND YEAR (PERCENT)

	Age at Beginning of Five-Year Interval																			
Year	0-14	65-79	10-79	0	1	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75
1957	42.09	1.31	68.07	4.21	16.26	11.47	10.16	12.34	11.03	9.87	7.40	4.93	3.19	2.03	2.47	2.32	1.02	0.58	0.58	0.15
1958	42.82	1.08	66.53	6.64	14.77	12.06	9.35	11.79	10.57	9.35	8.40	4.74	3.52	1.90	2.30	2.03	1.49	0.27	0.68	0.14
1959	42.44	1.06	66.58	5.17	15.12	13.13	9.02	12.33	8.62	11.27	7.69	5.84	3.05	1.86	1.99	2.39	1.46	0.27	0.66	0.13
1960	42.99	1.29	66.54	5.41	15.19	12.87	9.52	9.65	11.71	8.49	8.62	5.66	3.99	1.93	2.19	1.80	1.67	0.51	0.64	0.13
1961	42.70	1.47	67.24	4.17	13.99	14.60	9.94	10.67	10.80	8.83	8.22	6.26	3.56	2.45	1.60	1.96	1.47	0.74	0.37	0.37
1962	44.23	1.23	65.85	6.61	14.22	13.33	10.08	9.85	11.65	8.51	7.61	5.82	3.70	2.35	1.57	1.79	1.68	0.78	0.34	0.11
1963	44.95	1.37	65.47	5.58	15.16	13.79	10.42	9.26	11.26	8.63	7.26	6.42	3.58	2.74	1.37	1.79	1.37	0.95	0.21	0.21
1964	46.44	1.29	64.55	6.24	16.14	13.07	10.99	8.32	11.78	7.33	8.22	5.84	4.26	2.18	1.49	1.49	1.39	0.89	0.20	0.20
1965	46.59	1.38	63.28	7.47	15.59	13.65	9.87	9.32	9.23	9.69	6.18	6.18	4.15	2.68	1.75	1.57	1.29	0.92	0.09	0.37
1966	46.47	1.33	64.31	5.21	17.93	12.54	10.78	9.63	9.72	8.66	6.36	6.18	4.42	2.39	2.12	1.15	1.59	0.88	0.18	0.27
1967	46.72	1.51	63.53	4.71	17.73	14.03	10.25	9.83	8.57	9.33	6.55	5.88	4.29	2.69	1.85	1.26	1.51	1.01	0.25	0.25
1968	46.89	1.28	63.80	6.06	16.59	13.56	10.69	9.41	8.93	9.01	6.70	5.50	4.94	2.71	2.07	1.04	1.52	0.72	0.40	0.16
1969	47.19	1.52	63.28	5.16	17.00	14.57	10.47	9.79	8.27	9.33	5.77	6.30	4.55	3.26	1.82	1.14	1.06	0.91	0.46	0.15
1970	48.07	1.36	62.89	5.80	15.40	15.90	10.96	8.38	9.31	7.74	7.52	4.80	5.01	3.22	2.22	1.29	1.07	0.64	0.64	0.07
1971	46.99	1.57	62.91	4.71	16.53	15.85	9.90	9.49	9.22	8.06	6.83	5.05	4.92	3.48	1.91	1.57	0.89	0.89	0.55	0.14
1972	48.21	1.66	62.68	6.15	16.58	14.60	10.88	9.22	9.09	6.91	7.36	5.19	4.55	3.39	2.05	1.28	1.09	0.77	0.70	0.19
1973	48.42	1.71	62.30	5.48	16.81	15.41	10.72	9.38	8.53	7.13	7.06	5.36	4.26	3.84	1.95	1.46	0.91	0.85	0.55	0.30
1974	48.28	1.72	62.96	4.02	17.28	15.74	11.24	9.29	8.88	6.69	7.28	4.73	4.97	3.49	2.37	1.30	1.01	0.71	0.65	0.36
1975	48.32	1.91	63.97	5.78	15.26	14.98	12.29	9.88	7.58	7.58	6.12	5.95	3.87	3.82	2.53	1.52	0.95	0.95	0.51	0.45
1976	49.18	1.79	63.01	5.96	16.25	14.78	12.19	8.97	8.65	7.12	6.28	5.22	4.01	3.69	2.64	1.32	1.11	0.74	0.69	0.37
1977	48.74	1.96	62.86	4.72	16.33	16.08	11.61	9.95	8.24	7.29	5.33	5.88	4.07	3.52	2.56	1.51	0.95	0.80	0.65	0.50
1978	49.17	1.76	63.05	5.56	15.69	15.69	12.22	9.27	8.89	6.94	5.52	5.52	4.23	3.23	2.95	1.43	1.09	0.67	0.71	0.38
1979	49.02	1.64	63.22	5.01	17.02	14.75	12.24	9.70	9.01	7.06	5.10	5.51	3.87	3.78	2.64	1.78	0.91	0.73	0.50	0.41
1980	47.86	1.53	63.78	4.90	16.19	15.14	11.64	10.72	9.54	6.04	5.95	4.59	4.77	3.06	2.97	1.84	1.14	0.66	0.61	0.26
1981	48.11	1.45	63.58	4.89	15.51	16.01	11.70	10.66	8.79	7.13	5.68	4.81	4.15	3.40	2.82	1.95	1.04	0.71	0.46	0.29
1982	47.60	1.44	65.05	3.84	15.57	15.53	12.65	10.49	9.65	7.01	5.72	4.16	4.68	3.36	2.80	1.88	1.20	0.68	0.44	0.32
1983	46.89	1.54	65.93	4.08	14.64	15.35	12.82	10.76	9.18	7.76	5.74	4.31	4.35	3.48	2.69	2.26	1.03	0.83	0.44	0.28
1984	46.99	1.25	64.93	4.46	14.34	16.27	11.92	10.63	9.91	7.60	5.90	3.97	4.35	2.99	3.06	2.04	1.29	0.64	0.38	0.23
1985	46.51	1.48	65.81	4.50	13.90	15.78	12.32	10.48	10.05	8.36	4.97	4.72	3.53	3.75	2.45	2.34	1.37	0.83	0.32	0.32
1986	46.77	1.31	66.28	4.78	13.66	15.28	13.04	10.63	10.08	7.36	5.92	4.61	3.68	3.23	2.68	2.20	1.51	0.76	0.34	0.21
1987	46.24	1.38	66.10	5.05	14.15	14.70	12.34	11.26	9.94	8.01	5.55	4.59	3.12	3.64	2.69	2.13	1.44	0.85	0.30	0.23
1988	45.76	1.34	66.58	4.05	14.96	14.41	12.34	11.13	10.33	7.56	5.93	4.56	3.38	3.41	2.77	2.10	1.72	0.73	0.38	0.22
1989	46.38	1.52	66.78	4.41	14.53	14.28	13.16	10.49	10.03	8.17	5.71	4.75	3.04	3.45	2.39	2.48	1.58	0.99	0.34	0.19
1990	46.19	1.69	66.73	5.27	14.37	13.63	12.92	10.55	10.10	8.15	6.34	3.85	3.61	2.76	2.99	1.90	1.87	1.10	0.53	0.06
1991	45.90	1.81	66.46	5.02	14.81	13.71	12.36	11.28	9.82	8.24	5.61	4.63	3.58	2.82	2.54	2.06	1.72	1.24	0.48	0.08
1992	45.63	1.85	66.51	4.51	14.61	14.37	12.14	10.97	10.35	7.93	6.11	4.45	3.69	2.34	2.91	2.04	1.74	1.11	0.60	0.14
1993	44.77	2.03	67.02	3.72	14.92	14.34	11.79	10.89	10.27	8.29	6.12	4.79	3.77	2.54	2.70	2.11	1.71	1.28	0.56	0.19
1994	43.85	2.08	67.99	3.44	14.27	14.30	11.85	11.56	10.13	8.10	6.77	4.64	3.88	2.45	2.76	1.88	1.90	1.15	0.76	0.18
1995	43.35	2.67	68.03	4.18	12.70	15.09	11.37	11.60	10.01	8.30	6.64	5.26	3.14	3.02	2.21	2.34	1.46	1.48	0.81	0.38
57-95	46.41	1.64	65.36	4.81	15.14	14.70	11.77	10.44	9.68	7.89	6.23	4.94	3.92	3.08	2.51	1.88	1.39	0.90	0.50	0.24

SEX RATIO BY AGE BY YEAR

SAMPLE SEX RATIO BY YEAR AND AGE (MALES/FEMALES)

	Age at Beginning of Five-Year Interval																	
Year	0-79	0	1	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75
1957	0.85	1.55	0.92	1.10	0.83	0.73	0.43	1.01	0.67	0.88	0.86	1.00	0.53	0.38	1.00	0.75	1.25	0.00
1958	0.85	1.31	1.03	0.97	1.10	0.56	0.53	0.97	0.61	0.83	0.92	0.71	0.71	0.53	0.45	2.00	1.00	0.00
1959	0.86	1.03	1.21	0.94	1.13	0.62	0.60	0.78	0.66	0.64	1.22	0.57	0.87	0.50	0.36	2.00	0.80	1.00
1960	0.89	1.31	1.21	0.91	1.18	0.73	0.56	0.64	0.94	0.57	1.13	0.67	0.82	0.57	0.38	1.75	0.60	1.00
1961	0.88	1.09	1.31	0.90	1.05	0.68	0.64	0.63	1.03	0.49	1.21	0.60	1.00	0.56	0.33	1.33	0.33	1.00
1962	0.87	0.95	1.18	0.97	0.99	0.68	0.70	0.57	1.03	0.62	0.94	0.81	1.07	0.56	0.40	1.00	0.67	3.00
1963	0.88	1.02	1.08	1.08	0.88	0.92	0.56	0.67	1.04	0.57	0.85	0.92	0.85	0.71	0.54	0.56	1.50	2.00
1964	0.87	0.78	1.03	1.17	0.84	1.01	0.60	0.69	0.90	0.61	0.60	1.36	0.67	0.87	0.50	0.44	2.00	1.50
1965	0.87	0.89	0.95	1.14	0.85	0.94	0.67	0.58	0.76	0.94	0.53	1.31	0.53	0.88	0.64	0.30	7.00	0.75
1966	0.88	1.05	0.96	1.13	0.90	0.84	0.69	0.66	0.75	0.99	0.46	1.41	0.50	1.15	0.50	0.20	4.00	0.33
1967	0.87	1.07	0.92	1.07	0.96	0.84	0.71	0.72	0.60	1.06	0.61	1.00	0.86	1.07	0.44	0.33	2.33	0.67
1968	0.90	1.21	0.93	1.04	1.07	0.79	0.83	0.59	0.73	1.10	0.58	0.88	0.96	0.77	0.74	0.56	1.00	1.50
1969	0.89	1.09	1.02	0.96	1.09	0.76	0.88	0.63	0.78	0.93	0.62	0.67	1.21	0.60	1.07	0.58	0.67	2.00
1970	0.91	1.23	1.07	0.92	1.11	0.79	0.82	0.67	0.64	0.85	0.90	0.53	1.29	0.50	1.07	1.00	0.33	6.00
1971	0.90	0.99	1.11	0.96	1.11	0.80	0.76	0.71	0.68	0.82	0.96	0.47	1.43	0.43	1.15	0.77	0.25	3.50
1972	0.92	1.15	1.10	0.96	1.05	0.84	0.77	0.77	0.74	0.64	1.07	0.64	1.00	1.00	0.94	0.75	0.27	2.00
1973	0.92	0.99	1.11	1.03	0.97	0.96	0.74	0.87	0.59	0.75	1.07	0.60	0.91	1.00	0.73	0.93	0.44	0.80
1974	0.94	1.29	1.10	1.06	0.95	0.99	0.75	0.92	0.67	0.75	0.96	0.63	0.75	1.23	0.53	1.17	0.45	0.50
1975	0.94	0.97	1.14	1.14	0.93	0.97	0.76	0.90	0.69	0.71	0.83	0.94	0.56	1.33	0.59	0.76	0.67	0.38
1976	0.93	0.87	1.09	1.14	0.97	0.96	0.73	0.87	0.72	0.76	0.79	1.00	0.50	1.52	0.52	0.86	0.54	0.14
1977	0.93	0.93	1.02	1.15	0.97	0.92	0.77	0.86	0.80	0.77	0.63	1.06	0.65	1.03	0.95	0.81	0.46	0.10
1978	0.91	0.64	1.00	1.12	1.02	0.90	0.81	0.83	0.91	0.65	0.73	1.09	0.60	0.97	1.00	0.57	0.60	0.25
1979	0.90	0.83	0.82	1.19	1.04	0.88	0.82	0.84	0.94	0.74	0.68	0.94	0.60	0.74	1.30	0.38	0.91	0.33
1980	0.91	1.04	0.81	1.11	1.13	0.86	0.83	0.82	0.91	0.74	0.72	0.76	0.91	0.62	1.19	0.53	0.79	0.83
1981	0.91	1.03	0.84	1.04	1.12	0.90	0.83	0.76	0.93	0.75	0.78	0.72	0.97	0.53	1.32	0.53	0.91	0.86
1982	0.92	1.18	0.92	0.98	1.14	0.87	0.84	0.75	0.95	0.82	0.81	0.60	1.04	0.70	0.83	0.88	1.00	0.75
1983	0.93	1.12	1.01	0.89	1.11	0.96	0.85	0.82	0.88	0.94	0.71	0.70	1.05	0.58	0.92	0.81	0.64	1.29
1984	0.93	1.02	1.04	0.82	1.19	0.97	0.82	0.83	0.87	0.98	0.78	0.72	0.90	0.57	0.79	1.12	0.50	1.83
1985	0.92	0.94	1.03	0.85	1.08	1.02	0.79	0.90	0.82	0.96	0.82	0.74	0.78	0.82	0.63	1.17	0.67	1.11
1986	0.91	0.99	0.98	0.87	1.03	1.00	0.82	0.95	0.74	0.94	0.82	0.82	0.73	0.89	0.52	1.27	0.70	1.33
1987	0.91	0.93	0.96	0.91	0.99	1.03	0.78	0.94	0.75	0.96	0.89	0.81	0.56	0.98	0.70	0.85	1.44	1.14
1988	0.90	0.99	0.92	0.97	0.89	1.02	0.82	0.92	0.85	0.89	0.93	0.70	0.70	0.97	0.61	0.83	1.17	0.86
1989	0.91	1.02	0.93	0.99	0.80	1.11	0.86	0.88	0.89	0.88	1.02	0.77	0.70	0.81	0.59	0.66	1.45	0.67
1990	0.91	1.01	0.95	0.99	0.82	1.04	0.89	0.80	0.98	0.86	0.98	0.82	0.73	0.73	0.76	0.54	0.94	2.00
1991	0.91	1.08	0.98	0.96	0.86	0.97	0.91	0.81	1.05	0.74	0.95	0.82	0.80	0.71	0.79	0.45	1.06	1.33
1992	0.90	0.89	1.02	0.94	0.87	0.92	0.93	0.80	1.00	0.73	0.92	0.93	0.78	0.55	0.83	0.66	0.68	1.60
1993	0.92	1.17	1.00	0.96	0.97	0.84	0.93	0.85	0.94	0.83	0.83	0.98	0.69	0.71	0.83	0.60	0.71	1.29
1994	0.92	1.27	1.01	0.97	1.00	0.77	0.96	0.89	0.84	0.86	0.80	0.99	0.76	0.68	0.75	0.61	0.59	1.14
1995	0.92	1.00	1.10	0.96	1.00	0.78	0.93	0.91	0.78	0.94	0.81	0.90	0.84	0.70	0.76	0.64	0.59	0.93
57-95	0.90	1.03	1.01	1.00	0.99	0.91	0.82	0.83	0.84	0.82	0.83	0.81	0.77	0.75	0.75	0.72	0.83	1.07

METHOD OF ANALYSIS

DISCRETE TIME EVENT HISTORY ANALYSIS

All of the demographic indicators presented below were calculated using discrete-time event history analysis (Allison 1982; Allison 1984; Petersen 1995). Discrete-time event history analysis (DTEHA) is a relatively straightforward method to analyze event-based longitudinal processes.

The objective is to estimate the probability that an event will occur to a unit of analysis over a defined period of time - the *hazard* associated with that event. The model assumes that time can be broken down into sufficiently brief, discrete quanta or granules that the hazard of an event occurring during one granule is approximately constant. With this assumption, the *life* of a unit of analysis is divided into granules, and the state occupied by the unit of analysis at the start of each granule is defined. For each unit of analysis and for each granule that they *live* during the period of the study, a record is created in the analysis data set to record the state of the unit of analysis at the start of the granule and whether or not the event(s) of interest occurred during the granule. In practice, the units of analysis are usually people and the granules are often years or months, hence the synthetic unit of analysis is the person-month or person-year.

The resulting data set is a collection of item-granules representing the discretized history of the original units of analysis, or *items*. Ordinary logistic regression is used to estimate the probability of occurrence of the event(s) of interest over the period of one granule as a function of the states occupied by the units of analysis at the start of the granules. In practice,

the states are described by a collection of variables that can take any type of value, including continuous. Both constant and time-varying covariates may be addressed with this approach; a constant attribute takes the same value for all of the item-granules associated with a single unit of analysis, whereas a time-varying attribute takes on different values for each item-granule associated with a single unit of analysis. One of the most important time-varying covariates is the period of history during which a granules falls. For example, a person living through the period 1970 to 1975 contributes five person-years, and the "year" attribute of each records the calendar year over which the granule (year) was lived.

Two particular benefits result from the use of the DTEHA approach: 1) it is possible to obtain a good estimate of the base hazard, and 2) all of the benefits of logistic regression are available, including the ability to easily include dummy variables and form all sorts of interactions with them. Moreover because logistic regression is common, the method is easily understood and interpreted by most people with a basic education in statistics, which greatly facilitates the communication of results. Counteracting those is the one large drawback that usually prevents analysts from using this technique. The preparation of the analysis data set can be difficult, is tedious and is unusually prone to error.

The most common alternative is to analyze longitudinal data using the proportional hazards model proposed by Cox. While using different approaches, the two methods produce the same understanding of the underlying event histories, and I have chosen to use the DTEHA approach because it is possible for me to manipulate the data without too much trouble, and I prefer the flexibility of the logistic regression framework and the fact that it produces estimates

of the hazards for all defined states of the unit of analysis. The final product of DTEHA is a set of estimates of the hazard of occurrence of the event(s) of interests for each state that has been defined for the unit of analysis in the discrete-time event history data set. For example, if I were analyzing mortality, I would produce a set of estimates of the annual hazard of death at each age for males and females during each time-period over which the study population lived.

ANALYSIS FILES

Three discrete-time event history files were created to analyze the demography of the Gwembe Tonga. The first positions the individual as the unit of analysis and uses one year as the size of the granule. The result is a person-year file describing the discretized life histories of the sample of the Gwembe Study population used for this analysis. The states and events that are recorded pertain to individual people.

The second is organized around conjugal unions and years to create a union-year file describing the discretized history of couples living in conjugal relationships. In this case the states and events described pertain to couples in conjugal unions. For example, the birth of a child, the dissolution of the union, or the marrying of a cowife.

The third is a completely synthetic file that was created to estimate the annual hazard of forming a union for two people with specific attributes. This file is discussed in detail below.

The demographic indicators are calculated over the period 1957 to 1995. 1957 is the first complete year over which data are collected prospectively, and 1995 is the last year for which data were input to machine readable form.

PERSON-YEAR FILE

The three base tables of the clean, well-defined sample of the Gwembe Study population form the basis of the person-year file. A number of SQL (Appendix B) statements were used to associate, group, count and define various attributes and events pertaining to individuals in the sample. For each year lived by each individual between 1957 and 1995, a record was added to the person-year file and all of the variables defined to reflect the state of the reference person at the beginning of the year and whether or not any of the events of interest occurred during the year. The resulting person-year file has this structure:

TABLE 28: DESCRIPTION OF THE PERSON-YEAR TABLE

Name: tblIndividualYears
Records: 85,870 records representing 155,229 individual observed person years
Fields:

Field Name	Type	Description
ID	Number	Unique ID for this record
Frequency	Number	Number of records with these attributes*
Year	Number	Historical year of this granule
Village	Number	Village with which the person is associated
Sex	Number	The person's sex
Age	Number	The person's age
Dying	Number	True = the person dies this year
Single	Number	True = the person is single this year
Marrying	Number	True = the person marries this year
Married	Number	True = the person is married this year
Separating	Number	True = the person is separating this year
Marriages	Number	Number of spouses this person has
Unions	Number	Number of unions the person is initiating this year
Separations	Number	Number of unions the person is terminating this year
Fertile	Number	True = the person has had at least one birth
MaleChildBirth	Number	True = a male child born this year
MaleChildDeath	Number	True = a male child death this year
FemaleChildBirth	Number	True = a female child born this year
FemaleChildDeath	Number	True = a female child death this year
MaleBirthsThisYear	Number	Number of male children born this year
MaleDeathsThisYear	Number	Number of male children dying this year
FemaleBirthsThisYear	Number	Number of female births this year
FemaleDeathsThisYear	Number	Number of female deaths this year
MaleBirths	Number	Total male births to this person
SurvivingMaleBirths	Number	Total surviving male births to this person
FemaleBirths	Number	Total female births to this person
SurvivingFemaleBirths	Number	Total surviving female births to this person
LastBirthSex	Number	Sex of this person's last birth
LastBirthAlive	Number	True = this person's last birth is alive
LastBirthYears	Number	Years since this person's last birth
MotherInSample	Number	True = this person's mother is in the analysis sample
FatherInSample	Number	True = this person's father is in the analysis sample
MotherDead	Number	If dead, the number of years since the death of this person's mother
FatherDead	Number	If dead, the number of years since the death of this person's father

UNION-YEAR FILE

Like the person-year file, the union-year file was constructed from the three base tables of the clean, well-defined sample of the Gwembe Study population using a series of SQL and Visual Basic statements. For each year survived by each union between 1957 and 1995, a record was

* The raw item-granule file was aggregated at the record level to make a smaller analysis file. This was accomplished by identifying all groups of identical records, those that share the same values in all fields, and counting the number of records in

added to the union-year file and all of the state and event variables defined. The resulting union-year file has this structure:

TABLE 29: DESCRIPTION OF THE UNION-YEAR TABLE

Name: TblMarriedPairYears
Records: 20,287 records representing 20,344 observed married pair-years
Fields:

Name	Type	Description
ID	Number	Unique ID for this record
Frequency	Number	Number of records with these attributes*
Year	Number	Historical year of this granule
Wedding	Number	True = this pair is marrying this year[20]
Separation	Number	True = this pair is separating this year
Cowives	Number	Number of cowives in this union
MarryingCowife	Number	True = marrying a cowife this year
SeparatingCoWife	Number	True = separating from a cowife this year
MaleVillage	Number	Village with which the male is associated
MaleAge	Number	Age of the male
MaleDying	Number	True = male dies this year
FemaleVillage	Number	Village with which the female is associated
FemaleAge	Number	Age of the female
FemaleDying	Number	True = female dies this year
PairAge	Number	Age of the union
PairFertile	Number	True = this pair has had at least on child
PairMaleChildBirth	Number	True = male child birth this year
PairMaleChildDeath	Number	True = male child death this year
PairFemaleChildBirth	Number	True = female child birth this year
PairFemaleChildDeath	Number	True = female child death this year
PairMaleBirths	Number	Total number of mail births to this pair
PairSurvivingMaleBirths	Number	Total number of surviving male births to this pair
PairFemaleBirths	Number	Total number of female births to this pair
PairSurvivingFemaleBirths	Number	Total number of surviving female births to this pair
PairLastBirthSex	Number	Sex of the last birth to this pair
PairLastBirthAlive	Number	True = last birth to this pair is alive
PairLastBirthYears	Number	Years since the last birth to this pair

UNMARRIED UNION-YEAR FILE

The unmarried union-year file differs from the other two in that it is completely synthetic and does not represent the history of real people or unions. The aim is to estimate the hazard of

each group. That number is the *frequency* recorded in the analysis data set. The values of the other fields in the group for which the frequency is recorded are recorded in the remaining fields in the analysis data set.

[20] Some pairs marry and divorce in the same year, and because they are exposed to the risk of events specific to married pairs during that time, they are included in this data set and differentiated by a value in the *Wedding* field.

forming a union for two people with various characteristics. The challenge is to do this in a way that takes account of the proportional growth of the population - the scaling problem. As the population grows, the number of potential couples increases at a rate twice that of the growth rate of the population[21]. At the same time, the number of unions grows at a rate equal to that of the population. This results from the fact that the growth in the number of unions cannot exceed the growth in the population of the sex with the slowest growth rate. One way of estimating the hazard of union for a closed population of available men and women over time would be to use the population of potential unions as the *denominator*, and the number of unions formed during each granule of time as the *numerator*[22]. However because of the scaling problem just described, this is not appropriate. The solution is to multiply the number of potential couples by a correction factor that nullifies the scaling problem[23], or preferably, to construct a more appropriate population of potential couples. I chose the latter solution for two reasons: 1) it is less prone to error, and 2) it provides the basis for a clean, elegant algorithm for pairing people in the simulation described in Part 3.

For each granule of time, the following procedure was performed using Visual Basic and SQL to create the unmarried union-year records for the granule. Using the three clean, well-defined base tables, each available man is paired with each available women to form a total population

---

[21] If the population is growing at an annual proportional rate of r, and if the populations of available men and women are growing at the same rate r, then the population of potential couples is:

$$P_0^m e^{rt} \cdot P_0^f e^{rt} = P_0^m \cdot P_0^f \cdot e^{2rt},$$

and the annual proportional growth rate of the number of potential couples is $2r$.

[22] Denominator and numerator are used here to signify the pool of potential events, and the number of events that actually occur.

[23] This becomes unattractive when the growth of the male and female populations, or substantively relevant subpopulations of males and females, is not equal or changing rapidly.

of potential unions. That population of potential unions is divided into a number of subpopulations based on characteristics of the men, women and possibly of the potential pair. These include the ages of the men and women, the number of wives the men already have, the reproductive histories of the men and women and maybe whether the pair has ever been married. Each subpopulation is completely defined by those attributes, and the potential pairs within the subpopulation are assumed to be indistinguishable. Within each of these homogeneous subpopulations, or cells, the maximum number of possible unions is determining by comparing the number of men to the number of women in the cell. The maximum number of possible unions is equal to the minimum of those two numbers; for example, if there were 32 men and 39 women in a cell, the maximum number of unions would be 32. That completes the determination of the denominator. The numerator is calculated by determining the number of unions that actually formed in the cell. Once those are known, a number of records equal to the maximum number of possible unions is added to the analysis data set, the attributes of the cell are recorded in each of those records, and in a number equal to the number of unions that formed, it is recorded that a union was formed. This procedure eliminates the scaling problem by dynamically scaling the denominator to the real number of possible unions in each cell, not the total number of potential pairs. To see this point clearly, imagine a cell with two men and 100 women. There are 200 potential pairs in that cell, but only two possible unions[24].

[24] This assumes that a man can marry only one additional wife per granule. With the size of a granule set to one year, this is a reasonable assumption although it is violated in a trivial number of cases.

For the analysis of union formation, I chose to make a simple data set taking into account only the ages of the men and women and the number of wives currently married to each man. Much more is possible but not necessary for my purposes - namely a general understanding of the levels and trends in union formation and the generation of parameters for the simulation described in Part 3. Because the Tonga form polygynous families, all men older than ten and younger than 80 years are eligible for marriage in each granule, while only unmarried women older than ten and younger than 80 years are eligible. The resulting unmarried union-year file has this structure:

TABLE 30: DESCRIPTION OF THE UNMARRIED UNION-YEAR TABLE

Name: TblUnmarriedPairs
Records: **15,599** representing **50,235** individual observed cases
Fields:

Field Name	Type	Description
ID	Number	Unique ID for this record
Frequency	Number	Number of records with these attributes'
intYear	Number	Historical year of this granule
bytMaleWives	Number	Number of wives married to the man
intMaleAgeGroup	Number	Age group into which the man's age falls
intFemaleAgeGroup	Number	Age group into which the woman's age falls
blnWedding	Boolean	True = a union was formed during the granule

DEMOGRAPHIC CHARACTERISTICS

NUPTIALITY

The Tonga spend considerable time, money and emotion negotiating, initiating and dissolving marriages. Colson discusses the traditions and social structure of the marriage system in considerable detail in chapter four of The Social Organization of the Gwembe Tonga (Colson 1960b) and chapter five of The Social Consequences of Resettlement (Colson 1971b). In contrast I am concerned with a quantitative description of the dynamics of the marriage system with the goal of producing a set of parameters suitable for the construction of a general dynamic model of polygynous marriage.

This view of marriage is necessarily highly abstracted and cannot support the degree of detail and nuance that is possible with a qualitative description. That is sacrificed in order to gain a wide understanding of the dynamics of the system as a whole at the level of the whole population. Because this analysis is not concerned with the fine details that determine exactly when a marriage is called a "marriage" as opposed to an "elopement" or something else, I take a broad definition of a conjugal union and term it simply a *union.* From now on I will use the term union to refer to a conjugal relationship defined by the stable cohabitation of a man and woman.

To understand the dynamics of the marriage system, it is sufficient to estimate the annual male-female-age-specific hazard of union formation and male-female-age-duration-specific hazard of union dissolution. Together those hazards regulate the number, union-age and male-female-age composition of unions in the population.

UNION FORMATION

LEVELS

At the most fundamental level we need to know what proportions of the populations of men and women are married at any given point in time. Table 31 and Table 32 display that information. The overall proportion of men married at each age declined gradually between 1957 and 1995 with average levels of about 30 percent for men ages 20-24, about 60 percent for men aged 25-29, about 70 percent for men aged 30 to 59, about 75 percent for men aged 60 to 69 and about 60 percent for men over age 70. During the same time the average number of wives per married man dropped gradually from about 1.5 to about 1.3. At all ages 10-79 roughly 40 percent of men are married, and that has declined steadily from a level of about 45 percent to a level of about 35 percent between 1957 and 1995.

Table 32 displays similar information for females. During that same period between 1957 and 1995, about half of all women between ages 10 and 79 were married, and like the males, that proportion steadily declined from something close to 55 percent to about 48 percent. Over the period 1957 to 1995, roughly 25 percent of women 15-19 were married, about 70 percent between ages 20 and 24, 80 percent between ages 25 and 39, and a steadily decreasing number with age from about 75 percent between ages 40 and 44 to about 10 percent between ages 75 to 79. The average figures mask a more complicated trend. During the first decade or so of observation the proportion of women married at ages 40 and older increased steadily and the proportion married at young ages decreased substantially.

TABLE 31: SAMPLE PROPORTION MALES MARRIED BY YEAR AND AGE

Year	10-79	Age at Beginning of Five-Year Interval														Ave. Wives[25]
		10	15	20	25	30	35	40	45	50	55	60	65	70	75	
1957	0.44	0.00	0.00	0.33	0.62	0.79	0.80	0.84	0.93	0.67	0.67	0.57	0.67	0.40	0.00	1.55
1958	0.46	0.00	0.04	0.37	0.67	0.82	0.83	0.83	1.00	0.83	0.50	0.60	0.50	0.40	0.00	1.52
1959	0.45	0.00	0.03	0.38	0.68	0.82	0.79	0.89	0.88	0.85	0.56	0.50	1.00	0.50	0.00	1.54
1960	0.42	0.00	0.02	0.22	0.62	0.78	0.84	0.83	0.60	0.93	0.63	0.40	0.86	0.67	0.00	1.54
1961	0.44	0.00	0.07	0.18	0.67	0.77	0.84	0.86	0.75	0.92	0.67	0.50	0.88	1.00	0.33	1.52
1962	0.43	0.00	0.03	0.22	0.58	0.83	0.75	0.87	0.76	0.93	0.44	0.67	0.71	1.00	0.33	1.53
1963	0.43	0.00	0.00	0.33	0.60	0.81	0.74	0.86	0.75	0.82	0.58	0.57	0.80	0.67	0.50	1.50
1964	0.44	0.00	0.04	0.42	0.65	0.73	0.75	0.85	0.77	0.80	0.69	0.71	1.00	0.75	0.67	1.46
1965	0.44	0.00	0.00	0.42	0.64	0.69	0.78	0.88	0.74	0.70	0.80	0.56	0.67	0.86	0.67	1.47
1966	0.43	0.00	0.03	0.38	0.66	0.74	0.74	0.83	0.76	0.75	0.80	0.56	1.00	0.75	0.00	1.47
1967	0.43	0.00	0.01	0.35	0.64	0.77	0.78	0.77	0.78	0.68	0.88	0.38	1.00	0.71	0.50	1.45
1968	0.41	0.00	0.03	0.26	0.70	0.70	0.78	0.72	0.77	0.68	0.90	0.57	0.80	0.80	0.67	1.46
1969	0.41	0.00	0.04	0.29	0.71	0.73	0.73	0.73	0.72	0.79	0.78	0.73	0.71	0.75	0.75	1.43
1970	0.41	0.00	0.03	0.24	0.74	0.78	0.68	0.73	0.83	0.75	0.78	0.81	0.56	0.67	0.83	1.46
1971	0.41	0.00	0.02	0.28	0.64	0.78	0.70	0.68	0.83	0.78	0.80	0.80	0.70	1.00	0.71	1.44
1972	0.40	0.00	0.02	0.30	0.60	0.69	0.77	0.70	0.74	0.81	0.70	0.88	0.56	1.00	0.67	1.44
1973	0.40	0.01	0.01	0.31	0.58	0.78	0.73	0.69	0.74	0.79	0.75	0.73	0.77	0.75	0.75	1.46
1974	0.39	0.01	0.02	0.32	0.54	0.70	0.73	0.68	0.76	0.73	0.85	0.78	0.79	0.60	0.67	1.45
1975	0.37	0.00	0.03	0.23	0.55	0.71	0.72	0.70	0.69	0.84	0.78	0.70	0.77	0.67	0.33	1.43
1976	0.38	0.01	0.02	0.23	0.62	0.67	0.72	0.70	0.69	0.84	0.79	0.73	0.75	0.71	1.00	1.42
1977	0.38	0.01	0.02	0.23	0.65	0.66	0.71	0.80	0.68	0.70	0.81	0.72	0.77	0.67	1.00	1.39
1978	0.36	0.00	0.01	0.19	0.60	0.62	0.73	0.80	0.69	0.65	0.79	0.74	0.75	0.89	0.50	1.40
1979	0.35	0.00	0.02	0.20	0.59	0.65	0.71	0.76	0.68	0.71	0.66	0.81	0.83	0.80	0.67	1.40
1980	0.34	0.00	0.01	0.21	0.58	0.60	0.71	0.72	0.68	0.68	0.77	0.84	0.63	0.82	0.60	1.38
1981	0.35	0.00	0.02	0.31	0.55	0.67	0.66	0.72	0.71	0.68	0.72	0.85	0.67	0.80	0.67	1.37
1982	0.34	0.00	0.02	0.31	0.53	0.68	0.65	0.68	0.78	0.66	0.67	0.84	0.67	0.82	0.67	1.36
1983	0.34	0.00	0.02	0.29	0.58	0.70	0.61	0.71	0.74	0.64	0.64	0.75	0.71	0.86	0.78	1.36
1984	0.34	0.00	0.01	0.29	0.60	0.67	0.65	0.66	0.72	0.62	0.74	0.67	0.74	0.80	0.73	1.34
1985	0.34	0.00	0.02	0.20	0.63	0.67	0.62	0.69	0.70	0.62	0.64	0.75	0.70	0.50	0.90	1.33
1986	0.34	0.01	0.01	0.23	0.68	0.64	0.67	0.60	0.69	0.63	0.61	0.74	0.75	0.57	0.88	1.32
1987	0.35	0.01	0.01	0.28	0.66	0.67	0.66	0.56	0.66	0.72	0.63	0.61	0.73	0.69	1.00	1.31
1988	0.36	0.01	0.02	0.27	0.68	0.68	0.68	0.59	0.67	0.70	0.61	0.64	0.68	0.79	1.00	1.27
1989	0.36	0.01	0.02	0.27	0.64	0.73	0.64	0.62	0.62	0.72	0.60	0.73	0.62	0.88	0.75	1.26
1990	0.36	0.00	0.01	0.25	0.61	0.77	0.65	0.62	0.59	0.66	0.66	0.60	0.80	0.82	0.25	1.26
1991	0.36	0.00	0.02	0.25	0.63	0.78	0.64	0.67	0.54	0.65	0.67	0.60	0.75	0.83	0.25	1.27
1992	0.36	0.00	0.02	0.26	0.59	0.75	0.72	0.62	0.54	0.63	0.73	0.57	0.74	0.80	0.63	1.27
1993	0.35	0.00	0.02	0.24	0.55	0.75	0.68	0.65	0.53	0.63	0.70	0.57	0.66	0.80	0.67	1.27
1994	0.35	0.00	0.01	0.22	0.55	0.74	0.76	0.60	0.55	0.59	0.67	0.55	0.74	0.53	0.88	1.27
1995	0.35	0.00	0.02	0.23	0.53	0.69	0.79	0.65	0.55	0.54	0.62	0.64	0.55	0.74	0.79	1.26
57-95[26]	0.37	0.00	0.02	0.26	0.61	0.72	0.71	0.69	0.67	0.68	0.69	0.67	0.72	0.75	0.70	1.36

[25] Average wives per married man.

[26] Population-weighted average.

TABLE 32: SAMPLE PROPORTION FEMALES MARRIED BY YEAR AND AGE

		Age at Beginning of Five-Year Interval													
Year	10-79	10	15	20	25	30	35	40	45	50	55	60	65	70	75
1957	0.53	0.00	0.42	0.80	0.88	0.80	0.85	0.55	0.29	0.12	0.06	0.14	0.00	0.50	0.00
1958	0.55	0.03	0.51	0.82	0.87	0.77	0.83	0.50	0.29	0.12	0.07	0.09	0.00	0.20	0.00
1959	0.56	0.03	0.55	0.77	0.88	0.81	0.77	0.57	0.21	0.13	0.11	0.09	0.00	0.20	0.00
1960	0.54	0.03	0.36	0.82	0.85	0.81	0.80	0.58	0.47	0.12	0.07	0.15	0.00	0.20	0.00
1961	0.54	0.02	0.36	0.77	0.85	0.84	0.76	0.76	0.45	0.15	0.13	0.17	0.00	0.00	0.33
1962	0.53	0.01	0.34	0.76	0.84	0.85	0.75	0.76	0.52	0.21	0.13	0.07	0.00	0.00	0.00
1963	0.53	0.00	0.28	0.80	0.80	0.83	0.79	0.71	0.58	0.23	0.12	0.08	0.00	0.00	0.00
1964	0.54	0.01	0.38	0.77	0.80	0.84	0.80	0.65	0.64	0.27	0.13	0.14	0.00	0.00	0.00
1965	0.56	0.03	0.38	0.73	0.86	0.82	0.82	0.78	0.59	0.58	0.12	0.07	0.10	0.00	0.00
1966	0.55	0.03	0.40	0.81	0.82	0.81	0.80	0.76	0.59	0.54	0.15	0.11	0.10	0.00	0.00
1967	0.56	0.02	0.42	0.80	0.80	0.81	0.80	0.76	0.66	0.55	0.20	0.17	0.08	0.00	0.00
1968	0.55	0.02	0.40	0.78	0.81	0.79	0.83	0.77	0.65	0.50	0.23	0.16	0.11	0.00	0.00
1969	0.56	0.01	0.40	0.83	0.82	0.75	0.84	0.77	0.67	0.63	0.27	0.07	0.17	0.00	0.00
1970	0.57	0.01	0.36	0.87	0.80	0.80	0.82	0.77	0.78	0.61	0.44	0.07	0.11	0.11	0.00
1971	0.56	0.01	0.32	0.79	0.82	0.83	0.77	0.76	0.75	0.57	0.52	0.08	0.15	0.13	0.00
1972	0.55	0.01	0.36	0.80	0.82	0.81	0.80	0.76	0.75	0.63	0.35	0.18	0.08	0.09	0.00
1973	0.55	0.01	0.36	0.83	0.85	0.76	0.81	0.79	0.71	0.63	0.38	0.20	0.07	0.11	0.00
1974	0.54	0.01	0.38	0.77	0.84	0.76	0.79	0.80	0.76	0.58	0.45	0.18	0.00	0.18	0.00
1975	0.53	0.00	0.32	0.78	0.87	0.75	0.79	0.75	0.76	0.69	0.44	0.24	0.00	0.11	0.13
1976	0.52	0.00	0.26	0.76	0.85	0.81	0.79	0.72	0.73	0.68	0.52	0.29	0.07	0.08	0.14
1977	0.52	0.00	0.22	0.81	0.86	0.80	0.76	0.73	0.73	0.67	0.60	0.26	0.19	0.08	0.10
1978	0.50	0.00	0.15	0.74	0.82	0.84	0.78	0.76	0.75	0.66	0.53	0.35	0.29	0.07	0.13
1979	0.51	0.01	0.17	0.76	0.83	0.82	0.79	0.79	0.72	0.67	0.44	0.40	0.31	0.00	0.22
1980	0.49	0.01	0.16	0.66	0.86	0.83	0.78	0.79	0.71	0.68	0.57	0.38	0.20	0.07	0.17
1981	0.51	0.01	0.23	0.69	0.87	0.82	0.79	0.77	0.74	0.65	0.60	0.32	0.24	0.18	0.14
1982	0.51	0.01	0.23	0.71	0.83	0.84	0.79	0.76	0.74	0.63	0.57	0.43	0.12	0.27	0.13
1983	0.51	0.01	0.22	0.70	0.85	0.81	0.83	0.76	0.75	0.69	0.54	0.42	0.19	0.27	0.14
1984	0.51	0.01	0.17	0.71	0.82	0.85	0.82	0.74	0.77	0.67	0.56	0.41	0.29	0.20	0.00
1985	0.50	0.00	0.17	0.67	0.78	0.85	0.85	0.76	0.77	0.60	0.58	0.55	0.30	0.22	0.00
1986	0.50	0.01	0.19	0.69	0.79	0.83	0.84	0.78	0.76	0.63	0.58	0.55	0.36	0.10	0.17
1987	0.51	0.01	0.22	0.74	0.77	0.83	0.84	0.78	0.72	0.65	0.60	0.50	0.38	0.00	0.29
1988	0.50	0.01	0.20	0.73	0.79	0.80	0.82	0.76	0.73	0.62	0.59	0.48	0.30	0.08	0.29
1989	0.50	0.00	0.21	0.73	0.77	0.78	0.84	0.78	0.70	0.64	0.58	0.47	0.31	0.09	0.17
1990	0.50	0.01	0.22	0.70	0.79	0.77	0.82	0.78	0.70	0.64	0.55	0.48	0.41	0.11	0.00
1991	0.50	0.01	0.23	0.71	0.79	0.78	0.80	0.79	0.70	0.67	0.52	0.46	0.41	0.18	0.00
1992	0.49	0.00	0.21	0.67	0.78	0.78	0.80	0.79	0.71	0.64	0.48	0.45	0.32	0.32	0.00
1993	0.49	0.00	0.18	0.66	0.78	0.76	0.76	0.76	0.75	0.67	0.49	0.47	0.33	0.29	0.14
1994	0.48	0.00	0.19	0.67	0.75	0.75	0.75	0.78	0.76	0.65	0.54	0.45	0.36	0.21	0.14
1995	0.48	0.00	0.19	0.62	0.75	0.77	0.74	0.74	0.73	0.68	0.57	0.45	0.41	0.28	0.13
57-95	0.51	0.01	0.24	0.72	0.80	0.80	0.80	0.76	0.70	0.62	0.49	0.38	0.26	0.16	0.10

To begin understanding the relationship between the risk of union formation and age, a simple age-dependent logistic model is estimated using the individual Person-Year Data Set. The model is estimated for the two sexes individually using a fractional polynomial specification fit by the *fracpoly* routine provided with the STATA statistical software package:

$$\ln\left(\frac{p(age)}{1-p(age)}\right)=\beta_0\frac{1}{x^2}+\beta_1\frac{\ln(x)}{x^2}+c$$

where

$$x=\frac{(age+1)}{10}$$

Equation 6: Fractional Polynomial Specification of Age-Only Model

The probability of union formation by age *p(age)* was calculated from the coefficients estimated for the model specified in Equation 6 using STATA's logistic regression routine applied to the individual Person-Year Data Set. Because the unit of analysis is person year, those probabilities correspond to the annual hazard of union formation as a function of age.

The hazard of union formation by age is summarized in Figure 35 which displays the predicted annual hazard of union formation by age over the period 1957-1995. It is immediately obvious that the male and female curves are substantially different. This is the result of the fact that the Tonga's polygynous marriage system allows all males to be at risk to marry at all times while only unmarried females are at risk. Consequently the male denominator is much larger at all ages and the hazards smaller at younger ages. The fact that the male hazards are not smaller at all ages reveals the fact that men marry older and continue to marry much younger women into middle and old age. The annual hazard of union formation peaks for males at age 24 and for females at age 19-20 revealing a roughly six year differential in the maximum hazard of union formation. Remember that the hazards in Figure 35 relate to the formation of a union irrespective of the parity of that union, so the peak in these curves is

[27] Population-weighted average.

associated with the maximum risk of union formation at any parity - not the only the *first* union. The average age at first marriage over the period 1957-1995 is 25 years for males and 19 years for females.

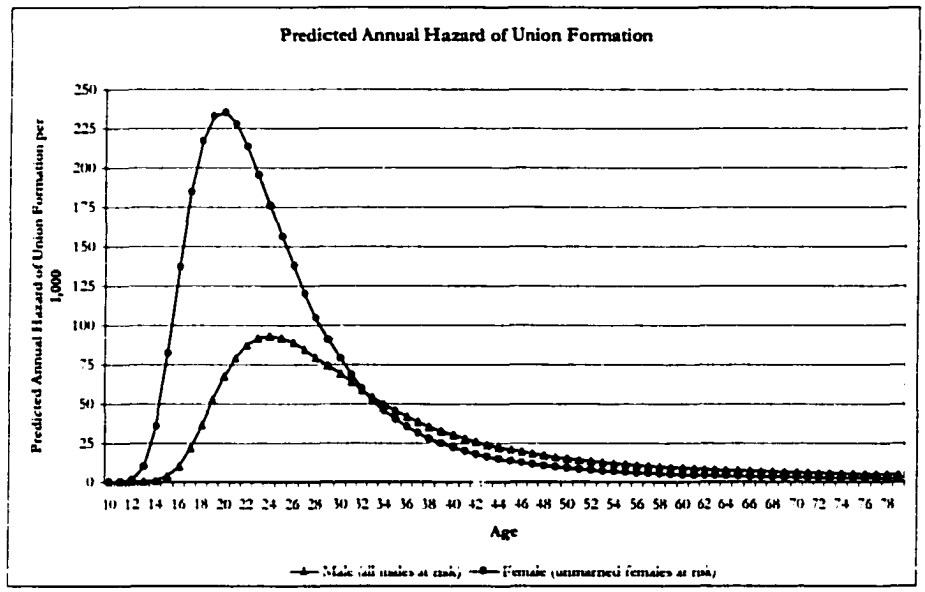

Figure 35: Predicted Annual Hazard of Union Formation by Age 1957-1995

The different shapes of the curves in Figure 35 hint at but largely mask the underlying complexity of the pairing process that creates unions. A more extensive analysis is undertaken to reveal how the hazard of union formation is related to both the ages of the male and female and the number of wives already married to the man – a two-sex hazard of union formation. The Unmarried Union-Year Data Set described above (Analysis Files) is used to estimate the two-sex hazard of union formation. The probability of union formation as a function of male age, female age, and male union parity is estimated using logistic regression, and because the

unit of analysis is an unmarried potential-union-year, the probability corresponds to the annual hazard of union formation for unmarried potential unions.

The model used to estimate the two-sex hazards of union formation is specified as:

$$\ln\left(\frac{p(a_i^m, a_j^f)}{1-p(a_i^m, a_j^f)}\right) = \beta_1 a_1^m + \cdots + \beta_n a_n^m + \beta_{n+1} a_1^f + \cdots + \beta_{2n} a_n^f + \beta_{2n+1} a_1^m a_1^f + \cdots + \beta_{2n+n^2} a_n^m a_n^f + c$$

Where p is the probability of union formation, a^m is a dummy variable that specifies the male age group, a^f is a dummy variable that specifies the female age group, $n+1$ is the total number of age groups, and $a^m a^f$ represents an interaction of the male and female dummy variables. The model is estimated using STATA's logistic regression routine for all unmarried potential-union-years, for only those involving unmarried men, for only those involving men with one wife, and for only those involving men with two or more wives.

Equation 7: Specification for Two-Sex Hazard of Union Formation Model

Table 33 and Figure 36 through Figure 39 display the annual hazards of union formation resulting from the analysis of the Unmarried Union-Year Data Set. All of the hazards presented relate to the period 1957-1995 as a whole since there are no period variables in the model.

The figures clearly reveal that new unions are most likely to be formed between women aged 15-24 and men aged 20-34 regardless of the number of wives the male might have, Figure 36. Unmarried men are most likely to marry between the ages of 20 and 34 and they are most likely to marry women between the ages of 15 and 24, and it is obviously this type of marriage that dominates the formation of new unions, Figure 37. The hazards themselves are of far

greater magnitude than the hazards for other male marriage orders, and the pattern clearly dominates the "All Marriage Orders" pattern displayed in Figure 36.

TABLE 33: ANNUAL HAZARD OF UNION FORMATION PER 10,000 BY MALE UNION PARITY, MALE AGE, AND FEMALE AGE FOR 1957-1995

Female Age	Male Age													
	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-79
	All Male Marriage Orders													
10-14	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
15-19	-NA-	89.10	489.81	182.59	52.89	33.80	14.08	24.17	7.65	-NA-	14.35	-NA-	-NA-	-NA-
20-24	-NA-	9.68	193.25	193.01	86.49	60.00	19.61	24.81	7.74	-NA-	-NA-	-NA-	-NA-	-NA-
25-29	-NA-	-NA-	9.96	59.78	60.58	47.98	33.99	26.99	8.37	-NA-	-NA-	-NA-	-NA-	-NA-
30-34	-NA-	-NA-	-NA-	18.53	55.84	47.92	36.25	28.41	17.36	22.80	15.11	-NA-	38.91	-NA-
35-39	-NA-	-NA-	-NA-	5.71	12.06	13.05	13.88	39.49	47.17	23.87	31.10	-NA-	-NA-	-NA-
40-44	-NA-	-NA-	-NA-	-NA-	-NA-	27.42	37.65	40.58	19.10	36.54	15.77	73.89	39.84	-NA-
45-49	-NA-	-NA-	-NA-	6.60	-NA-	7.47	-NA-	-NA-	20.02	61.43	16.67	-NA-	-NA-	-NA-
50-54	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	8.59	19.42	11.90	15.72	25.77	41.32	-NA-
55-59	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	10.20	-NA-	-NA-	-NA-	-NA-	-NA-
60-64	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	9.63	-NA-	-NA-	16.03	73.53	-NA-	-NA-
65-69	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	20.64	-NA-	-NA-	15.48	18.08	-NA-	-NA-	-NA-
70-74	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
75-79	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
	Male Unmarried													
10-14	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
15-19	-NA-	90.11	572.29	298.59	71.79	25.61	15.27	18.18	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
20-24	-NA-	9.96	255.25	302.63	115.91	78.84	46.15	55.56	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
25-29	-NA-	-NA-	17.61	85.47	96.85	58.56	67.23	18.83	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
30-34	-NA-	-NA-	-NA-	34.21	114.61	94.19	54.55	40.57	25.38	34.84	44.25	-NA-	-NA-	-NA-
35-39	-NA-	-NA-	-NA-	14.31	-NA-	-NA-	-NA-	68.34	85.96	-NA-	45.25	-NA-	-NA-	-NA-
40-44	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	70.92	24.21	-NA-	38.76	-NA-	161.29	-NA-	-NA-
45-49	-NA-	-NA-	-NA-	16.72	-NA-	22.78	-NA-	-NA-	-NA-	37.88	-NA-	-NA-	-NA-	-NA-
50-54	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	25.13	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
55-59	-NA-	-NA-	-NA	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
60-64	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	238.09	-NA-	-NA-
65-69	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
70-74	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
75-79	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-

TABLE 33 CONTINUED: ANNUAL HAZARD OF UNION FORMATION PER 10,000 BY MALE UNION PARITY, MALE AGE, AND FEMALE AGE FOR 1957-1995

Female Age	Male Age													
	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-79
	Male with One Wife													
10-14	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
15-19	-NA-	-NA-	206.49	113.15	35.59	21.49	19.92	14.43	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
20-24	-NA-	-NA-	84.03	107.53	74.49	38.97	-NA-	14.64	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
25-29	-NA-	-NA-	-NA-	54.84	27.73	50.30	12.39	16.16	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
30-34	-NA-	-NA-	-NA-	10.86	22.17	23.92	41.67	17.30	20.66	-NA-	-NA-	-NA-	-NA-	-NA-
35-39	-NA-	-NA-	-NA-	-NA-	28.37	15.50	33.28	-NA-	45.66	60.98	-NA-	-NA-	-NA-	-NA-
40-44	-NA-	-NA-	-NA-	-NA-	-NA-	49.59	17.92	40.73	23.15	-NA-	-NA-	-NA-	-NA-	-NA-
45-49	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	25.13	-NA-	-NA-	-NA-	-NA-	-NA-
50-54	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	23.26	-NA-	41.32	63.29	-NA-	-NA-
55-59	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	24.81	-NA-	-NA-	-NA-	-NA-	-NA-
60-64	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	25.06	-NA-	-NA-	43.29	-NA-	-NA-	-NA-
65-69	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	42.55	49.26	-NA-	-NA-	-NA-
70-74	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
75-79	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
	Male with Two or More Wives													
10-14	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
15-19	-NA-	-NA-	97.09	51.55	86.77	82.30	-NA-	48.54	29.59	-NA-	50.76	-NA-	-NA-	-NA-
20-24	-NA-	-NA-	-NA-	131.23	67.26	87.72	23.26	-NA-	30.03	-NA-	-NA-	-NA-	-NA-	-NA-
25-29	-NA-	-NA-	-NA-	-NA-	74.44	24.57	27.55	60.24	34.48	-NA-	-NA-	-NA-	-NA-	-NA-
30-34	-NA-	-NA-	-NA-	-NA-	27.03	24.69	-NA-	29.67	-NA-	41.15	-NA-	-NA-	147.06	-NA-
35-39	-NA-	-NA-	-NA-	-NA-	-NA-	27.03	-NA-	63.69	-NA-	-NA-	57.14	-NA-	-NA-	-NA-
40-44	-NA-	-NA-	-NA-	-NA-	-NA-	28.90	28.82	60.98	36.36	83.33	57.14	84.75	151.51	-NA-
45-49	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	36.50	168.07	59.17	-NA-	-NA-	-NA-
50-54	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	37.59	40.82	-NA-	-NA-	153.85	-NA-
55-59	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
60-64	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
65-69	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	69.93	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
70-74	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
75-79	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-

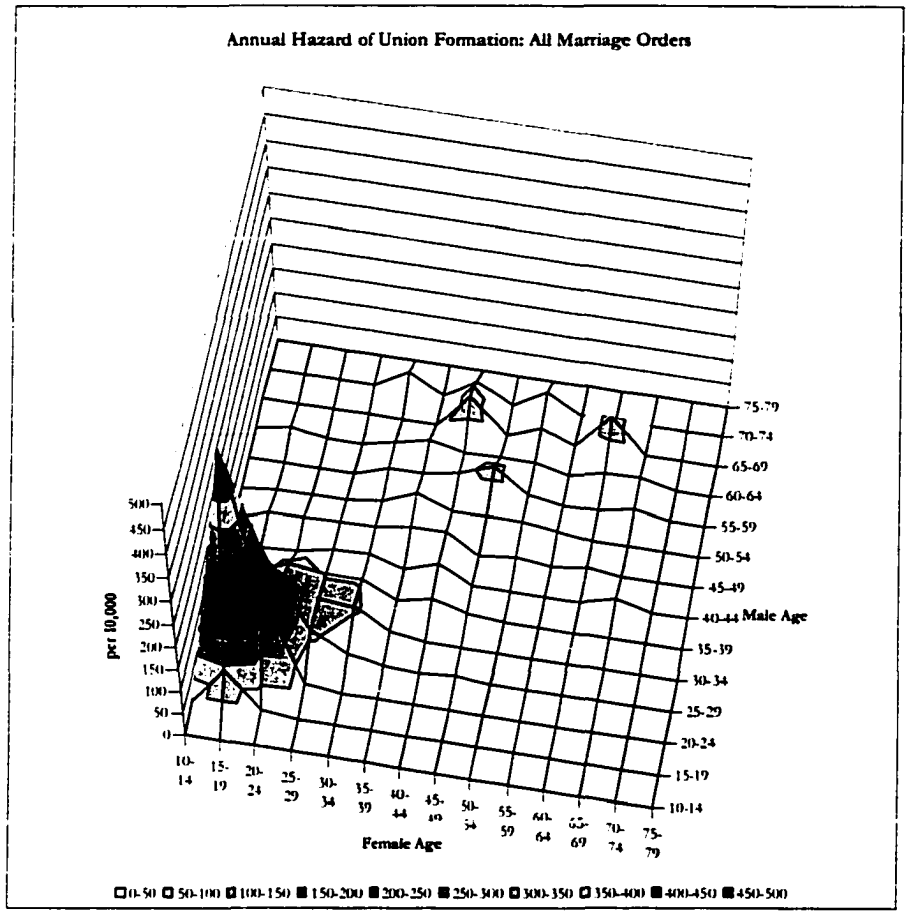

Figure 36: Annual Hazard of Union Formation: All Marriage Orders

Figure 38 and Figure 39 reveal that higher order polygynous unions are formed between a wider age range of both men and women. It is more likely for young men with one wife to marry another, especially in the 20-44 age range, and it is likely that they will marry young women, mostly younger than 25 years. However older men with one wife also may take another, but in general the second wife of older men is likely to be middle aged to elderly herself, Figure 38. In contrast, the likelihood that a man with two or more wives will acquire another is more diffuse with age for both the man and the women. The broadest peak is still

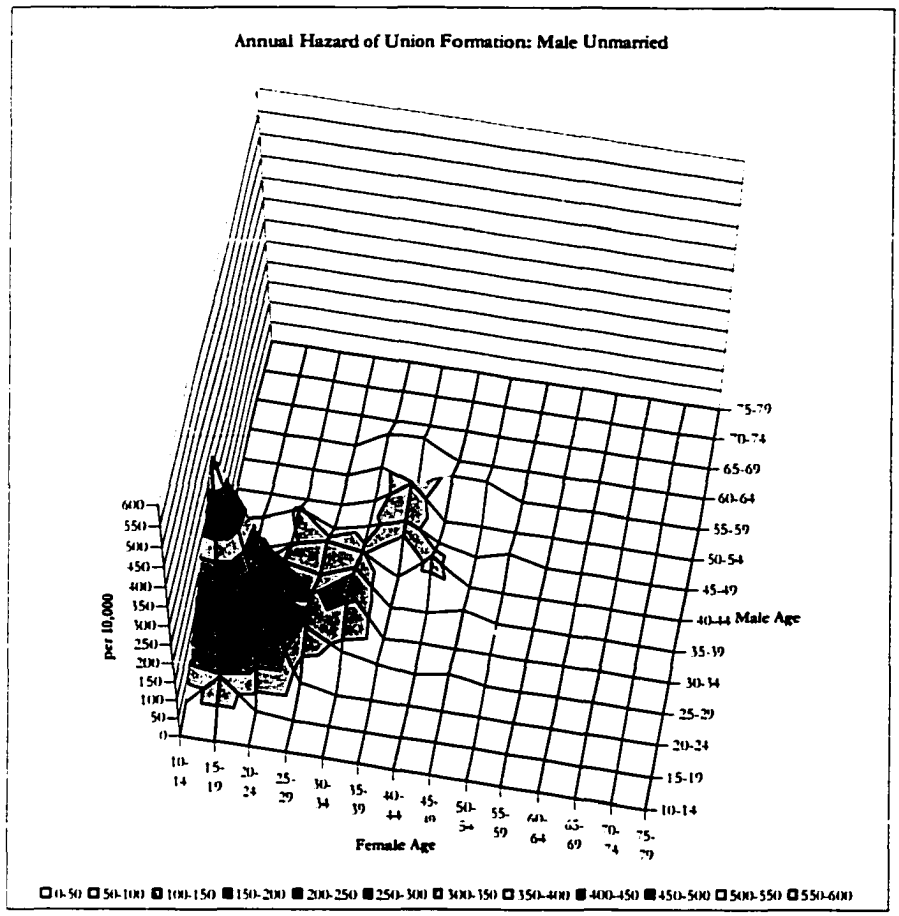

Figure 37: Annual Hazard of Union Formation: Male Unmarried

young for both men and women, but much wider than for lower order unions. Additionally, for higher order wives, men of all ages are more likely to form unions with young women, aged 50 or less for the most part with a substantial concentration below age 30, Figure 39.

All of this seems to indicate that there is a significant difference between the acquisition of a second and higher order wives. A second wife is taken by two different groups of men, young

men who most likely acquire her through a standard marriage arrangement and older men who most likely inherit her from a dead relative. The relatively advanced age of the women who

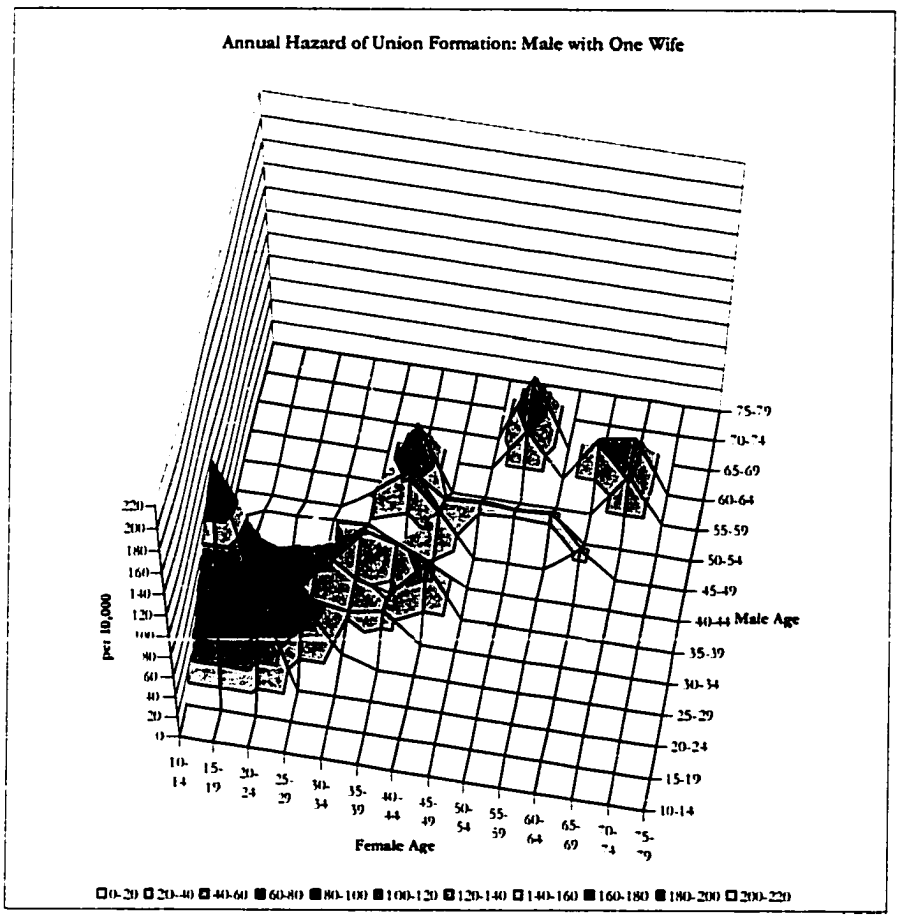

Figure 38: Annual Hazard of Union Formation: Male with One Wife

marry older men as second wives and the fact that it is not at all likely for older men to take a second wife supports this conclusion.

So it appears that men who will be polygynous through normal marriage embark on that future quite early in life and continue to marry younger women as they age, and that those who are

not going to be polygynous have a smaller likelihood of acquiring additional wives later in life, most likely though inheritance mechanisms.

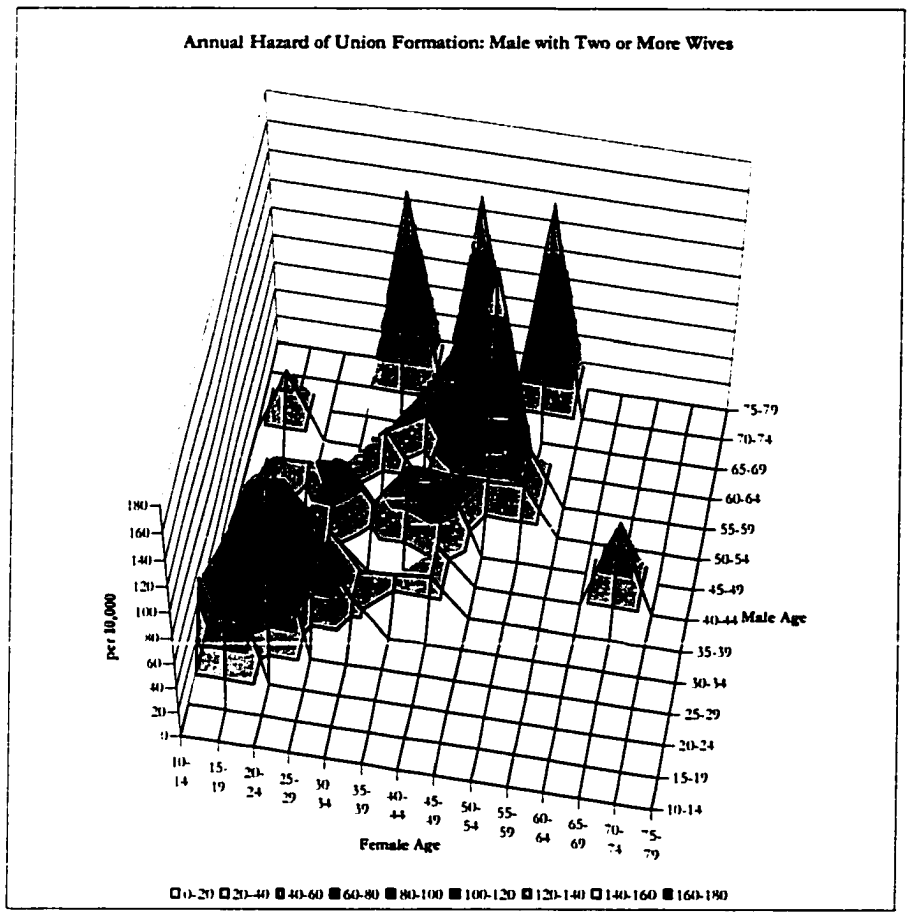

Figure 39: Annual Hazard of Union Formation: Male with Two or More Wives

TRENDS

The trend in the annual hazard of union formation is examined using a model that predicts the annual hazard of union formation as a function of period and male marital parity. Dummy

variables for male marriage parity and period are interacted to allow the independent estimation of a time series of odds ratios for each male marriage parity, Equation 8.

$$\ln\left(\frac{p(m_i,t_j)}{1-p(m_i^m,t_j^f)}\right)=$$
$$\beta_1 m_1+\beta_2 m_2+$$
$$\beta_3 t_1+\cdots+\beta_9 t_7+$$
$$\beta_{10} m_1 t_1+\cdots+\beta_{23} m_2 t_7+\varepsilon$$

Where m is a dummy variable for male marriage parity and t is a dummy variable for time period.

Equation 8: Specification for Trend in Union Formation Model

Table 34 displays the odds ratios from the logistic regression estimates of the union formation model specified in Equation 8. The reference time period is 1992-1995 so all of the odds ratios refer to that period. The odds ratios by time for all three male marriage parity categories are also plotted in Figure 40. The P-values listed in Table 34 result from testing the hypothesis that the coefficients for each pair of consecutive time periods are equal; they do *not* correspond to the P-values estimated for each of the coefficients themselves. A small value indicates that the difference between the two odds ratios is unlikely to occur by chance; so a change in the odds ratio from one period to another is "statistically significant" if the P-value is small.

TABLE 34: TREND IN PROBABILITY OF UNION FORMATION

Period	Unmarried		One Wife		2+ Wives	
	Odds Ratio	P-Value*	Odds Ratio	P-Value*	Odds Ratio	P-Value*
1957-1961	2.07	0.3270	2.38	0.8250	1.12	0.3590
1962-1966	1.66	0.4540	2.61	0.7635	1.94	0.5675
1967-1971	1.41	0.5589	2.93	0.1499	1.42	0.8620
1972-1976	1.58	0.3289	1.72	0.9906	1.55	0.3167
1977-1981	1.34	0.0609	1.71	0.5398	2.38	0.1819
1982-1986	0.99	0.5357	1.36	0.6031	1.25	0.8447
1987-1991	1.09	0.5620	1.64	0.1600	1.38	0.5520
1992-1995	1.00	-NA-	1.00	-NA-	1.00	-NA-

*P-value for test of the null hypothesis that each pair of consecutive odds ratios is equal

It is obvious that the odds of forming a union have been steadily decreasing over time for unmarried men and men with only one wife. In contrast the odds of forming a union for men with two or more wives seems to be relatively flat. From one period to another the changes in odds ratios are generally not statistically significant at the usual five or ten percent levels. Together with the fact that the differences in the odds ratios from one period to the next are relatively small, this indicates that change has been gradual, more like a gentle drift than a swift transition.

The gradual decline in the likelihood of union formation for married men fits well with the fact that the average number of wives per man has also been declining steadily between 1957 and 1995 as displayed in Figure 41. The average married man in 1957 had roughly 1.5 wives while in 1995 the average married man had about 1.3 wives.

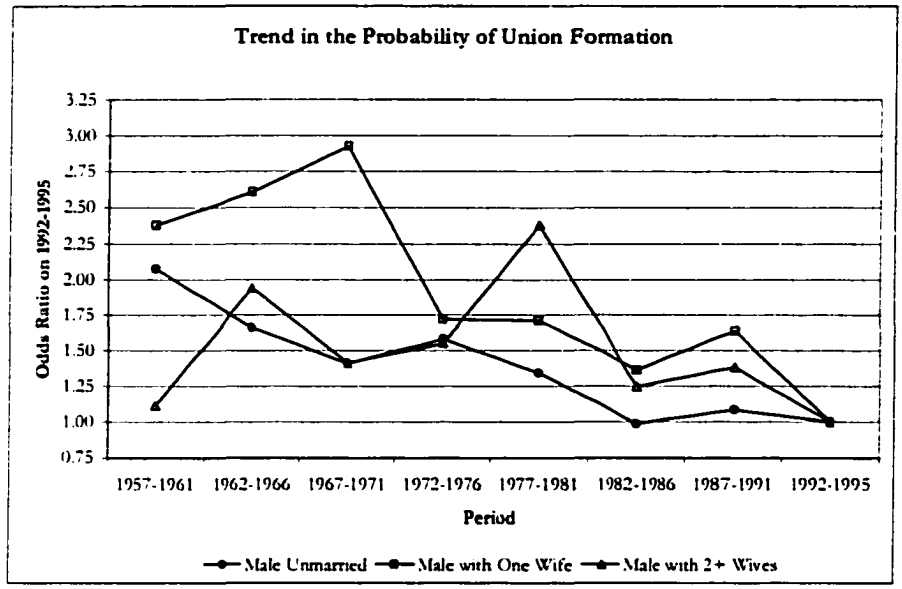

Figure 40: Trend in Probability of Union Formation: Male Unmarried

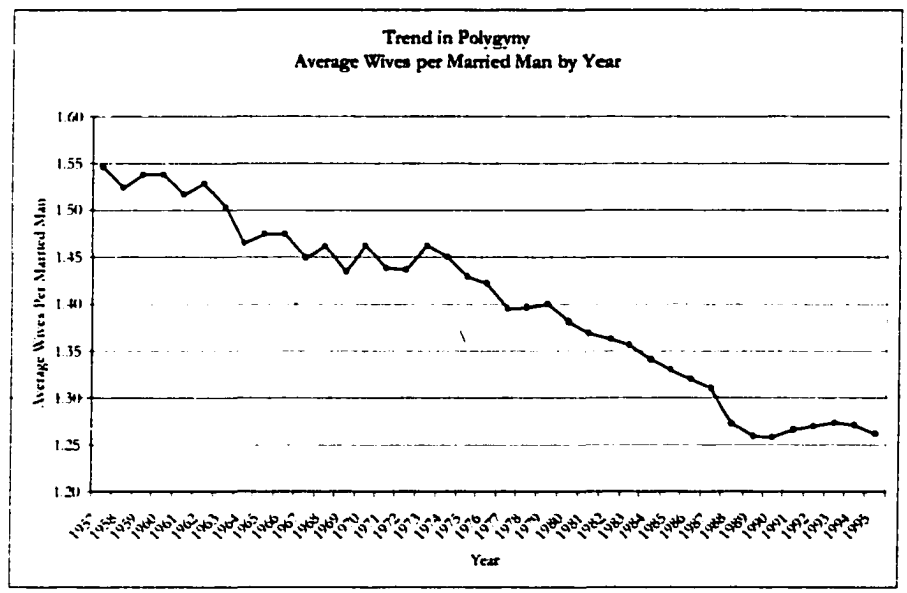

Figure 41: Trend in Polygyny

This has been a cursory examination of the very complex marriage system practiced by the Tonga. What emerges clearly is that most people spend most of their adult lives married, the likelihood of higher order marriages for men is not negligible, that a men who will be polygynous embark on that future early in their lives, and that the overall likelihood of union formation has been decreasing steadily for unmarried men and men with only one wife with the result that the average number of wives per man has also decreased steadily.

UNION DISSOLUTION

The dissolution of unions is examined along three major dimensions: 1) age, 2) achieved fertility of the pair, and 3) duration of the union. Scudder hypothesizes that the likelihood of separation is inversely related to the achieved fertility of the pair, and it is to test that hypothesis that achieved fertility is included in this analysis.

LEVELS

The overall likelihood of dissolving a union (excluding death) as a function of age and achieved fertility is examined through the use of dummy variable-specified logistic regression model (similar to that used to examine the likelihood of union formation by age) applied to the Union-Years Data Set, Equation 9. Successful reproduction is important to both men and women so it is likely that unions that produce no surviving children will have a higher likelihood of being dissolved.

$$\ln\left(\frac{p(a_i^m, a_j^f)}{1-p(a_i^m, a_j^f)}\right) =$$
$$\beta_1 a_1^m + \cdots + \beta_n a_n^m +$$
$$\beta_{n+1} a_1^f + \cdots + \beta_{2n} a_n^f +$$
$$\beta_{2n+1} a_1^m a_1^f + \cdots + \beta_{2n+n^2} a_n^m a_n^f +$$
$$\beta_{2n+n^2+1} d_1 + \cdots + \beta_{2n+n^2+m} d_m + c$$

p is the probability of dissolving a union, a^m is a dummy variable that specifies the male age group, a^f is a dummy variable that specifies the female age group, $n+1$ is the total number of age groups, $a^m a^f$ represents an interaction of the male and female dummy variables, d is a dummy representing the duration of the union, and m is the total number of union duration groups. The model is estimated using STATA's logistic regression routine for unions of all fertility orders, for unions with no surviving children, for unions with one to two surviving children, and for unions with three or more surviving children.

Equation 9: Specification for the Hazard of Union Dissolution Model

Table 35 and Figure 42 through Figure 45 display the annual hazards of separation for married unions resulting from the estimation of Equation 9. Substantial likelihoods of separation are restricted to the ages of men and women that are typical for married pairs.

Pairs with no surviving children face a dramatically increasing likelihood of separation as they age, and overall the likelihood of separation is highest over the largest range of ages for pairs with no surviving children. As the number of surviving children increases, the hazards of separation decrease and the range of ages over which the hazard of separation is substantial narrows. Young pairs with three or more surviving children appear to have a negligible likelihood of separation. In general the likelihood of separation decreases with the number of surviving children.

TABLE 35: SEPARATION PROBABILITIES PER 1,000 BY MALE AND FEMALE AGE

Female Age	Male Age													
	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-79
	All Fertility Orders													
10-14	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
15-19	-NA-	14.81	27.61	41.26	53.24	54.14	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
20-24	-NA-	-NA-	25.36	37.44	34.63	28.28	60.72	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
25-29	-NA-	-NA-	-NA-	31.39	37.95	45.85	46.74	48.41	76.47	86.24	-NA-	-NA-	-NA-	-NA-
30-34	-NA-	-NA-	-NA-	81.95	39.83	39.22	63.57	76.30	29.84	83.24	-NA-	-NA-	-NA-	-NA-
35-39	-NA-	-NA-	-NA-	-NA-	77.64	15.10	45.69	59.43	33.10	17.37	85.32	-NA-	-NA-	-NA-
40-44	-NA-	-NA-	-NA-	-NA-	-NA-	98.74	61.30	56.81	89.49	71.69	83.65	-NA-	-NA-	-NA-
45-49	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	38.53	62.11	41.77	172.15	-NA-	-NA-	-NA-
50-54	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	99.41	67.99	59.64	107.44	60.78	-NA-
55-59	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	59.56	113.98	102.10	-NA-	-NA-	-NA-
60-64	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	35.92	-NA-	-NA-
65-69	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
70-74	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
75-79	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
	No Children													
10-14	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	NA-	-NA-	-NA-	-NA-	-NA-	-NA-
15-19	-NA-	19.05	28.98	42.46	-NA-	41.64	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
20-24	-NA-	-NA-	32.20	50.69	41.11	32.39	43.49	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
25-29	-NA-	-NA-	-NA-	-NA-	21.81	73.11	76.53	50.19	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
30-34	-NA-	-NA-	-NA-	-NA-	91.54	33.71	82.82	-NA-	81.70	-NA-	-NA-	-NA-	-NA-	-NA-
35-39	-NA-	-NA-	-NA-	-NA-	262.83	-NA-	40.20	57.32	43.63	-NA-	113.82	-NA-	-NA-	-NA-
40-44	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	24.41	75.03	-NA-	147.60	-NA-	-NA-	-NA-
45-49	-NA-	-NA-	-NA-	-NA-	-NA-	327.05	-NA-	-NA-	137.33	57.16	284.90	-NA-	-NA-	-NA-
50-54	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	153.00	55.41	104.12	-NA-	-NA-	-NA-
55-59	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
60-64	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
65-69	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
70-74	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
75-79	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-

In addition to examining the dependence of the hazard of separation on age and achieved fertility, its dependence on the duration of the union is investigated. The annual hazard of separation is modeled as a function of duration of union for each of the three achieved fertility levels discussed above. The hazards resulting from that model are displayed in Table 36 and Figure 46 as odds ratios with the reference category being unions of duration one to four year.

TABLE 35 CONTINUED: SEPARATION PROBABILITIES PER 1,000 BY MALE AND FEMALE AGE

Female Age	Male Age													
	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-79
	One to Two Children													
10-14	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
15-19	-NA-	-NA-	25.23	39.14	23.54	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
20-24	-NA-	-NA-	-NA-	29.59	32.64	-NA-	54.05	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
25-29	-NA-	-NA-	-NA-	29.46	42.57	43.15	33.18	40.82	-NA-	144.84	-NA-	-NA-	-NA-	-NA-
30-34	-NA-	-NA-	-NA-	-NA-	17.00	32.44	66.32	65.05	25.88	121.69	-NA-	-NA-	-NA-	-NA-
35-39	-NA-	-NA-	-NA-	-NA-	-NA-	19.51	31.26	-NA-	39.05	-NA-	-NA-	-NA-	-NA-	-NA-
40-44	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	44.03	111.00	68.37	-NA-	-NA-	-NA-	-NA-
45-49	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	112.02	-NA-	-NA-	-NA-	-NA-
50-54	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
55-59	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
60-64	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
65-69	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
70-74	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
75-79	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
	Three or More Children													
10-14	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
15-19	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
20-24	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	213.87	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
25-29	-NA-	-NA-	-NA-	61.16	51.12	-NA-	-NA-	84.14	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
30-34	-NA-	-NA-	-NA-	-NA-	-NA-	40.99	46.66	66.72	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
35-39	-NA-	-NA-	-NA-	-NA-	-NA-	13.80	44.10	59.09	19.40	34.68	-NA-	-NA-	-NA-	-NA-
40-44	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	67.42	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
45-49	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	21.61	-NA-	-NA-	-NA-	-NA-	-NA-
50-54	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	243.90	-NA-	63.89	29.26	-NA-	-NA-	-NA-
55-59	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	98.32	-NA-	-NA-	-NA-
60-64	-NA	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
65-69	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
70-74	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
75-79	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-

Clearly as unions age they are less likely to break, and the level of achieved fertility does not appear to alter the duration-dependence of the separation hazard.

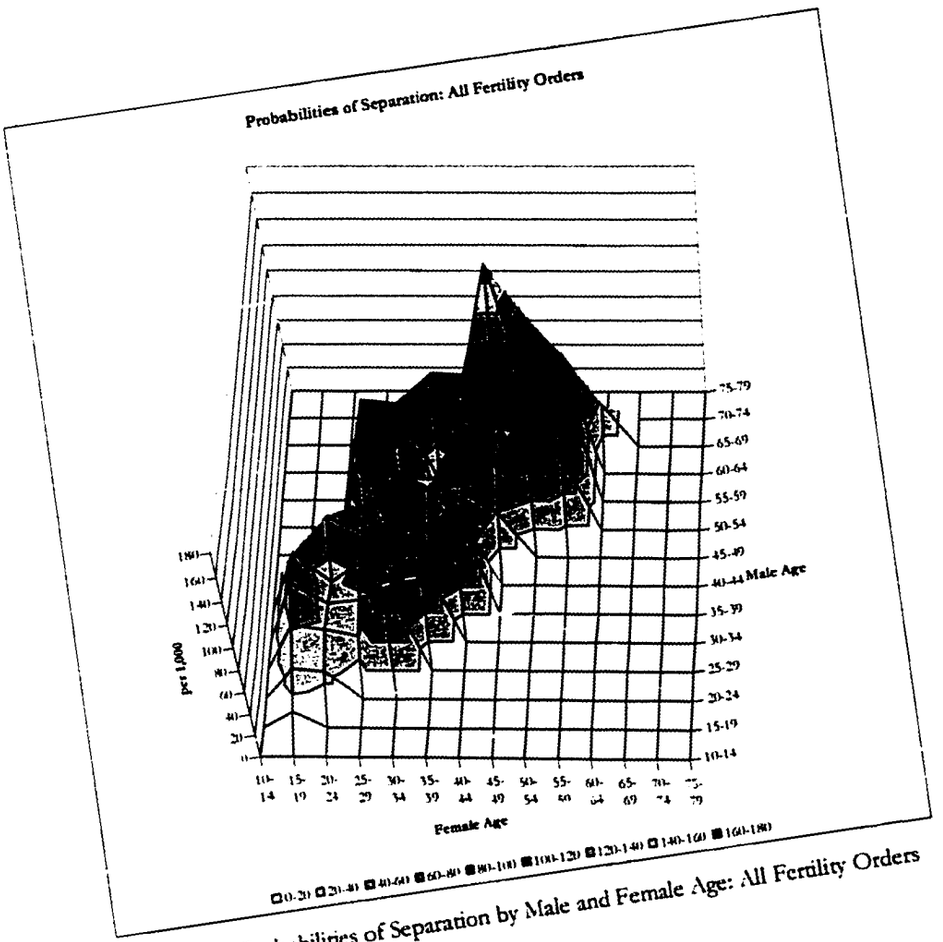

Figure 42: Probabilities of Separation by Male and Female Age: All Fertility Orders

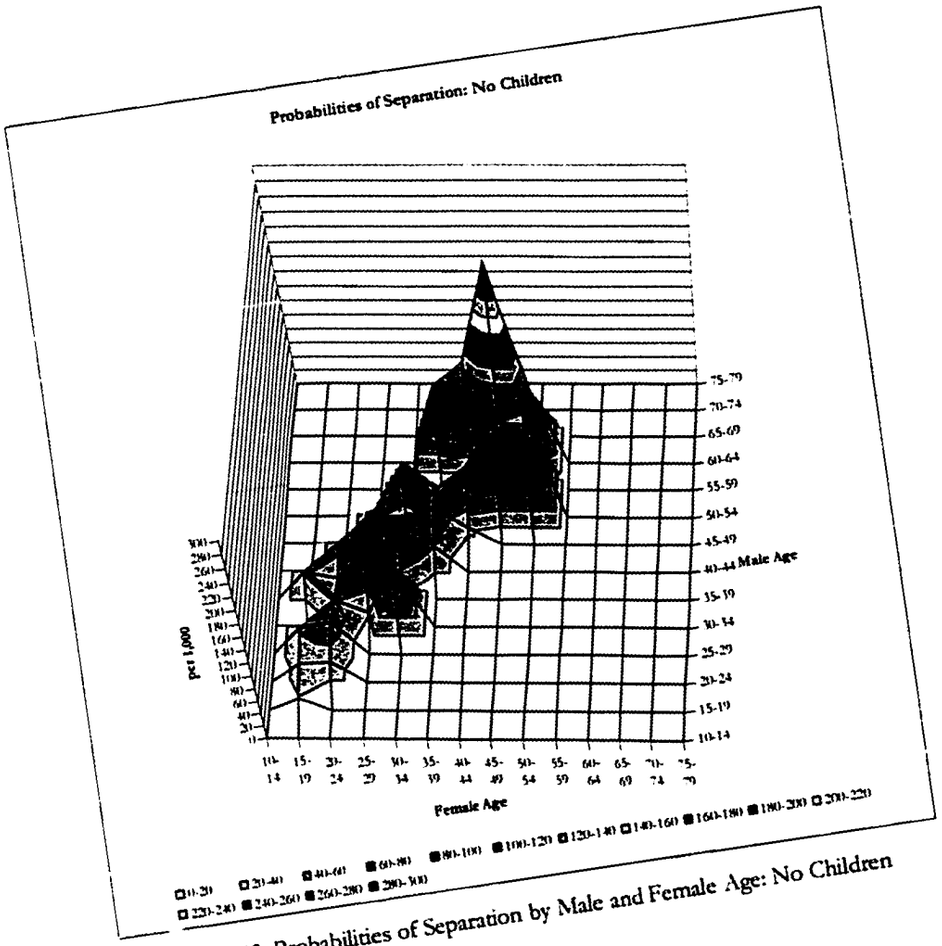

Figure 43: Probabilities of Separation by Male and Female Age: No Children

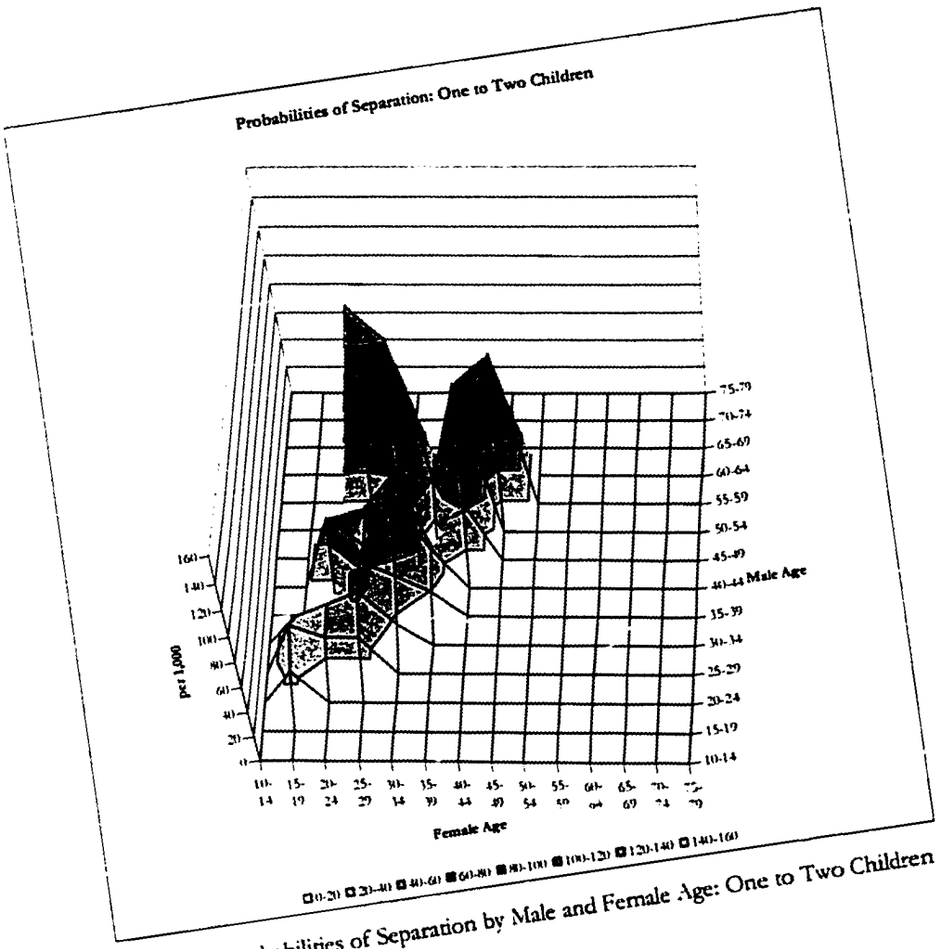

Figure 44: Probabilities of Separation by Male and Female Age: One to Two Children

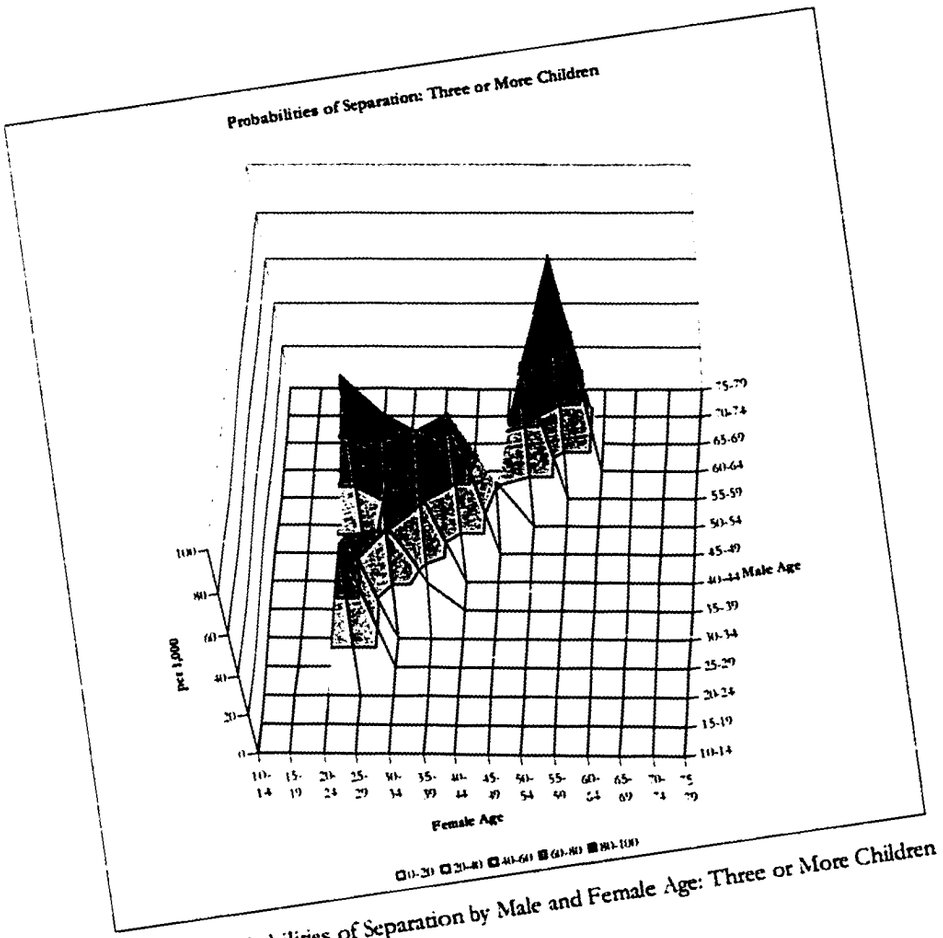

Figure 45: Probabilities of Separation by Male and Female Age: Three or More Children

TABLE 36: PROBABILITY OF SEPARATION BY DURATION OF MARRIAGE

ODDS RATIOS ON DURATION 1-4 YEARS

Years	Fertility Order: Number of Children			
	All	0	1-2	3+
1-4	1.0000	1.0000	1.0000	1.0000
5-9	0.6063	0.5530	0.8339	0.2018
10-14	0.2784	0.3715	0.3850	0.2342
15-19	0.2234	0.2192	0.1122	0.3023
20-24	0.2143	0.0976	0.3712	0.2566
25-29	0.0564	0.3930	0.0752	0.0448
30-34	0.0960	-NA-	0.2856	0.1979
35-39	0.1870	-NA-	0.0000	0.2815
40-44	0.1181	-NA-	-NA-	0.2303
45-49	0.0944	-NA-	-NA-	0.2405

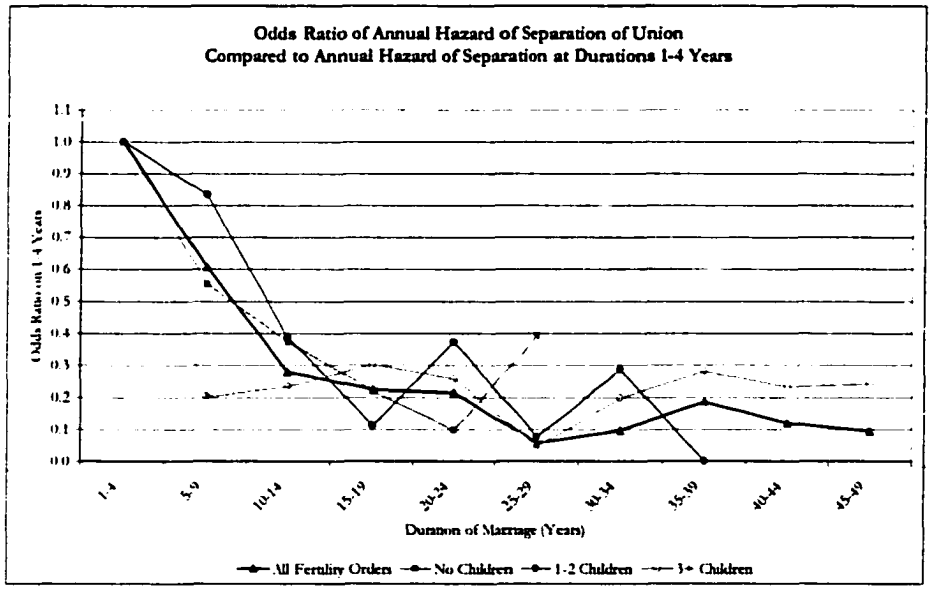

Figure 46: Annual Hazard of Separation by Duration of Union

TRENDS

The trend in the annual hazard of separation is examined with another simple model specification similar to that used to examine the trend in the hazard of union formation.

$$\ln\left(\frac{p(c_i,t_j)}{1-p(c_i,t_j^f)}\right)= \beta_1 c_1 + \beta_2 c_2 + \beta_3 t_1 + \cdots + \beta_9 t_7 + \beta_{10} c_1 t_1 + \cdots + \beta_{23} c_2 t_7 + \varepsilon$$

Equation 10: Specification for the Hazard of Separation Trend Model

Equation 10 describes the specification for that model. Dummy variables for the number of surviving children c and the time period t are interacted to allow the independent estimation of the hazards of separation for each period. The estimates are displayed as odds ratios in Table 37.

TABLE 37: TREND IN HAZARD OF SEPARATION

	No Children		1-2 Children		3+ Children	
Period	Odds Ratio	P-Value	Odds Ratio	P-Value	Odds Ratio	P-Value
1957-1961	1.84	0.6507	0.37	0.9389	0.48	0.4809
1962-1966	2.19	0.5353	0.39	0.2424	0.90	0.8490
1967-1971	1.74	0.5425	0.71	0.5590	0.79	0.9587
1972-1976	2.15	0.0930	0.93	0.9147	0.82	0.0792
1977-1981	1.21	0.4427	0.97	0.1542	1.98	0.5031
1982-1986	1.57	0.1598	0.50	0.0440	1.45	0.9053
1987-1991	1.01	0.9690	1.19	0.6910	1.37	0.5630
1992-1995	1.00	-NA-	1.00	-NA-	1.00	-NA-

As with the trend table for union formation, the P-values in Table 37 represent the likelihood that the difference in pairs of consecutive odds could have been produced by a random process. In general the individual consecutive pairs of odds ratios are not highly significantly

different from one another. However, the significance of the difference between the first and last are significant. Again, this indicates a gradual process that accumulates significant change over a longer period of time.

The striking result is that the odds of separation have increased by a factor of two for married pairs with children but have decreased, also by a factor of two, for married pairs with no children. This would seem to indicate that achieved fertility has become less important that in was, and that the likelihood of divorce has in general been increasing over time.

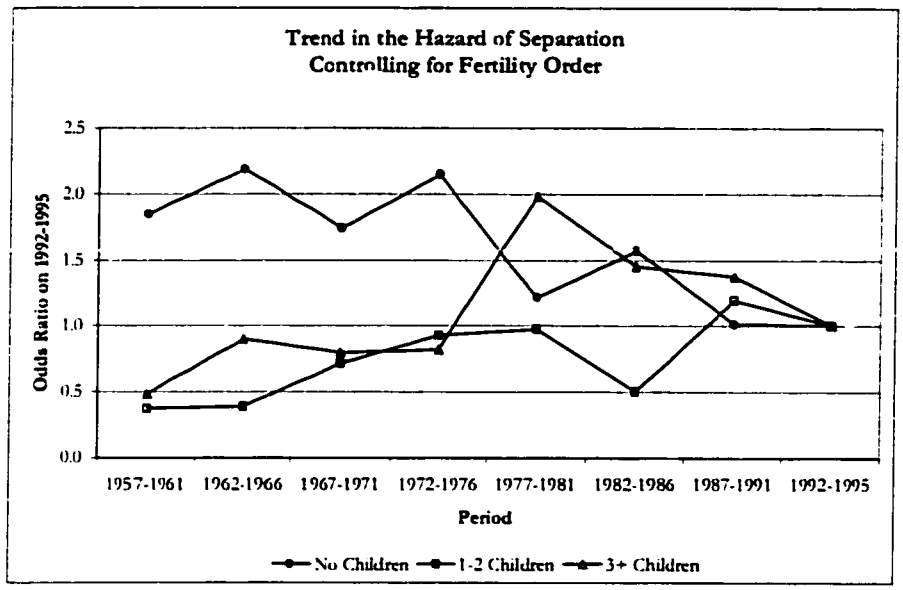

Figure 47: Trend in the Probability of Separation

FERTILITY

The Gwembe Tonga have maintained a high level of fertility from 1957 to 1995. The overall levels and trends are quantified and discussed briefly. As with the measures of nuptiality, these indicators are produced and examined mainly for the purpose of providing realistic parameters for the population simulation developed in Part 3. The only departure from this straightforward empirical approach is an effort to identify the potential impact of HIV/AIDS on fertility during the most recent periods. HIV/AIDS may reduce fertility, but because there are many other reasons why fertility may be falling in the most recent periods it is not possible to assign responsibility to AIDS alone. What is possible is to build a circumstantial case supported by evidence of the plausible impact of HIV on both fertility and mortality during the same period, and that is what I attempt to do.

The hazard of giving birth is calculated as a function of age and time for both men and women. Male fertility measures are rare because it is difficult to accurately assign paternity, and the Gwembe Tonga are no exception to this fact. Because of this, the male fertility measures must be understood as measures of *socially assigned* paternity. By that I mean that the measures reflect the fertility that has been legitimately assigned to males through their unions to females. In most cases this must closely match their biologic fertility, but the two are not necessarily identical. I find these measures interesting because they present a startling contrast to the female figures and clearly demonstrate that the reproductive histories and strategies of men and women are different. Additionally, they complete the description of the reproductive dynamics of the population that has been important for me as I thought about constructing the population simulation.

LEVELS

The overall level of fertility is identified by the boring yet classic age-specific fertility and total fertility rates. The unusual feature of these figures is how they are calculated. The *age-specific fertility rates* presented here are actually the annual hazards of giving birth by age. As such they cannot be interpreted exactly as standard age-specific fertility rates are even though they are qualitatively equivalent, they represent the number of children that an average woman could expect to deliver during a year. In this case the rates represent the probability that a woman of a certain age at the beginning of a period will give birth over the next year. This is a real probability measure which I find more appealing and easier to conceptualize under any circumstance. Moreover because they are calculated using the discrete time event history method, one is able to estimate all kinds of statistical models that control for various impacts other than age, such as period. This advantage is used here to examine the significance of the changes in age-specific fertility over time.

Equation 11 contains the basic specification for the hazard model used to estimate the female age-specific fertility rates.

$$\ln\left(\frac{p(a_i)}{1-p(a_i)}\right)=\beta_1 a_1+\cdots+\beta_8 a_8+c$$

Where p is the probability of giving birth during one year and a is a set of dummy variables representing age. The model is run on the Individual-Year Data set in various forms on both the total and married population of females.

Equation 11: Specification for Age-Specific Fertility Hazard Model

Equation 11 is the simplest form that simply estimates the age-specific annual hazards of giving birth over the entire period described by the data set 1957-1995. Additional specifications include dummy variables for time period and various interactions of age and time period.

A polygynous marriage system complicates the calculation of the male age-specific fertility rates because it is not only possible but indeed surprisingly common for polygynously married males to experience more than one birth in a year. Because of this it is not possible to use a standard binary logistic model to estimate the hazard of experiencing a birth for males. The straightforward but cumbersome solution is to use a multinomial logistic regression model that is able to jointly estimate the probability of experiencing zero, one, two, or more births in a given year. The multinomial specification uses the number of children born during a man-year as the dependent variable and dummies for the man's age and time period as the independent variables, Equation 12.

The probabilities resulting from an estimation of this model are combined into a weighted average according to Equation 13. To interpret this, think of a standard age-specific fertility rate as the number of children an average woman could expect to have during a year if she lived through the whole year (the basis for the concept of the total fertility rate). What Equation 13 does is produce the number of births that a man could expect to experience during a year. This can be conceptually manipulated in exactly the same way as a standard female age-specific fertility rate.

$$\Pr(i) = \frac{e^{A \cdot \beta^{(i)}}}{e^{A \cdot \beta^{(1)}} + e^{A \cdot \beta^{(2)}} + \cdots + e^{A \cdot \beta^{(n)}}}$$

$\Pr(i)$ is the annual hazard of having i children born to a man.

$$i \in \{1 \cdots n\}$$

where n is the greatest number of children born to a man during a year.

$$A \cdot \beta^{(i)} = [a_{15}, \cdots, a_{75}] \cdot [\beta_{15}, \cdots, \beta_{75}] + c$$

is the inner product of the vector of observed age dummies $A = \{a_{15}, \ldots, a_{75}\}$ and the vector of estimated coefficients $\beta = \{\beta_{15}, \ldots, \beta_{75}\}$ plus a constant c corresponding to i children born during one year.

Equation 12: Specification of the Age-Specific Multinomial Hazard Model of Male Fertility

$$P_i = 1 \cdot p_i(1) + 2 \cdot p_i(2) + \cdots + n \cdot p_i(n)$$

Where P represents the composite probability of birth, the $p(1)$ through $p(n)$ are the individual probabilities of giving birth to one to n children during a year, and the i increments the age group.

Equation 13: Weighted Average of Multiple Male Birth Probabilities

The male and female age-specific fertility rates are displayed in Figure 48 and Figure 49, and Figure 50 places the female period-specific age-specific fertility rates with respect the age-specific fertility rates measured by the 1992 DHS survey in Zambia. Table 38 and Table 39 contain the data on which the figures are based. The age-pattern of fertility appears reasonable for both males and females. Through time the female fertility rates fall closely around the values measured by the 1992 DHS in Zambia, and the 1957-1995 figures are pulled down by the fact that fertility during the most recent period between 1992 and 1995 is substantially

lower than it is during the previous periods. The fact that the age-pattern for females is very similar to that measured in the DHS survey (see Figure 50) supports the assertion that both the DHS and the Gwembe data are good measures of the underlying reality – or, both are equally bad and biased in exactly the same ways. Given the different methodologies employed in collecting the data and the widely different time periods surveyed, it is unlikely that they contain the same biases or errors, so I am confident that both are good. There is nothing to which to compare the male figures, but given the fact that the female figures are similar to other reliable data I am also confident that the male data are good. A striking feature of the male age-specific fertility rates is the fact that males continue to reproduce at significant rates through age 65, and there is measurable reproduction going on all the way through age 79.

The corresponding total fertility rates indicate the overall fertility is high in the Gwembe. Over the periods 1957-1995, the female total fertility rate calculated over ages 15 to 49 has been 5.5 while the female total marital fertility rate has been 7.9. The differences reflect the fact that not all women are married, and that when women are married they reproduce at a high rate. The figures for men are 8.1 for the total fertility rate and 14.3 for the total marital fertility rate. The substantially greater magnitude of the male total fertility rates results from the definition of the total fertility rate and demonstrates clearly why the total fertility rate is a poor measure of overall fertility.

The classic total fertility rate is the sum over ages 15 to 49 of the female age-specific fertility rates at each age. The sum is usually implemented in five-year age groups so that the seven five-year age-specific fertility rates are summed and multiplied by five to yield the number of

children that an average woman would expect to have if she lived all the way through her reproductive years and at each age bore children at a rate corresponding to the age-specific fertility rate prevailing at that age. This produces a nice age-standardized rate with a convenient interpretation. The problem is that it does not control for the size of the population at each age and therefore every age or age group is given the same weight in the calculation, even though there may be different numbers of women actually subject to the fertility rate at each age[28]. The male total fertility rates clearly identify this problem.

Males are able to continue siring children into old age with the consequence that they have non-zero age-specific fertility rates into old age. As the male population ages its size diminishes quickly so that the denominator of age-specific rates becomes dramatically smaller. Under these circumstances even a few reproductive men are able to generate comparatively large age-specific fertility rates. Summing up the male age-specific rates over a much larger age interval and including comparatively high rates at older ages results in a much larger total fertility rate for men.

That is in fact why the total fertility rates for men are so much higher, but it may still not satisfy the inquiring mind! Another feature of the total fertility rate as it is normally constructed is the fact that it is identified with a single sex and does not explicitly take into account the pairing of the sexes that is necessary to produce births. It is because of this fundamental discordance that the total fertility rates for men and women do not match: they are totally independent of one another and refer to completely different base populations.

[28] A solution to this problem may involve age-standardizing the age-specific fertility rates in a manner similar to the age-standardization of a crude death rate. This would solve the weighting problem but also destroy the neat interpretation.

This situation is exacerbated in a polygynous marriage system where old men are often married to and reproducing with young women.

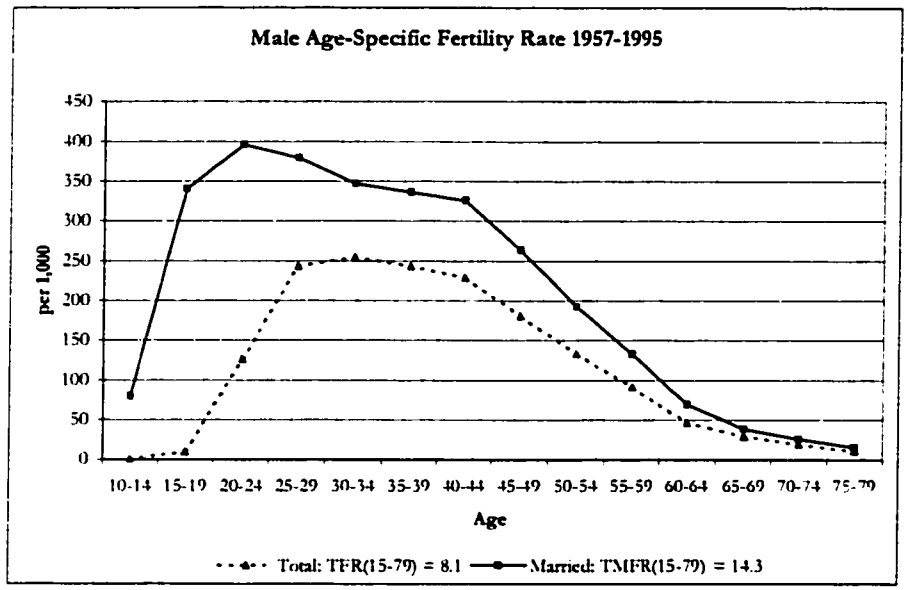

Figure 48: Male Age-Specific Fertility Rates 1957-1995

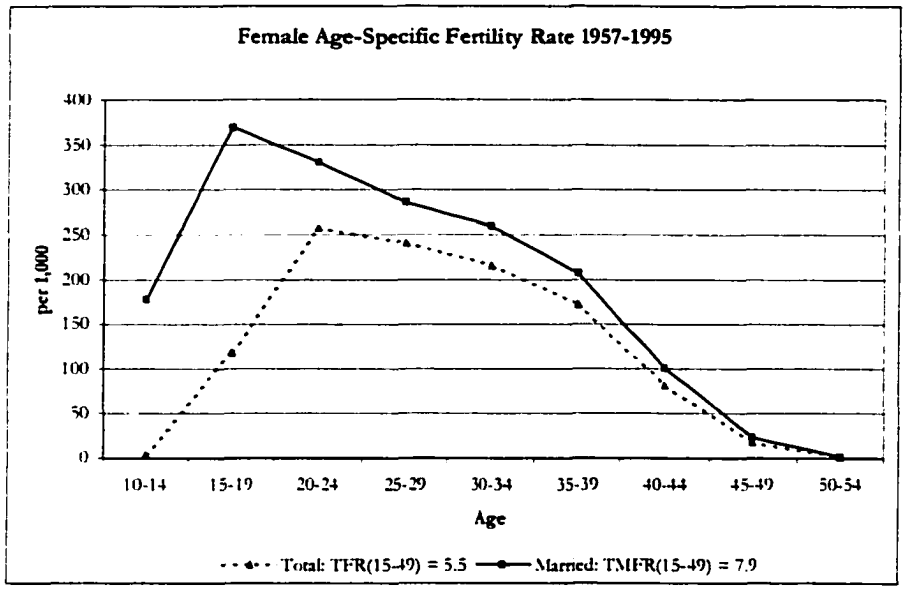

Figure 49: Female Age-Specific Fertility Rates 1957-1995

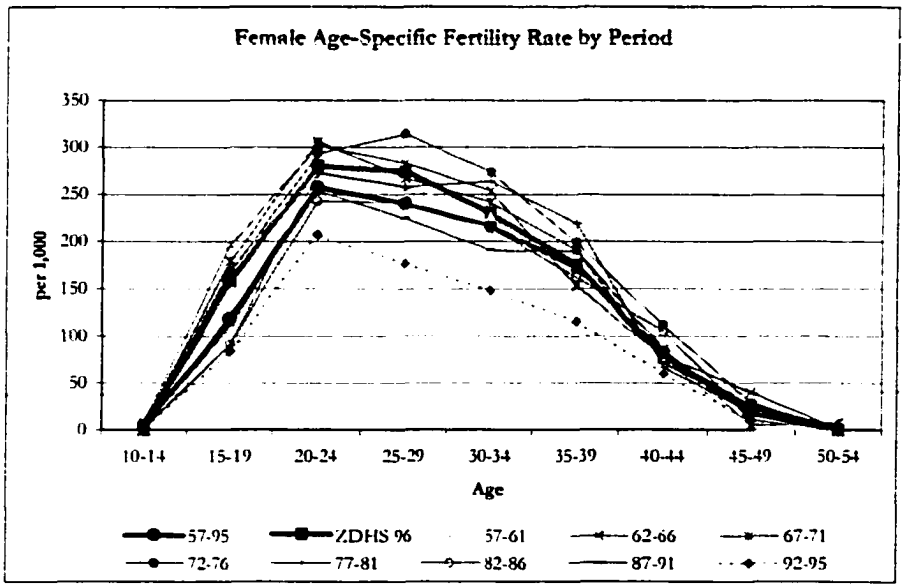

Figure 50: Female Age-Specific Fertility Rates by Period

TABLE 38: MALE AGE-SPECIFIC FERTILITY RATES BY PERIOD

Age	57-61	62-66	67-71	72-76	77-81	82-86	87-91	92-95	57-95
					Total				
10-14	0.00	0.00	0.00	1.04	0.72	0.00	0.00	0.00	**0.21**
15-19	10.60	9.69	16.23	11.86	10.11	5.85	8.69	11.99	**9.91**
20-24	118.18	190.20	138.00	160.88	117.50	128.61	131.81	92.10	**126.01**
25-29	273.36	301.82	321.05	246.68	258.90	232.18	247.58	186.22	**243.23**
30-34	330.58	344.72	360.93	312.34	230.35	194.36	246.54	201.62	**254.13**
35-39	335.77	365.96	278.26	353.66	262.53	217.71	160.57	191.25	**243.23**
40-44	482.27	451.13	288.14	255.01	320.24	194.90	137.14	108.23	**228.91**
45-49	314.81	340.14	345.32	242.80	177.51	191.95	114.91	61.50	**180.56**
50-54	278.69	103.45	202.61	255.32	137.34	95.09	110.75	77.92	**132.96**
55-59	200.00	187.50	129.63	124.14	107.14	67.63	75.34	42.65	**91.07**
60-64	80.00	105.26	132.35	52.63	45.80	24.39	42.11	19.51	**46.60**
65-69	38.46	95.24	57.14	98.36	0.00	28.30	9.80	0.00	**29.07**
70-74	0.00	41.67	0.00	40.00	21.74	27.78	25.64	0.00	**19.11**
75-79	0.00	0.00	45.45	58.82	0.00	0.00	0.00	0.00	**10.87**
TFR (10-79)	12.31	12.68	11.58	11.07	8.45	7.04	6.55	4.96	**8.08**
TFR (15-49)	9.33	10.02	8.74	7.92	6.89	5.83	5.24	4.26	**6.43**
					Marital				
10-14	0.00	0.00	0.00	200.00	200.00	0.00	0.00	0.00	**80.00**
15-19	333.33	375.00	461.54	266.67	294.12	250.00	346.15	407.41	**340.74**
20-24	354.84	430.89	386.36	486.84	420.77	378.57	427.03	321.74	**395.87**
25-29	407.41	445.09	453.85	403.28	416.44	359.70	365.62	319.33	**379.56**
30-34	418.85	439.02	471.37	419.93	337.14	285.05	334.91	269.84	**346.60**
35-39	401.79	485.88	364.71	466.67	360.54	331.41	237.29	253.28	**335.80**
40-44	558.33	517.54	400.00	363.64	416.00	282.23	213.62	168.38	**325.54**
45-49	377.78	450.45	440.37	335.26	245.69	257.51	179.28	108.91	**263.27**
50-54	326.92	127.66	271.93	318.58	194.97	144.93	161.14	130.43	**192.27**
55-59	333.33	272.73	155.56	159.29	142.86	103.70	119.57	63.38	**132.58**
60-64	153.85	173.91	191.49	68.18	57.14	32.61	66.67	33.90	**69.40**
65-69	47.62	117.65	40.00	133.33	0.00	39.47	13.70	0.00	**37.94**
70-74	0.00	52.63	0.00	55.56	27.03	38.46	31.75	0.00	**25.53**
75-79	0.00	0.00	62.50	90.91	0.00	0.00	0.00	0.00	**15.50**
TFMR (10-79)	18.57	19.44	18.50	18.84	15.56	12.52	12.48	10.38	**14.70**
TFMR (15-49)	14.26	15.72	14.89	13.71	12.45	10.72	10.52	9.24	**11.94**

TABLE 39: FEMALE AGE-SPECIFIC FERTILITY RATES BY PERIOD

Age	57-61	62-66	67-71	72-76	77-81	82-86	87-91	92-95	57-95
					Total				
10-14	2.76	9.45	5.78	3.04	2.30	2.98	2.91	1.11	**3.08**
15-19	173.30	195.74	179.03	167.29	90.25	91.17	110.30	83.92	**117.58**
20-24	273.87	301.85	306.12	292.75	272.73	242.54	253.81	206.83	**256.92**
25-29	258.33	282.76	267.02	314.14	257.94	241.65	224.26	176.19	**240.17**
30-34	226.23	253.48	241.54	273.20	263.59	213.53	190.13	147.24	**215.26**
35-39	153.85	152.10	190.08	198.24	219.13	161.24	189.04	115.07	**172.06**
40-44	76.34	78.05	85.71	110.81	66.81	104.20	71.17	59.89	**81.12**
45-49	0.00	40.00	4.88	28.75	18.77	11.14	22.99	10.13	**17.49**
50-54	0.00	0.00	7.63	0.00	0.00	0.00	0.00	0.00	**0.50**
TFR (10-54)	5.82	6.57	6.44	6.94	5.96	5.34	5.32	4.00	**5.52**
TFR (15-49)	5.81	6.52	6.37	6.93	5.95	5.33	5.31	4.00	**5.50**
					Marital				
10-14	0.00	111.11	250.00	166.67	181.82	230.77	266.67	0.00	**177.22**
15-19	322.75	396.45	381.97	402.99	352.66	343.07	417.53	320.24	**369.11**
20-24	336.48	365.16	354.17	347.22	337.08	318.63	327.15	301.57	**330.13**
25-29	294.87	339.83	320.43	355.90	297.34	280.00	275.29	214.74	**285.57**
30-34	272.36	298.66	291.78	338.50	298.50	246.03	228.72	190.35	**259.05**
35-39	186.75	187.76	227.12	246.54	272.32	187.11	218.70	136.94	**206.98**
40-44	128.21	106.67	103.31	144.88	84.03	130.12	77.46	68.40	**100.21**
45-49	0.00	68.49	6.90	38.63	21.98	14.71	26.95	13.75	**22.82**
50-54	0.00	0.00	13.33	0.00	0.00	0.00	0.00	0.00	**0.81**
TFMR (10-54)	7.71	9.37	9.75	10.21	9.23	8.75	9.19	6.23	**8.76**
TFMR (15-49)	7.71	8.82	8.43	9.37	8.32	7.60	7.86	6.23	**7.87**

To address the fact that the classic fertility measures do not properly recognize the link between males and females and therefore produce irreconcilable measures, another model is calculated that estimates a married pair's annual hazard of experiencing a birth as a function of male age[29], female age, and duration of the union. The model is another variation on the two-sex dummy variable models that were estimated to calculate the nuptiality hazards, Equation 14.

[29] As elsewhere in this document, male fertility recognizes "assigned" paternity rather than absolute biological paternity which is something about which I have no information.

$$\ln\left(\frac{p(a_i^m, a_j^f, t_k)}{1-p(a_i^m, a_j^f, t_k)}\right) =$$
$$\beta_1 a_1^m + \cdots + \beta_n a_n^m +$$
$$\beta_{n+1} a_{n+1}^f + \cdots + \beta_{2n} a_{2n}^f +$$
$$\beta_{2n+1} a_1^m a_1^f + \cdots + \beta_{2n+n^2} a_n^m a_n^f$$
$$\beta_{2n+n^2+1} t_1 + \cdots + \beta_{2n+n^2+m} t_m + c$$

Where p is the probability that the union will experience a birth, a^m is a dummy for the male's age, a^f is a dummy for the female's age, t is a dummy for the duration of the union, n is the maximum number of age groups for male and female, and m is the maximum number of duration-groups for the union. There are no duration/age interactions in this model (because there are not enough observations to support that many cells) so the age and duration effects are assumed to be completely independent.

Equation 14: Specification for Two-Sex Hazard of Experiencing a Birth Model

Figure 52 displays the hazards estimated by applying the model described by Equation 14 to the Union-Years Data Set. This figure clearly reveals where in the male-female age matrix reproduction is occurring. The maximum ridge in the hazard surface runs obliquely from young to old ages in such as way that the maximum likelihood of reproduction is occurring within pairs whose male is substantially older than the female. As it should be the likelihood of experiencing a birth is effectively zero for women older than 50. In contrast, men continue to have substantial likelihoods of experiencing a birth through age 64. A revealing feature of this surface is the age associativity that is present. For the most part hazards are substantial where the age difference between the male and female is about 5 to 10 years. However, women up to 50 years old face considerable likelihoods of experiencing a birth with men as young as 30 but not younger, and men aged 60 to 64 are not exposed to substantial likelihoods of experiencing a birth with women less than 40 years old. There is a much wider spread in the age of men with whom women reproduce at significant levels, and a somewhat narrower range of ages of women with whom men reproduce - especially as the man age. This is partly

a result of the shorter reproductive life of women, and partly due to the social processes governing marriage.

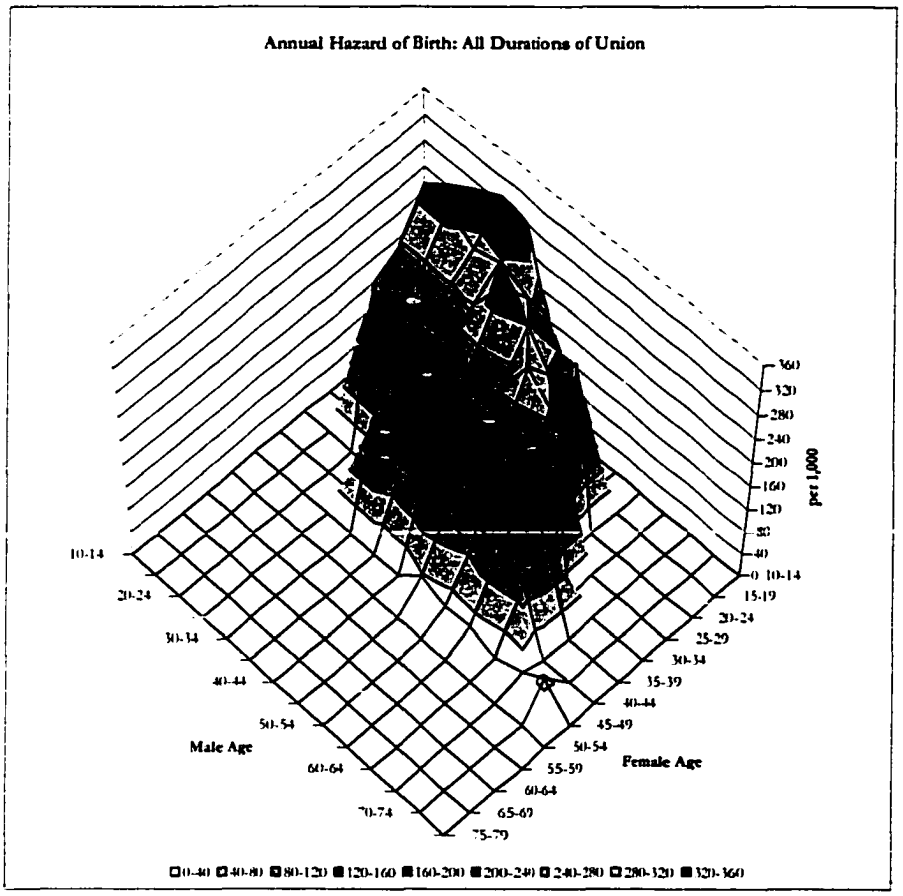

Figure 51: Annual Hazard of Birth for Pairs: All Duration of Union

Two more interesting results emerge from this model. The first answers the question of how the likelihood of experiencing a birth changes as a union ages. The likelihood of experiencing

a birth as a function of the age of union is plotted in Figure 52 as the odds ratio of the annual hazard at various durations after union formation compared to the likelihood of experiencing a birth during the first year of the union. Age effects are controlled for by the age dummies so what we see is a true duration-of-union effect with potential bias from period effects that are not controlled for in this model. The likelihood of experiencing a birth increases dramatically after the first year and more-or-less plateaus for the next twenty years of a union's life before dropping sharply.

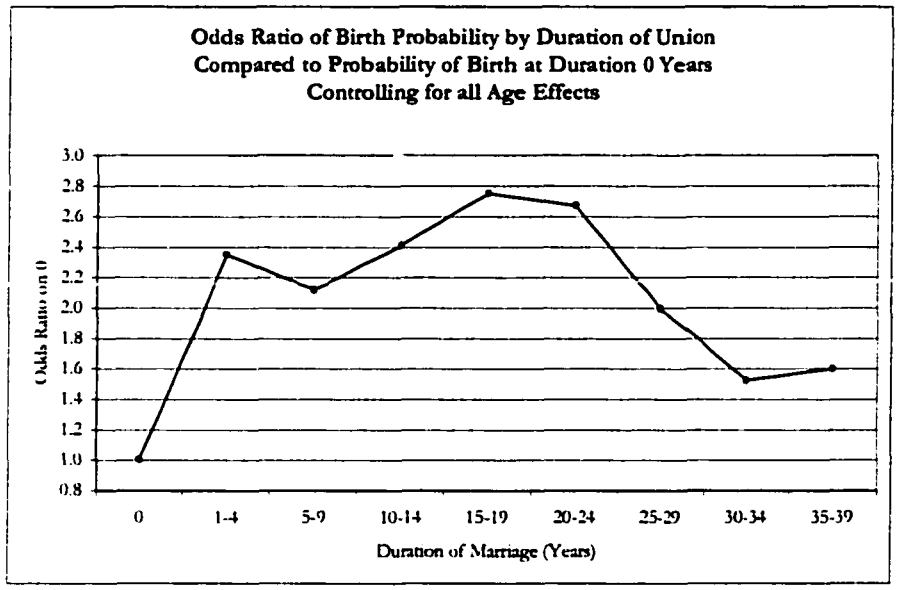

Figure 52: Odds Ratio of Birth Probability by Duration of Union

We may also ask how the likelihood of experiencing a birth is affected by the age of one's spouse? To answer this, the hazards produced by estimating Equation 14 can be manipulated to control for the age of one sex or the other to produce a novel measure similar to a total fertility rate. For each sex, the age-specific hazards of experiencing a birth can be summed

across all ages of the opposite sex to yield a sex-specific total fertility rate by age of the opposite sex. For example, if we take males aged 30 to 34 and sum the hazards of

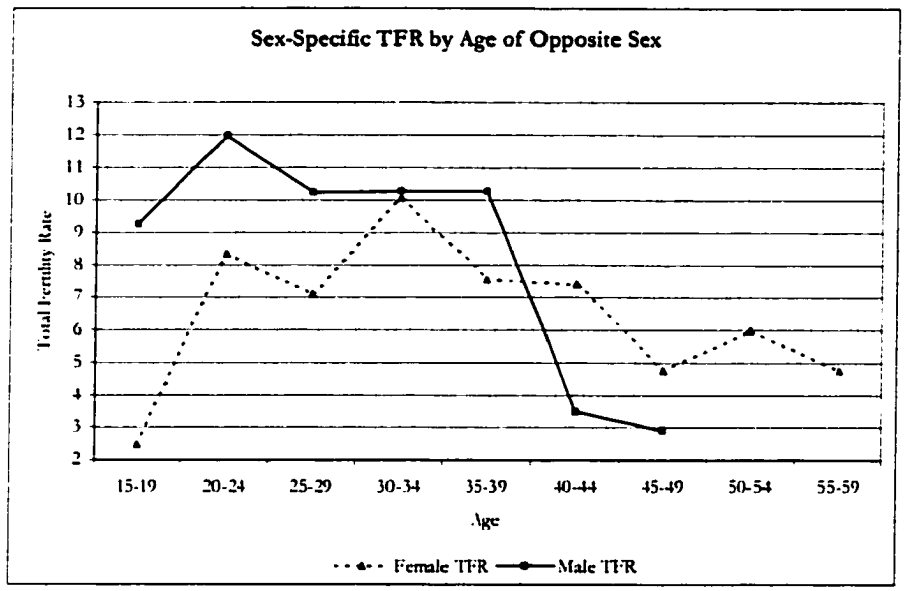

Figure 53: Sex-Specific TFR by Age of Opposite Sex

experiencing a birth across females of all ages, the result is a total fertility rate for an average female of any age paired with male who is between 30 and 34 - the number of children that a woman could expect to have if she survived through her entire reproductive life paired with a man aged 30 to 34 and experienced at all ages the fertility rates associated with being paired to a man aged 30 to 34. If we repeat this for all age groups of males, a series of total fertility rates is produced that reveals how females' reproductive potential changes with males' age. Total fertility rates calculated in this way are plotted in Figure 53.

The pattern of male total fertility rates as a function of female age is what we expect. Males fertility is high with women aged 15 to 39, low with women aged 40 to 49, and zero with women aged 50 and over. The interesting finding is how female fertility changes with male age. Females are reproductive with men from age 15 to 64 with a maximum reproductive potential when paired with men aged 30 to 34. The shape of the female curve indicates that males' reproductive potential increases more slowly than females' and diminishes more slowly over a wider range of ages. The total fertility of women paired with 55 to 59 year old men is still 5!

TRENDS

Since 1957 the fertility of the Gwembe Tonga has changed significantly. Unfortunately the data do not contain much of the information that is necessary to explain these changes; there is no systematic individual-level information on contraceptive use, attitudes toward fertility and family limitation, socio-economic standing, women's status or any of the other factors usually assumed to be proximate determinants of fertility. In the absence of those data, the best that can be done is to compare the trends in fertility to the known history of the Gwembe Valley over the same time period.

What is known from the ethnographic work of Colson, Scudder, Cliggett and Gillet-Netting is that there has never been widespread use of modern methods of contraception at any time during the period 1957 to 1995. Very small numbers of women have experimented with modern methods at various points in time, but continuous use has been prevented by the lack of continuous availability through the public health system, the prohibitive cost of acquiring

contraceptives on the open market and the resistance to family limitation that women face from their husbands. This means that whatever changes in fertility have taken place are largely the result of changes in behavior relating to marriage and sex, to a changing epidemiological profile, and to macro demographic changes in the population. An exhaustive analysis of what those changes may have been and what brought them about is largely the subject of a different investigation. Here it suffices to reveal that they have taken place and to correlate them to the global changes that have taken place in the economy, the environment, and the epidemiological profile of the population during the same period.

This work is primarily concerned with understanding the HIV/AIDS pandemic in Africa from the perspective of a population scientist. Toward that end, there is one aspect of the trends in fertility that is examined in detail: namely, has there been a measurable falloff in fertility during the most recent periods when HIV/AIDS has begun to affect the population, and if so are the largest reductions observed for age groups that would be most affected by HIV/AIDS? In most significantly affected African populations, HIV infects the male and female populations so that the age-pattern of infection is significantly younger for females. There are no HIV prevalence figures available for the Gwembe, but if we assume that the epidemic there is similar to the epidemic in other polygynous areas of Africa, then we would expect the age-pattern of the decline in fertility to be slightly younger for females than for males and to match closely the ages at which each sex is typically most affected by the epidemic. Reductions of this type are indeed revealed by the data, and they are measured with a great deal of certainty. This is still a flimsy case at best - open to many other explanations. However, it is given some support by the congruent finding the mortality has increased substantially in the same age

groups for both men and women over the same time, again measured with great certainty. There are few explanations that are able to account for both the reductions in fertility and the increases in mortality as well as HIV/AIDS.

Two methods are used to investigate the trends in fertility, and both are (surprise!) event history formulations. The first was outlined above during the discussion of age-specific fertility and total fertility rates, Equation 11 and Equation 12. Age-specific fertility rates were calculated for each of the five-year period between 1957 and 1995 for both men and women, and total fertility rates were calculated from those. Figure 54 contains the series of male total fertility rates, and Figure 56 contains the series of total fertility rates for females.

The drawback to this method is that it is not possible to associate standard errors with the differences in the total fertility rates from one period to the next so it is not possible to say with certainty that the change from one period to the next is not the result of a random process. The best way to solve this would be to interact the period and age variables and use the standard errors on the period variables and the interactions as indicators of the likelihood that the observed changes were not the result of random processes. This method was attempted but failed due to the lack of sufficient observations. The final number of variables included in a dummy variable regression with fourteen age groups for each sex and eight period variables and all of the appropriate interactions requires a great number of observations and events – a number not reached by the relatively small size of the Gwembe data set.

Instead the data are divided into age groups through time, and for each age group the likelihood of experiencing a birth is modeled as a function of time period. For example all

female person-years lived between the ages of 20 and 24 over the entire period 1957 to 1995 are selected into a sample, and the likelihood of experiencing a birth for those person years is modeled as a function of the time period in which the person-years were lived. The model is implemented as a dummy variable logistic regression:

$$\ln\left(\frac{p(t_i)}{1-p(t_i)}\right)=\beta_1 t_1+\cdots+\beta_7 t_7+c$$

Where p is the probability of experiencing a birth and t_i is a time period dummy variable. This model is run separately for males and female on samples of the Person-Year Data Set defined by age.

Equation 15: Specification for the Fertility Trend Hazard Model

With this specification it is possible to test the null hypothesis that two consecutive period coefficients are equal. If P-value produced by this test indicates the likelihood that the difference in the two coefficients is the result of a random process. If the P-value is significant, we can state with some certainty that there was a real change measured with these data, and furthermore if we trust the data, we can be confident that a real change took place in the population.

Table 38 and Table 39 (above) contain the results of the age-specific fertility rate calculations and the series of total fertility rates that they produce. Those total fertility rates are plotted in Figure 54 and Figure 56. It is immediately evident that there has been a decline in fertility from 1972 onward. The decline is steady and substantial for both males and females; from a total marital fertility rate of about nineteen for males to about eleven and from about ten for females to about six. Movements in the total and marital rates parallel each other very closely during all periods indicating that changes in marriage patterns probably do not explain the

changes in fertility. The fertility of married people declines in the same manner as the fertility of the whole population demonstrating that reproductive behavior within marriage is similar to reproductive behavior as a whole.

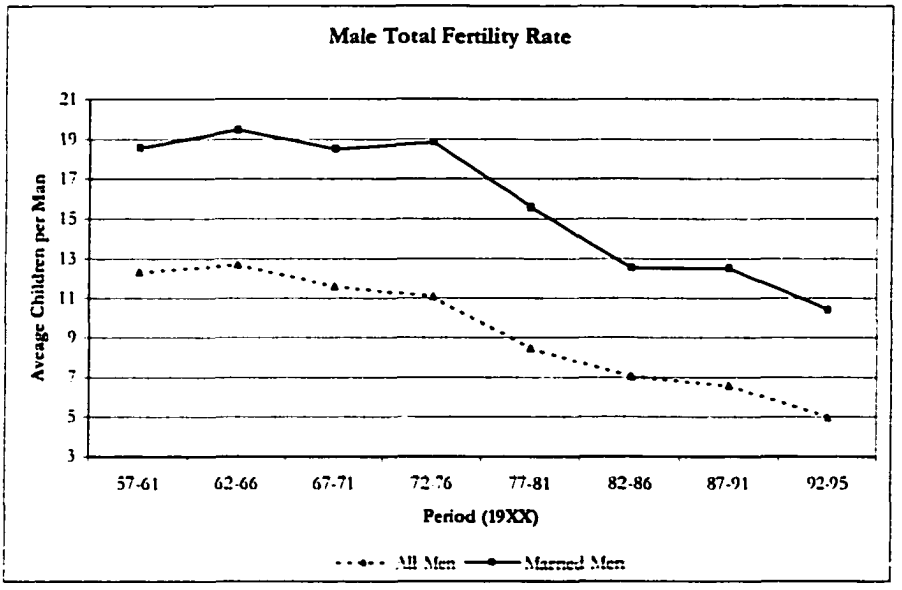

Figure 54: Trend in Male Total Fertility Rates

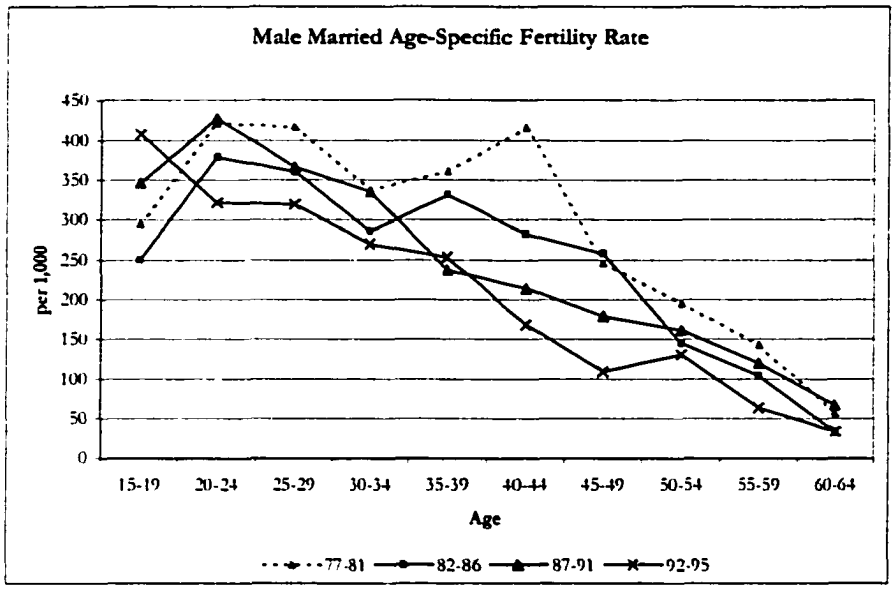

Figure 55: Male Married Age-Specific Fertility Rates by Period

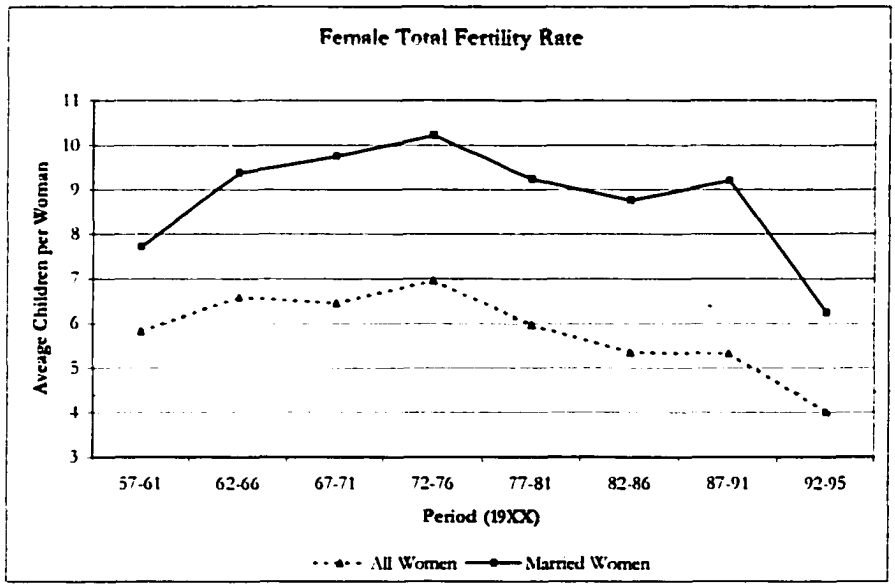

Figure 56: Trend in Female Total Fertility Rates

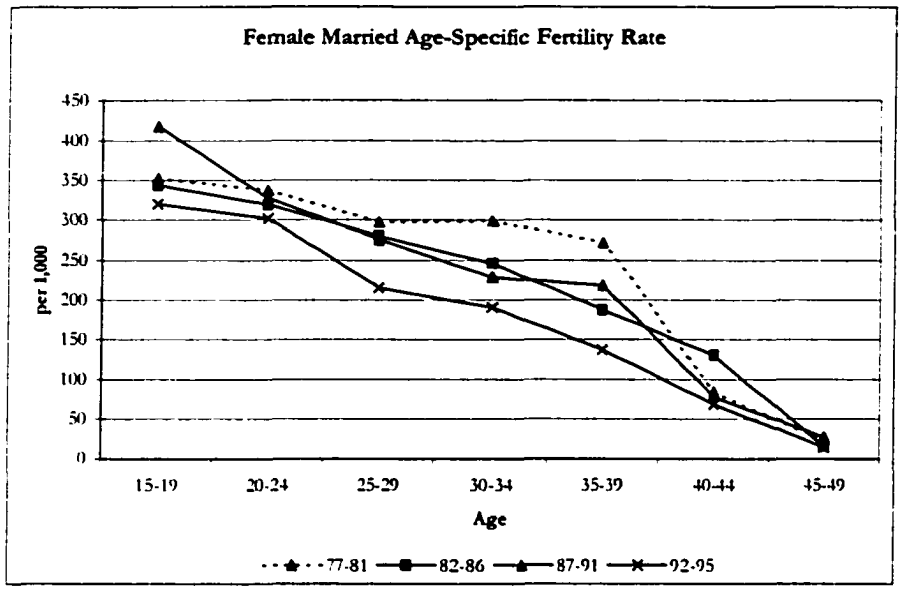

Figure 57: Female Married Age-Specific Fertility Rates by Period

The question becomes what may have driven these changes and through what proximate determinant, or combination of proximate determinants, were they actually effected? As mentioned above, the data necessary to answer those questions is not presently available. However, the slight increase prior to 1972 occurred during the most auspicious period of Gwembe history recorded in these data. During that time the Zambian economy as a whole was healthy, and the local economy was booming fueled by the great productivity of the lake Kariba fisheries while the lake consumed the initial bolus of biomass provided by the flooding. The nearly concurrent collapse of the price of copper on the world market and the oil crisis during the mid seventies contributed to the initiation of a permanent downturn in the Zambian economy, and once the initial fertility of the lake was tapped out the local economy also turned down. During the eighties and early nineties the situation was compounded by a series of serious droughts. Together these influences provided the kind of macro environment

in which fertility is perhaps more expensive and less rewarding in all ways. Precisely how these influences may have exerted downward pressure on fertility is not known, although it is certain that the provision of medical services deteriorated, the average income of individuals and especially households decreased substantially, and the overall health and nutrition level of the population suffered considerably. The only exception to this generally unsatisfying picture is the sharp downturn in fertility of both sexes during the 1992 to 1995 period. It is likely that this decrease is associated with HIV/AIDS.

EVIDENCE OF HIV

If the Tonga were reducing fertility to limit the size of their families, we would expect fertility to remain high at younger ages and start to drop first at older ages, and as time goes by, for the reduction to appear at progressively younger ages until there was a relatively narrow and young age band in which all fertility was taking place. Alternatively if they were using contraceptive techniques to increase their control over the spacing of births, we would expect fertility to decrease mildly at all ages with no particular dependence on age. Last if HIV/AIDS is affecting fertility, we expect fertility to remain more or less constant at all ages except those at which HIV is prevalent. Assuming that the HIV epidemic in the Gwembe is similar to that measured elsewhere, prevalence for females is highest in the twenties and thirties and for males in the thirties and forties. Support for this conclusion comes from an examination of the trend in fertility by age. Figure 55 and Figure 57 present married age-specific fertility rates for males and females for the periods 1977-1981, 1982-1986, 1987-1991 and 1992-1995.

It is clear that there is no substantial trend in male fertility outside of the 35 to 49 age group, and that between those ages there has been a steady and substantial decrease in fertility, especially in the 40 to 44 age group that is most likely to be affected by HIV infection. For females the picture is similar except younger. There is no real change in female fertility outside of the 25 to 39 age group, and between those ages fertility has fallen consistently.

This is a nice finding, but it could still be a serendipitous confluence of stochastic changes in age-specific fertility. To rule out that possibility, the results of the statistical analysis of the trends (described above) are presented in Table 41 and Table 43 and plotted in Figure 58 through Figure 74 for males and females respectively.

The tables list the annual hazards of experiencing a birth during each period, the odds ratio of the period hazards compared to the most recent period 1992 to 1995, and the P-value generated from comparing each pair of consecutive period coefficients. A small P-value indicates that it is unlikely that the difference observed in the annual hazard from one period to the next is the result of a random process. The figures display the trend in the annual hazard of birth for each age group. All of the numbers are calculated for both males and females and for both the total and married populations of males and females.

The figures clearly reveal that the likelihood of experiencing a birth has decreased substantially during the most recent two periods from 1987 to 1995 for men aged 40 to 49, and that men aged 25 to 34 have also experienced a sharp reduction in the likelihood of experiencing a birth during the most recent period from 1992 to 1995. Men aged 50 to 59 also experienced a noticeable reduction during 1992 to 1995. The anomalous age group is aged 35 to 39 for

which men appear to have experienced a sharp decline from 1972 to 1992 and a leveling off between 1992 and 1995. To determine which of these changes is statistically significant we examine the P-values in Table 41. Table 40 summarizes the P-values for the trends in male annual hazard of birth. The changes from the 1987 to 1991 period to the 1992 to 1995 period are highly significant in the 25 to 29, 35 to 39, 40 to 44 and 45 to 49 age groups. The trend in the 30 to 34 and 40 to 44 age groups has been statistically significant for some time.

TABLE 40: SIGNIFICANCE LEVELS OF MALE TREND IN ANNUAL HAZARD OF BIRTH 1977-1995

	Significance Level (Percent)					
	77-81 → 82-86		82-86 → 87-91		87-91 → 92-95	
Age Group	Total	Married	Total	Married	Total	Married
25-29	40%	18%	54%	88%	Better than 1%	2%
30-34	5%	5%	Better than 1%	3%	8%	4%
35-39	23%	77%	23%	10%	32%	84%
40-44	2%	9%	5%	11%	11%	11%
45-49	88%	75%	1%	6%	1%	4%
50-54	10%	15%	33%	38%	21%	42%
55-59	44%	66%	40%	46%	96%	90%

The statistical tests provide confidence in the fact that the observed changes are not the result of a random process and provide confidence in the fact that the sets of person years from which the estimates are made are large enough to measure changes of this magnitude. Consequently, the changes at ages 40 to 44 and 45 to 49 are both substantial and significant.

TABLE 41: MALE TREND IN ANNUAL HAZARD OF BIRTH BY AGE

Period	i	Annual Hazard* Total	Annual Hazard* Married	OR** Total	OR** Married	P-Value** Total	P-Value** Married	Annual Hazard* Total	Annual Hazard* Married	OR** Total	OR** Married	P-Value** Total	P-Value** Married
		Ages 10-14						Ages 15-19					
57-61	1	0.0000	0.0000	3.1E-07	-NA-	1.0000	-NA-	0.0000	0.3333	0.88	0.73	0.3904	0.3948
62-66	2	0.0000	0.0000	3.1E-07	-NA-	1.0000	-NA-	0.0097	0.3750	0.40	0.24	0.1247	0.1777
67-71	3	0.0000	0.0000	3.1E-07	-NA-	1.0000	-NA-	0.0162	0.4615	1.36	1.25	0.2731	0.2877
72-76	4	0.0010	0.2000	1.1E+10	1.2E+10	0.7962	1.0000	0.0119	0.2667	0.77	0.53	0.8533	0.8632
77-81	5	0.0007	0.2000	7.3E+09	1.2E+10	1.0000	1.0000	0.0101	0.2941	0.84	0.61	0.2465	0.7635
82-86	6	0.0000	0.0000	3.1E-07	8.3E-07	1.0000	1.0000	0.0058	0.2500	0.48	0.48	0.3591	0.4838
87-91	7	0.0000	0.0000	3.1E-07	8.3E-07	1.0000	1.0000	0.0087	0.3462	0.72	0.77	0.3520	0.6460
92-95	8	0.0000	0.0000	1.00	1.00	-X-	-X-	0.0120	0.4074	1.00	1.00	-X-	-X-
		Ages 20-24						Ages 25-29					
57-61	1	0.1182	0.3548	1.22	0.99	0.0116	0.1702	0.2734	0.4074	1.61	1.44	0.4038	0.4444
62-66	2	0.1902	0.4309	2.38	1.57	0.0224	0.2599	0.3018	0.4451	1.90	1.70	0.8853	0.6683
67-71	3	0.1380	0.3864	1.51	1.17	0.1070	0.0177	0.3211	0.4538	1.85	1.56	0.1552	0.7808
72-76	4	0.1609	0.4868	2.03	2.11	0.0027	0.0352	0.2467	0.4033	1.48	1.48	0.6152	0.6934
77-81	5	0.1175	0.4208	1.24	1.31	0.1109	0.8697	0.2589	0.4164	1.59	1.58	0.3955	0.1830
82-86	6	0.1286	0.3786	1.57	1.35	0.7500	0.5224	0.2322	0.3597	1.42	1.30	0.5361	0.8769
87-91	7	0.1318	0.4270	1.51	1.50	0.0010	0.0110	0.2476	0.3656	1.52	1.33	0.0000	0.0200
92-95	8	0.0921	0.3217	1.00	1.00	-X-	-X-	0.1862	0.3193	1.00	1.00	-X-	-X-
		Ages 30-34						Ages 35-39					
57-61	1	0.3306	0.4188	1.39	1.35	0.1350	0.1278	0.3358	0.4018	1.52	1.37	0.7876	0.3526
62-66	2	0.3447	0.4390	1.87	1.86	0.4197	0.5621	0.3660	0.4859	1.64	1.77	0.5589	0.4020
67-71	3	0.3609	0.4714	1.61	1.66	0.4517	0.5269	0.2783	0.3647	1.45	1.47	0.1396	0.1113
72-76	4	0.3123	0.4199	1.41	1.47	0.1654	0.4670	0.3537	0.4667	1.91	2.02	0.0259	0.0430
77-81	5	0.2303	0.3371	1.13	1.29	0.0494	0.0501	0.2625	0.3605	1.28	1.37	0.2293	0.7676
82-86	6	0.1944	0.2850	0.84	0.94	0.0047	0.0262	0.2177	0.3314	1.04	1.29	0.2508	0.0957
87-91	7	0.2465	0.3349	1.23	1.29	0.0810	0.0430	0.1606	0.2373	0.85	0.97	0.3220	0.8390
92-95	8	0.2016	0.2698	1.00	1.00	-X-	-X-	0.1912	0.2533	1.00	1.00	-X-	-X-
		Ages 40-44						Ages 45-49					
57-61	1	0.4823	0.5583	5.64	4.38	0.0146	0.0105	-NA-	-NA-	6.10	4.35	0.5731	0.3432
62-66	2	0.4511	0.5175	2.73	1.97	0.9699	0.3217	0.3401	0.4505	7.61	6.40	0.8823	0.9630
67-71	3	0.2881	0.4000	2.76	2.66	0.6076	0.6845	0.3453	0.4404	7.92	6.31	0.2759	0.4485
72-76	4	0.2550	0.3636	2.46	2.42	0.1768	0.4652	0.2428	0.3353	6.02	5.15	0.0771	0.0662
77-81	5	0.3202	0.4160	3.21	2.81	0.0215	0.0925	0.1775	0.2457	4.09	3.36	0.8800	0.7475
82-86	6	0.1949	0.2822	2.08	2.00	0.0458	0.1100	0.1920	0.2575	3.96	3.12	0.0104	0.0632
87-91	7	0.1371	0.2136	1.41	1.44	0.1130	0.1100	0.1149	0.1793	2.17	1.97	0.0120	0.0380
92-95	8	0.1082	0.1684	1.00	1.00	-X-	-X-	0.0615	0.1089	1.00	1.00	-X-	-X-
		Ages 50-54						Ages 55-59					
57-61	1	0.2787	0.3269	5.31	3.77	0.0130	0.0150	0.2000	0.3333	3.30	4.10	0.6728	0.8886
62-66	2	0.1034	0.1277	1.20	0.86	0.0834	0.0505	0.1875	0.2727	4.23	4.48	0.3081	0.1494
67-71	3	0.2026	0.2719	3.19	2.63	0.7029	0.9122	0.1296	0.1556	2.34	1.89	0.9403	0.8308
72-76	4	0.2553	0.3186	3.59	2.73	0.0343	0.0883	0.1241	0.1593	2.44	2.13	0.5929	0.6604
77-81	5	0.1373	0.1950	1.87	1.58	0.0973	0.1530	0.1071	0.1429	1.94	1.77	0.4366	0.6570
82-86	6	0.0951	0.1449	1.11	0.99	0.3307	0.3861	0.0676	0.1037	1.39	1.45	0.4049	0.4603
87-91	7	0.1107	0.1611	1.50	1.30	0.2050	0.4190	0.0753	0.1196	0.98	1.06	0.9630	0.8950
92-95	8	0.0779	0.1304	1.00	1.00	-X-	-X-	0.0427	0.0634	1.00	1.00	-X-	-X-

TABLE 41 CONTINUED: MALE TREND IN ANNUAL HAZARD OF BIRTH BY AGE

Period	t	Annual Hazard*		OR**		P-Value**		Annual Hazard*		OR**		P-Value**	
		Total	Married	Total	Married	Total	Married	Total	Married	Total	Married	Total	Married
		Ages 60-64						Ages 65-69					
57-61	1	0.0800	0.1538	4.37	5.18	0.7390	0.8767	0.0385	0.0476	1.4E+10	1.3E+10	0.4433	0.4405
62-66	2	0.1053	0.1739	5.91	6.00	0.6841	0.8591	0.0952	0.1176	3.6E+10	3.5E+10	0.5958	0.3591
67-71	3	0.1324	0.1915	7.67	6.75	0.1448	0.0951	0.0571	0.0400	2.1E+10	1.1E+10	0.4870	0.2396
72-76	4	0.0526	0.0682	2.79	2.09	0.8403	0.7966	0.0984	0.1333	3.7E+10	4.0E+10	0.0000	0.0000
77-81	5	0.0458	0.0571	2.41	1.73	0.3640	0.4165	0.0000	0.0000	7.6E-10	7.6E-10	0.0000	0.0000
82-86	6	0.0244	0.0326	1.26	0.96	0.9102	0.7234	0.0283	0.0395	9.9E+09	1.1E+10	0.3536	0.3524
87-91	7	0.0263	0.0417	1.37	1.25	0.6460	0.7440	0.0098	0.0137	3.4E+09	3.7E+09	0.0000	0.0000
92-95	8	0.0195	0.0339	1.00	1.00	-X-	-X-	0.0000	0.0000	1.00	1.00	-X-	-X-
		Ages 70-74						Ages 75-79					
57-61	1	0.0000	0.0000	1.7E-05	1.1E-07	1.0000	1.0000	0.0000	0.0000	6.1E-06	1.7E-05	1.0000	1.0000
62-66	2	0.0417	0.0526	8.3E+09	7.9E+09	1.0000	1.0000	0.0000	0.0000	6.1E-06	1.7E-05	1.0000	1.0000
67-71	3	0.0000	0.0000	1.7E-05	1.1E-07	1.0000	1.0000	0.0455	0.0625	9.6E+09	9.4E+09	0.8515	0.7830
72-76	4	0.0400	0.0556	7.9E+09	8.4E+09	0.6617	0.6035	0.0588	0.0909	1.3E+10	1.4E+10	1.0000	1.0000
77-81	5	0.0217	0.0270	4.2E+09	4.0E+09	0.8607	0.7998	0.0000	0.0000	6.1E-06	1.7E-05	1.0000	1.0000
82-86	6	0.0278	0.0385	5.4E+09	5.7E+09	0.9472	0.8734	0.0000	0.0000	6.1E-06	1.7E-05	1.0000	1.0000
87-91	7	0.0256	0.0317	5.0E+09	4.7E+09	0.0000	0.0000	0.0000	0.0000	6.1E-06	1.7E-05	1.0000	1.0000
92-95	8	0.0000	0.0000	1.00	1.00	-X-	-X-	0.0000	0.0000	1.00	1.00	-X-	-X-

* The annual hazards listed in this table are calculated as the parity-weighted average of the annual hazard of 1, 2, or three births defined in Equation 13. These should be interpreted as the number of births an average male of a given age could expect to have in a give period.

** The odds ratios and P-values listed in this table refer to the coefficients that predict the annual hazard of a single birth. Consequently, the odds ratios are close to but not exactly equal to what would result from calculating them from the annual hazards that are listed.

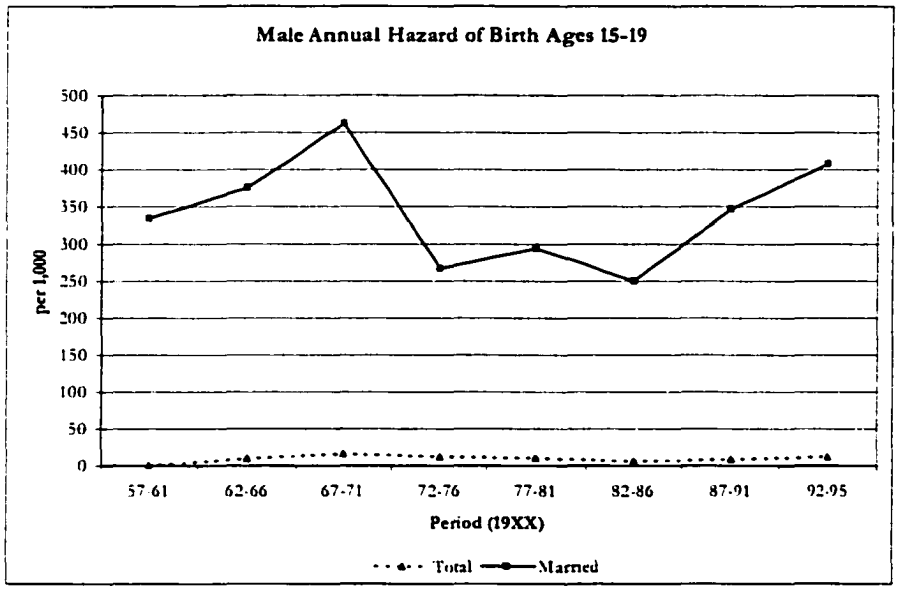

Figure 58: Male Annual Hazard of Birth Ages 15-19

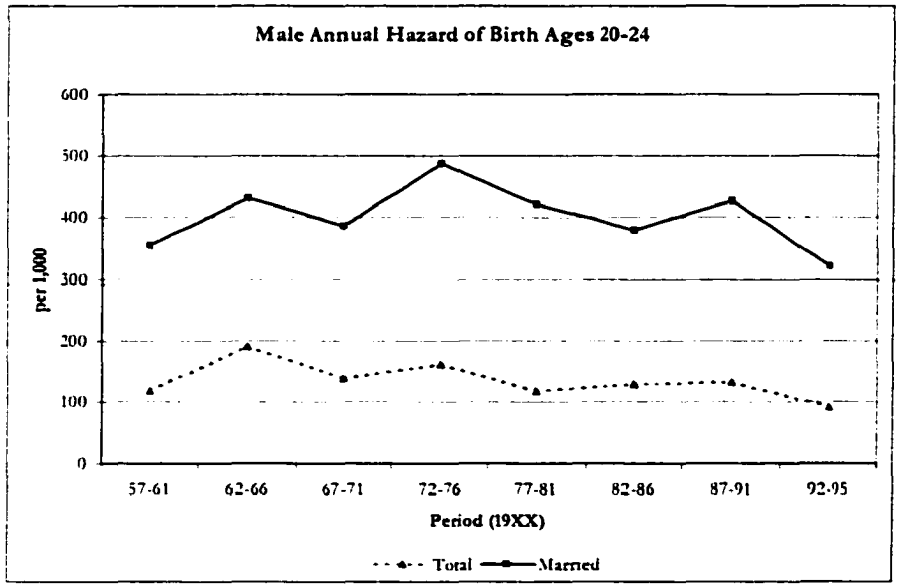

Figure 59: Male Annual Hazard of Birth Ages 20-24

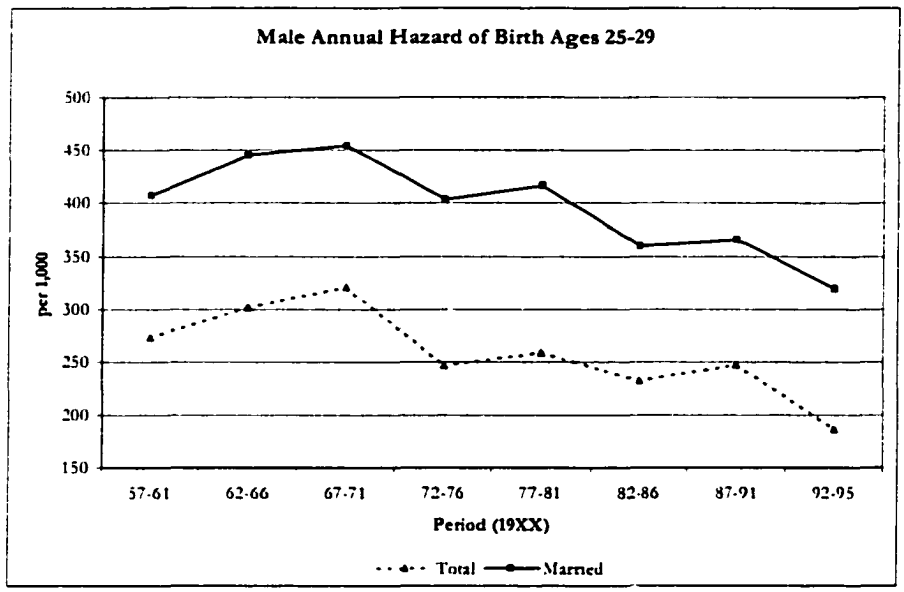

Figure 60: Male Annual Hazard of Birth Ages 25-29

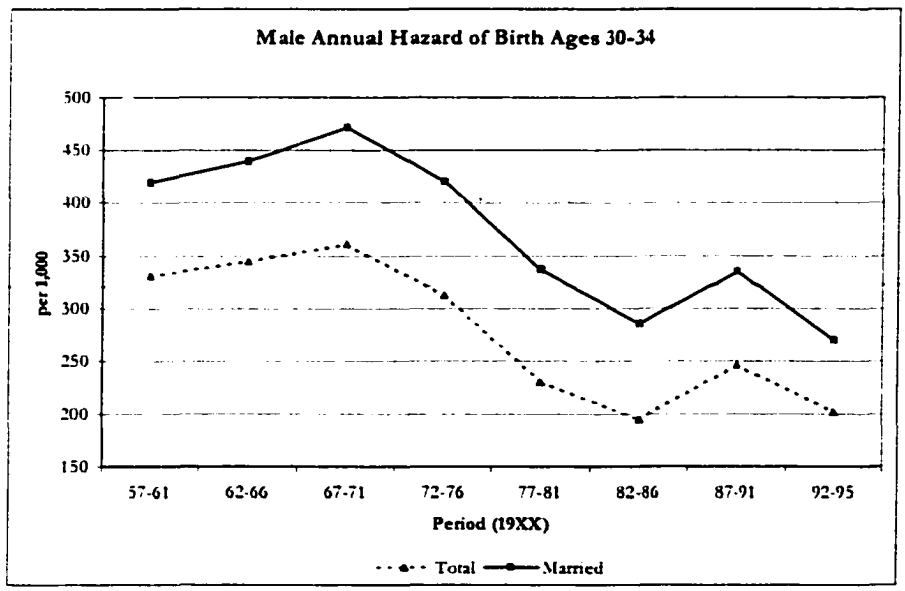

Figure 61: Male Annual Hazard of Birth Ages 30-34

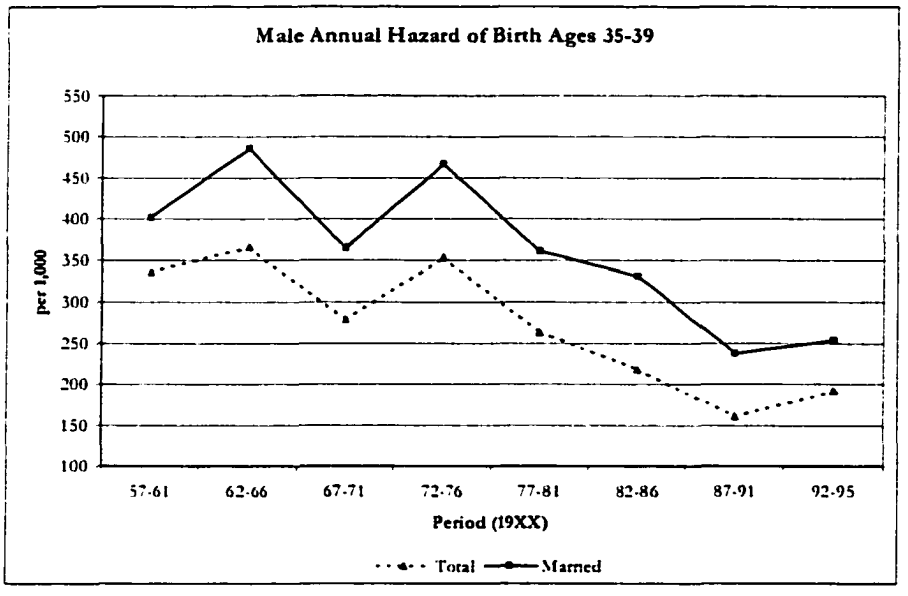

Figure 62: Male Annual Hazard of Birth Ages 35-39

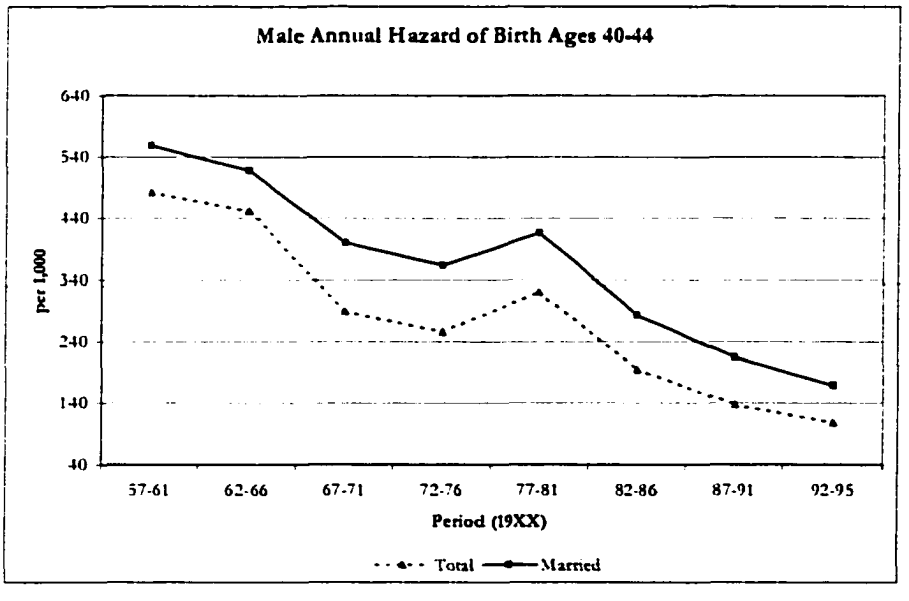

Figure 63: Male Annual Hazard of Birth Ages 40-44

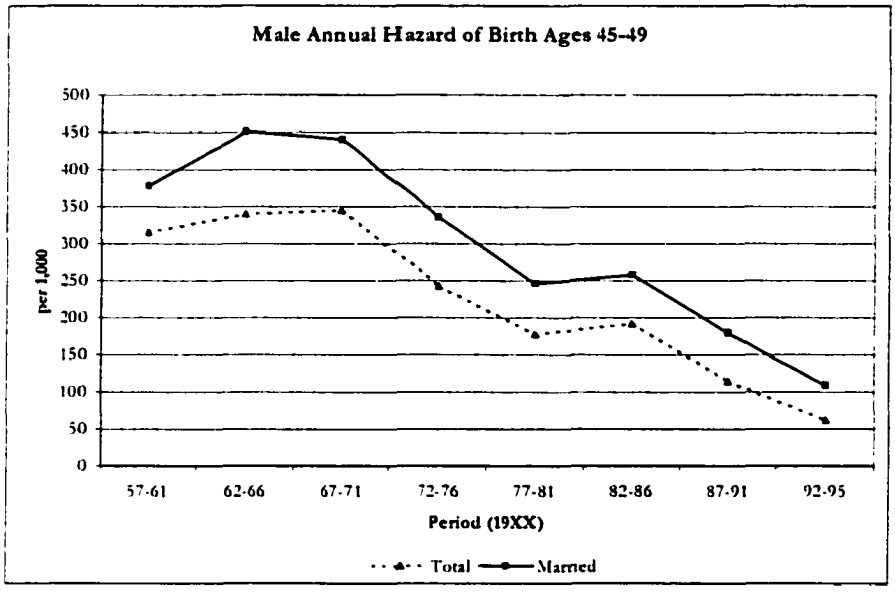

Figure 64: Male Annual Hazard of Birth Ages 45-49

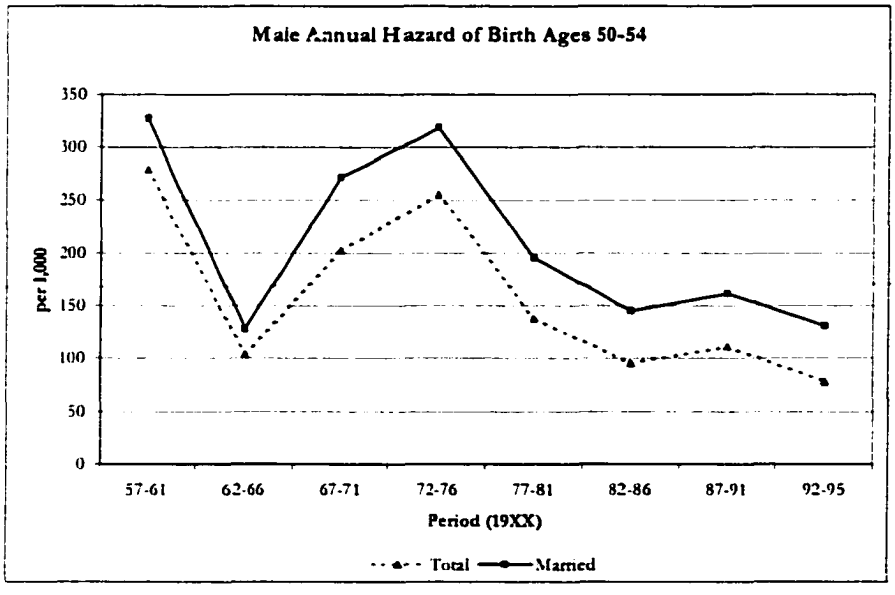

Figure 65: Male Annual Hazard of Birth Ages 50-54

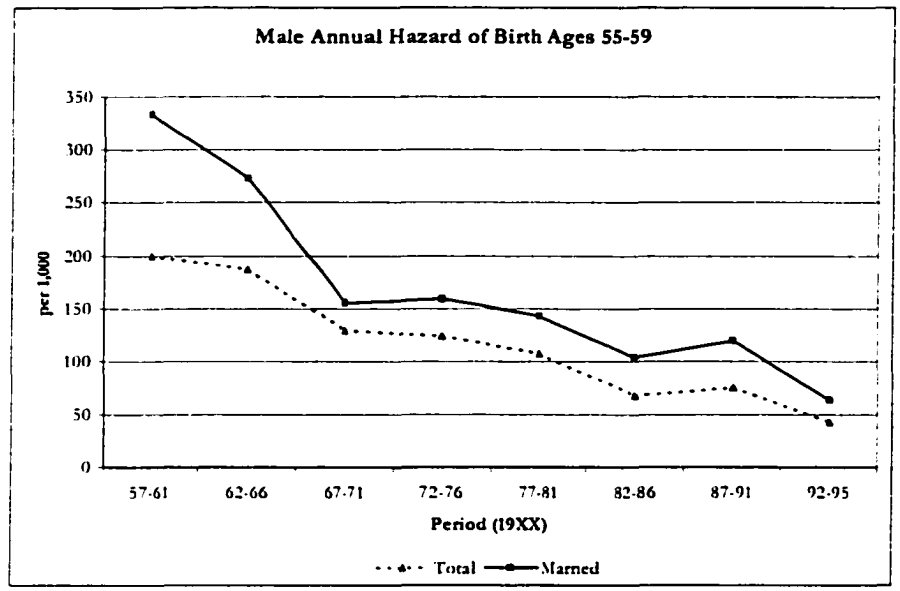

Figure 66: Male Annual Hazard of Birth Ages 55-59

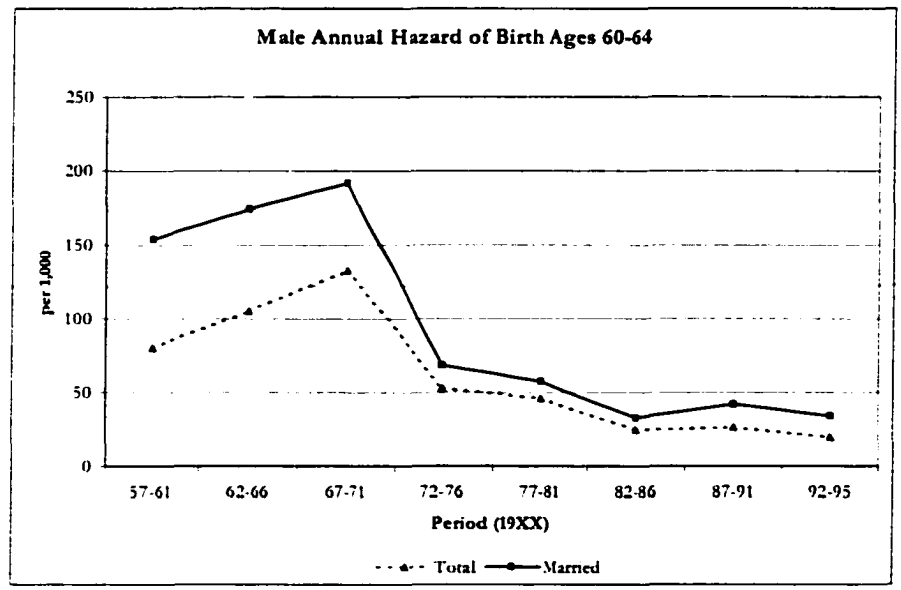

Figure 67: Male Annual Hazard of Birth Ages 60-64

A similar examination of the changes in the female likelihood of experiencing a birth is presented in Table 42, Table 43 and Figure 68 through Figure 74. The main finding is that there are substantial and highly significant reductions in the likelihood that females aged 20 to 39 will experience a birth between the 1987 to 1991 period and the 1992 to 1995 period. The reductions are of particularly large magnitude in the 25 to 39 age groups and particularly in the 35 to 39 age group.

TABLE 42: SIGNIFICANCE LEVELS OF FEMALE TREND IN ANNUAL HAZARD OF BIRTH 1977-1995

	Significance Level (Percent)					
	77-81 → 82-86		82-86 → 87-91		87-91 → 92-95	
Age Group	Total	Married	Total	Married	Total	Married
20-24	10%	43%	48%	68%	Better than 1%	7%
25-29	43%	47%	32%	82%	Better than 1%	Better than 1%
30-34	3%	5%	23%	45%	1%	7%
35-39	1%	Better than 1%	19%	20%	Better than 1%	Better than 1%
40-44	4%	4%	5%	1%	45%	61%
45-49	37%	50%	17%	26%	15%	25%

TABLE 43: FEMALE TREND IN ANNUAL HAZARD OF BIRTH BY AGE

Period	i	Annual Hazard		OR		P-Value		Annual Hazard		OR		P-Value	
		Total	Married	Total	Married	Total	Married	Total	Married	Total	Married	Total	Married
		Ages 10-14						Ages 15-19					
57-61	1	0.0028	0.0000	2.48	-NA-	0.2598	-NA-	0.1733	0.3228	2.29	1.01	0.3876	0.1469
62-66	2	0.0095	0.1111	8.55	1.8E+07	0.4618	0.4337	0.1957	0.3964	2.66	1.39	0.4828	0.7688
67-71	3	0.0058	0.2500	5.21	4.7E+07	0.3997	0.6904	0.1790	0.3820	2.38	1.31	0.5612	0.6311
72-76	4	0.0030	0.1667	2.74	2.8E+07	0.7310	0.9376	0.1673	0.4030	2.19	1.43	0.0000	0.2630
77-81	5	0.0023	0.1818	2.07	3.1E+07	0.7208	0.7690	0.0903	0.3527	1.08	1.16	0.9368	0.8268
82-86	6	0.0030	0.2308	2.68	4.2E+07	0.9677	0.8269	0.0912	0.3431	1.10	1.11	0.0758	0.0530
87-91	7	0.0029	0.2667	2.62	5.1E+07	0.2390	0.0000	0.1103	0.4175	1.35	1.52	0.0090	0.0070
92-95	8	0.0011	0.0000	1.00	1.00	-X-	-X-	0.0839	0.3202	1.00	1.00	-X-	-X-
		Ages 20-24						Ages 25-29					
57-61	1	0.2739	0.3365	1.45	1.17	0.3507	0.4198	0.2583	0.2949	1.63	1.53	0.4410	0.2129
62-66	2	0.3019	0.3652	1.66	1.33	0.8762	0.7319	0.2828	0.3398	1.84	1.88	0.5788	0.5567
67-71	3	0.3061	0.3542	1.69	1.27	0.5979	0.8138	0.2670	0.3204	1.70	1.72	0.0751	0.2412
72-76	4	0.2927	0.3472	1.59	1.23	0.3624	0.7027	0.3141	0.3559	2.14	2.02	0.0221	0.0345
77-81	5	0.2727	0.3371	1.44	1.18	0.1017	0.4322	0.2579	0.2973	1.63	1.55	0.4326	0.4673
82-86	6	0.2425	0.3186	1.23	1.08	0.4820	0.6798	0.2417	0.2800	1.49	1.42	0.3239	0.8220
87-91	7	0.2538	0.3272	1.30	1.13	0.0020	0.1980	0.2243	0.2753	1.35	1.39	0.0020	0.0020
92-95	8	0.2068	0.3016	1.00	1.00	-X-	-X-	0.1762	0.2147	1.00	1.00	-X-	-X-
		Ages 30-34						Ages 35-39					
57-61	1	0.2262	0.2724	1.69	1.59	0.4134	0.4997	0.1538	0.1867	1.40	1.45	0.9569	0.9795
62-66	2	0.2535	0.2987	1.97	1.81	0.6964	0.8481	0.1521	0.1878	1.36	1.46	0.1949	0.2635
67-71	3	0.2415	0.2918	1.84	1.75	0.2523	0.1582	0.1901	0.2271	1.80	1.85	0.7699	0.5612
72-76	4	0.2732	0.3385	2.18	2.18	0.7087	0.1871	0.1982	0.2465	1.90	2.06	0.4140	0.4065
77-81	5	0.2636	0.2986	2.07	1.81	0.0308	0.0483	0.2191	0.2723	2.16	2.36	0.0123	0.0021
82-86	6	0.2135	0.2460	1.57	1.39	0.2307	0.4509	0.1612	0.1871	1.48	1.45	0.1897	0.2013
87-91	7	0.1901	0.2287	1.36	1.26	0.0120	0.0680	0.1890	0.2187	1.79	1.76	0.0000	0.0000
92-95	8	0.1472	0.1903	1.00	1.00	-X-	-X-	0.1151	0.1369	1.00	1.00	-X-	-X-
		Ages 40-44						Ages 45-49					
57-61	1	0.0763	0.1282	1.30	2.00	0.9543	0.6278	-NA-	-NA-	-NA-	-NA-	-NA-	-NA-
62-66	2	0.0780	0.1067	1.33	1.63	0.7565	0.9158	0.0400	0.0685	4.07	5.28	0.0520	0.0328
67-71	3	0.0857	0.1033	1.47	1.57	0.2748	0.1538	0.0049	0.0069	0.48	0.50	0.0892	0.0976
72-76	4	0.1108	0.1449	1.96	2.31	0.0260	0.0161	0.0288	0.0386	2.89	2.88	0.3915	0.2772
77-81	5	0.0668	0.0840	1.12	1.25	0.0370	0.0418	0.0188	0.0220	1.87	1.61	0.3691	0.5031
82-86	6	0.1042	0.1301	1.83	2.04	0.0546	0.0131	0.0111	0.0147	1.10	1.07	0.1694	0.2634
87-91	7	0.0712	0.0775	1.20	1.14	0.4510	0.6110	0.0230	0.0270	2.30	1.99	0.1520	0.2500
92-95	8	0.0599	0.0684	1.00	1.00	-X-	-X-	0.0101	0.0137	1.00	1.00	-X-	-X-

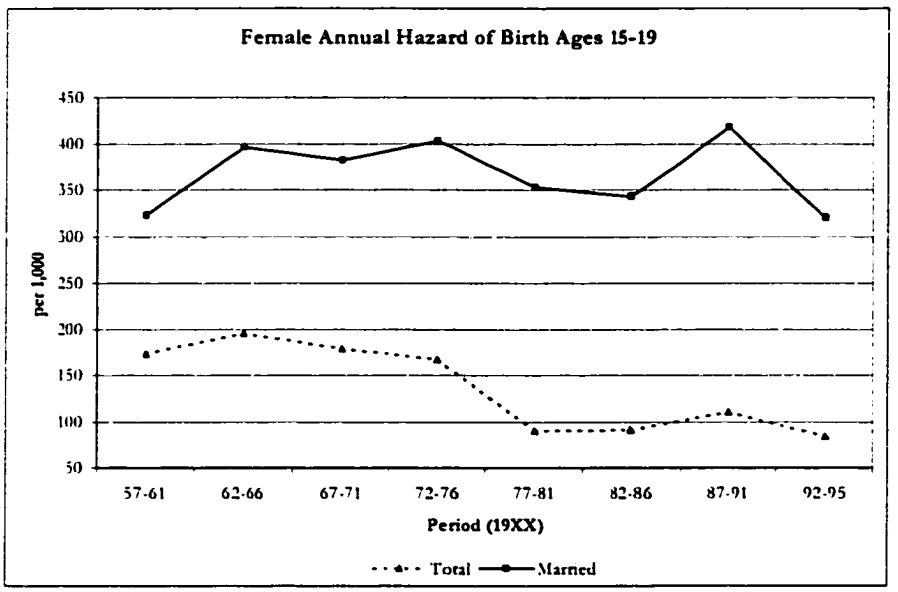

Figure 68: Female Annual Hazard of Birth Ages 15-19

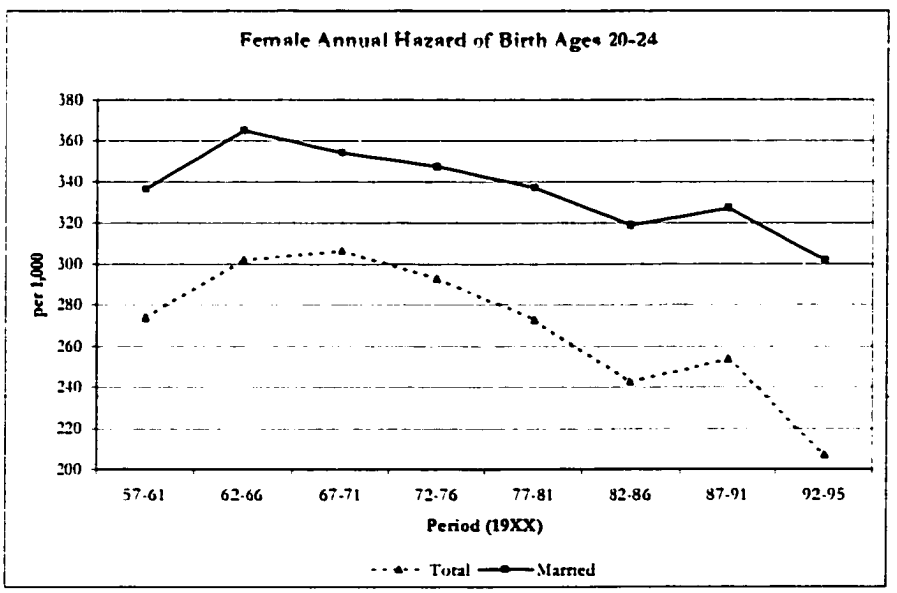

Figure 69: Female Annual Hazard of Birth Ages 20-24

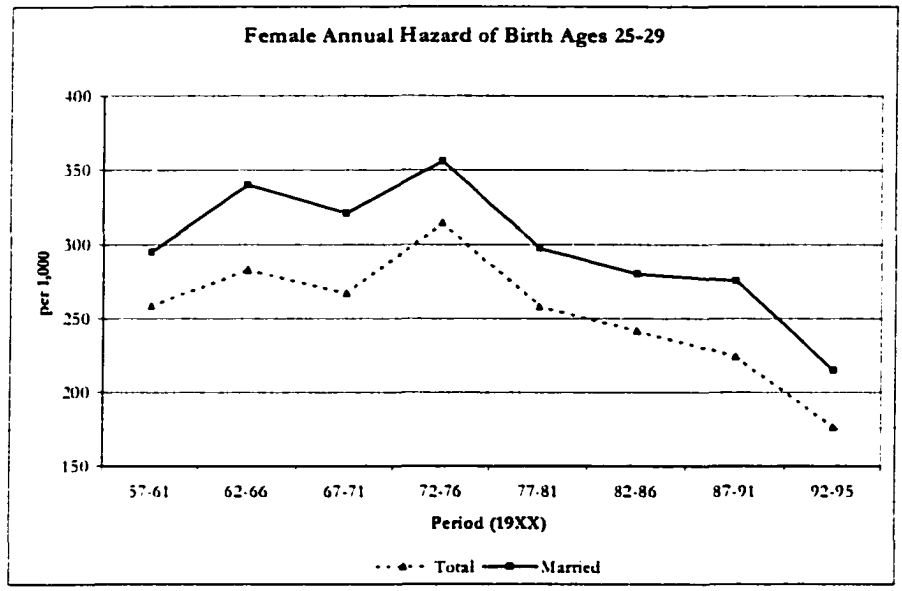

Figure 70: Female Annual Hazard of Birth Ages 25-29

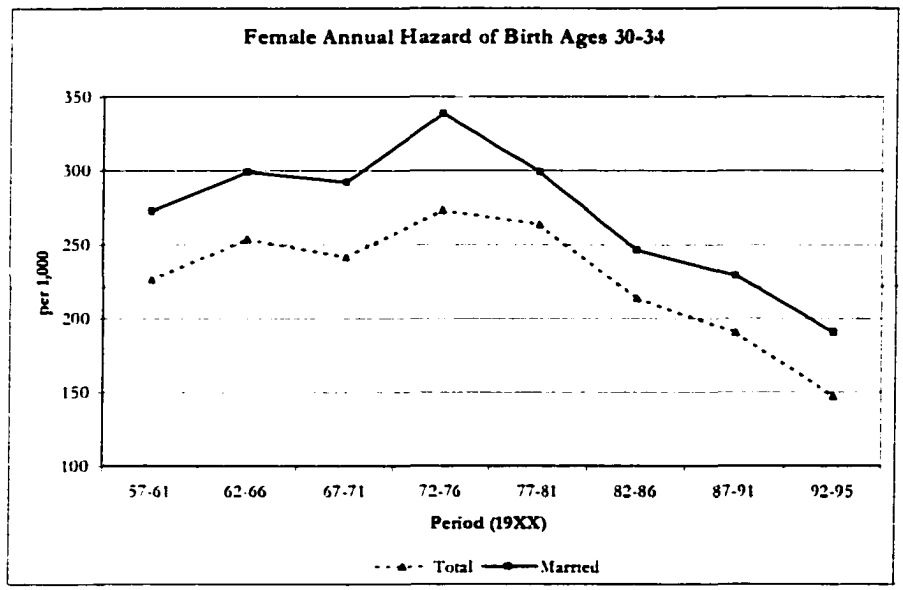

Figure 71: Female Annual Hazard of Birth Ages 30-34

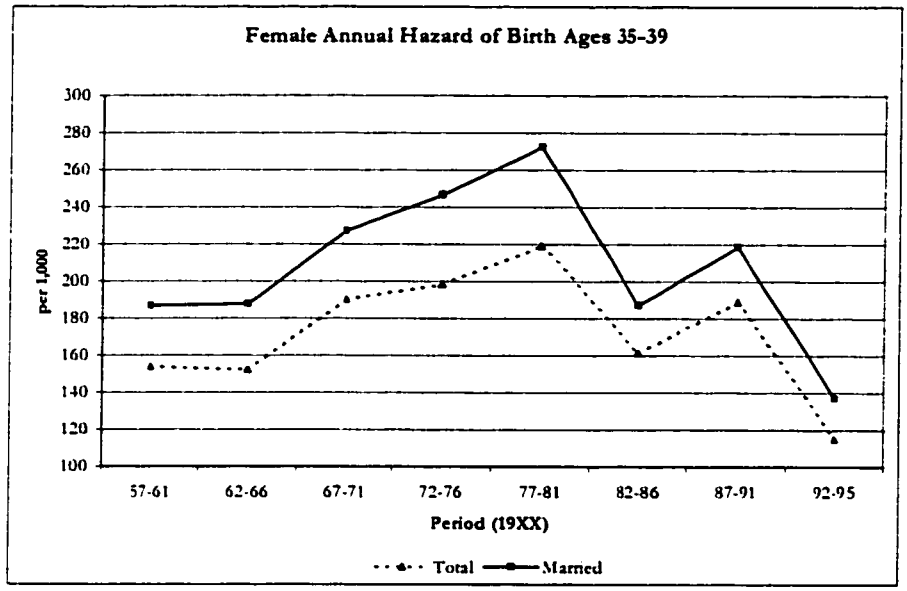

Figure 72: Female Annual Hazard of Birth Ages 35-39

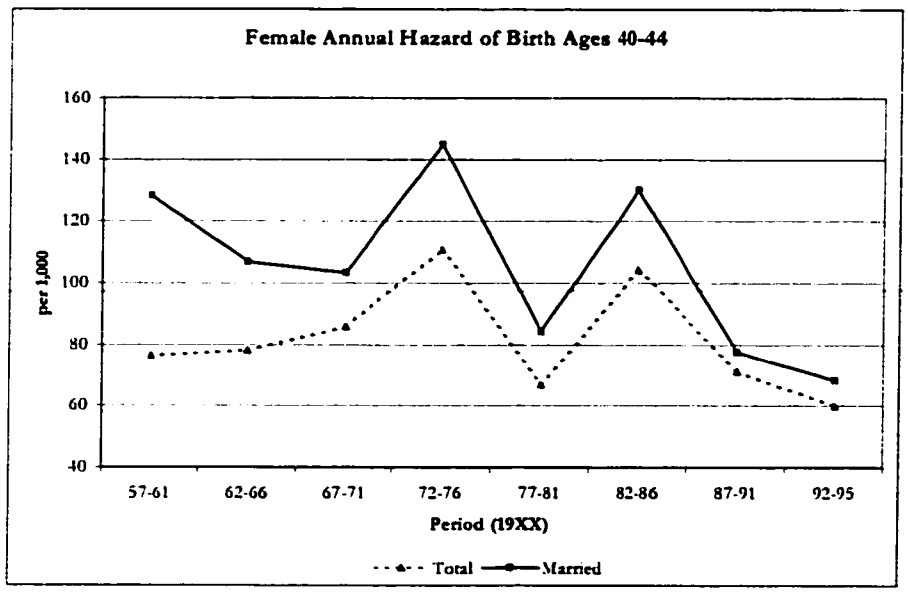

Figure 73: Female Annual Hazard of Birth Ages 40-44

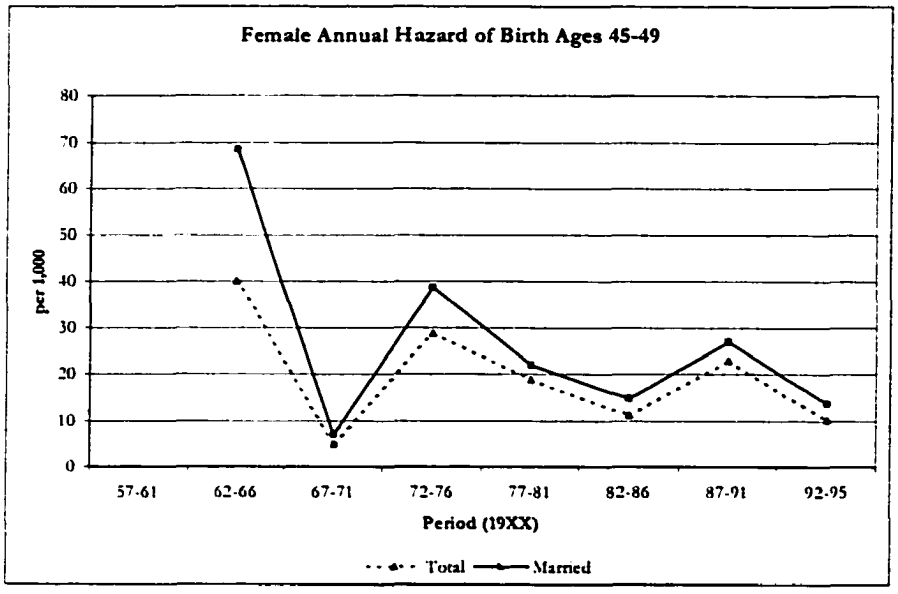

Figure 74: Female Annual Hazard of Birth Ages 45-49

MORTALITY

The risk of death faced by the Tonga is surprisingly low given the challenging environment in which they live. Between 1957 and 1995 the expectation of life at birth was 48 years for males and 52 years for females. The likelihood that an infant would die before its first birthday was 0.119 for males and 0.104 for females. What is most interesting is the trend in the likelihood of dying at various ages. The chances of survival improved steadily in most age groups from 1957 until the late eighties to early nineties. After that mortality conditions have generally deteriorated, and some age groups have seen all of the advances of the previous two or three decades erased within five to ten years. It is likely that these reversals are the result of a generally deteriorating environment and economy, the impact of a long series of serious droughts, and finally HIV/AIDS.

The analysis presented for mortality is similar to the analysis of fertility. The overall levels are demonstrated and discussed followed by a discussion of the trends experienced by various age groups.

LEVELS

The overall level of mortality is summarized by life tables calculated over the period 1957 to 1995, Table 44 for males and Table 45 for females. Figure 75 plots the life table probability of dying ${}_nq_x$ from the life tables, and Figure 76 plots the life table survivors l_x. The life tables appear reasonable, the age-pattern of the risk death has the normal shape, and overall level of mortality is modest. Noteworthy features of the tables are the relatively high infant and child mortality, the slight bulge in mortality between ages 35 and 49, and the stochastic nature of the figures at older ages. The stochasticity at older ages is simply the result of relatively few death

events at those ages in the data set. It is not the aim of this section to produce an exhaustive comparison of these mortality patterns with other existing patterns; for that please refer to Part 1. Here we are primarily interested in the age-specific trends that may provide some indication of when and to what extent the HIV/AIDS epidemic may have affected the population.

All of the life tables presented in this section are calculated in the following way. The annual hazard of death for an individual of a given sex and age living during a defined period is obtained by estimating a dummy variable logistic regression model with the following form:

$$\ln\left(\frac{p(a_i, s)}{1-p(a_i, s)}\right) = \beta_0 s + \beta_1 a_1 + \cdots + \beta_n a_n \beta_{n+1} a_1 s + \cdots + \beta_{2n} a_n s + c$$

Where p is the probability of experiencing death, a_i is a set of dummy variables for age, s is a dummy variable for sex, and $a_i s$ is a set of interactions between age and sex. The model is estimated on samples of the Person-Year Data Set defined along time and age. Because the unit of analysis is a person-year, the resulting probabilities are annual hazards of experiencing a death. This model allows for statistical tests of the influence of sex as a whole and the independent influence of sex at each age. The age dummies and the main sex effect are always highly significant. The interaction effects are highly significant as a group but not necessarily significant as individual effects.

Equation 16: Specification for the Hazard Model of Mortality

The annual hazards of experiencing a death are assumed to correspond to the life table probabilities of death. The annual hazards are aggregated to produce the life table probabilities of death ${}_nq_x$ over the standard five-year, and in some cases ten-year, age groups. The remainder of the life table is constructed in the usual fashion as described by Preston et al. (Preston, Heuveline, and Guillot 2001). Because life tables are somewhat sensitive to the choice of ${}_na_x$ for the first age interval 0 to 1, this value must be chosen with care. The values

used are those suggested by Coale and Demeny (Coale and Demeny 1966), 0.33 for males and 0.35 for females. Finally, the life tables must be closed in a rational fashion. This involves choosing a reasonable value for the expectation of life of those who live to the oldest age reported in the data. This is accomplished by fitting the Himes, Preston, Condran (Himes, Preston, and Condran 1994) standard survival curve for older ages to the life table for ages 45 and older[30]. The predicted values of the survival curve for ages older than the oldest recorded age in the data are then used to extrapolate the life table survival curve beyond the oldest age reported in the data up through age 100. Using these extrapolated values it is possible to calculate the expectation of life for those who survive to the oldest age reported in the data assuming that they face age-specific risks of death that are similar to old people living in populations with their approximate level and slope of mortality for ages 45 and older[31]. When the life tables are presented the cells with extrapolated values are shaded.

---

[30] The usual approach is to assume that ${}_nL_x$ is the ratio of l_x to ${}_nm_x$ which makes the implicit assumption that either ${}_nM_x$ or ${}_na_x$ is known. The event history approach does not provide either thus necessitating another approach.

[31] To accomplish the fit, the Brass logit of both the life table and the old age standard survival curves is calculated for ages 45 and older. An OLS regression of the logit of the life table survival values against the logit of the standard survival values is performed to estimate the α and β parameters of the Brass logit relational model of mortality. Using the linear relationship defined by this regression and the constant and slope coefficients that it produces, the logits of the standard survival curve are transformed into the values that as a group most closely match the logits of the life table survival values. The inverse logits of these values for ages older than the oldest age reported in the data are used to finish the life table. The resulting expectation of life at the oldest age reported in the data is used as the expectation of life for those who survive to the oldest age reported in the data.

TABLE 44: MALE LIFE TABLE 1957-1995

Age	l_x	$_nq_x$	$_nd_x$	$_nm_x$	$_nL_x$	T_x	e_x
0	100,000	0.119380	11,938	0.129759	92,002	4,809,905	48.10
1-4	88,062	0.119393	10,514	0.031743	331,220	4,717,904	53.57
5-9	77,548	0.036785	2,853	0.007495	380,608	4,386,684	56.57
10-14	74,695	0.016540	1,235	0.003336	370,388	4,006,075	53.63
15-19	73,460	0.010530	774	0.002117	365,366	3,635,687	49.49
20-24	72,686	0.011832	860	0.002380	361,282	3,270,321	44.99
25-29	71,826	0.023005	1,652	0.004655	355,001	2,909,040	40.50
30-34	70,174	0.037705	2,646	0.007686	344,255	2,554,039	36.40
35-39	67,528	0.054139	3,656	0.011129	328,501	2,209,784	32.72
40-44	63,872	0.052541	3,356	0.010792	310,971	1,881,283	29.45
45-49	60,516	0.043619	2,640	0.008918	295,982	1,570,312	25.95
50-54	57,877	0.085159	4,929	0.017789	277,062	1,274,330	22.02
55-59	52,948	0.087810	4,649	0.018368	253,116	997,268	18.83
60-64	48,299	0.166817	8,057	0.036399	221,351	744,152	15.41
65-69	40,242	0.137141	5,519	0.029447	187,411	522,801	12.99
70-74	34,723	0.242933	8,435	0.055304	152,526	335,390	9.66
75-79	26,287	0.438311	11,522	0.112266	102,632	182,864	6.96
80-84	**14,765**	**0.558961**	**8,253**	**0.155155**	**53,194**	**80,232**	**5.43**
85-89	**6,512**	**0.714745**	**4,655**	**0.222444**	**20,924**	**27,038**	**4.15**
90-94	**1,858**	**0.852047**	**1,583**	**0.296893**	**5,331**	**6,114**	**3.29**
95-99	**275**	**0.930470**	**256**	**0.347992**	**735**	**783**	**2.85**
100+	**19**	**1.000000**	**19**	**0.400000**	**48**	**48**	**2.50**

Shaded area contains extrapolated values

TABLE 45: FEMALE LIFE TABLE 1957-1995

Age	l_x	$_nq_x$	$_nd_x$	$_nm_x$	$_nL_x$	T_x	e_x
0	100,000	0.103592	10,359	0.111071	93,267	5,164,944	51.65
1-4	89,641	0.114698	10,282	0.030419	338,000	5,071,677	56.58
5-9	79,359	0.033999	2,698	0.006917	390,051	4,733,677	59.65
10-14	76,661	0.014270	1,094	0.002875	380,570	4,343,627	56.66
15-19	75,567	0.010147	767	0.002040	375,919	3,963,056	52.44
20-24	74,800	0.017334	1,297	0.003497	370,760	3,587,138	47.96
25-29	73,504	0.020453	1,503	0.004133	363,760	3,216,378	43.76
30-34	72,000	0.029760	2,143	0.006042	354,645	2,852,617	39.62
35-39	69,858	0.023808	1,663	0.004819	345,130	2,497,972	35.76
40-44	68,194	0.037743	2,574	0.007694	334,538	2,152,842	31.57
45-49	65,621	0.028145	1,847	0.005709	323,486	1,818,304	27.71
50-54	63,774	0.060809	3,878	0.012543	309,174	1,494,818	23.44
55-59	59,896	0.052174	3,125	0.010714	291,666	1,185,645	19.80
60-64	56,771	0.074202	4,212	0.015412	273,322	893,979	15.75
65-69	52,558	0.180129	9,467	0.039592	239,123	620,656	11.81
70-74	43,091	0.293640	12,653	0.068834	183,822	381,533	8.85
75-79	30,438	0.441591	13,441	0.113344	118,586	197,711	6.50
80-84	**16,997**	**0.645810**	**10,977**	**0.190759**	**57,542**	**79,125**	**4.66**
85-89	**6,020**	**0.801010**	**4,822**	**0.267228**	**18,045**	**21,583**	**3.59**
90-94	**1,198**	**0.912238**	**1,093**	**0.335455**	**3,258**	**3,538**	**2.95**
95-99	**105**	**0.966665**	**102**	**0.374192**	**272**	**280**	**2.67**
100+	**4**	**1.000000**	**4**	**0.400000**	**9**	**9**	**2.50**

Shaded area contains extrapolated values

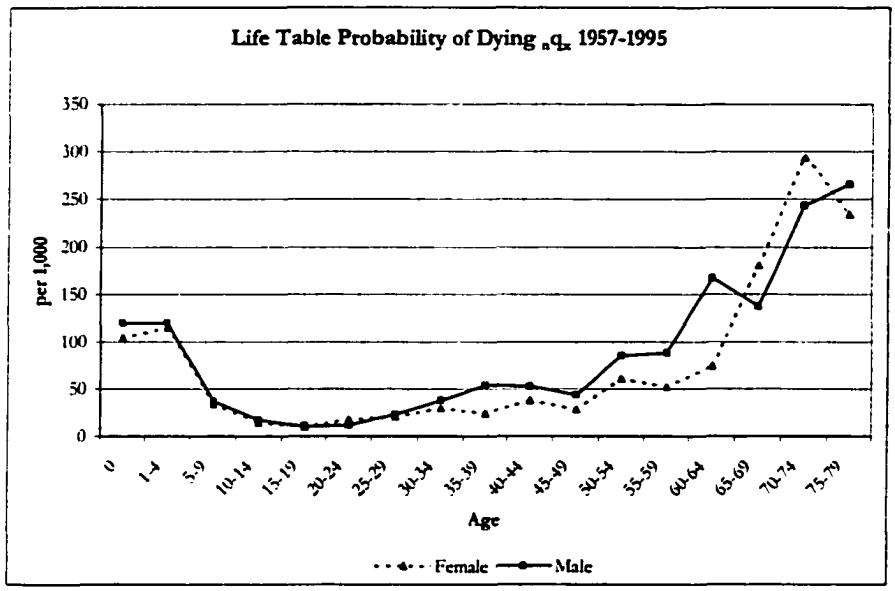

Figure 75: Life Table Probability of Dying 1957-1995

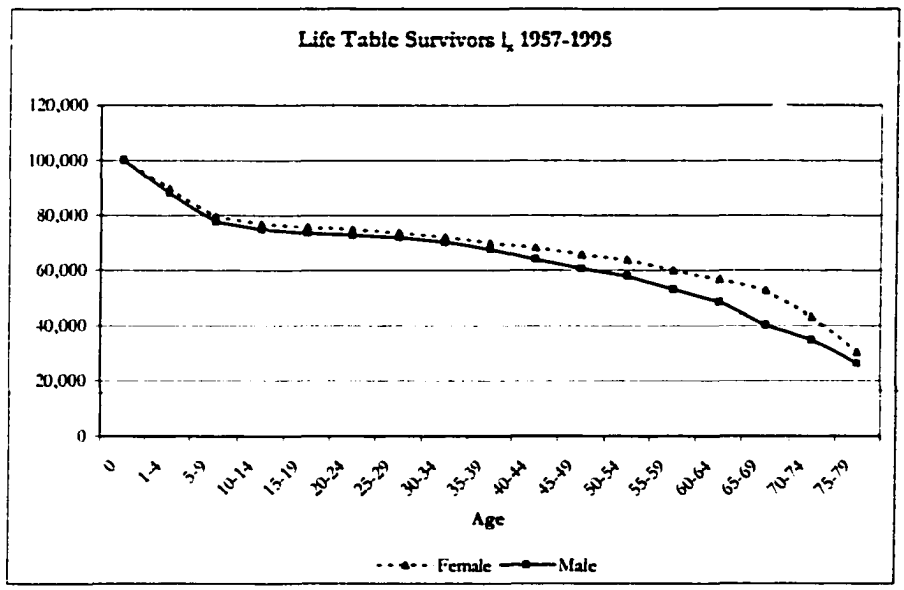

Figure 76: Life Table Survivors 1957-1995

TRENDS

Mortality conditions improved steadily from 1957 until 1972 after which conditions remained constant for about twelve years until 1986. After 1986 overall mortality definitely began to deteriorate.

TABLE 46: LIFE EXPECTANCY AT BIRTH

Period	Female			Male		
	Measured	Predicted	Difference	Measured	Predicted	Difference
57-61	37.95	37.56	0.39	37.82	39.62	-1.79
62-66	44.61	45.30	-0.69	47.83	45.14	2.68
67-71	47.26	49.82	-2.56	47.84	48.38	-0.54
72-76	58.90	53.04	5.86	53.84	50.67	3.16
77-81	52.52	55.53	-3.01	48.94	52.45	-3.51
82-86	58.84	57.56	1.28	51.74	53.91	-2.17
87-91	54.90	59.28	-4.39	47.19	55.14	-7.95
92-95	50.66	60.77	-10.11	46.20	56.20	-10.00

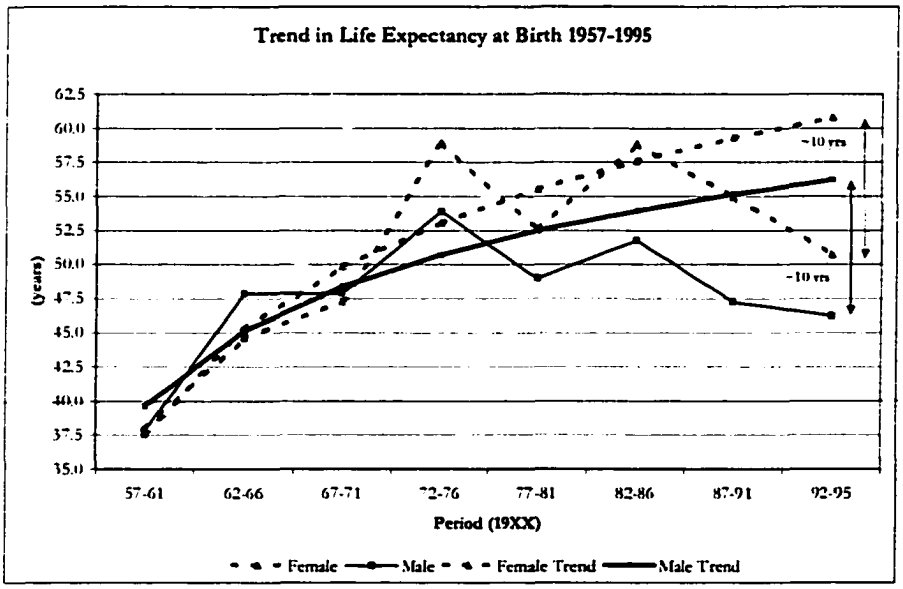

Figure 77: Trend in Life Expectancy at Birth

The trend in overall mortality is examined by calculating sex-specific life tables for five-year periods from 1957 to 1995, Table 48 and Table 49. Table 46 and Figure 77 summarize the trend in the expectation of life at birth, and Table 47 summarizes the age-specific trends in the life table probability of dying, ${}_nq_x$.

The decrease in the expectation of life at birth between 1982 and 1995 is substantial. If mortality had continued to improve at the same rate after 1982 as it had before 1982 it is clear that the expectation of life in 1995 would be far greater than what was actually measured. The question is by how much? To answer that, the trend in the expectation of life at birth is fit with an exponential curve between 1957 and 1981, and the trend embodied in the fit is extrapolated until 1995. The predicted expectation of life at birth calculated in this fashion is compared to the actual expectation of life at birth to estimate the number of years of life expectancy lost since 1982, Table 46. Both males and females have lost roughly *ten* years of life expectancy!

Having established that overall mortality has worsened considerably since 1982, we now turn to an examination of the trend in mortality at different ages. The objective is to determine whether or not the deterioration is concentrated at certain ages, or whether it has occurred in a uniform fashion at all ages. In the absence of cause-specific mortality figures, the clear elucidation of a specific cause responsible for these changes is not possible. However, the age pattern of the changes may provide some circumstantial clues. Most notably, if HIV/AIDS has played a major role, we should be able to see substantial impacts for age groups typically affected by HIV/AIDS: infants and young children and young to middle-aged adults.

TABLE 47: LIFE TABLE PROBABILITY OF DYING $_nq_x$ BY AGE AND PERIOD

Age	Period							
	57-61	62-66	67-71	72-76	77-81	82-86	87-91	92-95
	Female							
0	0.212435	0.117460	0.105714	0.104255	0.103448	0.082616	0.092426	0.094527
1-4	0.165041	0.175663	0.144201	0.114162	0.120479	0.064985	0.104934	0.112207
5-9	0.050393	0.051010	0.064398	0.045516	0.034679	0.030745	0.025750	0.015709
10-19	0.049556	0.048938	0.022632	0.027630	0.028634	0.012866	0.015488	0.028120
20-29	0.051534	0.030346	0.033924	0.014837	0.039621	0.017071	0.033386	0.065937
30-39	0.110991	0.029540	0.012338	0.009611	0.049617	0.058249	0.046557	0.095190
40-49	0.135221	0.087279	0.019065	0.043066	0.035269	0.039443	0.098176	0.100819
50-59	0.119610	0.059684	0.171219	0.090707	0.077932	0.059684	0.131613	0.154508
60-69	0.346620	0.454606	0.459672	0.062285	0.232999	0.229505	0.129677	0.163874
70-79	0.734765	0.820026	0.871826	0.269736	0.564795	0.542514	0.463201	0.545310
	Male							
0	0.203320	0.129693	0.134518	0.096907	0.093686	0.120861	0.119593	0.108865
1-4	0.189474	0.167304	0.136074	0.072266	0.113645	0.125303	0.095736	0.126976
5-9	0.063004	0.071959	0.065501	0.039058	0.033588	0.029584	0.026689	0.023341
10-19	0.044143	0.011267	0.024027	0.017316	0.037302	0.015386	0.029901	0.037531
20-29	0.000000	0.000000	0.034699	0.036626	0.041516	0.025437	0.038815	0.042653
30-39	0.124363	0.017809	0.045412	0.040617	0.050571	0.057763	0.107929	0.168161
40-49	0.228759	0.035145	0.026349	0.096851	0.099842	0.039083	0.111723	0.134615
50-59	0.181275	0.152354	0.216913	0.067770	0.172587	0.107035	0.210612	0.192833
60-69	0.179651	0.291682	0.178058	0.351419	0.373697	0.266880	0.365818	0.195121
70-79	0.499537	0.614093	0.547631	0.613612	0.713939	0.540319	0.729374	0.534338

TABLE 48: MALE LIFE TABLES IN FIVE-YEAR PERIODS

Age	l_x	$_nq_x$	$_nd_x$	$_nm_x$	$_nL_x$	T_x	e_x
1957-1961							
0	100,000	0.203320	20,332	0.235385	86,378	3,782,167	37.82
1-4	79,668	0.189474	15,095	0.052326	288,482	3,695,789	46.39
5-9	64,573	0.063004	4,068	0.013011	312,694	3,407,307	52.77
10-19	60,505	0.044143	2,671	0.004514	591,692	3,094,613	51.15
20-29	57,834	0.000000	0	0.000000	578,338	2,502,921	43.28
30-39	57,834	0.124363	7,192	0.013261	542,376	1,924,583	33.28
40-49	50,641	0.228759	11,585	0.025830	448,490	1,382,207	27.29
50-59	39,057	0.181275	7,080	0.019934	355,167	933,717	23.91
60-69	31,977	0.179651	5,745	0.019738	291,044	578,550	18.09
70-79	26,232	0.499537	13,104	0.066584	196,801	287,506	10.96
80-89	13,128	0.813194	10,676	0.137039	77,903	90,705	6.91
90-99	2,452	0.977971	2,398	0.191378	12,532	12,802	5.22
100+	54	1.000000	54	0.200000	270	270	5.00
1962-1966							
0	100,000	0.129693	12,969	0.142035	91,311	4,782,560	47.83
1-4	87,031	0.167304	14,561	0.045644	319,002	4,691,249	53.90
5-9	72,470	0.071959	5,215	0.014929	349,313	4,372,248	60.33
10-19	67,255	0.011267	758	0.001133	668,764	4,022,934	59.82
20-29	66,497	0.000000	0	0.000000	664,975	3,354,170	50.44
30-39	66,497	0.017809	1,184	0.001797	659,054	2,689,196	40.44
40-49	65,313	0.035145	2,295	0.003577	641,655	2,030,142	31.08
50-59	63,018	0.152354	9,601	0.016492	582,173	1,388,487	22.03
60-69	53,417	0.291682	15,581	0.034148	456,264	806,314	15.09
70-79	37,836	0.614093	23,235	0.088620	262,186	350,050	9.25
80-89	14,601	0.898909	13,125	0.163276	80,386	87,864	6.02
90-99	1,476	0.993395	1,466	0.197375	7,429	7,478	5.07
100+	10	1.000000	10	0.200000	49	49	5.00
1967-1971							
0	100,000	0.134518	13,452	0.147842	90,987	4,783,865	47.84
1-4	86,548	0.136074	11,777	0.036502	322,639	4,692,877	54.22
5-9	74,771	0.065501	4,898	0.013544	361,612	4,370,238	58.45
10-19	69,874	0.024027	1,679	0.002432	690,342	4,008,626	57.37
20-29	68,195	0.034699	2,366	0.003531	670,116	3,318,284	48.66
30-39	65,828	0.045412	2,989	0.004647	643,338	2,648,168	40.23
40-49	62,839	0.026349	1,656	0.002670	620,112	2,004,830	31.90
50-59	61,183	0.216913	13,271	0.024330	545,476	1,384,718	22.63
60-69	47,912	0.178058	8,531	0.019546	436,463	839,243	17.52
70-79	39,381	0.547631	21,566	0.075412	285,977	402,779	10.23
80-89	17,815	0.846624	15,082	0.146808	102,735	116,802	6.56
90-99	2,732	0.985150	2,692	0.194147	13,865	14,067	5.15
100+	41	1.000000	41	0.200000	203	203	5.00

Shaded areas contains extrapolated values

TABLE 48 CONTINUED: MALE LIFE TABLES IN FIVE-YEAR PERIODS

Age	l_x	$_nq_x$	$_nd_x$	$_nm_x$	$_nL_x$	T_x	e_x
			1972-1976				
0	100,000	0.096907	9,691	0.103636	93,507	5,383,595	53.84
1-4	90,309	0.072266	6,526	0.018744	348,185	5,290,088	58.58
5-9	83,783	0.039058	3,272	0.007967	410,734	4,941,905	58.98
10-19	80,511	0.017316	1,394	0.001747	798,136	4,531,169	56.28
20-29	79,117	0.036626	2,898	0.003731	776,677	3,733,033	47.18
30-39	76,219	0.040617	3,096	0.004146	746,709	2,956,357	38.79
40-49	73,123	0.096851	7,082	0.010178	695,820	2,209,647	30.22
50-59	66,041	0.067770	4,476	0.007015	638,032	1,513,827	22.92
60-69	61,565	0.351419	21,635	0.042633	507,478	875,795	14.23
70-79	39,930	0.613612	24,502	0.088520	276,793	368,317	9.22
80-89	15,429	0.907299	13,998	0.166066	84,294	91,524	5.93
90-99	1,430	0.994467	1,422	0.197799	7,191	7,230	5.06
100+	8	1.000000	8	0.200000	40	40	5.00
			1977-1981				
0	100,000	0.093686	9,369	0.099961	93,723	4,894,348	48.94
1-4	90,631	0.113645	10,300	0.030123	341,926	4,800,625	52.97
5-9	80,332	0.033588	2,698	0.006832	394,912	4,458,699	55.50
10-19	77,633	0.037302	2,896	0.003801	761,854	4,063,787	52.35
20-29	74,737	0.041516	3,103	0.004240	731,861	3,301,933	44.18
30-39	71,635	0.050571	3,623	0.005188	698,233	2,570,072	35.88
40-49	68,012	0.099842	6,790	0.010509	646,168	1,871,839	27.52
50-59	61,222	0.172587	10,566	0.018889	559,385	1,225,671	20.02
60-69	50,656	0.373697	18,930	0.045957	411,906	666,285	13.15
70-79	31,726	0.713939	22,650	0.111027	204,006	254,380	8.02
80-89	9,075	0.945058	8,577	0.179168	47,871	50,374	5.55
90-99	499	0.997966	498	0.199188	2,498	2,503	5.02
100+	1	1.000000	1	0.200000	5	5	5.00
			1982-1986				
0	100,000	0.120861	12,086	0.131510	91,902	5,173,865	51.74
1-4	87,914	0.125303	11,016	0.033419	329,624	5,081,962	57.81
5-9	76,898	0.029584	2,275	0.006006	378,803	4,752,338	61.80
10-19	74,623	0.015386	1,148	0.001551	740,491	4,373,535	58.61
20-29	73,475	0.025437	1,869	0.002576	725,405	3,633,045	49.45
30-39	71,606	0.057763	4,136	0.005948	695,379	2,907,640	40.61
40-49	67,470	0.039083	2,637	0.003986	661,513	2,212,261	32.79
50-59	64,833	0.107035	6,939	0.011309	613,632	1,550,748	23.92
60-69	57,893	0.266880	15,451	0.030798	501,681	937,116	16.19
70-79	42,443	0.540319	22,933	0.074032	309,765	435,435	10.26
80-89	19,510	0.857655	16,733	0.150157	111,437	125,670	6.44
90-99	2,777	0.987501	2,742	0.195062	14,059	14,233	5.12
100+	35	1.000000	35	0.200000	174	174	5.00

Shaded areas contains extrapolated values

TABLE 48 CONTINUED: MALE LIFE TABLES IN FIVE-YEAR PERIODS

Age	l_x	$_nq_x$	$_nd_x$	$_nm_x$	$_nL_x$	T_x	e_x
			1987-1991				
0	100,000	0.119593	11,959	0.130010	91,987	4,718,798	47.19
1-4	88,041	0.095736	8,429	0.025137	335,305	4,626,810	52.55
5-9	79,612	0.026689	2,125	0.005410	392,748	4,291,505	53.91
10-19	77,487	0.029901	2,317	0.003035	763,288	3,898,757	50.31
20-29	75,170	0.038815	2,918	0.003958	737,115	3,135,469	41.71
30-39	72,253	0.107929	7,798	0.011409	683,535	2,398,354	33.19
40-49	64,454	0.111723	7,201	0.011833	608,539	1,714,819	26.61
50-59	57,253	0.210612	12,058	0.023540	512,242	1,106,280	19.32
60-69	45,195	0.365818	16,533	0.044771	369,285	594,037	13.14
70-79	28,662	0.729374	20,905	0.114805	182,093	224,752	7.84
80-89	**7,757**	**0.950111**	**7,370**	**0.180993**	**40,718**	**42,659**	**5.50**
90-99	**387**	**0.998353**	**386**	**0.199342**	**1,938**	**1,941**	**5.02**
100+	**1**	**1.000000**	**1**	**0.200000**	**3**	**3**	**5.00**
			1992-1995				
0	100,000	0.108865	10,886	0.117430	92,706	4,619,973	46.20
1-4	89,114	0.126976	11,315	0.033896	333,823	4,527,267	50.80
5-9	77,798	0.023341	1,816	0.004723	384,451	4,193,444	53.90
10-19	75,982	0.037531	2,852	0.003825	745,565	3,808,993	50.13
20-29	73,131	0.042653	3,119	0.004358	715,710	3,063,428	41.89
30-39	70,011	0.168161	11,773	0.018360	641,248	2,347,718	33.53
40-49	58,238	0.134615	7,840	0.014433	543,183	1,706,471	29.30
50-59	50,398	0.192833	9,718	0.021341	455,392	1,163,288	23.08
60-69	40,680	0.195121	7,937	0.021621	367,112	707,896	17.40
70-79	32,742	0.534338	17,496	0.072914	239,947	340,784	10.41
80-89	**15,247**	**0.841147**	**12,825**	**0.145169**	**88,345**	**100,837**	**6.61**
90-99	**2,422**	**0.984209**	**2,384**	**0.193782**	**12,301**	**12,493**	**5.16**
100+	**38**	**1.000000**	**38**	**0.200000**	**191**	**191**	**5.00**

Shaded areas contains extrapolated values

TABLE 49: FEMALE LIFE TABLES IN FIVE-YEAR PERIODS

Age	l_x	$_nq_x$	$_nd_x$	$_nm_x$	$_nL_x$	T_x	e_x
			1957-1961				
0	100,000	0.212435	21,244	0.246468	86,192	3,794,964	37.95
1-4	78,756	0.165041	12,998	0.044971	289,030	3,708,773	47.09
5-9	65,758	0.050393	3,314	0.010339	320,508	3,419,743	52.00
10-19	62,445	0.049556	3,095	0.005082	608,974	3,099,235	49.63
20-29	59,350	0.051534	3,059	0.005290	578,209	2,490,261	41.96
30-39	56,292	0.110991	6,248	0.011751	531,676	1,912,053	33.97
40-49	50,044	0.135221	6,767	0.014503	466,602	1,380,377	27.58
50-59	43,277	0.119610	5,176	0.012722	406,886	913,775	21.11
60-69	38,100	0.346620	13,206	0.041929	314,972	506,889	13.30
70-79	24,894	0.734765	18,291	0.116147	157,484	191,916	7.71
80-89	**6,603**	**0.978521**	**6,461**	**0.191589**	**33,723**	**34,432**	**5.21**
90-99	**142**	**0.999816**	**142**	**0.199927**	**709**	**709**	**5.00**
100+	**0**	**1.000000**	**0**	**0.200000**	**0**	**0**	**5.00**
			1962-1966				
0	100,000	0.117460	11,746	0.127170	92,365	4,461,227	44.61
1-4	88,254	0.175663	15,503	0.048144	322,010	4,368,862	49.50
5-9	72,751	0.051010	3,711	0.010469	354,477	4,046,852	55.63
10-19	69,040	0.048938	3,379	0.005017	673,507	3,692,375	53.48
20-29	65,661	0.030346	1,993	0.003081	646,650	3,018,868	45.98
30-39	63,669	0.029540	1,881	0.002998	627,284	2,372,218	37.26
40-49	61,788	0.087279	5,393	0.009126	590,916	1,744,934	28.24
50-59	56,395	0.059684	3,366	0.006152	547,123	1,154,018	20.46
60-69	53,029	0.454606	24,107	0.058834	409,756	606,895	11.44
70-79	28,922	0.820026	23,717	0.138991	170,635	197,139	6.82
80-89	**5,205**	**0.990825**	**5,157**	**0.196363**	**26,265**	**26,504**	**5.09**
90-99	**48**	**0.999958**	**48**	**0.199983**	**239**	**239**	**5.00**
100+	**0**	**1.000000**	**0**	**0.200000**	**0**	**0**	**5.00**
			1967-1971				
0	100,000	0.105714	10,571	0.113514	93,129	4,726,156	47.26
1-4	89,429	0.144201	12,896	0.038851	331,923	4,633,028	51.81
5-9	76,533	0.064398	4,929	0.013308	370,343	4,301,105	56.20
10-19	71,604	0.022632	1,621	0.002289	707,941	3,930,762	54.90
20-29	69,984	0.033924	2,374	0.003451	687,967	3,222,821	46.05
30-39	67,610	0.012338	834	0.001241	671,926	2,534,854	37.49
40-49	66,776	0.019065	1,273	0.001925	661,390	1,862,928	27.90
50-59	65,502	0.171219	11,215	0.018725	598,948	1,201,538	18.34
60-69	54,287	0.459672	24,954	0.059685	418,100	602,591	11.10
70-79	29,333	0.871826	25,573	0.154555	165,463	184,491	6.29
80-89	**3,760**	**0.993903**	**3,737**	**0.197576**	**18,913**	**19,028**	**5.06**
90-99	**23**	**0.999976**	**23**	**0.199990**	**115**	**115**	**5.00**
100+	**0**	**1.000000**	**0**	**0.200000**	**0**	**0**	**5.00**

Shaded areas contains extrapolated values

TABLE 49 CONTINUED: FEMALE LIFE TABLES IN FIVE-YEAR PERIODS

Age	l_x	$_nq_x$	$_nd_x$	$_nm_x$	$_nL_x$	T_x	e_x
			1972-1976				
0	100,000	0.104255	10,426	0.111834	93,223	5,889,991	58.90
1-4	89,574	0.114162	10,226	0.030268	337,846	5,796,768	64.71
5-9	79,348	0.045516	3,612	0.009315	387,713	5,458,922	68.80
10-19	75,737	0.027630	2,093	0.002802	746,905	5,071,209	66.96
20-29	73,644	0.014837	1,093	0.001495	730,979	4,324,303	58.72
30-39	72,552	0.009611	697	0.000966	722,029	3,593,325	49.53
40-49	71,854	0.043066	3,094	0.004401	703,070	2,871,295	39.96
50-59	68,760	0.090707	6,237	0.009502	656,413	2,168,225	31.53
60-69	62,523	0.062285	3,894	0.006429	605,757	1,511,812	24.18
70-79	58,629	0.269736	15,814	0.031179	507,214	906,056	15.45
80-89	42,814	0.601886	25,769	0.086100	299,297	398,841	9.32
90-99	17,045	0.915990	15,613	0.169000	92,385	99,544	5.84
100+	1,432	1.000000	1,432	0.200000	7,160	7,160	5.00
			1977-1981				
0	100,000	0.103448	10,345	0.110906	93,276	5,251,824	52.52
1-4	89,655	0.120479	10,802	0.032050	337,018	5,158,549	57.54
5-9	78,854	0.034679	2,735	0.007058	387,432	4,821,531	61.15
10-19	76,119	0.028634	2,180	0.002905	750,293	4,434,099	58.25
20-29	73,939	0.039621	2,930	0.004042	724,747	3,683,807	49.82
30-39	71,010	0.049617	3,523	0.005088	692,482	2,959,060	41.67
40-49	67,487	0.035269	2,380	0.003590	662,965	2,266,578	33.59
50-59	65,106	0.077932	5,074	0.008109	625,694	1,603,613	24.63
60-69	60,032	0.232999	13,988	0.026372	530,387	977,919	16.29
70-79	46,045	0.564795	26,006	0.078706	330,420	447,532	9.72
80-89	20,039	0.915913	18,354	0.168974	108,620	117,112	5.84
90-99	1,685	0.996072	1,678	0.198435	8,458	8,491	5.04
100+	7	1.000000	7	0.200000	33	33	5.00
			1982-1986				
0	100,000	0.082616	8,262	0.087304	94,630	5,884,491	58.84
1-4	91,738	0.064985	5,962	0.016792	355,030	5,789,861	63.11
5-9	85,777	0.030745	2,637	0.006245	422,291	5,434,831	63.36
10-19	83,139	0.012866	1,070	0.001295	826,047	5,012,541	60.29
20-29	82,070	0.017071	1,401	0.001722	813,693	4,186,494	51.01
30-39	80,669	0.058249	4,699	0.006000	783,193	3,372,801	41.81
40-49	75,970	0.039443	2,996	0.004024	744,716	2,589,608	34.09
50-59	72,973	0.059684	4,355	0.006152	707,957	1,844,891	25.28
60-69	68,618	0.229505	15,748	0.025926	607,440	1,136,934	16.57
70-79	52,870	0.542514	28,683	0.074445	385,285	529,494	10.02
80-89	24,187	0.904290	21,872	0.165060	132,511	144,209	5.96
90-99	2,315	0.994676	2,303	0.197882	11,636	11,698	5.05
100+	12	1.000000	12	0.200000	62	62	5.00

Shaded areas contains extrapolated values

TABLE 49 CONTINUED: FEMALE LIFE TABLES IN FIVE-YEAR PERIODS

Age	l_x	$_nq_x$	$_nd_x$	$_nm_x$	$_nL_x$	T_x	e_x
			1987-1991				
0	100,000	0.092426	9,243	0.098334	93,992	5,489,643	54.90
1-4	90,757	0.104934	9,524	0.027686	343,982	5,395,650	59.45
5-9	81,234	0.025750	2,092	0.005217	400,940	5,051,668	62.19
10-19	79,142	0.015488	1,226	0.001561	785,292	4,650,728	58.76
20-29	77,916	0.033386	2,601	0.003395	766,157	3,865,435	49.61
30-39	75,315	0.046557	3,506	0.004767	735,619	3,099,278	41.15
40-49	71,809	0.098176	7,050	0.010324	682,837	2,363,660	32.92
50-59	64,759	0.131613	8,523	0.014088	604,972	1,680,823	25.96
60-69	56,236	0.129677	7,292	0.013867	525,894	1,075,851	19.13
70-79	48,943	0.463201	22,671	0.060281	376,079	549,956	11.24
80-89	26,273	0.840336	22,078	0.144928	152,337	173,877	6.62
90-99	4,195	0.986502	4,138	0.194673	21,257	21,540	5.13
100+	57	1.000000	57	0.200000	283	283	5.00
			1992-1995				
0	100,000	0.094527	9,453	0.100716	93,856	5,065,874	50.66
1-4	90,547	0.112207	10,160	0.029719	341,869	4,972,018	54.91
5-9	80,387	0.015709	1,263	0.003167	398,779	4,630,149	57.60
10-19	79,124	0.028120	2,225	0.002852	780,119	4,231,370	53.48
20-29	76,899	0.065937	5,070	0.006818	743,642	3,451,251	44.88
30-39	71,829	0.095190	6,837	0.009995	684,102	2,707,609	37.70
40-49	64,992	0.100819	6,552	0.010617	617,153	2,023,507	31.13
50-59	58,439	0.154508	9,029	0.016744	539,245	1,406,354	24.07
60-69	49,410	0.163874	8,097	0.017850	453,613	867,109	17.55
70-79	41,313	0.545310	22,528	0.074973	300,487	413,495	10.01
80-89	18,785	0.898942	16,886	0.163287	103,414	113,008	6.02
90-99	1,898	0.994596	1,888	0.197850	9,543	9,594	5.05
100+	10	1.000000	10	0.200000	51	51	5.00

Shaded areas contains extrapolated values

Age-specific trends in mortality are examined using two different techniques. First, the age-specific time series of life table probabilities of dying ${}_nq_x$ are collected together from Table 48 and Table 49 and displayed in Table 47. Second, the event history approach used to examine trends in fertility is applied to the mortality data. Within each age-group the probability of death is modeled as a function of sex and time period according the following specification:

$$\ln\left(\frac{p(t_i, s)}{1 - p(t_i, s)}\right) = \beta_0 s + \beta_1 t_1 + \cdots + \beta_7 t_7 \beta_8 t_1 s + \cdots + \beta_{14} t_{14} s + c$$

Where p is the probability of experiencing a death, s is a dummy variable for sex, t is a set of dummy variables for time period, and ts is a set of interactions between time and sex. This model is estimated on the Person-Year Data Set for person-years lived over specified age ranges. With this specification it is possible to test the null hypothesis that consecutive period coefficients are equal. A small P-value generated by that test indicates that it is unlikely that the change observed between the consecutive periods in question is the result of a random process

Equation 17: Specification for Age-Specific Hazard Model of Mortality

Table 50 contains estimates of the annual hazards of death by period for each five-year age-group along with the odds ratios comparing each period to the most recent 1992 to 1995 period and the P-values produced testing the null hypothesis that each pair of consecutive period effects is the same. Figure 78 through Figure 87 display the age-specific trends in the annual hazard of death.

The annual hazard of death for infants and children aged one to four improved steadily during the early periods from 1957 through the 1970s and then remained constant with no consistent substantial movement up or down. After a brief deterioration during 1962 to 1966, the hazard

of death for children five to nine has improved steadily right through the most recent period 1992 to 1995. The low hazard of death for ten to nineteen year old children has remained low and roughly constant throughout the period of observation.

The 20 to 29 age group is the youngest to manifest the mortality reversal noted above. The annual hazard of death has been consistently low for young adults in their twenties, but the beginning of an upward trend is evident during the most recent to periods between 1987 and 1995. This change is statistically significant for females (at the 7% level) and very mildly significant (at the 30% level) for males between the 1987 to 1991 and the 1992 to 1995 periods.

A similar situation is observed for adults aged 30 to 39. After a dramatic improvement immediately following relocation of the population, the annual hazard of death remained relatively constant until the most recent two periods when it increased dramatically for both sexes. Both the improvement after relocation and the most recent increase are statistically significant for females (at the 9% and 8% levels respectively) while none of the changes for males are statistically significant from one period to the next. Changes in the 40 to 49 age group mirror those in the 30 to 39 age group with initial improvement followed by a long plateau and a recent deterioration. None of the changes in this age group are highly statistically significant from one period to the next. At ages older than 49, no real trends are evident, and there appears to have been little meaningful change over time measured at older ages.

TABLE 50: TREND IN ANNUAL HAZARD OF DEATH BY AGE

		Age 0						Ages 1-4					
		Annual Hazard		OR		P-Value		Annual Hazard		OR		P-Value	
Period	i	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
57-61	1	0.2124	0.2033	2.58	2.44	**0.0044**	0.6232	0.0441	0.0512	1.53	1.79	0.7900	0.5578
62-66	2	0.1175	0.1297	1.27	1.43	0.6307	0.6325	0.0471	0.0447	1.64	1.55	0.3358	0.9783
67-71	3	0.1057	0.1345	1.13	1.49	0.9462	0.2577	0.0382	0.0359	1.31	1.23	0.2512	0.2040
72-76	4	0.1043	0.0969	1.11	1.03	0.9664	0.9238	0.0299	0.0186	1.02	0.63	0.7787	**0.1815**
77-81	5	0.1034	0.0937	1.11	0.99	0.2281	**0.0625**	0.0316	0.0297	1.08	1.01	**0.0035**	**0.0106**
82-86	6	0.0826	0.1209	0.86	1.32	0.5283	0.5985	0.0167	0.0329	0.56	1.13	**0.0203**	**0.0055**
87-91	7	0.0924	0.1196	0.98	1.30	0.8940	0.6010	0.0273	0.0248	0.93	0.84	0.6880	0.3610
92-95	8	0.0945	0.1089	1.00	1.17	-X-	-X-	0.0293	0.0334	1.00	1.14	-X-	-X-
		Ages 5-9						Ages 10-19					
57-61	1	0.0103	0.0129	3.28	4.13	0.9830	0.8715	0.0051	0.0045	1.78	1.58	0.9847	0.3052
62-66	2	0.0104	0.0148	3.32	4.74	0.6085	0.5877	0.0050	0.0011	1.76	0.40	0.2825	0.2569
67-71	3	0.0132	0.0135	4.23	4.30	0.3726	0.7616	0.0023	0.0024	0.80	0.85	0.7820	0.6265
72-76	4	0.0093	0.0079	2.95	2.52	0.4966	0.8303	0.0028	0.0017	0.98	0.61	0.9507	0.4033
77-81	5	0.0070	0.0068	2.23	2.16	0.7598	0.9908	0.0029	0.0038	1.02	1.33	0.1973	0.9156
82-86	6	0.0062	0.0060	1.98	1.90	0.6532	0.8963	0.0013	0.0015	0.45	0.54	0.7723	0.5641
87-91	7	0.0052	0.0054	1.65	1.71	0.2940	0.5700	0.0016	0.0030	0.55	1.06	0.2430	0.5760
92-95	8	0.0032	0.0047	1.00	1.49	-X-	-X-	0.0028	0.0038	1.00	1.34	-X-	-X-
		Ages 20-29						Ages 30-39					
57-61	1	0.0053	-NA-	0.78	-NA-	0.4791	-NA-	0.0117	0.0132	1.18	1.33	**0.0939**	0.6430
62-66	2	0.0031	-NA-	0.45	-NA-	0.8821	-NA-	0.0030	0.0018	0.30	0.18	0.4715	0.2763
67-71	3	0.0034	0.0035	0.51	0.52	0.3339	0.4400	0.0012	0.0046	0.12	0.46	0.8590	0.9332
72-76	4	0.0015	0.0037	0.22	0.55	0.2145	0.3997	0.0010	0.0041	0.10	0.41	**0.1237**	0.2703
77-81	5	0.0040	0.0042	0.59	0.62	**0.1733**	0.6836	0.0051	0.0052	0.51	0.52	0.7601	0.9718
82-86	6	0.0017	0.0026	0.25	0.38	0.2510	0.7571	0.0060	0.0059	0.60	0.59	0.6453	**0.1890**
87-91	7	0.0034	0.0040	0.50	0.58	**0.0740**	0.3050	0.0048	0.0114	0.48	1.14	**0.0840**	0.6210
92-95	8	0.0068	0.0043	1.00	0.64	-X-	-X-	0.0100	0.0182	1.00	1.85	-X-	-X-
		Ages 40-49						Ages 50-59					
57-61	1	0.0144	0.0256	1.37	2.46	0.5697	0.2660	0.0127	0.0198	0.76	1.19	0.5523	0.7350
62-66	2	0.0091	0.0036	0.86	0.34	**0.1773**	0.4883	0.0061	0.0164	0.36	0.98	0.3177	0.6051
67-71	3	0.0019	0.0027	0.18	0.25	0.4737	0.7454	0.0186	0.0242	1.12	1.46	0.3729	0.6160
72-76	4	0.0044	0.0101	0.41	0.96	0.8030	0.8115	0.0095	0.0070	0.56	0.42	0.8354	0.2977
77-81	5	0.0036	0.0105	0.34	0.99	0.8814	0.2918	0.0081	0.0188	0.48	1.13	0.6958	0.7896
82-86	6	0.0040	0.0040	0.38	0.37	**0.1061**	0.8671	0.0061	0.0113	0.36	0.67	**0.1549**	0.9053
87-91	7	0.0103	0.0118	0.97	1.12	0.9490	0.7770	0.0140	0.0234	0.84	1.41	0.6780	0.6390
92-95	8	0.0106	0.0144	1.00	1.36	-X-	-X-	0.0166	0.0212	1.00	1.28	-X-	-X-
		Ages 60-69						Ages 70-79					
57-61	1	0.0417	0.0196	2.41	1.11	0.6077	0.8890	0.1034	0.0870	2.52	2.08	-NA-	-NA-
62-66	2	0.0588	0.0339	3.46	1.94	0.9765	0.6078	0.0438	0.0420	-NA-	-NA-	-NA-	-NA-
67-71	3	0.0597	0.0194	3.52	1.10	**0.0322**	**0.0234**	0.0000	0.0233	0.00	0.52	**0.0000**	**0.0000**
72-76	4	0.0064	0.0424	0.36	2.45	**0.1948**	0.2793	0.0122	0.0952	0.27	2.30	**0.0323**	**0.0131**
77-81	5	0.0262	0.0457	1.49	2.65	0.9765	0.6138	0.1058	0.0317	2.59	0.72	0.7509	0.7548
82-86	6	0.0257	0.0306	1.46	1.75	0.2577	**0.1639**	0.0920	0.0375	2.21	0.85	0.7022	**0.1617**
87-91	7	0.0138	0.0445	0.77	2.58	0.6390	**0.1630**	0.0761	0.0962	1.80	2.33	0.2860	0.6960
92-95	8	0.0177	0.0215	1.00	1.22	-X-	-X-	0.0438	0.0420	1.00	0.96	-X-	-X-

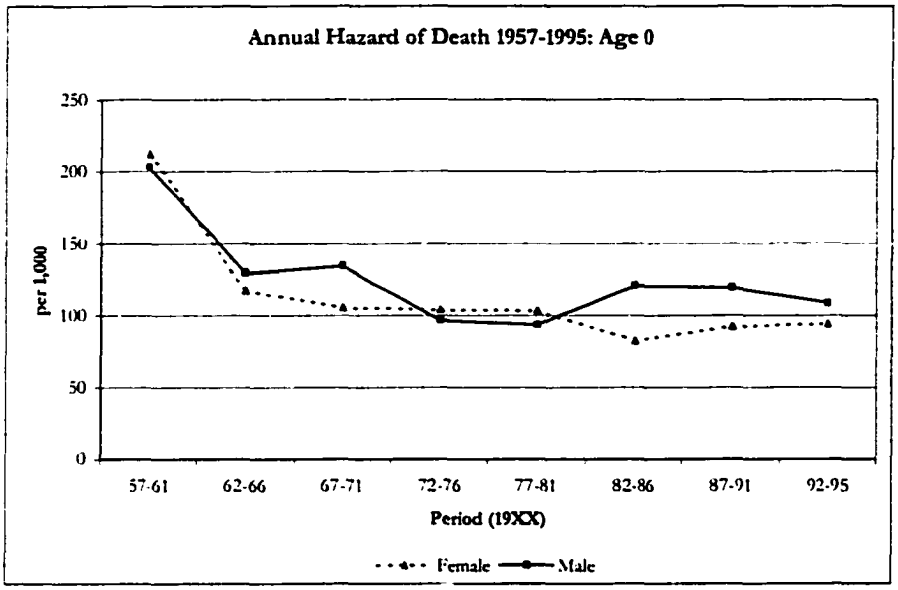

Figure 78: Annual Hazard of Death 1957-1995 Age 0

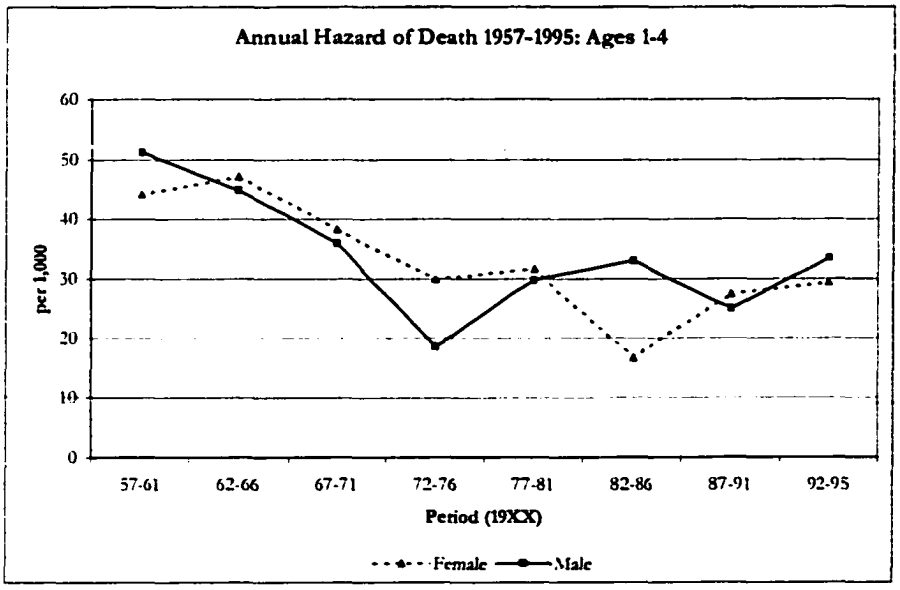

Figure 79: Annual Hazard of Death 1957-1995 Ages 1-4

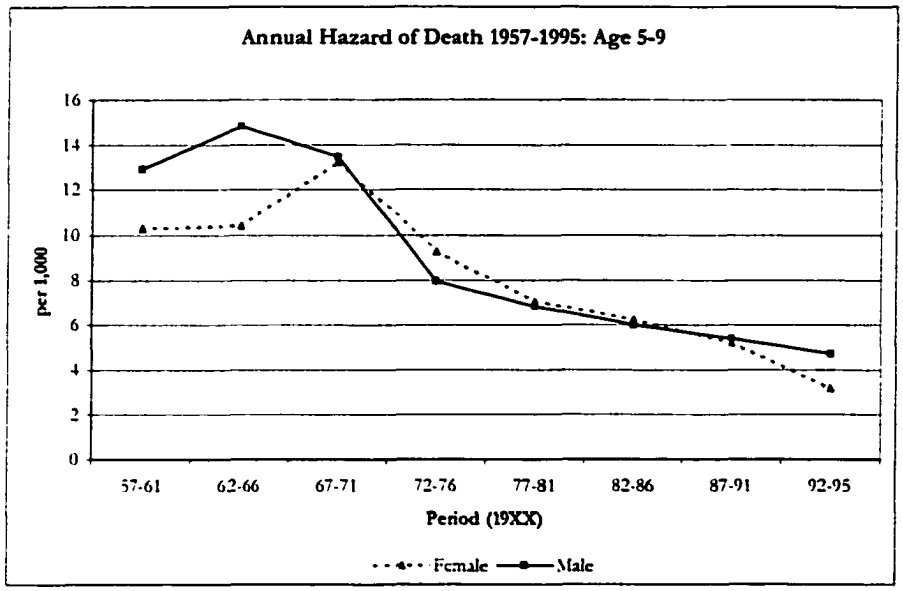

Figure 80: Annual Hazard of Death 1957-1995 Ages 5-9

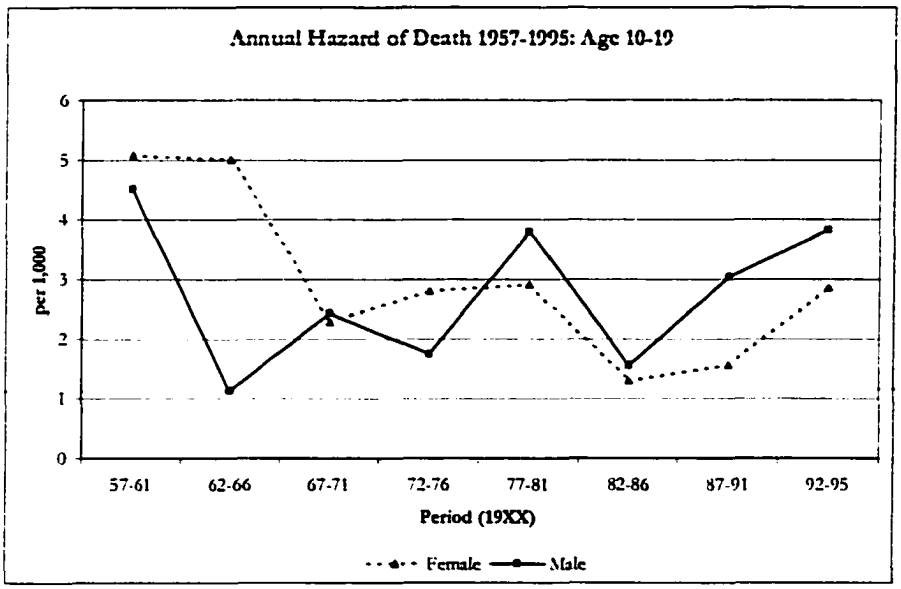

Figure 81: Annual Hazard of Death 1957-1995 Ages 10-19

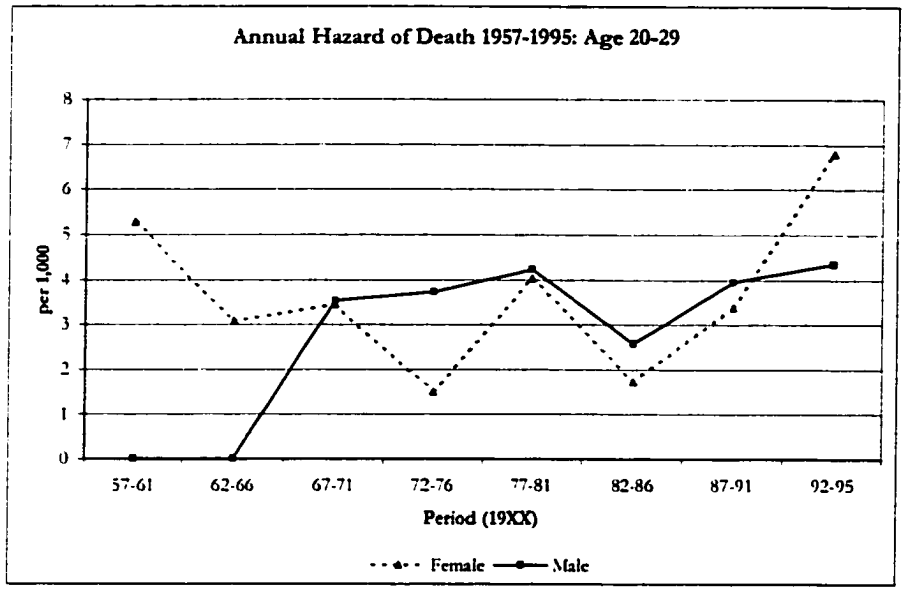

Figure 82: Annual Hazard of Death 1957-1995 Ages 20-29

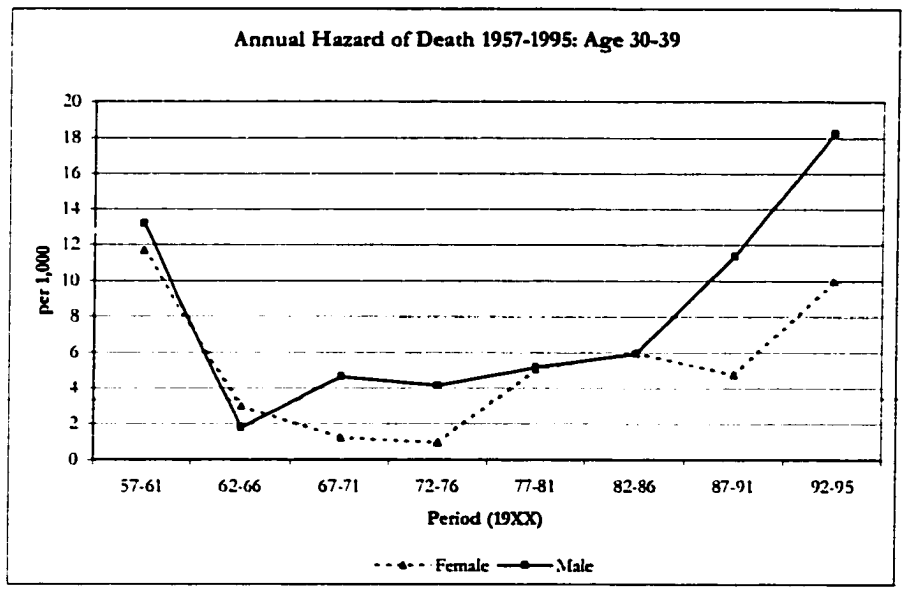

Figure 83: Annual Hazard of Death 1957-1995 Ages 30-39

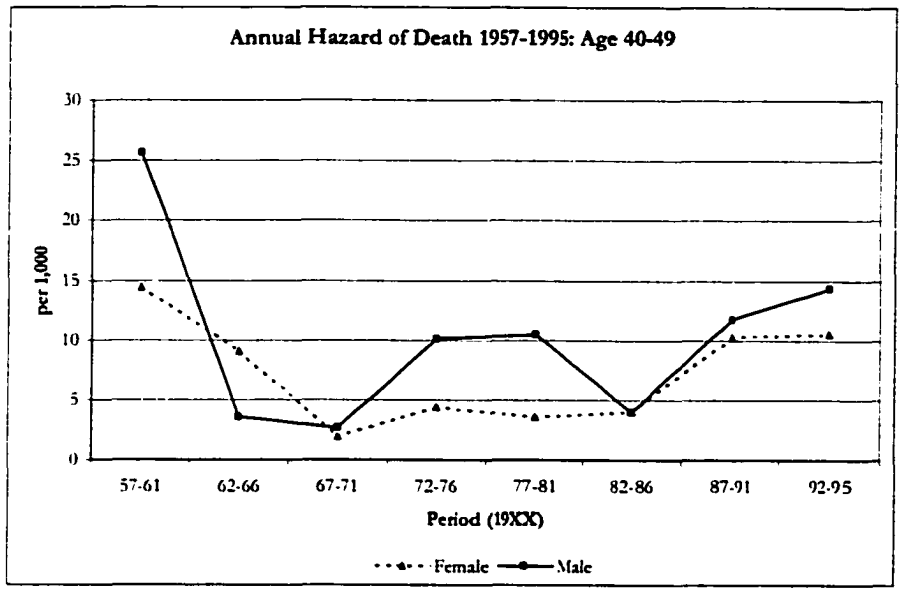

Figure 84: Annual Hazard of Death 1957-1995 Ages 40-49

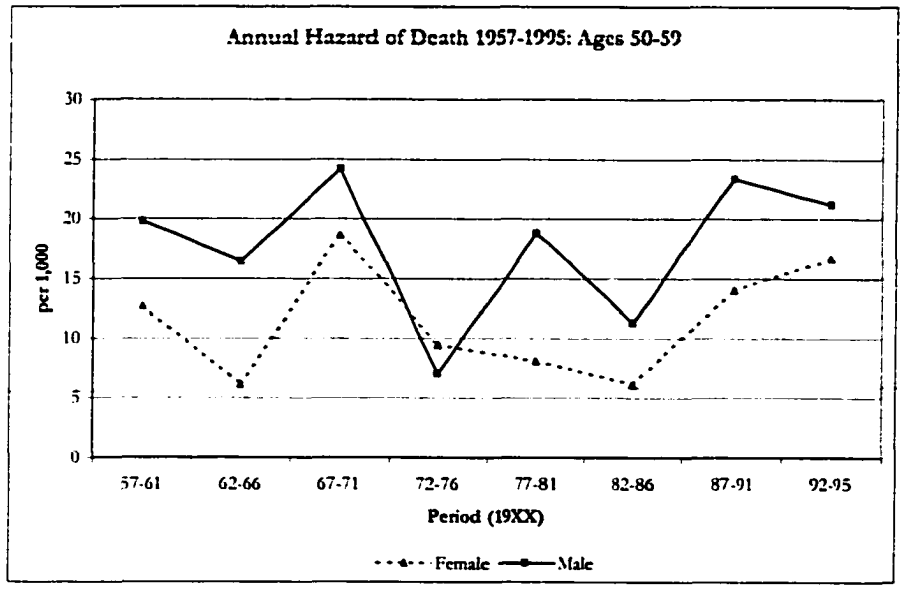

Figure 85: Annual Hazard of Death 1957-1995 Ages 50-59

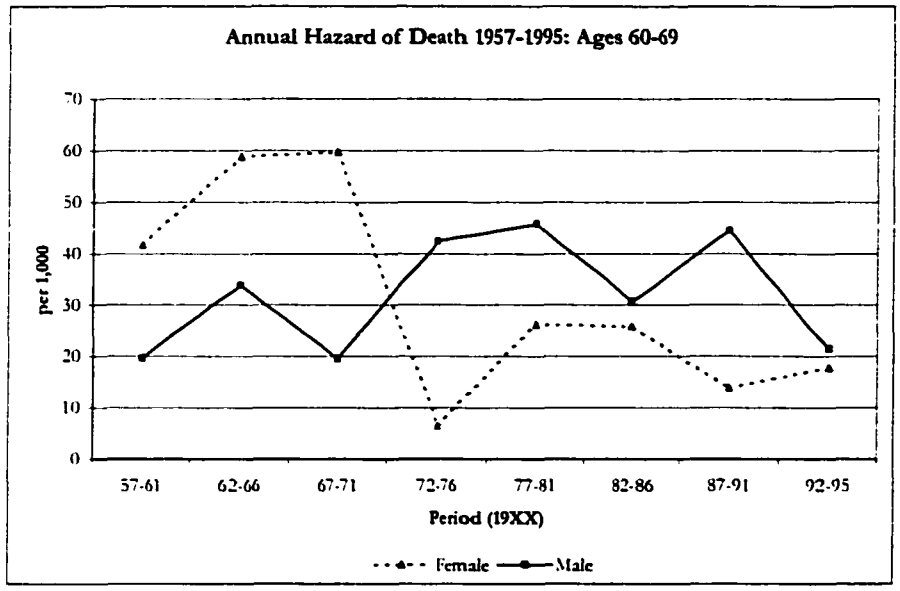

Figure 86: Annual Hazard of Death 1957-1995 Ages 60-69

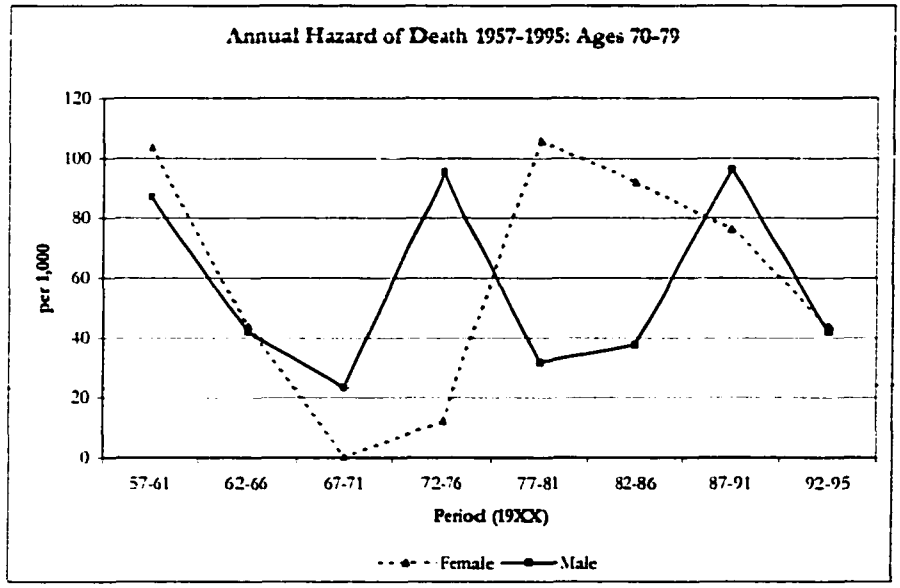

Figure 87: Annual Hazard of Death 1957-1995 Ages 70-79

EVIDENCE OF HIV

The only evidence of HIV that these data can potentially reveal is a time-specific increase in the probability of death occurring near the time that it is known that the AIDS epidemic began to have an impact in Zambia. That time is the late 1980s to the early 1990s, and the impacts we expect to find are substantially increased and growing risks of death to infants, young children aged one to four, and middle-aged adults.

The results of the statistical tests for trends presented in the previous section are disappointing in that they did not clearly reveal HIV-like changes in the risk of death. It is possible that the data are broken down into cells so small that statistically significant changes could not be measured. To address that potential problem another approach is adopted. The annual hazard of death experienced during a defined age group is modeled as a polynomial function of a continuous variable coding calendar year – the so-called fractional polynomial model that has already been used to examine the age pattern of fertility above. The advantage of this approach is that it estimates the magnitude and significance of the trend using all of the data for a given age group instead of breaking it down into smaller cells. The model is specified as a binary logistic regression as follows:

$$\ln\left(\frac{p(year)}{1-p(year)}\right)=\beta_0\alpha+\beta_1\gamma+\varepsilon$$

Where p is the probability of experiencing a death and *year* is a continuous variable representing calendar year. The α and γ variables represent transformations of the year variable. STATA's *fracpoly* routine estimates the model with a wide variety of transformations and chooses those that maximize the quality of the overall fit. The transformations chosen in that fashion for each age group are listed below in Table 51.

Equation 18: Specification for the Fractional Polynomial Hazard Model of Mortality Through Time

TABLE 51: TRANSFORMATIONS SUBSTITUTED FOR α AND β IN EQUATION 18: $x = \frac{year}{10}$

Age Group	Male		Female	
	α	γ	α	γ
0	x^{-2}	$x^{-0.5}$	x^{-2}	$x^{0.5}$
1-4	x^{-2}	$\ln(x)$	x^{-1}	x^{3}
20-24	x^{-2}	$x^{0.5}$	x^{-2}	$x^{-0.5}$
25-29	x^{-1}	x^{3}	x^{-1}	x^{3}
30-34	x^{-2}	$x^{-0.5}$	x^{-2}	$x^{0.5}$
35-39	x^{-2}	$\ln(x)$	x^{-2}	$x^{-0.5}$
40-44	x^{-2}	$x^{-0.5}$	x^{-2}	$x^{-0.5}$
45-49	x^{-2}	$\ln(x)$	x^{-2}	$\ln(x)$

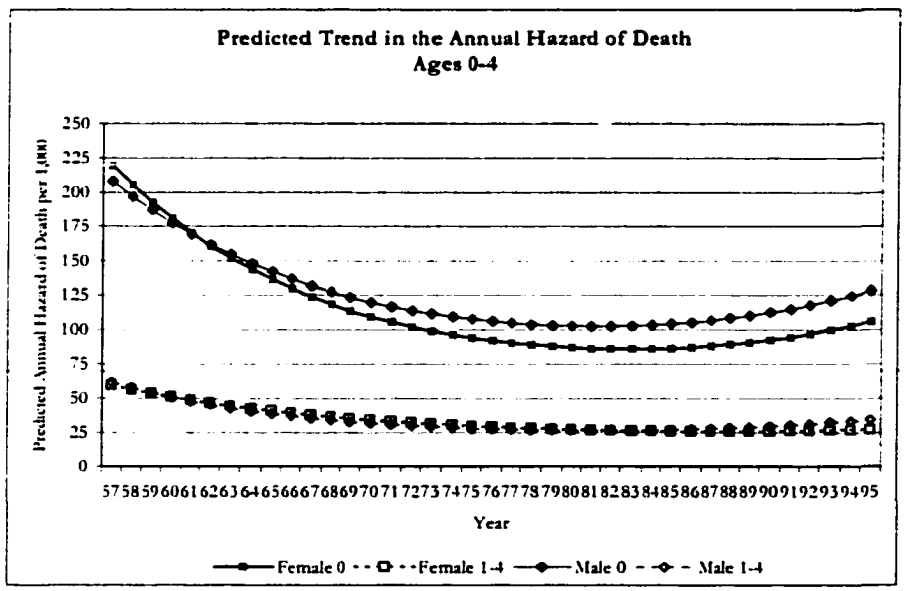

Figure 88: Predicted Trend in Annual Hazard of Death Ages 0-4

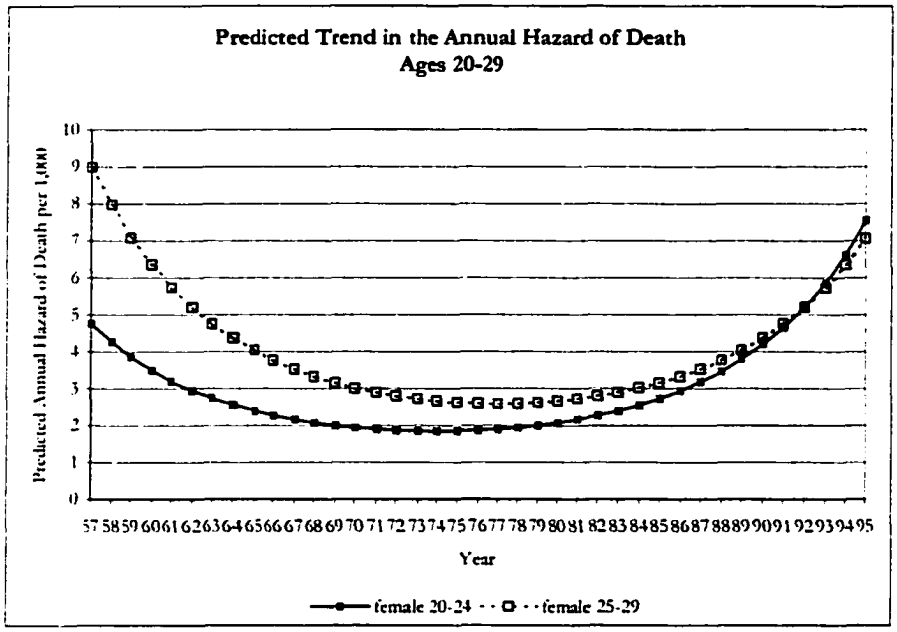

Figure 89: Predicted Trend in Annual Hazard of Death Ages 20-29

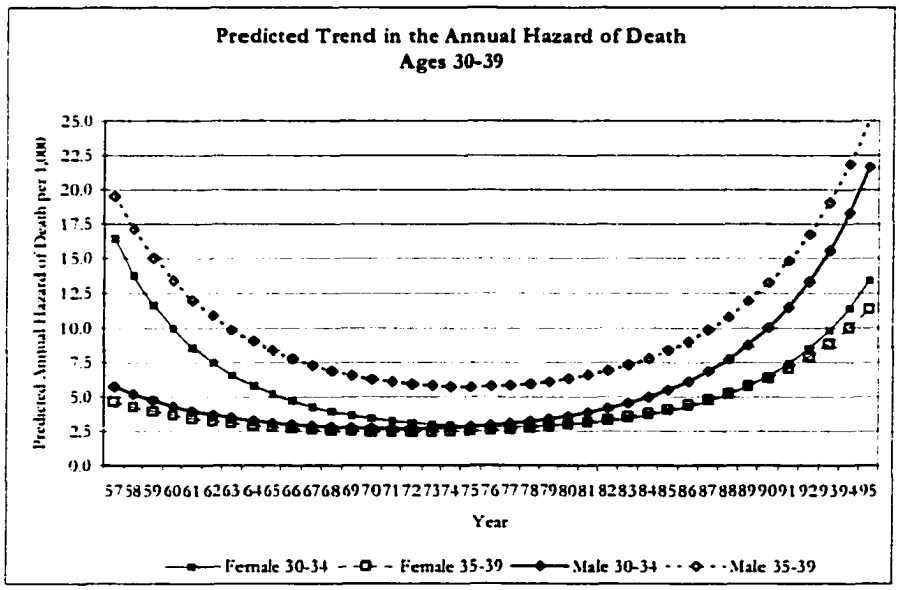

Figure 90: Predicted Trend in Annual Hazard of Death Ages 30-39

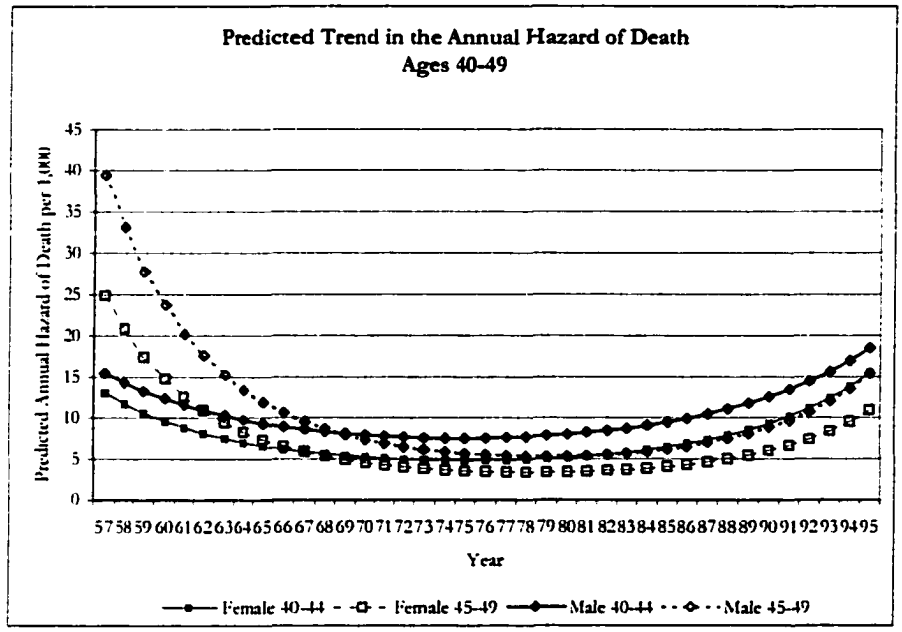

Figure 91: Predicted Trend in Annual Hazard of Death Ages 30-39

The model is estimated using STATA's *logistic* routine in conjunction with the *fracpoly* routine applied to the Person-Year Data Set. The unit of analysis is a person-year which makes the resulting predicted probabilities refer to the annual hazard of death as a function of year. Table 52 contains the P-values corresponding to the coefficients on the transformed year variables for all of the models estimated.

TABLE 52: P-VALUES FOR THE YEAR EFFECTS RESULTING FROM ESTIMATION OF EQUATION 18

Age Group	Male		Female	
	α	γ	α	γ
0	**0.001**	**0.001**	**0.001**	**0.001**
1-4	**0.001**	**0.001**	**0.047**	**0.051**
5-9	0.910	0.940	**0.096**	**0.093**
10-14	*0.143*	*0.142*	0.342	0.345
15-19	0.338	0.336	0.704	0.706
20-24	*0.203*	*0.203*	**0.074**	**0.073**
25-29	0.997	0.982	**0.068**	**0.068**
30-34	**0.048**	**0.047**	**0.004**	**0.004**
35-39	**0.015**	**0.014**	*0.200*	*0.197*
40-44	*0.196*	*0.195*	**0.109**	**0.109**
45-49	**0.038**	**0.038**	**0.055**	**0.056**
50-54	0.462	0.460	*0.110*	*0.107*
55-59	**0.069**	**0.069**	0.793	0.791
60-64	0.538	0.539	0.248	0.258
65-69	0.303	0.300	0.866	0.872
70-74	0.617	0.617	0.209	0.212
75-79	-NA-	-NA-	0.765	0.767

Bold ⇒ significant to highly significant
Italic ⇒ mildly significant

So what does all this mean? The transformations of the year variable give the trend its shape, and the coefficients on the transformed year variable determine how much each transformation contributes to the final shape of the trend. The significance tests on those coefficients test the null hypothesis that they are equal to zero and thus indicate how likely it is that the value they take could have been the result of a random process. Because the coefficients represent how much of each component shape is necessary to create the final shape, if they are non-negligible in magnitude (all of them are) and statistically significant, we can comfortably assume that the shape of the predicted curve represents something real – in this case, the trend in the annual hazard of death in a given age group.

The P-values presented in Table 52 reveal a nice age-pattern. The trends fit for ages 0 to 4 and generally for ages 20 through 49 are significant, meaning that a meaningful relationship

between the annual hazard of death and time period exists. The significance of the time period effects for other ages is negligible or not at all significant. These findings indicate that there probably are long-term trends in the annual hazard of death for ages 0 to 4 and 20 to 49, while there are probably not any significant long term trends in the annual hazard of death at other ages.

Having established real trends for ages 0 to 4 and 20 to 49, the next question is how exactly the annual hazards have changed over time. Figure 88 through Figure 91 display the shape of the trends in annual hazards for those ages. According to the predicted trends mortality conditions improved for both infants and young children through the middle 1970s. After that the risk of death faced by young children remained constant while the risk of death faced by infants began to increase during the late 1980s and early 1990s, exactly the time when the AIDS epidemic began to grow in Zambia. This is what we expect as an HIV/AIDS begins to have an effect. The ongoing downward trend in the risk of death is halted for young children and actually reversed for infants.

The trend for young adults is only significant for females, and the shape indicates a substantial reversal of what was a steep downward trend in the risk of death. The turnaround coincides with the time when the AIDS epidemic began to grow in Zambia, and the magnitude is impressive. This supports the finding elsewhere hat the age-pattern of HIV infection is typically younger for females.

Both males and females exhibit significant and substantial trends at ages 30 to 39. In both cases the risk of death has increased substantially from the mid 1980s onward. At the risk of

becoming boring, it is worth noting again that this reversal coincides with the onset of the AIDS epidemic in Zambia. At the beginning of the time period, there appears to have been a dramatic improvement in mortality conditions for males 35 to 39 and females 30 to 34. This supports Colson and Scudder's assertion that the forced relocation that occurred in 1957-58 was most stressful for middle-aged adults who bore the brunt of the emotional, physical and financial responsibility of successfully moving their households.

The trends for both males and females aged 40 to 49 are similarly significant and important. As with the 30 to 39 age group, there was an immediate improvement after relocation followed by a long plateau and a deterioration beginning in the late 1980s when AIDS started to have an effect.

Taken as a whole, these results firmly support the conclusion that there have been significant and substantial increases in the risk of death in precisely the age groups that may be affected by HIV/AIDS. Additionally, there have not been significant or substantial changes in other age groups that are not typically affected by the disease. The conclusion is that HIV/AIDS may have began to kill the Gwembe Tonga in significant number during the late 1980s and early 1990s.

Part 3

A TWO-SEX STOCHASTIC MICROSIMULATION OF A POPULATION WITH HIV

BACKGROUND

MODELS OF HUMAN POPULATION

Perhaps the simplest model of human population is the parsimonious two-parameter exponential growth model that predicts the overall size of a population as a function of its original size and its proportional growth rate. From there, models become quickly more complex and elaborate. For example, Lotka's (Lotka 1939) elegant formal description of a stable population describes how population renewal and depletion processes work together to determine the overall size and age-structure of the population and provides the basis for a deeper understanding of fundamental population dynamics. Lotka's model is specified with respect to a one-sex population and draws insightful relationships between the age-specific rate of reproduction, age-specific survival probabilities, and the growth rate of the population. It does not attempt to describe the vital dynamics of another sex, the pairing processes that must operate to bring the sexes together to reproduce, any of the biological processes that underpin reproduction and survival or any other important processes affecting population dynamics. However, by sacrificing that additional richness, Lotka's model remains tractable and the relationships he derives can be manipulated analytically to yield additional insights.

More differentiated models that take into account both sexes and attempt to model behavioral and biological processes quickly become mathematically intractable and only able to yield numerical results. Some models of this type, such as Anderson and Garnett's models of populations with HIV (Garnett and Anderson 1994b; Garnett and Anderson 1996b) are specified at the population level by a set of differential equations. In most cases it is not possible to formulate analytical solutions to these equations, and consequently numerical techniques are used to investigate specific conditions. These usually proceed by specifying the parameter set so that it broadly reflects the reality that one wants to investigate and then using a powerful computer to numerically solve the differential equations or simulate the population governed by those equations as it evolves through time. Beyond not being able to "solve" these systems or even manipulate them analytically to a large extent, the computational complexity and power required to conduct the simulations is considerable, and finally and most importantly, a large number of behavioral and biological aspects of the system that relate directly to individuals cannot be modeled explicitly because the systems are defined relative to the population as a whole.

This leads us to the last general category of population model – the *individual-level, stochastic* model. These are specified with respect to individual people who are stepped through time and exposed to the risk or experiencing various events during each time step. Whether or not the events take place during a given time step is determined by comparing each individual's probability of experiencing each event for which they are at risk to a random number; if the random number is less than the probability of the event, the event takes place. The most prominent and also perhaps the longest-lived example of this type of model is the SOCSIM

microsimulator developed by Eugene Hammel and Kenneth Wachter and supported by the Demography Department at Berkeley (Wachter Accessed 2001). SOCSIM contains a full demographic model of a closed population; meaning that marriage (male/female pairing), reproduction (birth), death, and migration are all modeled well with numerous behavioral influences taken into account. It is "closed" in the sense that all partners must be drawn from within the existing population and cannot come from outside (or to generated). SOCSIM has been successfully used to examine a range of issues relating to household dynamics and the evolution of kinship networks over time, see for example (Lin 1995; Wachter 1995; Wachter, Knodel, and VanLandingham 2000).

Population models have been reviewed and critiqued by many, see (Coale and Trussell 1996; Palloni and Glicklich 1991; Srikantan 1982). The various reviews identify two broad themes along which models can be differentiated: 1) the substantive issues they are designed to address, and 2) the methodology they employ. Srikantan (Srikantan 1982) identified four categories of model: 1) structural equation models, 2) macro models, 3) micro models, and 4) analytical models. This classification appears to be primarily based on the technique employed by the model. Palloni and Glichlich (Palloni and Glicklich 1991) divide models into three broad groups: 1) extrapolation and curve fitting models, 2) back calculation models, 3) biological and behavior models. This scheme relies on both the technique and the substantive nature of the models to classify them. Last Coale and Trussell (Coale and Trussell 1996) review population models specifically in the context of Demography and draw broad distinctions between models that are purely mathematical, those that summarize empirical observations, those that are used to predict future population outcomes, and those that aim to

elucidate the operation of population processes at the individual level. The emphasis here is on the substantive nature of the model and what it is used to explore.

CLASSIFICATION OF POPULATION MODELS

Adding to these classifications, I believe there are three important dimensions on which models can be classified.

1. Unit of analysis:
   a) macro, or
   b) micro.
2. Mode of operation:
   a) deterministic, or
   b) stochastic.
3. Substantive aim of to which the model is directed:
   a) to identify and manipulate an empirical regularity, and/or
   b) to measure some aspect of a population process, and/or
   c) to understand and summarize the past, and/or
   d) to predict or project some measure of a population into the future, and/or,
   e) to understand complex, inter-related population processes, and/or
   f) to test individual and population-based interventions.

Macro models are specified with respect to populations while micro models are specified at the level of the individual. A deterministic model is one whose operation is predictable while a stochastic model operates in a probabilistic manner that is not predictable on the level of the unit of analysis. Macro models can be constructed to operate in either deterministic or stochastic modes or both, while truly micro models can only operate stochastically.

Macro models are well suited to explore substantive questions falling into all six categories enumerated above. Their one critical limitation is that they are not able to describe or examine truly individual-level characteristics such as behavioral and biological attributes. Mainly

because of their complexity and lack of analytical results, micro models are best suited to examine those questions that rely heavily on specific individual-level attributes, again usually behavioral and biological. Micro models are not well-suited to examine the past, which usually involves cumulating some sort of parameter; or to predicting the future, because they must actually generate an entire future population. The predictive value of micro models is also limited by the large number of parameters that they require as inputs, each of whose future trend must also be predicted. The combined uncertainty in the predictions of all of the parameters is substantial and does not compare favorably with simpler models.

MACRO AND MICRO MODELS

MACRO DETERMINISTIC MODELS

The unit of analysis in large part determines the structure of the model and the types of reality that it can model. "Macro" models are specified with reference to groups of individuals and make the implicit assumption that all individuals within a group are homogenous in all respects. A direct consequence is that all interactions take place between homogenous groups – not between individuals. The movement of "individuals" from one group to another, or between states, is usually governed by a set of differential equations that determine what fraction of a group in a given state will move to another state over a specified period of time. These movements can be modeled as either deterministic or stochastic processes, although in practice the specifications are almost always deterministic. However, the models as a whole do not have to be deterministic in the traditional sense, but may comprise a chaotic system. The analytical specification is a big advantage of macro models; the expressions governing the model can sometimes be solved to yield a set of equations describing the behavior of the

solution (a "steady" or "equilibrium" condition) to the model – some aspect of the model that is not specified but can be derived from those aspects that are specified. These models are often too complex to solve, but even in those cases, the governing expressions can often be manipulated to yield additional non-trivial insights into aspects of the model that are not explicitly specified. Macro models are ideal for studying processes that can be specified on the population level and do not explicitly rely on individual-level behavior, preferences or biology. Their standard mathematical specification also makes them comparatively easy to understand, implement and interpret.

Macro models can model the past to understand the present through various techniques including cumulation and iterative estimation of the parameters necessary to correctly cumulate the past to arrive at the present, and they can also model the future in a predictive sense based on a set or reasonable parameters. Their complexity may range from very simple to extremely complex; one of their advantages is that they can ignore individual-level behavior, and in comparison to individual-level models, are generally more simple. An important facet of this comparative simplicity is the fact that they do not need as much information to run and can operate on a much smaller parameter set than most individual-level models. For example, they are often specified with respect to one sex only, and they can model whole subgroups of a population as a single unit. The results that they produce usually give a broadly realistic but not very specific understanding of a process. These kinds of results are good for overall predictions of population size and growth, the general outlines of age structures and gross movements between various groups. What they are not able to produce are detailed understanding of systems involving individual-level behavioral and biological processes.

Finally, macro models are implemented in a wide array of technologies ranging from simple calculator-based calculations to spreadsheets to sophisticated customized software, reflecting the wide range of complexities that are possible with macro models.

EXAMPLES OF MACRO MODELS ADDRESSING HIV/AIDS

A small selection of macro models addressing HIV/AIDS is discussed below with an attempt to present a range of substantive foci and methodologies. A now somewhat dated, but nonetheless comprehensive and important, review of population models examining the impact of HIV/AIDS is contained in the special UN volume The AIDS Epidemic and its Demographic Consequences. Proceedings of the United Nations/World Health Organization Workshop on Modelling the Demographic Impact of the AIDS Epidemic in Pattern II Countries: Progress to Date and Policies for the Future, New York, 13-15 December 1989 (United Nations Department of International Economic and Social Affairs and World Health Organization Global Programme on AIDS 1991). Much of the material below is based on this publication, and the fact that not a lot more work has been done in this field since that time is somewhat surprising. All of the models below in this section are macro models and thus do not model individuals, individual coital acts, individual-based transmission of HIV, or individual behavior and biology as it relates to transmission and disease history within individuals. As a result, all of these models are limited in their ability to examine the primary processes responsible for the transmission and incubation of HIV, and they are fundamentally different from the type of model that I have built.

HEUVELINE

Heuveline (Heuveline 2001) extends the cohort-component population model, most commonly used as the basis for population projections, to model a population with HIV. To the usual age-specific cohort-component model, he adds sex-specificity and a number of HIV-related states through which the population is moved. The population is divided into HIV-infected and non-HIV-infected, and as time steps forward, the non-infected are infected according to a parameter-dependant incidence rate (both horizontal and vertical). HIV status affects both fertility and mortality, and some degree of duration dependence is supported. A relatively parsimonious set of parameters governs the HIV-related behavior of the model. Realistic values for these parameters are obtained through a maximum likelihood procedure that identifies the parameter set most likely to reproduce a range of incidence ratios and fertility impairment indices. The resulting parameters allow the model to reproduce, as a group, the set of observations over which the parameters are estimated, and consequently to calibrate the model to behave in a fashion similar to the populations that provided the observations. Heuveline suggests using the model to examine the evolution of and relationship between population-level indices of the HIV epidemic and the overall size and composition of the population. For example, the future course of mortality and fertility rates, overall population size, growth rates, and the relationships between the prevalence of various subgroups within the population.

This model is relatively simple and elegantly specified, and as a result is widely useful and quite general. Its parameter set is static (non time-evolving) which may limit its ability to predict the

future, but nonetheless, it is a useful tool for understanding and estimating the overall impact of HIV/AIDS on a population.

Although none of his models appear to be published in accessible sources, Artzrouni has also pursued a similar approach to Heuveline's beginning during the late 1980s and early 1990s. He also developed multi-state variations of the standard cohort-component model, but he does not appear to have taken the work as far as Heuveline.

IWG – The United States Interagency Working Group on AIDS Models and Methods

The IWG model was created by the United States Interagency Working Group on AIDS models and Methods beginning in the late 1980s (Stanley, Seitz, Way, Johnson, and Curry 1991). The model has been continuously refined since then, but a detailed description of these refinements could not be found as of this writing.

The IWG model is a deterministic model specified using a set or partial differential equations. The level of analysis is the population rather than the individual. It takes into account a bewildering array of behavioral and biological processes affecting the transmission and incubation of the HIV, and it is uniquely comprehensive in this regard.

All important modes of transmission in the developing world setting are addressed: horizontal, vertical and through blood transfusions, and the progression of the disease within infected groups is also modeled as a duration-since infection-dependent process.

Demography is fully modeled as two-sex and age-differentiated, and the population is split into urban and rural groups that can modeled differently. However, the IWG model does not attempt to correctly model couples, opting rather to divide males and females into single and paired groups with various infection statuses. This is done to avoid the complexity of properly modeling a polygynous marriage system, and in our view is a major weakness in the demographic component of the IWG model. As with the Anderson models (below), the IWG model resorts to a relatively ad-hoc method to balance sexual partnership formation – an unappealing aspect of the model that potentially introduces an unknown bias into the demography of the modeled population. Unlike most of the other models, an attempt to address polygyny is made, although it is rudimentary at best. Additionally, the model does not appear to take into account the direct impact of HIV infection on fertility, nor does it tightly couple conception (fertility) and the transmission of HIV because individual coital acts are not modeled.

This model is unique in that it attempts to address the impact of migration on the growth and development of and HIV epidemic. Local circular and permanent migration are addressed as is international migration.

It is also among a small number that make a comprehensive attempt to model STD cofactors associated with the spread of the HIV. The STDs are spread through the same sexual contacts that transmit AIDS, and the presence of STDs modifies the transmission rates of the HIV.

The IWG model also models sexual activity within partnerships as a function of the duration of the partnership assuming that longer term partnerships are associated with less casual sex on the part of their members.

As with all the macro models described in this section, the IWG model does not correctly recognize or model the absolute link between conception (fertility) and, the transmission of HIV and other STDs, and the affect of condoms that limit all three simultaneously. All of these processes are mediated through individual sexual contacts and cannot be de-coupled if the aim is to satisfactorily model both the demography and the epidemiology of a population with HIV.

The IWG model has been used to study a wide range of issues revolving around understanding the inter-relationship of the processes that contribute to an HIV epidemic. Because of its detail, it can be used to study the impact of interventions, to make medium range projections, to study changes in the age structure and dependency ratios, and to examine the impact of the differences between urban and rural populations.

The initial results presented in the UN publication (Stanley et al. 1991) outline a familiar three scenario analysis that clearly demonstrates that an HIV epidemic is difficult to initiate and maintain, but once a threshold is passed, an epidemic grows quickly and has substantial impacts on the demography of the affected population. In hindsight, the "worst case scenario" that is presented appears to be the one that is playing out in Southern Africa a decade later. That scenario predicts a flattening of population growth and an overall

prevalence of about 40 percent prevalence, not too different from what appears to be building in Zimbabwe, Zambia, Botswana and South Africa.

ANDERSON, GARNETT & GREGSON

Anderson, Garnett, Gregson and their colleagues have developed and refined a significant macro model (and many variations on it) of an HIV infected African population (Anderson, May, and McLean 1988; Anderson, Ng, Boily, and May 1989; Anderson 1991a; Anderson 1991b; Anderson, May, Boily, Garnett, and Rowley 1991; Anderson, Ng, Rowley, and McLean 1991; Garnett and Anderson 1993a; Garnett and Anderson 1993b; Garnett and Anderson 1995; Garnett and Anderson 1994b; Garnett and Anderson 1996b; Gregson, Garnett, and Anderson 1994a; Gregson, Garnett, and Anderson 1994b; Gregson, Garnett, Shakespeare, Foster, and Anderson 1994). In broad terms, their work seeks to understand the demographic impact of HIV/AIDS over the medium and long term and to identify the parameters or processes that are most sensitive in determining the progression and overall impact of an HIV epidemic in populations where the modes of transmission are predominantly heterosexual and from mother to child.

The most refined of the Anderson models are specified as deterministic, semi-continuous age and sex-specific epi-demographic models of human populations with HIV using a set or partial differential equations. They take into account a range of behavioral and biological processes at the population level. Individuals are divided into various sexual activity classes that determine their overall propensity to engage in sex, and males and females are paired to form unions in which both reproduction and transmission of HIV occur based on various

partnership acquisition schemes. Individuals are not modeled so neither are individual coitus events or the direct transmission of HIV between individuals. Additionally, because individuals are not modeled, the pairing of males and females requires various sophisticated "fudges" in order to insure that the number of new couples matches the number of available males and females and their union formation preferences. This limitation is unsatisfying as it makes necessary a range or relatively arbitrary techniques to satisfy the partner balancing requirement. Transmission of HIV from mother to child and the natural history of HIV disease within infected individuals is modeled. Numerical techniques are used to simulate populations under different assumptions regarding the patterns of sexual mixing and sometimes other disease and demography-specific parameters as well.

Their early work made the dire prediction that the HIV/AIDS epidemic in Africa had the potential to bring overall population growth rates down to zero or even below zero. These and other similar warnings elicited widespread criticism at the time they were made and may have done some harm to the credibility of modeling exercises of this kind. Although growth rates have not yet turned negative in any country as of this writing, they have been substantially reduced in countries such as Zimbabwe, Botswana and Kenya. However, this is probably not the sole result of the HIV epidemics in those countries, but a complicated result of the interaction between the HIV epidemic and secular declines in fertility – interactions that the Anderson models have not fully addressed.

What these models have done well is to clearly identify the fact that the growth and form of an HIV epidemic is sensitively dependent on the nature of the sexual mixing in the population.

How much interaction occurs between males and females of different sexual activity classes, different ages, and how potential pairs decide to form are all powerful predictors of the overall evolution and magnitude of an epidemic. In particular, they have clearly demonstrated that the form of sexual mixing is critical; whether it is assortative or dissassortative, whether partners choose each other based on the raw numbers of available partners or on the proportion of available partners compared to the total population of a given type.

Additionally, these models have examined the role of the duration of the incubation period on the overall development of an HIV epidemic. They have conclusively demonstrated that the duration since infection-dependence of the incubation of the HIV within an infected individual is critical in determining the pace and penetration of the spread of an HIV epidemic. They also confirm the transparent importance of the transmission likelihood in governing the development of and epidemic. In the final analysis, the Anderson models have conclusively shown that the most important parameters governing the development and spread of an HIV epidemic are those that are under social or behavioral control: partnership formation, frequency of sexual contacts within partnerships, and transmission probabilities per sexual contact. Additionally, various biological parameters such as the individual progression of the disease and associated changes in infectivity are also important.

Anderson and his colleagues have been building sophisticated models for a long time, and over the course of the 1990s they developed a powerful and useful model. However, their early results were based on much simpler models that did not fully elaborate the age, sex and behavior heterogeneity within the modeled population, and as a result, their early predictions

of a an imminent demographic collapse for populations with endemic HIV were exaggerated and possibly contributed to an overall skepticism of modeling as an appropriate tool to examine complex processes involving social, epidemiological, demographic and biological processes. This serves as a cautionary tale and reminds us to fully acknowledge and as faithfully as possible reflect the full complexity of the systems I choose to model. The early work of John (below) and others using more age, sex and behavior class differentiated models questioned the alarming findings of the Anderson group, and time (and the later Anderson-derived models) have born out the more sober findings of the more differentiated models.

At this time Anderson and his group are beginning to address other issues such as the evolution of the demographic structure of a population with HIV, changes in the dependency ratios and number of orphans. They are also pursuing an interesting new line of work modeling the evolution of genetic variation within the population of HIV virions within a single infected person (Anderson 2001).

JOHN

John (John 1991) develops one of the earliest macro models with a fully specified age and sex-specific demography and several behavioral classes. The model examines a wider range of modes of transmission than most including intravenous drug use and blood transfusions. It also includes explicit modeling of extra-marital sex and bases heterosexual transmission on a per-coitus transmission probability. An acknowledged limitation is the fact that an individual's infectivity is constant with duration since infection. Another substantial limitation is the lack of an HIV-related impact on fertility, a result of HIV infection that was not well-documented

at the time when this model was formulated. Altogether, the model appears to reproduce the demography of a population quite well and obtains realistic estimates of the progression and growth of an HIV epidemic. The key finding from this fully age and sex specified epi-demographic model is that it is unlikely that an HIV epidemic is able to turn overall population growth rates negative or even very close to zero in a rapidly growing population.

PALLONI

Palloni adopts a variation of the Demographer's multistate life table (Palloni and Lamas 1991). His variation adds duration-dependence to the progression of the disease and corresponding infectivity of individuals in addition to age-dependency for most processes. This is a macro model that defines a number of "compartments" through which groups of individuals flow according to various transition probabilities, many of which are duration-dependent.

The model is relatively simple, differentiating three stages in an individual's infection with HIV: healthy, infected/non-symptomatic and AIDS. Transitions to the absorbing state, death, occur from all three states, and groups of individuals move through the states in a duration since infection-dependent fashion. Modes of transmission include horizontal, vertical and blood transfusion related infections. A lot of thought is given to the manner in which men and women are divided among sexual activity classes, prostitution is included as a sexual activity class, and a great deal of consideration is given to how people in different sexual activity classes interact.

As with other macro models, balancing male/female pair formation is handled with an somewhat arbitrary process, and "steady" unions are not allowed to end in any other way

except death of a spouse. These limitations seriously inhibit the model's ability to realistically model the diverse heterogeneous and complex sexual mixing patterns that exist in large parts of Africa, and other parts of the developing world where HIV/AIDS is a significant problem.

The description of the model is vague concerning the renewal of the population – fertility. It is not clear how fertility is handled or whether or not it is linked to the sexual activity that is modeled to produce the transmission of HIV. It appears that fertility processes are not linked to the transmission of HIV or to the HIV status of the partners contributing to a reproductive relationship.

The major innovation in this model is the duration-dependence of the transitions between the uninfected, infected/asymptomatic and AIDS stages of the disease. The results appear to indicate that there is a significant threshold somewhere in the parameter space that pushes an epidemic from being non-sustaining to full blown and serious with relatively small movement through the parameter space. All but the most "infective" of the simulations presented result in dying epidemics, while the highly infective simulation generates a catastrophic epidemic that results in negative population growth and very substantial prevalence of around 80 percent.

Some of the other models hint at a threshold effect, so some form of a threshold probably does exist. However, the severity of the high infectivity simulation castes some doubt on the model, and with hindsight, no epidemic of that severity has yet to develop – a reality that leads one to question the overall usefulness of this modeling approach.

BULATAO

Bulatao constructs an interesting model that is a synthesis of the standard cohort-component model (discussed above under Heuveline) and an epidemiological model of the transmission and incubation of the HIV in the infected population. This model is specified at the population level and does not attempt to model the dynamics of union formation and separation. As such it shares the major limitation of many of the macro models in that it is not able to capture the behavioral component governing the choice of sexual partners. However, like the Heuveline model it retains an elegant simplicity which frees it from the substantial data requirements of more sophisticated models and allows it to be powerful and tractable at the same time.

Although much of the behavioral complexity is removed from this model, it does comprehensively manage the transmission and progression of HIV between people of various sexual orientations. It is rare in including homosexual and bisexual sexual orientation classes and in managing the transmission of HIV between members of those groups. It models heterosexual, vertical, needle-born, and blood transfusion-mediated infections. Two different schemes for modeling the progression of HIV in the infected population are adopted: non-staged progression with a constant infectivity and staged progression with step-wise varying infectivity. The progression to AIDS and eventually death is modeled as a logistic function of the time since infection. The population can be further divided into various levels of susceptibility based on genetic variation. The transmission of the HIV is governed by an "attraction matrix" that governs the "contact" between various subgroups of the population defined by age, sex, sexual orientation and potentially other characteristics. The

transmissibility of the virus depends on the type of sexual activity that occurs between people of various subgroups, and transmission cofactors such as condom use and presence of STDs are acknowledged.

The epidemiological processes described above are updated on an annual basis on a population whose overall demography is updated on a five-year basis by the standard cohort-component methodology. At the end of each five-year period the two "projections" are reconciled and dovetailed to create the base population for the next five year's worth of projection.

The Bulatao model appears to have produced reasonable descriptions of the HIV epidemics as they have subsequently unfolded. The three typical high/medium/low scenarios are presented, and in none of them do growth rates become negative. Life expectancies are reduced by up to ten years or so in the worst case (which is about what has happened in Southern Africa a decade later), and prevalence rates are predicted from nearly zero to as high as 55 percent or so (also in line with what has come to pass). The model clearly demonstrates the importance of "staged infectivity" in determining the epidemic, with a more nuanced staging leading to more slowly growing epidemics. Judged in hindsight, this appears to be a good model whose tradeoffs between complexity and tractability were well chosen.

Bulatao demonstrates a use for models of this type that is not commonly recognized; namely that useful relationships can be identified by comparing the large number of indices that a model of this type is able to generate; for example, the relationship between seroprevalence and percent condom use.

BONGAARTS

Bongaarts (Bongaarts 1989) constructs a model similar to the Bulatao model above that mixes the Demographer's traditional cohort component projection model with an epidemiological submodel that divides the population along various epidemiological and behavioral lines. This is a fully age and sex differentiated macro model whose epidemiological component is specified as a set of linear differential equations. Similar to the Bulatao model, it appears to produce reasonable descriptions of an AIDS epidemic, and it does not predict a massive population collapse through negative growth.

BROUARD

Brouard constructs a very simple "two renewal equation" model (Brouard 1991) in which the two renewal processes modeled were the birth of new individuals and the "birth" of new HIV infections within the population of individuals. He argues that there is insufficient information on which to build any more complex models, and that the substantial lack of information relating to the two key parameters of infectivity and sexual partnership change (or acquisition) is so great that modeling exercises are essentially not valid.

The model he presents clearly does not capture in any realistic way the behavior of an AIDS epidemic, even more clearly with a decade's worth of hindsight. His criticism of the lack of information was certainly valid at the time it was made, but his vision for modeling, even with the lack of all necessary input data, is limited. Most of the other models presented in the U.N's publication (United Nations Department of International Economic and Social Affairs and World Health Organization Global Programme on AIDS 1991) were able to provide

significant insight beyond the fact that not all of the necessary parameters were known with sufficient confidence.

LESSONS FROM MACRO MODELS

The previous examples are by no means a comprehensive survey of macro models and have been chosen mainly for their focus on HIV/AIDS and the existence of relatively lucid descriptions of how they operate. Nonetheless, it is clear that overly simplified models are of limited value when the underlying processes are complex and inter-related. The Brouard model is a good example of a very simple model that failed to capture the behavioral and biological nuances of the system and therefore produced results of little value. At the other end of the spectrum are the very complex models best exemplified by the IWG model. Those models are so complex that it has not been possible (or maybe not desirable) to fully explicate their inner workings, thereby leaving the reader with many questions relating to what their results really mean, and from which set or parameters or modeled processes differences in outcomes are produced. The utility of models of this type is limited by there overbearing complexity and opacity. The Anderson models are a fortifying case study that appear to have attained a useful middle ground. The original models in that series were clearly unable to adequately capture the nature of the processes that they hoped to model, and the results they produced were correspondingly exaggerated and of diminished value. However, as the models were carefully enhanced and further differentiated over time, they became complex enough to adequately capture the nature of the system without becoming completely unwieldy and indescribable. There are several lucid descriptions of the Anderson models that allow the

reader a reasonable glimpse of their inner workings; this in contrast to the comprehensive paucity of revealing description of the IWG models.

It is clear from the small collection of macro models described here that there is no one "correct" methodology or procedure. Very different methodologies and even creative syntheses of rather different methodologies created similar, realistic results. It appears that the best predictor of success, given a reasonable level of detail to begin with, is the skill with which the modeler is able to implement the methodologies most familiar to him; Demographers use some variant of the cohort-component projection method while epidemiologists use some form of compartmentalized flow model specified with partial differential equations, and those inclined to more actuarial disciplines apply multistate life tables and other variations of state transition models. It does not appear that any methodology is superior, and indeed mixtures of two or more appear to work as well as purely one or another.

MICRO STOCHASTIC MODELS

In contrast to macro models, "micro" models are specified with reference to individuals, with the assumption that each individual can be different and that individuals interact with each other directly – not through groups of identical individuals. An immediate consequence is that the potential complexity of a micro model is enormous. Any one individual has the potential to interact with every other individual in the modeled population, and the potential heterogeneity of the population of individuals grows very rapidly as the number of individual-level attributes increases. Because interactions between individuals are the focus of models of this sort, and because individuals are indivisible units, micro models cannot be expressed

analytically in terms of what *fraction* of an individual moves from one state to another or combines with what *fraction* of another individual. Consequently the models must be specified in terms of stochastic processes that describe the probability that a whole individual will transition from one state to another or that whole individuals will join in social relationships.

Micro models are the modeling methodology that is appropriate when it is necessary to examine systems in which individual behavioral and biological processes play an important role. These include diseases whose natural history is substantially differentiated between individuals and/or whose modes of transmission depend on individual-level attributes or individual behavior. Sexually transmitted diseases fall into this category because their transmission is the result of an activity that is regulated by individual behavior and social pressures exerted on an individual by a larger group – sex. AIDS in particular meets both criteria. The disease is transmitted through sex which is a highly differentiated individual-level behavior, and the natural history of the HIV is different within individuals and highly dependant on individual behavior and health over the usually long course of its incubation. Another example of a category of system that requires micro models is the modeling of kinship, households and other social aggregates whose composition depends on individual-level behavior and biological processes. Much work has gone into this type of model over the past three decades at Berkeley resulting in the SOCSIM individual and household simulator, more below.

Micro models are always complex and require the specification of a comparatively vast number of parameters. This is a substantial disadvantage and means that they can only be specified in

the small number of circumstances where sufficient empirical information is available to define literally tens or hundreds of parameters – in some cases even thousands! Their complexity also limits their generality and sometimes makes their behavior hard to understand and interpret. For this reason they are not the method of choice for predictions or for the estimation of macro indices or statistics. Rather, they are more suited to understand the inner workings of a complex system, or to make predications regarding the future composition or rates of change and transformation within a population. This does not rule out their ability to make accurate predications of the overall future size and growth rate of a population, but such an endeavor requires correctly guessing many parameters as they evolve into the future, and as that number of parameters increases, the probability that one is correctly guessing all of them diminishes steeply.

What micro models are able to offer is finely nuanced understanding of the relationships shared by various processes and how the detailed composition of a population will change over time as a function of a large parameter set whose individual affects may be interacting in complex ways. This is particularly true when one process may affect many others in different ways, and all of those feed back to the original process to modify it. A good example is HIV/AIDS; infection with the HIV increases the probability of death, decreases the probability of successful reproduction, creates a probability of infecting the next generation, modifies one's standing in society and modifies various social interactions in ways that affect the transmission of the disease to others, one's ability to form reproductive unions, to travel, to work, to obtain health care etc. All of these effects work to change the overall size and composition of the population that in turn alters the number of people susceptible to

infection. The system is a very complicated one in which individual and aggregate-level processes interact through recursive feedback loops.

In order to handle this type of complexity, all micro-level models of which I am aware are implemented using customized software.

EXAMPLES OF MICRO MODELS ADDRESSING HIV/AIDS

In addition to the U.N's volume on population models relating to HIV (United Nations Department of International Economic and Social Affairs and World Health Organization Global Programme on AIDS 1991), in 1991 the National Academy of Sciences published two volumes (Citro and Hanushek 1991a; Citro and Hanushek 1991b) examining the role of microsimulation in social policy formulation. All of these volumes provide a comprehensive review of micro modeling techniques and good examples as of 1991.

WACHTER & HAMMEL – SOCSIM

The Wachter/Hammel micro simulator (Anonymous SOCSIM Accessed 2001a; Anonymous SOCSIM Accessed 2001b; Wachter Accessed 2001) is likely the most well known individual-level, stochastic population simulator. It is well designed and implemented, it can trace its life back in time for almost three decades, and most impressively, it has led to a long list of informative publications. Its current name is SOCSIM although it started life as POPSIM and matured through numerous versions before attaining its present implementation. Although the reference implementation of SOCSIM does not appear to explicitly model any aspect of health or disease, including HIV, it is described here because no discussion of population microsimulators is complete without it. Moreover, SOCSIM has been tweaked recently to

allow it to address some aspects of the HIV/AIDS epidemic in Thailand, work done in association with Knodel (Wachter, Knodel, and VanLandingham 2000).

SOCSIM transparently reveals the disciplinary affiliations of its two primary architects. It is a comprehensive model of a closed population with a great deal of sophistication surrounding both demographic and anthropological considerations. Its strongest feature is its faithful reproduction of household dynamics and its meticulous accounting for intrafamily relationships between individuals. Naturally, it has been applied in numerous circumstances to study the distributions of individuals within families and the evolution of family (household) structure over time as it is influenced by demographic factors, see for example (Wachter 1997; Wachter, Blackwell, and Hammel 1997), and for a complete selected bibliography (Anonymous SOCSIM Accessed 2001b). SOCSIM also allows an unlimited degree of heterogeneity (including location) to be built into the simulated population by allowing individuals to be members of different "groups" that share common characteristics.

The unit of analysis is the individual, and individuals are stepped through time on a monthly basis facing the risk of a number of events including death, marriage, migration and childbirth. The marriage module in SOCSIM uses a two-stage process to pair eligible males and females from within the simulated population. They first initiate a marriage "search" that effectively places them at risk to form a union with any of the members of the opposite sex who have also initiated a marriage search. Those who are searching are then paired through a random process contingent on the likelihood that each type of pair will form based on the individual characteristics of the potential spouses. SOCSIM is one of the only microsimulators for which

adequate, lucid documentation exists regarding the precise mechanism through which unions are formed, and it appears to be a very sensible mechanism that does not fall prey to the scaling problem and for which parameters can be obtained empirically.

SOCSIM allows several ways of expressing and modeling fertility but unfortunately does not provide a means to explicitly model coital events and their relationship to conception, fertility or disease transmission processes. This excludes SOCSIM from being able to effectively model the spread of sexual transmitted diseases, including HIV. Although, it is not likely to be too difficult to add this kind of facility to SOCSIM in the manner in which it is implemented in most of the other microsimulators that address AIDS – namely a fertility-independent allocation of coital events to partners in a union that are then used to generate disease transmissions based on per coital act transmission probabilities. However, as is argued below, this is not a satisfactory solution to the problem because it decouples fertility and the transmission of disease – it is a fundamental weakness of all of the models discussed here.

All transitions, including mortality, childbirth, migration and the events involved with unions, are handled through a clever and realistic competing risks procedure that draws waiting times from hazard distributions for each event and schedules the set of events for which each individual is at risk based on those waiting times. The shortest waiting time determines the event that occurs when two or more events compete.

SOCSIM has been used at least once to address an AIDS-related question, and true to its pedigree it addressed a question concerning the impact of HIV deaths on the household in Thailand. Wachter and Knodel (Wachter, Knodel, and VanLandingham 2000) use SOCSIM

to simulate a cohort of Thai adults (and their children) from age 50 onward in order to study the pace and degree to which the Thai elderly can expect to loose their own children as a result of HIV-related mortality. Because SOCSIM naturally maintains all of the intrafamily relationships that are necessary to study intrafamily dependency, it is an ideal tool to estimate the degree to which the elderly will have to care for and live without the support from their own children.

AUVERT – SIMULAIDS

Over the past decade Auvert and his colleagues have refined a comprehensive individual-level stochastic model of a population with HIV, resulting in *SimulAIDS* (Auvert, Buonamico, Lagarde, and Williams 2000; Auvert 1991). The most recent iteration is built around a standard individual-level stochastic framework that is minimally able to represent the demography, behavior and biology necessary to produce an HIV epidemic.

Individual coitus acts are not modeled so that conception (fertility) and the transmission of the HIV are decoupled. Additionally, the use of condoms is not modeled (although it was in an earlier version) so the simultaneous contraceptive and transmission-blocking effects of condom use are not linked or addressed. The mortality of the population is assumed to follow a standard life table, and after progressing through a staged HIV incubation process, infected individuals die one year after developing full blown AIDS. The probability of transmitting the HIV between males and females is assumed to be asymmetric and dependent on the type of union that the partners share, mainly to what sexual activity category the partners belong. STD cofactors are modeled and modulate the transmission probabilities of those coinfected

with an STD. Vertical transmission does not appear to be modeled. Once infected, individual's progress through a staged incubation process that effectively takes into consideration the duration since infection-dependent nature of many of the impacts of infection.

Of critical interest is how the model handles the pairing of individual males and females into sexually active unions. This is done by dividing the men and women into "long-term", "short-term" and "one-off" sexual activity groups. The rules governing how these groups mix are male-centric; meaning that the demand generated by men is met by women. The degree of sexual activity enjoyed by the couples depends on the sexual activity classes to which the couples belong; essentially, each couple is assigned a number of coital events per month that corresponds to the degree of sexual activity brought to the union by the partners.

This approach to modeling pairing and coitus is a substantial compromise between reality and ease of implementation, with the later receiving most of the weight. It is clear that in the real world women in almost all societies are able to exert substantial influence over the form of most of their sexually active unions, and it is not reasonable to assume that unions are entirely formed in a male-centric way. Also, the male-centric approach runs into the balancing problem described above in Examples of Macro Models Addressing HIV/AIDS. If there are insufficient females to "meet" the male demand, then some arbitrary correction must be made to allow the system to proceed, most likely that a randomly selected group of males go unsatisfied. This is at best an inelegant solution to the problem, and at worst an inaccurate portrayal of how things really work. In most circumstances this would not be a problem, but

where the central dynamics of the system under study depend sensitively (see above) on the exact nature of the sexual pairing between men and women, it is important to reproduce as faithfully as possible the dynamics and preferences evident in the real system.

Given that, the model is useful and has been positioned to answer one of the most fundamental questions surrounding the HIV/AIDS situation in the developing world; namely, to what extent to behavioral and biological processes contribute to the development of an AIDS epidemic, and when compared which type of process is more important, or do they share roughly equal responsibility? The clear finding is that behavioral parameters play a major role in shaping an AIDS epidemic, and that the behavioral factors alone can account for all of the observed variability in HIV prevalence in Sub-Saharan Africa.

VAN DER PLOEG – STDSIM

STDSIM is a comprehensive individual-level, stochastic simulation of a population with AIDS (Van der Ploeg, Van Vliet, De Vlas, Ndinya-Achola, Fransen, Van Oortmarssen, and Habbema 1998). It was developed primarily as a decision support tool for health care professionals, and as such it contains a number of elements that are unique. For example, it contains components to model the impact of the health care system in general and a separate component that specifically manages "interventions". This broadening of the scope of the model makes it more of a practical tool than many of the others discussed here.

Not surprisingly individual coital acts are not modeled, so like most of the other models discussed here, STDSIM does not tightly couple conception and the transmission of HIV.

However, almost every other aspect of the fundamental population and epidemiological dynamics are modeled well.

There are separate modules governing:

- demography,
- sexual behavior,
- STD transmission (including HIV),
- natural history of each STD,
- health care, and
- interventions.

The demography module includes complete modeling of fertility, mortality and migration. The sexual behavior module is described in a working paper that was not accessible for review (Van Vliet 1995) at this writing. This is unfortunate because it is arguably the most important component of any *sexually transmitted* disease model. What is revealed concerning this module is that partner selection is age-based and individuals are sexual activity graded, but it is not clear if an individual's sexual activity category affects their partner acquisition strategy. Sexual activity within unions is based on of the male, and it is again not clear whether an individual's sexual activity category affects his coital frequency. Because of the significant sensitivity to the exact form of sexual mixing demonstrated by Anderson et al. (Garnett and Anderson 1994b), the ambiguity concerning this critical module needs to be elucidated before any results are taken too seriously.

The model allows for both horizontal and vertical transmission, and the progression of all of the included STDs (HIV, gonorrhea, chlamydia, syphilis, and chancroid) is well modeled in a duration since infection-dependent manner. A unique twist to the natural history component

is that the treatment and immune response affects are taken into account in determining an individual's susceptibility to infection or re-infection after treatment.

The health care component governs how successful the *health care system* is in identifying and treating curable STDs. In comparison the intervention component specifically models:

- condom use and condom use programs,
- specific improvements in the health care system,
- syphilis screening programs,
- mass STD treatment programs, and
- information and education programs that might have an impact on health behavior.

The detail of how these components are implemented is largely hidden from the reader making it impossible to adequately evaluate their utility. However, the concept of including these modules in a model is undoubtedly a very good one and should be noted by other modelers.

The modelers make a number of important, general points regarding their choice of design. They argue that the micro, individual structure is the best for confronting complex systems in which many of the contributing processes are inter-dependent affecting each other as time progresses. This results from the fact that micro, stochastic simulations are able to "take stock" at the beginning of each time step and adjust their transition probabilities according to the state distribution of all or any subset of the individuals in the simulation. Complex time-evolving interactions are thus relatively easy to handle. A convincing argument is also made for the role of individual-level models in decision support and intervention identification. The virtual population created by the simulator can be subjected to numerous different

interventions and many variations of each intervention at a negligible cost, thereby providing planners with valuable and very cost effective guidance as they allocate precious time and resources to real situations. It is also interesting to note that the STDSIM group is the only one to clearly delineate the full set of factors contributing to an HIV epidemic; whereas most modelers restrict themselves to the population and its endogenous interactions, the STDSIM group has properly identified the environment in which the population is living as a major contributing factor. Thus, they explicitly include the health care system and specific interventions. It is a trivial extension of this line of thought to also include the overall socio-economic, ecological and epidemiological context in which the population lives. A truly useful model would include some component to handle economic, ecological and macro ecological considerations as well.

CHICK, ADAMS & KOOPMAN

Chick, Adams and Koopman (Chick, Adams, and Koopman 2000) explore the connection between macro deterministic models specified using continuous differential equations and individual-level micro models implemented as stochastic simulations. Their work does not address HIV/AIDS specifically, but rather provides unique insight into the relationship between these modeling methodologies. An interesting result relating to the epidemiologist's basic reproduction number R_0 is also obtained; namely that the conventionally defined R_0 is not an appropriate measure of the number of secondary infections associated with each primary infection when the disease transmission process occurs in the context of a relationship of finite duration.

This article develops and solves an analytic model of the susceptible-infectious-susceptible (SIS) disease model in a homogenously mixing population. A population with well defined characteristics corresponding to those that can be described by the analytic model is generated through stochastic, individual-level simulation. The properties of that population are then compared to the analytic results predicted by the deterministic analytic model, and they are found to be satisfactorily similar. Specifically, a multiple simultaneous partner situation is examined that corresponds loosely to what is thought to be a prevalent form of sexual mixing.

The emphasis on concurrent partnerships and infection – re-infection disease processes makes this work relevant to the study of HIV/AIDS and the common STD cofactors that significantly alter HIV transmission rates. Additionally, the added theoretical investigation of the two different modeling approaches and the reassuring finding that they are closely complimentary is valuable.

HIV MODELS – MACRO OR MICRO?

Population models used to examine the impact of HIV/AIDS on the population fall into all of the categories defined above. However, it is our position that because the transmission of HIV in the developing world is predominantly confined to sexual contact, and because sexual contact is governed by the prevailing culture, the spread of HIV is predominantly the result of behavioral mechanisms – not biological mechanisms. This does not downplay the importance of biology in the process but correctly places the priority on the behavior that provides the potential for the biological processes to run their course. Given that, I must place priority on examining and developing models that are capable of manipulating both behavioral and

biological processes within a population, and the only type of model that can accomplish that is some variation of the individual-level, stochastic simulation.

Moreover to address HIV/AIDS, I believe that it is *critical* to minimally model the effect of HIV infection on: 1) mortality, fertility, sexual mixing patterns, and household dynamics. Because sexual intercourse mediates both conception and the transmission of the HIV, and because barrier methods introduced to prevent either conception or the transmission of the HIV affect both, it is *absolutely necessary* to effectively link the sex-mediated processes in the model. The easiest way to accomplish this is to explicitly model sexual intercourse within unions. None of the models reviewed above took this approach, and as a result all fail to properly associate these inextricably linked processes. Any model that wishes to properly model the impact of the HIV on a *whole* population, and to assess the impacts of interventions on the *whole* population must **completely link** the transmission of the HIV to conception. Accomplishing this link is one of the unique features of the model I have built.

Taken altogether, I seek to create a simulator that is capable of modeling a complex set of behavioral and biological processes at the individual level, and specifically required to model individual sexual encounters. Additionally, the simulator must model the complete demography of an entire population and provide a flexible means through which to incorporate individual-level disease transmission and progression. This can only be accomplished with a stochastic, individual-level model that is built around a very flexible internal representation of individuals and their connections to each other.

The simulator presented below addresses all of these issues from a methodological perspective. It is by no means a *finished* product or even one that is satisfyingly useful in a substantive sense. Rather it should be viewed as a successful ***proof of concept*** that clearly vindicates the basic idea of creating an individual-level population simulator that completely links reproduction and the transmission of sexually transmitted diseases, proves that the technological innovations work well, and that the whole thing is tractable and cheap enough to be useful. This is the starting point from which a practically useful simulator can be created.

AIM – WHAT DOES THE SIMULATOR DO?

The primary purpose of the model described here is to provide a tool to understand the impact of the HIV epidemic in Africa at both the individual and the population levels, and so far as possible, to properly associate the individual and population-level processes that contribute to the spread and incubation of the disease. The aim is to simulate an entire population of individuals who marry, divorce, reproduce and die in ways similar to a polygynous African population with endemic HIV. Modeling the individual-level processes in the context of a whole population allows the individual-level behavior to be influenced by the state of the population as a whole and in turn generates a population whose size and composition is derived directly from the actions of the individuals who compose it. The entire model is fully time differentiated – meaning that all durations are specified to the smallest unit of time available within the model and all important duration-dependent processes are modeled as a function of duration[32]. Moreover, the parameter set is also time-specific allowing the parameter values to change over time and thereby providing the ability to model populations going through various types of macro-demographic transitions.

The model may be used in at least two conceptually distinct ways. First with a relatively complete set of well-measured and reliable parameters it can be used to project a real population into the future. Numerous projections can be done to investigate different assumptions regarding the future course of population and epidemiological processes such as fertility, mortality, and various AIDS prevention and treatment programs. And since the

[32] This includes individual age, duration since infection with HIV, duration of marital unions and extramarital affairs, durations of the components of the interbirth interval, and the virtual historical time through which the simulated population is living.

model is at the individual-level and therefore must be stochastic in nature, it is possible to calculate the variance in any metric by performing the same simulation a large number of times and pooling the results into one data set.

Second, with or without a relatively complete set of empirical parameters the model can be used to investigate the relative importance and sensitivity of various processes and parameters as they contribute to the growth of an HIV epidemic and to examine a wide range of outcomes for any set of input parameters. Viewed in this way the model may be used to explore the entire parameter space in order to gain a holistic understanding of how an HIV epidemic perpetuates and spreads, how the disease interacts with the population over time, and how the structure of the population changes as the epidemic matures:

- What patterns of networking and mixing are necessary to perpetuate an epidemic, and how do different networking and mixing patterns affect the dynamics and penetration of the disease into a population?
- Which parameters have the most sensitive impact on the velocity and saturation of the epidemic?
- What combination of parameter values is necessary to elicit negative population growth?
- Are there different combinations of parameter settings that can accomplish the same thing?
- What are the characteristic dynamics of the epidemic?
- Are there characteristic times over which similar processes take place?
- Are there sharp bifurcations or nodes in the behavior of the system that entail precipitous changes in dynamics?
- How does the structure of families change over time?
- How many orphans are created and how quickly?
- What is the ratio of young orphans to grandparents as an epidemic matures?
- What is the classic dependency ratio over time?
- By how much and how fast does the size of the economically productive population change?
- By how much and how fast do lost wages accumulate?

- Proportionately how many people would be eligible for treatment as the epidemic matures?
- How many people need to be treated successfully to stabilize an epidemic and eventually eradicate the disease?
- How many new HIV cases are prevented through the successful treatment of a single case?
- How much would it cost to implement a treatment program?
- How much would a treatment program save in terms of lives, money, lost wages, and future treatment?
- How do mortality and fertility rates of the population change over time as the epidemic grows?
- What is the relationship between the incidence and prevalence of HIV as an epidemic matures?
- Are there individuals who form particularly active nodes in the network that spreads HIV?
- Are treatment and prevention programs targeted at those high infectivity nodes effective, practical and affordable; and if so, how do they compare with non-targeted programs?
- How do the simultaneous effects of HIV on mortality and fertility interact with one another, and to what degree are these synergistic in their overall affect on the population?
- Broadly defined, what is the relative importance of social factors compared to biological factors in the spread of an HIV epidemic?
- Based on a rough answer to the previous point, on what types of research and program development should scarce money, time, and manpower be focused?

The second view of the model encourages the analyst to explore all possibilities in order to gain an understanding of both likely and unlikely scenarios and what differentiates them. An example of this would be to create a set of model AIDS epidemics that span the most likely scenarios for various types of populations; a set of models that could be used to characterize and classify real epidemics, and to provide an indication of the likely future course of a real epidemic under various assumptions. This type of approach and understanding should contribute to the identification of the most practical, affordable, and effective strategies to limit the future spread of the disease and alleviate the suffering of those already affected. Because

the whole population is modeled at the individual level it is possible to implement *virtual* treatment and prevention programs targeted at individuals and to calculate the costs and benefits of each over various periods of time at both the individual and population levels. Although it is unlikely that any implementable model is able to accurately predict the impact of a treatment and prevention program on a real population, a model of this type will be able to differentiate between those strategies that are likely to be effective or not effective and to provide an idea of the relative costs and benefits of implementing various forms of treatment and prevention at both the individual and the population level. Although not entirely satisfying, that information could save substantial time and money in the development of practical programs to combat HIV/AIDS.

Another use to which the model may be put is to examine and measure the biases that an AIDS epidemic may bring about in common indirect demographic measures such as the Brass method of estimating child mortality, the sibling and cousin methods for estimating adult mortality, and any others that rely on indirect information to infer population-level indicators. If a valid parameter set were known, the model could be used to simulate a number of epidemics and estimate the actual biases introduced into the measures by the disease. Without a full valid parameter set the model could be used to generate a large number of virtual populations that correspond to a wide range of potential real situations, and from those the general nature of the bias that might be encountered under different circumstances could be identified and potentially quantified.

Finally the model may be used to create large amounts of well-defined data describing a population with or without AIDS. Those kind of data are necessary during the process of implementing and debugging software to store, manipulate and analyze longitudinal data that attempt to describe a whole population and all of the demographically important connections between its members. Well-defined data of this type are difficult to obtain but absolutely essential to the creation of new software.

ARCHITECTURE

The model described here contains the components necessary to adequately describe an entire population of males and females who:

- Enter into and terminate polygynous marriages
- Enter into and terminate unlimited simultaneous extramarital affairs
- Engage in sexual intercourse within marriages and extramarital affairs
- Based on the number of sexual intercourse events between individuals, create conceptions in females and transmit HIV from an infected to uninfected people
- If female and impregnated, carry a conception through gestation and birth facing the risk of miscarriage throughout
- If HIV positive, incubate the HIV and develop AIDS over a prescribed period of time following infection
- If HIV positive, experience varying levels of health and infectivity as the HIV incubates
- If female and HIV positive, infect offspring at birth
- Die

Additionally, an individual's HIV status affects the individual's:

- Ability to infect an uninfected individual
- Ability to marry
- Ability to sustain a marriage
- Ability to engage in an extramarital affair
- Ability to sustain an extramarital affair
- Ability to engage in sexual intercourse
- If female, ability to conceive
- If female, ability to carry a conception to term
- Ability to survive

The model is implemented as an individual-level, two-sex, stochastic state-transition machine. The duration of each time step is one month. At the beginning of each month the current status of each individual or entity is identified, and that status determines for what they are at

risk during the current month. For each event for which an individual or entity is at risk, the appropriate monthly hazard (between zero and one) of occurrence is retrieved based on the individual's or entity's characteristics, and a random number (between zero and one) is drawn to determine whether or not the event occurs during the current month. If the random number is less than the hazard, the event occurs and that fact and its repercussions are recorded in the database.

The implementation of the model is broken down into conceptual and organizationally distinct units. The data storage and manipulation is handled by a relational database management system, the logic of the model is implemented in a programming language, and the overall implementation is divided into conceptual modules that interact with each other.

From here onward, the model and the logical and physical machinery necessary to implement it are referred to variously as the simulator, the simulation and other variations of the root *simulate*. This is done because there are a number of demographic and epidemiological sub models within the simulator with the result that the term *model* rapidly becomes ambiguous when describing the simulator and its various components.

RELATIONAL DATABASE IMPLEMENTATION

The simulator is implemented as a relational database (see Appendix A) using the Microsoft Access Database Management System (DBMS)[33]. The operations on the database are written

[33] Microsoft Access is not a particularly fast, efficient or scalable database management system. At the time the implementation was started I did not have access to either more sophisticated and scalable database management systems or to computers powerful enough to take advantage of a better database management system. The finished simulator clearly demonstrates the limitations of Microsoft Access and because of those limitations is not able to live up to its full potential. That problem has already been rectified with the purchase of a dual-processor Sun Microsystems UNIX workstation that is more than capable of running a number fast relational database management systems with essentially unlimited scalability, including IBM's DB2 and

in the Microsoft Access version of the Structured Query Language (SQL, see Appendix B), and the logic of the simulator is written in Embedded Microsoft Visual Basic and runs within the Microsoft Access DBMS. Together the relational database and SQL provide a storage facility and logic that are ideal to capture, store and manipulate the information that describes individuals and the connections between them over time[34]. Visual Basic is a full-featured programming language that is able to implement the additional logic required by the simulator.

DATA REPRESENTATION

Information describing the individuals and their relationships in the simulated population is stored in a set of tables and relationships that form the core structure of the data – what is typically termed the *data model*. The data model for the simulator is relatively simple and derived directly from the data model used to store and manipulate the Gwembe population data, described above. The core tables are described in Table 53 below:

TABLE 53: POPULATION SIMULATOR BASE TABLES

Table Name	Description
tblALives	Describes individuals
tblAUnions	Describes marital and extramarital unions between males and females
tblALnkLivesUnions	Links individuals to their unions
tblAFertilityEvents	Describes the history of fertility events for female individuals
tblAPregnancies	Describes pregnancies

Oracle's Oracle 8 relational database management systems. The process of implementing the simulator in ANSI C on IBM's DB2 relational database management system is ongoing at this time.

[34] The relational architecture of this model stems directly from experience designing database systems for longitudinal demographic surveillance projects in Africa. The data storage and manipulation requirements for the virtual populations created in these simulations are very similar to those encountered when recording the details of a real population over time.

TABLE 54: DESCRIPTION OF THE TBLALIVES TABLE

Name: TblALives
Fields:

Name	Type	Description
lngID	Number	Unique ID for this individual
bytSex	Number	Value corresponding to the sex of this individual: male, female
lngPregnancyID	Number	ID of the pregnancy that produced this individual
lngDOB	Number	Date of birth of this individual
lngDOD	Number	Date of death of this individual
lngDOI	Number	Date of infection with HIV of this person
lngInfectedByID	Number	ID of the person who infected this person with HIV
bytSexDrive	Number	Value corresponding to this person's sex drive: five categories, 1 through 5

TABLE 55: DESCRIPTION OF THE TBLAUNIONS TABLE

Name: TblAUnions
Fields:

Name	Type	Description
lngID	Number	Unique ID for this union
bytType	Number	Type of the union: marriage or affair
lngDOU	Number	Date union initiated (date of union)
lngDOS	Number	Date of separation
BytEndEvent	Number	Type of ending (separation) event: divorce, death, end of affair

TABLE 56: DESCRIPTION OF THE TBLALNKLIVESUNIONS TABLE

Name: TblALnkLivesUnions
Fields:

Name	Type	Description
lngID	Number	Unique ID for this link
lngLifeID	Number	ID of the life to which the associated union is linked
lngUnionID	Number	ID of the union to which the associated life is linked

TABLE 57: DESCRIPTION OF THE TBLAFERTILITYEVENTS TABLE

Name: TblAFertilityEvents
Fields:

Name	Type	Description
LngID	Number	Unique ID for this fertility event
lngUnionID	Number	ID of union with which this event is linked
lngFemaleID	Number	ID of female with whom this event is linked
bytEvent	Number	Value corresponding to the type of event that occurred: birth, miscarriage, conception leading to birth, conception leading to miscarriage, end of breast feeding, recovery after birth or miscarriage
lngDOE	Number	Date of the event

TABLE 58: DESCRIPTION OF THE TBLAPREGNANCIES TABLE

Name: TtblAPregnancies
Fields:

Name	Type	Description
LngID	Number	Unique ID for this pregnancy
lngUnionID	Number	ID of union with which this pregnancy is associated
lngFemaleID	Number	ID of female with whom this pregnancy is associated
lngEndEventID	Number	ID of fertility event that ended this pregnancy
lngDOC	Number	Date of conception leading to this pregnancy
lngDOE	Number	Date on which this pregnancy ended
bytEndType	Number	Value indicating the type of end for this pregnancy: birth, miscarriage or currently pregnant

Figure 92 displays the relationships between the base tables that store all of the data used for the simulation. Each table is represented with a T-diagram that displays the name of the table and all of the fields that are contained in the table. In each table there is one field with an asterisk to the left of the vertical line; that field is the primary key for the table and contains a unique value that identifies each row of the table.

The lines between fields in different tables indicate that there is a correspondence between the field values in the two tables. An arrowhead indicates a something-to-*many* relationship while a blunt end indicates a something-to-*one* relationship. For example, the blunt-to-arrowhead line

leading from lngID in tblALives to lngLifeID in tblALnkLivesUnions indicates that each life may be associated with many different links to unions, but that each link to a union is associated with just one life. Correspondingly, the blunt-to-blunt line between lngID in tblAPregnancies to lngPregnancyID in tblALives indicates that each life is produced by just one pregnancy and each pregnancy produces just one life – a one-to-one relationship (this version of the simulator does not support twinning).

The relationship table tblALnksLivesUnions provides the mechanism to store an unlimited number of unions for each individual and also to link both of the participating individuals to the union that they share. The flexibility of this structure allows the storage and manipulation of polygynous unions. Women's fertility histories are captured and linked to both women and the unions during which the events take place by the fertility events and pregnancies tables. Children are linked to their mothers and the unions that created them by the pregnancies table. The one-to-one relationship between the lngPregnancyID field in the lives table and the lngID field in the pregnancies table links children to the pregnancy that created them while the relationship between the lngID field in the lives table and the lngFemaleID field in the pregnancies table links women to the pregnancies that they have experienced. The remainder of the relationships should be self explanatory.

SQL queries defined on this data structure are able to execute all of the operations necessary to conduct and analyze the simulations. Virtual families can be reconstructed and all of the standard demographic counts and rates can be calculated in a straightforward fashion using simple temporal logic.

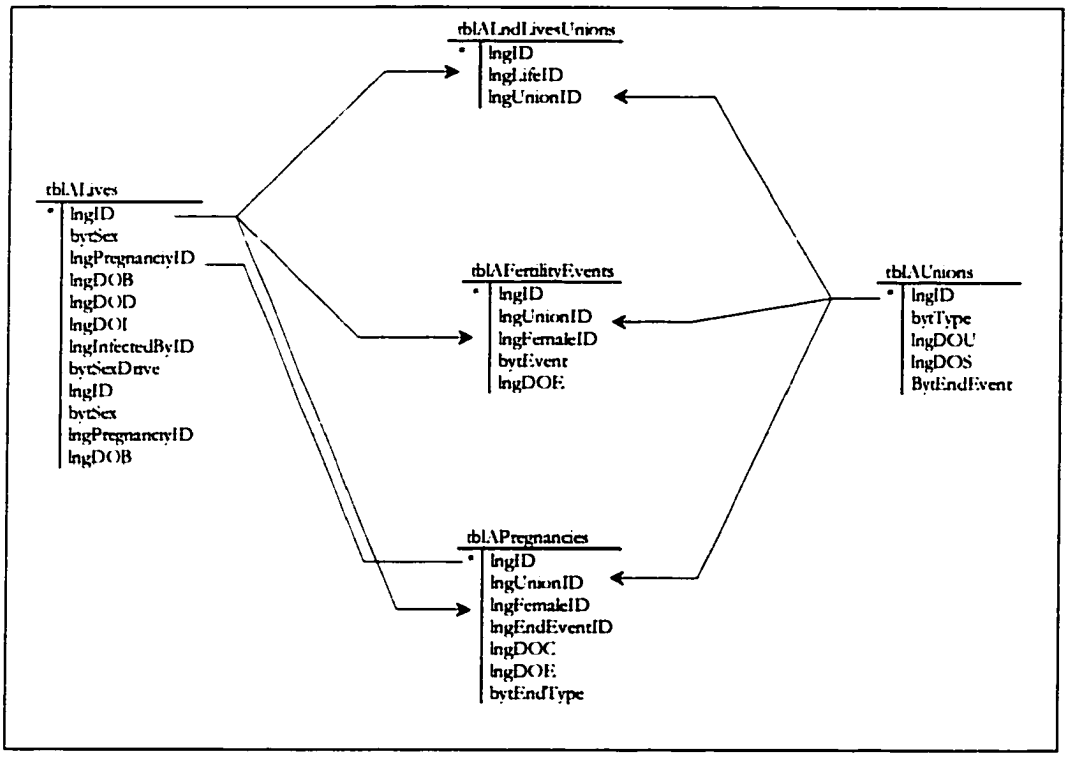

Figure 92: Simulator Data Model

RECORD OF HISTORY

One of the primary aims of the simulation is to record the entire simulated history of both the individuals in a virtual population and the connections that they form amongst themselves as virtual time progresses. That is accomplished through the use of the three tables that store the duration of the lives, unions (marital and non-marital), and pregnancies that are generated as the simulation unfolds. The relationship table between the lives and unions tables and the other relationships between individual tables record the links between entities (lives, unions, or pregnancies) that prevailed over the valid duration of the entities themselves. The fertility events table simply keeps track of the sequence of fertility events that each woman

experiences. The end result is a very simple temporal database that can adequately store and manipulate the *history* of a simulated population. The final product resulting from a run of the simulation is a **complete history** of the virtual individuals who lived during the period of time over which the simulation was run.

INDIVIDUAL-LEVEL

The simulation operates at the *individual* level by creating and manipulating individual lives – or *virtual* people. All of the operations conducted by the simulation operate on individuals: for example, the formation of a union between two individuals, dissolution of a union, an intercourse event, a conception, a miscarriage, a birth, the cessation of breast feeding, or the transmission of the HIV between two people. Reasons to simulate individuals rather than populations are both substantive and methodological.

The individual level readily allows the simulation of families and the interpersonal ties that are crucial for the transmission of HIV. The ability to simulate families as collections of individuals permits the analyst to study the intra-family dependency ratios, to compare families with and without an HIV-death, and to reconstruct in detail the formation of families and their potential destruction by HIV. The individual level also permits the reconstruction of transmission networks through which HIV is passed and the identification and characterization of potentially important nodes in those transmission networks.

The typical alternative to building the model at the individual level is to specify the model at the population level. A *population-level* model is defined relative to various aggregations of individuals that as a group occupy different states such as 'HIV negative', 'twenty years old',

'female' or 'unmarried'. The dynamics of such a model are designated using differential equations that specify the fraction of each group moves to another group or state during a period of time.

There are three strong advantages to a population-level model specified in this way: 1) the actual specification is concise and easy to read if one is familiar with differential equations, 2) the implementation of the simulation is accomplished using relatively standard numerical simulation techniques well known to mathematicians and engineers, and 3) it is *sometimes* possible to solve the collection of differential equations to derive relationships that are not known a priori. The allure of the third point is powerful but elusive. As the complexity of the model grows the likelihood that it can be solved in closed form diminishes rapidly, and in practice most models that are sufficiently descriptive to be of interest cannot be solved satisfactorily, and without a solution much of the advantage of the differential equation specification is lost.

A clear disadvantage to the differential equation specification is that it can be difficult to bring closure to processes that fundamentally occur between individuals and not between populations; examples being sexual intercourse and the formation of male-female unions. The differential equation specification would have a defined fraction of males of age x marrying a defined fraction of females age y (for various combinations of x and y); however, it is rare that the number of males and females specified in this way is equal as it must be to properly model the formation of marital unions. Various solutions to solve problems like this are invented to

force parity on pairing processes, but all appear unsatisfactorily *ad hoc* when compared to reality.

Another disadvantage of the differential equation models is that their specification does not necessarily record history, and certainly does not record the history of individuals. As a result it is not possible to examine such things as HIV transmission networks or the evolution of specific families or individuals.

Finally, for most researchers and analysts in medicine, public health and demography the differential equation models have the disadvantage of being virtually indecipherable due to the fact that they are build using foreign concepts and written in a foreign language: namely differential calculus. Although this is not an inherent disadvantage of these kinds of models, it is a practical disadvantage that limits the extent to which they are constructed and utilized in the above-mentioned fields. Furthermore, even when good models of this type exist the results they produce are under utilized because the process through which they are produced is inaccessible to many researchers.

In contrast the individual-level state-transition machine described here is built on relatively simple concepts that are directly derived from the statistical principles underlying much of the analysis carried out by researchers in medicine, public health and demography – namely state transition probabilities defined over specified periods of time, or hazards. The structure of the model is more complex than most, but it is all built on that simple principle. It is hoped that this makes the model more accessible to a wide range of researchers and endows the results that it produces with more credibility in their eyes.

Compared to the population-level differential equation models, the individual-level state transition machine:

- explicitly models individuals and the interactions between individuals instead of groups of individuals and the interactions between groups of individuals
- is harder to specify and describe in concise terms
- cannot be solved explicitly
- is non-deterministic or *stochastic* which allows the calculation of the likelihood of various outcomes
- may require more parameters
- is easier to extend and add onto
- requires more storage space to run
- requires roughly the same amount of computation to run
- produces *both* population and individual histories instead of only population histories
- is conceptually simpler and more accessible
- may be more complicated in its practical implementation
- is more flexible and easier to manipulate
- ... ...

An individual-level model of the type described here operates at the level of individual people by modeling and manipulating the interactions between individuals. In so doing it creates *histories* for each of the individuals in the simulated population that can be analyzed individually or at an arbitrary level of aggregation. The individual histories provide the basis to examine the formation and evolution of interconnected groups of individuals such as families and HIV transmission networks. Operating at the individual-level allows the explicit modeling of disease processes that depend substantially on the dynamics of interpersonal contact. AIDS is a good example of a disease that is passed between individuals during intimate interpersonal contact that is purposeful and largely under the control of the persons contacting each other. To the extent that people control their contact and that the rules governing those contacts are

known, the transmission of HIV is something that is able to be understood, modeled and controlled. This clearly differentiates HIV from environmentally transmitted diseases such as malaria or schistosomiasis or other health problems that operate more generally on groups of people such as malnutrition or disorders associated with aging. In those cases models operating at the population level are more appropriate as there is little or no individual-level variation or individual contribution involved with the transmission or acquisition of the disease. HIV is clearly different, and there is a clear advantage to being able to model individuals and their behavior.

POPULATION-LEVEL

In so far as a collection of individuals composes a population, the individual-level state transition machine described here models a population. The population that is modeled is simply the collection of all the virtual individuals who are created by the simulation. All of the population level measures that are familiar to demographers may be calculated over the population of virtual individuals who are created by the simulation over a specified period of time; including per period growth rates indexed by sex, age or any other individual attribute, and period and cohort fertility and mortality rates indexed by sex and age. The composition of the virtual population over time may be quantified in an arbitrary way limited only by the degree of variability that is recognized and recorded between the individuals composing the population.

RESULTS

The *result* produced by the simulator is a collection of individual histories and histories of relationships between the individuals. The histories consist of sequences of events associated

with individuals that define state-time intervals that are in turn associated with individuals or relationships between individuals. For example, an individual is alive between their date of birth and date of death, and they may be associated with various other individuals of the opposite sex through marriage during the time when they are alive. Each of those unions is valid between its date of union an date of separation and is associated with the two individuals who participate in each union. The *result* is analyzed in exactly the same way as an empirical longitudinal data set describing the basic demographic dynamics of a real population. Both individual event histories and aggregate population-level measures may be produced from the *result*.

MODULAR CONSTRUCTION & EXTENSABILITY

The simulator is divided into modules, each of which is responsible for the dynamics of a particular process such as reproduction or death. This partitions the simulator in a way that allows the logic of each process to be self-contained and easy to develop, maintain and extend if necessary. It also clearly separates the storage and manipulation of the states and state-time intervals that are associated with each process and provides a clear interface between them. The definition of this interface is valuable in that it provides a clear avenue through which the different processes may influence each other. Moreover this structure is naturally extensible allowing the easy addition of more self-contained modules.

ESSENTIAL MODULES

The essential modules govern the fundamental demographic dynamics of a population and include:

- fertility – manages all aspects of reproduction except pairing and intercourse
- mortality – manages all aspects of death
- nuptiality – manages pairing into and separating from marital unions
- extramarital affairs – manages pairing into and separating from extramarital unions
- sexual intercourse – manages intercourse

The one additional module implemented in the simulator at this time is the HIV module that regulates the progression of AIDS within an individual and the effect that an individual's HIV status has on all of the other processes that involve the infected person.

ADDING A MODULE

Adding a module is a relatively straightforward process that involves the definition of the level at which the new process operates (individual, intra-individual, group etc.), the new states that are necessary to describe the reality that is modeled, the storage requirements for those states, the logic necessary to govern the transfer to and from those states, and last how the occupation of the new states will affect the transfer probabilities to and from already modeled states.

Additional modules that are being considered for future versions of the simulator include those necessary to model Tuberculosis, Malaria, Migration, hereditary transmission of traits, and environmental influences. The first three are desirable additions to improve the ability to model processes that interact with and impact the development of an HIV/AIDS epidemic or pandemic. Together the last two form the basis for modeling an evolving population.

ACCOUNTING FOR PARAMETERS AND TIME

The parameters necessary for each module are discussed below. There are a large number of parameters necessary to describe even a static population – on the order of hundreds. Because

the real populations that this model aims to describe are undergoing rapid demographic transformation at this time, it is necessary to have the ability to change the parameter values over time in order to properly model the changing demographic reality. For example, countries like Zimbabwe and South Africa are going through fertility transitions at this time, and until recently underlying mortality was also trending downward. Those changes are occurring simultaneous with the rapidly growing AIDS pandemic and form the foundation upon which AIDS is affecting the population.

GRANULARITY OF TIME-STEP

Time enters into the structure of the simulator in a number of ways. There is a fundamental time step or *granularity* to the time that is modeled. The notion of a granule represents the shortest period of time over which change is measured or modeled. As mentioned above, the state transition machine works on a granularity of one month: all transitions occur over the period of one month, the hazards representing the transition probabilities are referenced to the period of one month, and the simulator steps through virtual time one month at a time. There are other reasonable choices for the base granularity including 'day', 'quarter', or 'year'; however, the processes modeled have dynamics that are roughly on the time scale a month making *month* the natural choice for the granularity of the time step. Conception, pregnancy, birth, and the other fertility-related processes all operate at the level of months, a life can be readily measured in units of a month, and it is not unreasonable to quantify the probability of dying during the period of a month, and likewise for marriages and affairs. The only process included in the model that is not well-described on the time scale of a month is sexual intercourse. It seems more natural to measure and apply the probability of engaging in sexual

intercourse over the period of one day or one week. The solution to this problem is discussed below in the section describing the intercourse module.

The other notions of time inherent in the model are the granularity of the variation in parameters with time and age. Most of the parameters take constant values over the standard five-year age groups used by demographers and also over five-year periods of simulated time. These granularities were chosen more out of convenience in order to keep the number of parameters from growing to a truly untenable number. It is also reasonable to assume that at most ages and times, the parameter values are not changing so rapidly as to require a finer level of granularity in their specification.

TIME VARYING PARAMETERS

The majority of the parameters in the model can vary with time. All parameters that do vary with time are constant over five-year periods, and currently the simulator stores and utilizes forty sets of five-year parameter values. This number is arbitrary and can be expanded if necessary; however, it is thought that 200 year's worth of simulation is probably sufficient for most purposes[35].

ODDS & ODDS RATIOS

The parameter sets are stored in terms of a base set and odds ratios that define modifications to the base parameter set. The values for all of the hazards that prevail in the first period are stored as they are. For each of the 39 additional periods that that the simulator can model,

[35] In point of fact 200 year's worth of simulation is substantially beyond the storage and computational capacity of the Microsoft Access database management system!

there is a set of odds ratios that describe by how much the base hazards are to be multiplied for each period. The odds ratio scheme has the significant advantage that it never produces invalid hazards – less than zero or greater than one. The odds ratio is also intuitively appealing and commonly understood by a wide range of researchers. The disadvantage is that the odds transformation is not linear, but for relatively small deviations it is sufficiently linear to be acceptable. Moreover, the analyst can easily ensure that the desired hazards are indeed produced by the odds transformation by working the transformation backwards.

The real reason why this scheme is chosen is to ensure valid hazards at all times.

SIMULATION OF INTERCOURSE EVENTS

Sex is the activity that confers a risk to both conception and the transmission of HIV, and consequently modeling intercourse events is necessary to properly link reproduction and the transmission of HIV. Furthermore, since HIV affects a woman's fecundity and most likely also affects individuals' likelihood of engaging in sexual intercourse there is substantial interaction between being HIV positive and sexual intercourse and the repercussions of sexual intercourse. It is fundamental to this model that individual sexual intercourse events and the their consequences are modeled and simulated.

REPRODUCTION

A woman's probability of conceiving during a month depends strongly on the number of intercourse event to which she is exposed. The interbirth interval model of fecundability (Bongaarts and Potter 1983) is used to generate a monthly probability of conception as a function of the number of intercourse events to which a woman is exposed and a number of

other parameters describing various components of the interbirth interval, see the description of the fertility module below.

TRANSMISSION OF HIV

The probability of transmitting HIV from an infected individual to an uninfected individual is specified and modeled per individual intercourse act. Data describing empirical per intercourse HIV transmission probabilities from Rakai, Uganda (Gray 2000) are used to calibrate this component of the model

COMPLETE LINKAGE OF REPRODUCTION AND TRANSMISSION OF HIV

The overall aim of this simulation is to model an entire population as an HIV epidemic unfolds. Because the HIV is a sexually transmitted disease, the exposure to infection is directly proportional to the exposure to conception, and consequently basic reproduction of both the disease and its host population are tightly linked. Augmenting and modifying that link is the fact that HIV infection affects fecundability and the likelihood of intercourse. In sum, it would be unwise to unlink reproduction and the transmission of HIV, and the only good way to link them is to model individual sexual intercourse events.

The following sections describe each of the modules in detail and provide the basic parameters used by each.

MORTALITY MODULE

The mortality module governs the rate with which individuals in the simulated population die. An individual's probability of death is a function of their sex, age, and HIV status.

Once during each month of the simulation, each individual's hazard of death is calculated based on their sex, age at the beginning of the month, and HIV status at the beginning of the month. A random number between 0.0 and 1.0 is drawn for each individual and compared to their hazard of death for that month, and if the random number is less than their hazard of death, they die during that month. That transition is recorded by placing the value of the month in the Date of Death (DOD) field of their record in the Lives table.

After someone has died, they are no longer at risk for any event of any kind. If the dead person was participating in any marital unions at the time of his or her death, those unions are terminated and it is recorded that the terminating event was a death. If the surviving spouse is female she is immediately at risk to form another union starting in the month following the one in which her spouse died.

LOGIC

The mortality rate is conceptualized and implemented as a monthly hazard of dying – the probability of death faced over the duration of one month. The underlying hazard of death is specified as a function of sex and age, and if an individual is HIV positive the hazard of death is increased in proportion to the current viral load possessed by the individual, see HIV Module below.

Because values of the hazard of death are restricted to the range zero to one, the method used to augment the hazard of death for HIV positive individuals must ensure that the hazard of death does not exceed one. This is accomplished by adding a fraction of the difference between one and the underlying hazard of death (the hazard of survival) that is proportional to the individual's viral load. Put another way, an HIV positive individual's hazard of survival is diminished by an amount that is proportional to their viral load, Equation 19.

$$\gamma_{as} = \gamma^{u}_{as} + \left(1 - \gamma^{u}_{as}\right) \cdot \nu \cdot k$$

Equation 19: Monthly Hazard of Death

Where γ_{as} is the monthly hazard of death for a given age (a) and sex (s), γ^{u}_{as} is the underlying monthly hazard of death, ν is the current viral load possessed by the individual (restricted to values between zero and one), and k is a scaling factor that takes values between zero and one and mediates the strength of the impact of the viral load. If k is zero then there is no effect of HIV on mortality, as k increases the impact of the viral load (and thereby the fact of being HIV positive for a given duration) increases. The scaling factor is included to give the model more flexibility, but under most circumstances it is set to a value of one.

The monthly hazard of death that is applied to an individual is a composition of the underlying hazard of death and the infected duration-specific impact of HIV. The infected duration component of mortality is solely a function of the length of time that a person has been infected with HIV (operating through the viral load parameter) and is totally independent of age. This is almost certainly an incorrect assumption, but I believe it is better than assuming a non-infected duration-specific impact of HIV that is instead dependent on age. Moreover, I

have not been able to locate reliable data that would allow the calculation of the age/infected duration interaction that is necessary to correct this assumption.

EMPIRICAL MORTALITY HAZARDS

The empirical monthly hazards of death are derived from the annual hazards of death estimated using the Gwembe data. The underlying age-patterns of death should not be affected by HIV-related mortality because the simulation will add the effect of HIV for those who are infected. This is a problem where the Gwembe data are concerned because there is no information on cause of death and therefore it is not possible to subtract HIV-related mortality from all-cause mortality (a difficult task even if the data were available because of the unknown but likely very important interactions between HIV and the other disease-related causes of death). The solution to this problem lies in the fact that much of the Gwembe data were collected before HIV had a significant affect on the population – all the data collected between 1957 and about 1985 are almost entirely free of HIV-related mortality. Estimates of mortality using that data correspond to all-cause mortality in the absence of HIV. Additionally, it is possible to specify a fractional polynomial logistic regression model that does not readily capture changes in the age-pattern of mortality over short age intervals. Both of these observations are used to calculate an essentially HIV-free age-pattern of mortality for males and females.

The data used to estimate the monthly hazard of mortality are contained in the familiar individual Person-Year Data Set discussed at length above in Part 2. All male and female person years exposed to death between the ages of zero and 79 over the period 1957 to 1995

are used[36]. The fractional polynomial logistic regression model used to estimate the annual probability of dying is specified in Equation 20.

$$\ln\left(\frac{p({}_5a)}{1-p({}_5a)}\right)=\beta_0 x^{0.5}+\beta_1\left(\ln(x)\cdot x^{0.5}\right)+\beta_2 s+c$$

$$x=\frac{{}_5a+4}{10}$$

Equation 20: Fractional Polynomial Logistic Model of the Annual Probability of Death

Where $p({}_5a)$ is the annual probability of dying between ages a and a+5 (standard five-year age groups), s is a dummy variable specifying sex, and c is the constant. This model is fit using STATA's *fracpoly* and *logistic* routines.

Because the unit of analysis is the person-year the resulting annual sex/age-specific annual probabilities of death correspond to sex/age-specific annual hazards of death. Another benefit of the fractional polynomial specification is that the annual hazard of death is a continuous function of age, and that provides the means to extrapolate the annual hazard of death beyond age 79. This is done to create annual hazards of death for ages 80 to 119. The extrapolated values must not be viewed as *real* measurements of any kind, but rather reasonable values based on the estimated values between ages 0 and 79. They are necessary to *close* the simulation gracefully at ages 80 and older.

The annual hazards of death are transformed into monthly hazards of death according to Equation 21, below.

[36] The fractional polynomial logistic regression model fits all of the person years for all ages and periods simultaneously so that

$$\gamma_{as} = 1 - (1 - \Gamma_{as})^{\frac{1}{12}}$$

Equation 21: Transformation of Annual Hazard to Monthly Hazard

Where γ_{as} is the monthly hazard of death and Γ_{as} is the annual hazard of death.

The underlying monthly sex/age-specific hazards of death used in the simulations are displayed in Table 59, below.

TABLE 59: UNDERLYING MONTHLY HAZARD OF DEATH

Age	Male	Female
0	0.011362	0.009975
1-4	0.002163	0.001884
5-9	0.000785	0.000683
10-14	0.000475	0.000414
15-19	0.000378	0.000329
20-24	0.000354	0.000308
25-29	0.000372	0.000324
30-34	0.000423	0.000368
35-39	0.000511	0.000444
40-44	0.000646	0.000562
45-49	0.000846	0.000736
50-54	0.001141	0.000993
55-59	0.001573	0.001370
60-64	0.002210	0.001926
65-69	0.003150	0.002747
70-74	0.004538	0.003962
75-79	0.006585	0.005759
80-84	0.009589	0.008407
85-89	0.013952	0.012274
90-94	0.020183	0.017840
95-99	0.028860	0.025671
100-104	0.040540	0.036349
105-109	0.055614	0.050338
110-114	0.074163	0.067830
115+	0.095896	0.088649

the data not affected by HIV substantially outweighs the relatively short, recent period when HIV has affected the population.

Figure 93 displays the monthly hazard of death used in the simulation, and for comparison's sake also displays the corresponding annual hazard of death and the equivalent life table probability of dying over a five-year age interval, ${}_nq_x$.

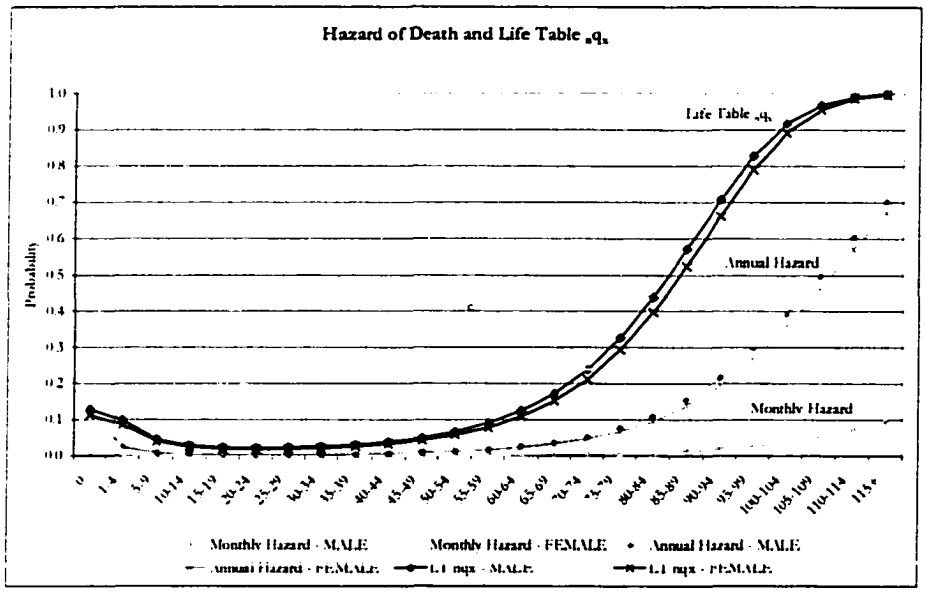

Figure 93: Hazard of Death and Life Table ${}_nq_x$

PARAMETER STRUCTURE

The simulation requires parameters sufficient to specify the monthly hazard of death by sex and age over five-year age intervals and five-year time intervals. This is accomplished by providing the model with the annual odds of death compared to the zero to one age group (the transformation from annual to monthly hazards is accomplished by the simulation). Additionally, the simulation is provided with the odds of death for each five-year age group during each five-year period subsequent to the initial period compared to the initial period. In this manner the odds of death used by the simulation can vary independently by sex, age (five-

year groups), and time period (five-year periods) over a total period of 205 years (40 five-year periods subsequent to the initial period). The advantage of specifying this variation using odds is that the resulting annual hazards are naturally restricted to the necessary range between zero and one. One disadvantage is that the transformation from odds to probabilities is not linear, but for relatively small changes it is sufficiently linear be useful.

The mortality parameter set allows the monthly hazard of death to be specified independently by sex, five-year age group, and five-year period. The number of parameters used to do this is 2,050.

DESIRABLE ADDITIONS

The literature is relatively vague in its description of the relationship between an individual's viral load and their probability of death. An HIV positive individual's probability of death appears to be more tightly linked to their CD4 cell count which in turn depends on their viral load. To properly understand these relationships it is necessary to know how the immune system and the HIV interact with each other, and in turn how the immune system interacts with opportunistic infects as it is slowly compromised by HIV. I was not able to find a good description of any of these relationships in the literature.

I formulated and tested a simple differential equation model of the interaction of viral load and CD4 count that appears to fit the existing data from the West relatively well. However, without significant work to ensure that it accurately models the HIV-immune system interaction in Africa, and lacking the data necessary to examine that question, I have postponed the use of the viral load/CD4 model until the next generation of the simulation.

INTERCOURSE MODULE

The intercourse module manages the number of intercourse events experienced by both married couples and people engaged in extramarital unions.

LOGIC

Intercourse is assumed to occur between exactly one male and one female. Male-female pairs who are at risk for intercourse are identified as those who are engaged in either married or extramarital unions during the current month. It is assumed that there are an average of 26 days during the month when the couple is at risk for intercourse, four or five days being removed for the period of menstruation. For each of those 26 days, the couple is exposed to an empirically estimated daily hazard of intercourse. A random number is drawn to determine if an intercourse event takes place; if the random number is less than the daily hazard of intercourse, then an intercourse event takes place between the two individuals participating in the union. The number of intercourse events that take place during the course of the month are cumulated to yield a total number of intercourse events for each couple for the current month. That number is recorded temporarily and used by both the fertility and HIV modules. This is a stochastically determined value and varies considerably from month to month and couple to couple even if nothing else changes.

EMPIRICAL INTERCOURSE PROBABILITIES

The empirical daily hazards of intercourse are estimated from the Gwembe fertility data based on the observed interbirth intervals. The general method is to work backwards from observed interbirth intervals to the number of intercourse events that were necessary to produce them.

The overriding assumption being made throughout is that there is negligible use of effective means of contraception, and for the Gwembe until very recently that is a tenable assumption. The observed intervals bear this out as they are very short indeed and correspond to a fertility regime that is very high.

METHOD

The interbirth interval is broken down into its constituent elements, one of which is the waiting time to conception. Assuming that the probability of conception f is constant, the waiting time to conception is the inverse of f. It is possible to write f as a function of the other parameters of the average interbirth interval including the average duration of the interbirth interval. Consequently, if the average interbirth interval is known and if the other parameters can be assigned reasonable values, it is possible to estimate f.

f is in turn a function of four other parameters: 1) the probability that the cycle is ovulatory, 2) the probability that insemination occurs during the fertile period in the middle of the cycle, 3) that the insemination leads to fertilization, and 4) that the fertilization leads to a successfully implanted embryo. The values of first and last two are known to be roughly 0.95, 0.95 and 0.5. The value of the second can be written as a straightforward combinatorial function of the number of randomly distributed intercourse events during the month and the number of days during the month that the woman is exposed to intercourse. That expression can be rearranged and combined with the known values of the other three probabilities to yield the number of intercourse events per month as a function of f.

Substituting in the value of f calculated from the data provides an estimate of the number of intercourse events per month.

See Bongaarts and Potter (Bongaarts and Potter 1983) for a deeper discussion of all of the basic ideas manipulated here.

CALCULATION OF FECUNDABILITY: f

The interbirth interval can be broken down into the following segments:

- A fertile period during which the woman is exposed to intercourse but is not yet pregnant: w the "waiting time to conception"
- A sterile period following conception and prior the termination of the pregnancy: g "gestation"
- A sterile period following the termination of the pregnancy: p

The duration of the fertile waiting time to conception is determined by the woman's intrinsic fecundability, the frequency with which she has intercourse with a viable male, and the effectiveness of her contraception. The duration of this period for a population of women is approximately:

$$w = \frac{1.5}{f(1-e)}$$

Equation 22: Average Waiting Time to Conception

f is the monthly probability of conception or *fecundability* and is assumed to be constant through time. e is the effectiveness of contraception and takes values from zero (not effective) to one

(completely effective). The value of 1.5 in the numerator reflects the inherent heterogeneity among women with respect to fecundability.

The average duration of the gestation period is very nearly equal to nine months for pregnancies that end in a birth and equal to several months for pregnancies that end in a miscarriage. The sterile period following the end of a pregnancy is short for pregnancies ending in a miscarriage. For pregnancies ending in a birth it is a direct function of the intensity and duration of breastfeeding.

The sterile period following conception, composed of gestation and the period immediately following the end of the pregnancy, is assigned the variable S for pregnancies ending in a birth and S^* for pregnancies ending in a miscarriage. The variance in S is relatively small while the variance in S^* can be larger compared to the mean duration.

The average interbirth interval for a pregnancy ending in a birth I_b can be written as the sum of the waiting time to conception and the sterile period following conception:

$$I_b = \frac{1.5}{f(1-e)} + S$$

Equation 23: Average Duration of Interbirth Interval Ending in a Birth

The average interbirth interval for a pregnancy ending in a miscarriage can be written using the same expression substituting S^* for S.

If we assign k to be the ratio of the number of miscarried conceptions to the number of conceptions that result in a live birth, we can write the average interbirth interval I as the sum of the average interval terminating in a birth plus the average interval terminating in a miscarriage, or:

$$I = \frac{1.5}{f(1-e)} + S + k\left(\frac{1.5}{f(1-e)} + S^*\right)$$

Equation 24: Average Interbirth Interval

Rearranging this equation to isolate f yields:

$$f = \frac{1.5(1+k)}{(I - S - kS^*)(1-e)}$$

Equation 25: Probability of Conception During One Month

CALCULATION OF THE AVERAGE NUMBER OF INTERCOURSE EVENTS PER MONTH: n

The probability that at least one intercourse event will coincide with the roughly two-day fertile period of a woman's menstrual cycle is:

$$p = 1 - \frac{(M-n)(M-n-1)}{(M^2 - M)}$$

Equation 26: Probability that Intercourse Coincides with Fertile Period

where M is the total number of days during a woman's cycle that she is exposed to intercourse and n is the number of randomly distributed intercourse events occurring during those M days.

Knowing that f is a function of: 1) the probability that the cycle is ovulatory 0.95, 2) the probability that insemination occurs during the two-day fertile period during the middle of the cycle p, above, 3) the probability that insemination leads to fertilization 0.95, and 4) the probability that fertilization leads to a successfully implanted embryo 0.5 yields,

$$f = (0.95)^2 0.5p = 0.45p$$

Equation 27: Fecundability as a Function of p

Rearranging this and substituting in the expression for p yields:

$$n = M - \frac{1}{2} - \frac{1}{2}\left(1 + 4\left(1 - \frac{f}{0.45}\right)\left(M^2 - M\right)\right)^{\frac{1}{2}}$$

Equation 28: The Number of Intercourse Events During a Month as a Function of f and M

Substituting in the expression for f yields:

$$n = M - \frac{1}{2} - \frac{1}{2}\left(1 + 4\left(1 - \frac{1.5(1+k)}{0.45(I - S - kS^*)(1-e)}\right)\left(M^2 - M\right)\right)^{\frac{1}{2}}$$

Equation 29: The Number of Intercourse Events During a Month as a Function of Basic Parameters

To obtain the daily hazard of intercourse H, the value of n estimated using Equation 29 is divided by the number of days of exposure M,

$$H = \frac{n}{M} = 1 - \frac{1}{2M} - \frac{1}{2M}\left(1 + 4\left(1 - \frac{1.5(1+k)}{0.45(I - S - kS^*)(1-e)}\right)(M^2 - M)\right)^{\frac{1}{2}}$$

Equation 30: The Number of Intercourse Events During a Month as a Function of Basic Parameters

Equation 30 indicates the H is a function of k, e, S, S^*, M, and I. k is assumed to take a value of 0.35, S 13.4 months, S^* 4 months, and M 26 days. All of these are relatively standard values obtained from Bongaarts and Potter (Bongaarts and Potter 1983) except for S, the mean sterile period following a birth. The value of 13.4 is composed of nine months for gestation and 4.4 months of exclusive breastfeeding during which sterility is maintained. The later was estimated by the 1996 DHS survey of Zambia for women living in the Southern Province of the country. The only remaining variable is I, the interbirth interval.

The predicted empirical interbirth intervals as a function of female and male age are obtained by estimating a dummy variable linear ordinary least squares regression on the individual Person Years data set. The model is simple and very similar to all of the dummy variable specified logistic regression models discussed in the previous part. The independent variable is the duration of the most recent closed interbirth interval and the independent variables are dummies representing male and female age independently in five-year age groups with 15-19 as the comparison group. Male/female age/age interaction dummy variable are also included for ages 20 and older to make each age/age cell fully independent.

The predicted interbith intervals are substituted into Equation 30 along with the assumed values for the other variables to yield the estimated daily hazard of intercourse as a function of male and female age. The result is a table of daily hazards of intercourse with one cell for each (male age, female age) pair in a five-year grid over ages 15 to 79. The reader may wonder what happened to the cells for which females are older than age 50. The estimated coefficients resulting from the regression were used to extend the predicted interbirth intervals into cells for which women are aged 50 or more. Although it is impossible for women to reproduce and have birth intervals after menopause, the model assumes that the trend in interbirth interval with female age would continue if females were able to have children at older ages. Since we are actually interested in their frequency of intercourse and not their fertility, this is not wrong.

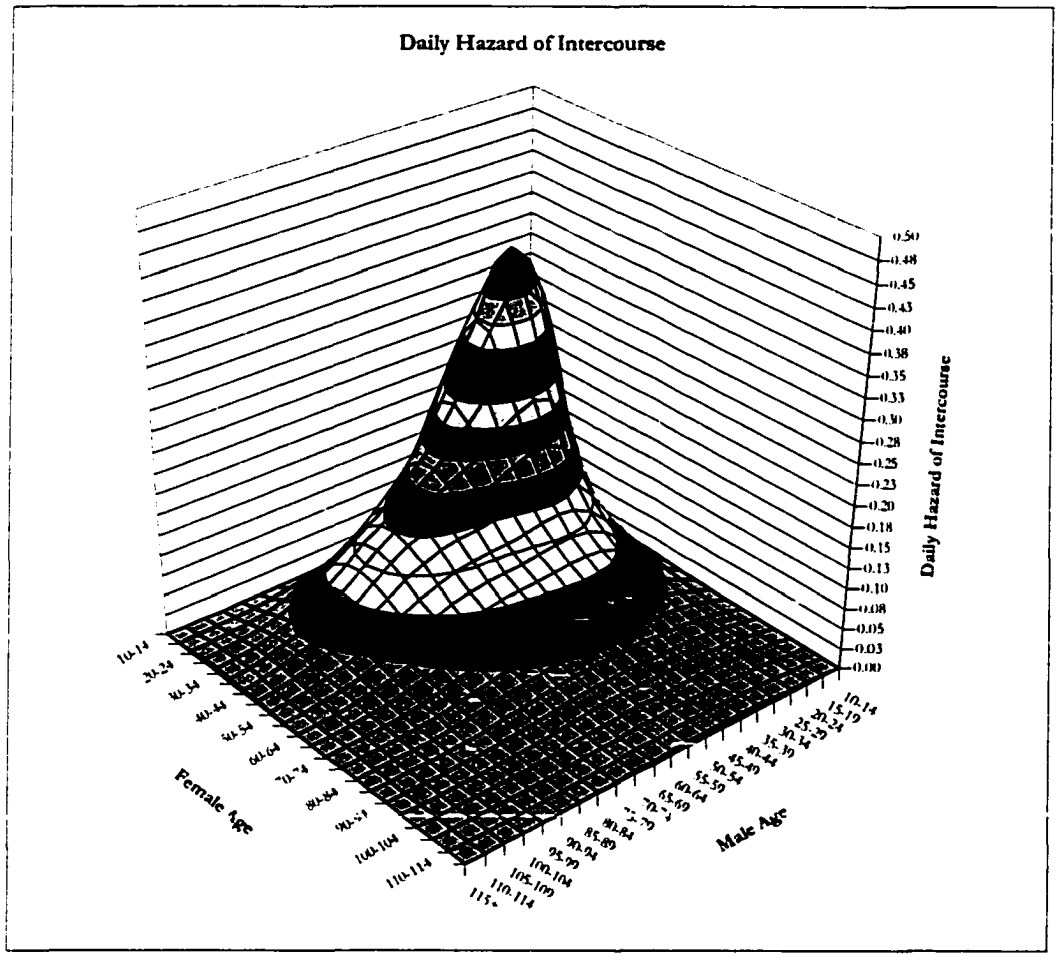

Figure 94: Daily Hazard of Intercourse

It is, however, unsatisfactory and is used only in the absence of a reasonable alternative. The result is a fairly uneven surface that reflects a good deal of stochasticity in the estimates. The surface is smoothed using a two-dimensional moving average to produce a reasonably smooth surface that does not represent substantial uneven change form one cell to the next. Finally, the hazards are extrapolated using an exponential fit to extend into cells where men or women

are aged 80 and older. The estimated daily hazards of intercourse used by the simulator are displayed in Figure 94.

It is worth nothing here that the age pattern of the fertility of the simulated population depends on two things: the male age/female age pattern of pairing which exposes couples to intercourse, and the male age/female age pattern of the daily hazard of intercourse. The fecundity of the female is solely a function of her exposure to intercourse and other fixed factors, so any female age pattern of fertility arises from that and not from an explicit modeling of the age-dependence of a female's fecundity. This is another refinement to the fertility model that needs to be made in the future.

PARAMETER STRUCTURE

The parameters used for the intercourse module are very simple, and they are not allowed to vary with time like most of the other parameters. It is assumed that the libido of individuals and couples of a certain age does not change dramatically over time, and even if it did, I have no idea where to find data that would shed some light on how it might change.

The daily hazards of intercourse are stored in a two dimensional table that is accessed by the intercourse module in order to calculate the number of intercourse events experienced by a couple whose members are of a certain age.

FERTILITY MODULE

The fertility module manages all aspects of female reproduction and the linking of children to the union from which they are created, and hence to their parents.

LOGIC

The logic employed in the fertility module is based on the interbirth interval model of human reproduction (Bongaarts and Potter 1983). The interbirth interval is divided into periods associated with:

1. the waiting time to conception,
2. the period of gestation,
3. the period of recovery following a birth, and
4. the period of recovery following a miscarriage.

The events that sequence these components of the interbirth interval are:

1. a conception leading to a birth, or
2. a conception leading to a miscarriage,
3. a birth, or
4. a miscarriage,
5. recovery and end of breastfeeding and return of fecundity after a birth, or
6. recovery and return of fecundity after a miscarriage.

The duration of each component of the interbirth interval is determined stochastically based either on a monthly hazard of occurrence or on a specified mean value with some degree of variation around. This uses the same interbirth interval model presented above in the

intercourse module, but in reverse. The interbirth interval is specified by Equation 24 and is repeated here for reference:

$$I = \frac{1.5}{f(1-e)} + S + k\left(\frac{1.5}{f(1-e)} + S^*\right)$$

Equation 31: Average Interbirth Interval

Where I is the average interbirth interval, f is the monthly the woman's fecundability (monthly hazard of conception), S is the average period of sterility associated with a birth, k is the ratio of conceptions leading to a miscarriage to the number of conceptions leading to a birth, and S^* is the period of sterility associated with a miscarriage (see above for an explanation of this relationship).

The value of k is assumed to be 0.35, and the average values of S and S^* are assumed to be 21 (nine months of gestation and twelve months of breastfeeding and recovery) and six (two months of gestation and four months of recovery) months respectively, and the variance around the non-gestation components of S and S^* is 1.8 (standard deviation of 1.33).

FECUNDABILITY AND CONCEPTION

Every month, the number of intercourse events experienced by a woman (see the Intercourse Module) is used to calculate her fecundability f based on Equation 27. A random number between zero and one is drawn to determine whether or not she will conceive during the month; if the random number is less than her fecundability for the month, she conceives. If

conception occurs, another random number is drawn to determine if the conception will terminate in a miscarriage or a birth based on the value of $\acute{e}$.

If a conception leads to a birth, the period gestation is assumed to nine months followed by a variable period of breastfeeding and recovery with a mean duration of twelve months (SD = 1.8 months). If the conception leads to a miscarriage, the period of gestation is assumed to be two months followed by a variable period of recovery with mean four months (SD = 1.8 months). After recovery or the end of breastfeeding the women is assumed to be fecund again and is exposed to conception if she is exposed to intercourse. To account for menopause, a woman's fecundability declines linearly between ages 40 and 50, and after age 50 a woman is not able to conceive at all.

When a conception leads to a birth the sex of the child is determined stochastically at the time of birth, and the proportion of males to females is assumed to be the relatively standard 102 that is empirically observed in Black populations. When a birth occurs, a new life is created in the table tblALives, the pregnancy that gave rise to the birth is closed, and the appropriate links are created between the new life and both the birth event and pregnancy to which it is related.

The current version of the simulator does not model or allow multiple births – twins, triplets etc.

MARRIED & UNMARRIED FERTILITY

Because the simulator forms and dissolves extramarital affairs, women are exposed to sexual intercourse both within and outside of marriage. As a result, there is the potential for reproduction inside and outside of marriage. The intercourse module exposes all couples to the risk of intercourse, and the fertility module uses the number of intercourse events experienced by each couple, married or unmarried, to calculate the woman's fecundability and hence her fertility.

The alert reader will notice that a problem arises here due to the discrepancy between the granularity of time used to model intercourse and the granularity of time used to model the other fertility events; namely, it is not possible to accurately determine paternity when a woman has experienced intercourse with more than one man during a month. This is a serious problem with the simulator that must be solved in future versions. The current version implements a half-way fix by employing a probabilistic algorithm to determine paternity. Briefly, the probability that paternity is assigned to an individual man is equal to the number of times he has experienced intercourse with a woman during a month divided by the total number of intercourse events the woman has experienced with any man during the month. On average this scheme assigns paternity correctly. However, on an individual basis, it does not, and in order to model genetic heretability, as I would like to do in the future, this issue must be solved correctly.

REPRODUCTIVE HISTORIES

The simulator stores the reproductive histories of all women in the table tblAFertilityEvents. For each woman this table records the complete sequence of all the fertility events she

experiences in her life from the time she starts being exposed to intercourse sometime in her late teens until menopause between ages 40 and 50. Each record in this table stores the type of event that occurred, the date on which it occurred, the ID of the woman to which it occurred, and when appropriate the ID of the union during which it took place – and hence to the male partner.

Another table in the simulator stores specific information about pregnancies. This table exists primarily as a relationship to link individual children to their mothers, the union that the mother was in at the time of conception, and the fertility events that led to their birth. Additionally it stores the date of conception and end date of the pregnancy along with the type of termination, birth or miscarriage, and the Ids of the events that define the pregnancy. This allows the straightforward analysis of pregnancies and miscarriages.

CONTRACEPTION

In its current version the simulator allows for non-age-specific contraception through the adjustment of e in Equation 31. As e increases from zero to one, the effectiveness of contraception increases from none to complete, and a woman's fecundability is reduced to zero. Correspondingly as e increases and fecundability decreases, the waiting time to conception increases to infinity as e reaches a value of one.

FETAL WASTAGE

Fetal wastage are modeled in a very crude way through the k parameter. The ratio of conceptions leading to miscarriage to the number of conceptions leading to a birth is k, and the duration a gestation leading to a miscarriage is fixed. As described below, if the woman is

infected with HIV the ratio of miscarriages to births increases with her viral load. In this system, the sex of a child is assigned at birth, and the ratio of

Reality is substantially more complex with this, and deserves to be modeled more carefully. In particular, the sex of an embryo is determined at the time of conception, and the ratio of male to female embryos at the time of conception, the primary sex ratio, is heavily weighted toward males who face a substantially higher risk of intrauterine death throughout gestation. The sex and duration-specific differential in intrauterine death is understood well enough to be modeled, and it is likely to be strongly affected by the HIV status of the mother. Although I have not found empirical data to reveal if there is a differential effect of the mother's HIV status on the male and female fetus, it is would not be surprising to discover that there is given the already existing substantial differential. Correctly modeling this in a population with HIV would lead to a sex ratio at birth, secondary sex ratio, and an overall level of fetal wastage that more realistically represents reality when women are infected with HIV.

Another birth outcome which is not modeled correctly at this time is the phenomenon of still births. Given that infection with the HIV is very likely to increase the incidence of still birth in addition to increasing the levels of fetal wastage, this aspect of the simulator needs to be given some attention.

Future work on the simulator will put significant emphasis on improving the modeling of conception, gestation and birth outcomes. The main impediment to implementing a more sophisticated model of biological reproduction is not a lack of theoretical models or modeling

machinery but rather the lack of empirical data from which reasonable parameters can be obtained.

VERTICAL TRANSMISSION OF HIV

If a woman is HIV positive, she exposes any child that she bears to the risk of infection at the time of its birth. The vertical transmission rate is a function of the mother's HIV status and takes the following form:

$$V = V_b + (1 - V_b) \cdot (\varphi_m V_{max})$$

Figure 95: Vertical Transmission Probability

Where V is the vertical transmission probability, V_b is the base vertical transmission probability, φ_m is the viral load of the mother at the time she is giving birth, and V_{max} is the maximum amount of the base probability of *not* transmitting HIV that can be added to the base probability by the viral load of the mother. See the HIV Module below for a discussion of the viral load parameter.

The values used for V_b and V_{max} are 0.2 and 1.0 respectively. This means that an HIV positive mother always has a 0.2 probability of infecting a child to whom she gives birth. Added to that is a fraction of the 0.8 probability that she does not infect her newborn. That fraction depends on her viral load (which in turn depends on the duration of her infection) and on the parameter V_{max}. The viral load is expressed as a number between zero and one corresponding to the duration of an individual's infection with HIV. V_{max} is also restricted to take a value between zero and one which corresponds to the maximum of the base probability that the

mother does *not* infect her newborn (1-V_u) that can be added to V_b by the viral load parameter. If V_{max} is one, then a very, very sick mother whose viral load is also near one will have a very high probability of infecting her newborn because the majority of (1- V_u) will be added to V_b to yield a high probability of infecting her newborn. If V_{max} is near zero, the same very sick mother will have a probability of infecting her newborn that is very close to V_b, 0.2 in the current version of the simulator. Consequently V_{max} regulates the maximum effect that can be exerted by the mother's HIV infection, and since the severity of a mother's HIV infection changes over time, together φ and V_{max} regulate how infective a mother is at various times during the progression of her disease.

A major component of vertical transmission that is not modeled in the current version of the simulator is transmission during breastfeeding. This is not difficult to do and there is sufficient empirical information to create a good parameter set. A preliminary model of lactational transmission has been constructed and tested, but lack of time to completely debug it has prevented its inclusion in this version of the simulator.

PARAMETER STRUCTURE

All of the parameters governing biological reproduction – the length of the interbirth interval – are allowed to vary by time period except for the age/age-specific daily hazards of intercourse, discussed above. The parameters are stored in a table with 40 rows and one field for each parameter. Each row contains a vector of parameters that describes one five-year period.

DESIRABLE ADDITIONS

There are a number of highly desirable improvements to be made to the fertility module. Foremost among them is the necessity to harmonize the granularity of time used to model the different processes contributing to fertility. Because of the potential for one woman to experience intercourse with more than one man over the course of a month, it would be wise to model the experience of intercourse at the granularity that is finer than one month. The choice of granularity should be on the order of the duration of time over which it is unlikely that a woman will experience intercourse with more than one man. This obviously not on the order of one month, semi-monthly, or weekly. So it must be on the order of one day or a fraction of a day. It seems very unlikely that a woman will experience intercourse with more than one man during the period of a few hours, unlikely during the period of a day, and potentially likely over the period of two or more days. Hence, a granularity of one day seems appropriate, and future versions of the simulator will model all fertility processes with a granularity of one day.

Harmonizing the time granularity of fertility processes on one day will solve the problem of assigning paternity since it will be explicitly assumed that women do not have sex with more than one man in the course of a single day. In the event that a woman is engaged in more than one union on a given day, she will be exposed to her partners randomly and stop being exposed once she has experienced an intercourse event with one of them or after being exposed to all of them without experiencing an intercourse with any of them.

Once paternity can be accurately assigned, it will become possible to model genetic heretability. That will require the definition of individual characteristics that affect behavior and interact with the environment and are tightly associated and stored with individuals. A genetic algorithm governing the selection, mixing, and transmittal of the traits to offspring at the time of conception will also need to be identified, and finally, a rudimentary "environment" will need to defined with which the traits are able to interact. This will be a substantial addition to the model as it will affect most of the modules so it will not appear immediately in future versions of the simulator. However, all modifications will be made in ways that facilitate the introduction of a " Genetics Module".

Finally, there needs to be substantial attention given to the modeling of conception and gestation, as discussed briefly above. Beginning with conception and the cardinality of the conceptus, the primary sex ratio, through gestation and the sex-specific mortality of the embryo and fetus that takes place during that time, to still birth mortality, perinatal and neonatal mortality, and finally birth, all aspects of gestation could be modeled with substantially more precision. Some thought will need to be invested in how much detail is necessary to increase the utility of the simulator, especially in light of HIV/AIDS and its potential interaction with malaria during gestation, and when the level of detail is no longer producing additional benefit – the perpetual conundrum of modeling.

HIV MODULE

The HIV module governs the transmission of the HIV from infected to uninfected people and the progression of the disease in infected individuals. The goal is to identify a parameter or a small set of parameters that accurately reflect the health status and infectivity of an HIV-positive person as time progresses. This is critical to maintaining the correct temporal dynamics of any model that seeks to include HIV as a component.

Surprisingly, there is relatively little empirical information to bring to bear on this topic. There are two indicators of the disease that are widely discussed and measured, although neither has been adequately quantified in Africa. The "viral load" is a measure of the number of virion particles per unit of infected blood, and it is well known that the viral load increases sharply after infection to peek after several weeks and then drop precipitously to nearly immeasurable levels. The viral load typically stays very low for a substantial period of time and then begins to increase very slowly but at an accelerating pace as the infected individual becomes progressively more ill, manifests more symptoms of the disease and eventually develops full blown AIDS. The viral load is directly related to the infectivity of the infected person but much more loosely related to his or her overall health.

The other commonly followed indicator is the "CD4 count" which is the number of CD4 cells per unit of infected blood. The CD4 cell is a specific immune cell that expresses the CD4 protein on its surface and is a primary target of the HIV. A high CD4 count indicates that the immune system is functioning well and that the HIV has not been able to destroy many CD4 cells. As the disease progresses the CD4 count falls indicating that the immune system is

gradually losing the battle to suppress the HIV. The CD4 count is closely related to an infected person's overall health and the ability of their immune system to ward of both the HIV and other infections. As such it is a sensitive marker of health and fitness in a way that the viral load is not. It is known that the CD4 count stays steady or surges just after infection during the initial burst of replication of the virus, and that it usually decreases very slowly thereafter at a rate that depends on a number of factors including the severity of the initial infection, the underlying strength of the infected's immune system, the overall health of the infected at the time of infection and thereafter, age and a number of other possible factors including, obviously, the presence of therapeutic agents.

To adequately model the history of the disease in an individual and the individual's potential to infect others at each time during their illness, it is necessary to track the course of both indicators. Within a single infected body there is a raging battle between the immune system and the virus with the CD4 cells playing a primary role. Knowing that, it should be possible to construct a model of the interaction between the CD4 cells, the virus particles and the parts of the body that serve as reservoirs for either the CD4 cells and/or the virus particles. That model should provide a reasonable prediction of the number of both CD4 cells and virus particles in circulation as the disease progresses. I have not found a reliable model of this type in the literature so I built one myself based on a relatively cursory reading of the literature.

I will not spend a lot of time describing this model because I have not yet incorporated it into the simulator, but I will describe it enough for the reader to see where this is going in the near future. The HIV and CD4 cells circulating in the blood are conceived as populations of

individuals whose vital dynamics depend on the size of the "other" population. The replication and death rates of both depend on the number of the other that is currently circulating. Additionally, the ability of CD4 cells to kill the HIV and prevent its replication depend on the parameters describing the underlying strength of the infected person's immune system. Since most therapeutic agents inhibit viral replication, they enter the model by affecting the virus' ability to replicate. The model is expressed using a set of relatively simple differential equations and is capable of producing time-dependant trajectories of CD4 counts and viral loads that span most of what appears to be empirically observed.

The drawback to this model is that there are not reliable empirical data from Africa to calibrate it. I also want to be able to endow it with some degree of stochastic variation, and to find a computationally efficient way of implementing it. In the absence of acceptable solutions to any of those challenges, I have chosen a very crude duration-dependent indicator of HIV diseases progression consisting of a simple "viral load" parameter that evolves over time since infection to provide a very rough indication of infectivity and health as time progresses.

LOGIC

In its current form, the HIV module is a very blunt tool. It seeks only to represent the qualitative reality that immediately after persons are infected they are mildly ill and relatively infective, that for a long time before they become terminally ill, they are relatively healthy and not very infective, and that toward the end of their illness and life they become rapidly more ill and concurrently more infective.

The manner in which this is done is arbitrary and does not correspond closely to the underlying biology. This results from the general scarcity of sufficient empirical data to calibrate a more sophisticated model, and from the exigency of completing a working model in a reasonable period of time. The solution does capture the most important underlying aspect of the progression of the disease; namely its broad dependence on time since infection.

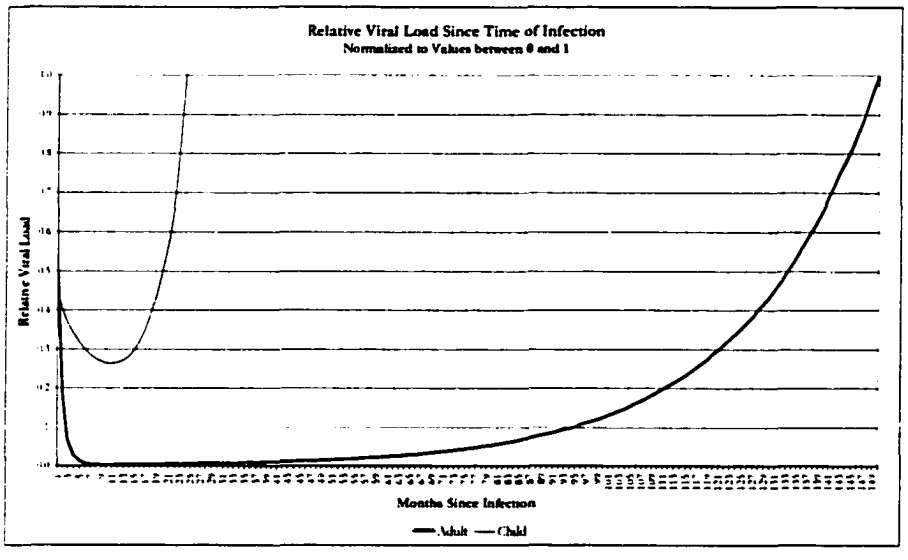

Figure 96: Viral Load by Time Since Infection

INCUBATION & DURATION DEPENDENCE

A fully parametric curve with the general shape of a elongated swish like the Nike® shoe company logo is generated to represent the "viral load" of an infected individual as a function of time. Two different time trajectories are calculated for the viral load; one for children who are infected and one for adults. Pediatric AIDS patients progress through the disease at a much faster pace than adults and face a substantially higher risk of death at all times since infection. The pediatric viral load curve is calibrated so that most of the pediatric cases die

before their second birthday, or 24 months after infection. The adult viral load curve is calibrated to allow an average adult infected with HIV to live ten years. During that time, they are relatively infective immediately after infection, not very infective over the next eight years or so, and progressively more infective during the last two to three years of their lives. The curves used in the simulator are displayed in Figure 96.

The curves are normalized to take values between zero and one, with zero corresponding to not sick at all and not at all infective and one corresponding to extremely sick, completely infective and dead. The various intermediate states between those two are represented by values between zero and one and occur at times after infection that fall on the curves displayed in Figure 96. There is nothing special about the range zero to one, and it does not represent any empirically measurable quantity; it is simply chosen in order to make the viral load parameter easy to use when the degree to which an individual has progressed through the disease must be used by other processes modeled by the simulator.

TRANSMISSION

There are only two modes through which the HIV is transmitted in the simulator: heterosexual transmission through intercourse and perinatal vertical transmission from mother to child. Homosexual and blood-sharing forms of transmission play a minor role in the spread of the HIV in Africa and are therefore not included in the model.

What would be useful to include in the simulator is some form of migration and prostitution. It is thought that mobile workers moving to and from home and work locations and possibly between various work locations have contributed greatly to the spread of the HIV around

Africa. Exacerbating this situation is the fact that most of the mobile workers are men and visit prostitutes when they are away from home (usually living in urban areas) – effectively creating a reservoir of the HIV within the community of prostitutes. These two processes work synergistically to provide a reservoir for the virus and an efficient mechanism to move it quickly into all the regions from which migrant workers are coming. Once the virus has moved to the sending communities it is able to spread by more conventional means through those communities.

Given the likely importance of migrant labor and prostitution in the spread of the HIV, it is critical to build these two components into future versions of the simulator. Given the outline of the process above, it is likely that the spread of the HIV is a two-step process with the first step occurring on a relatively short time scale, the movement from the prostitutes into the community at large, and the second step, the spread within the community, occurring on a longer time scale. It is the second step that the current version of the simulator attempts to model.

EMPIRICAL TRANSMISSION PROBABILITIES

Empirical per intercourse transmission probabilities are obtained from the Rakai project in Uganda (Gray 2000). A recent study conducted at Rakai and still being prepared for publication examines 174 monogamous discordant couples to ascertain the per intercourse probability of transmission. The average number of intercourse events per month is 8.9 (similar to the average for younger people used in the simulation), and the overall probability

of transmission per intercourse is 0.0014. The study controls for age and viral load and finds some variability with both.

Per intercourse transmission rates by age and viral load are exactly the data needed to calibrate a more sophisticated model of transmission as a function of time since infection. Unfortunately, these data came to light after the basic simulator was constructed, and there has not been time to rework the HIV incubation and transmission submodel in time for inclusion in this work.

However, the overall average per intercourse transmission probability used by the simulator calculated over infected durations of zero to eight years is set so that it matches the value of 0.0014 found at Rakai. This ignores the age dependence, and since the "viral load" disease progression indicator that is used is relatively arbitrary, it does not properly model the real viral load dependence either. In spite of that, to a first order approximation the simulator does reproduce a reasonable per intercourse transmission probability.

INTERCOURSE EVENTS & TRANSMISSION

At the beginning over every simulated month, the Intercourse Module produces the number of intercourse events experienced by each couple that survives into the month. The HIV Module identifies discordant couples and uses the number of intercourse events that they will experience to determine whether or not the uninfected partner will become infected. The probability that the uninfected partner will be infected during the month is:

$$I = 1 - (1 - p_t)^C$$

Figure 97: Probability of Infection

Where C is the number of intercourse events during the month and p_t is the per intercourse probability of transmission. The per intercourse probability of transmission used by the simulator is:

$$p_t = V \cdot V_{max} \cdot (1 - B)$$

Figure 98: Simulator Per Intercourse Probability of Transmission

Where V is the viral load disease progression indicator used by the model, V_{max} is the maximum per intercourse transmission probability, and B is an indicator variable used to specify the effectiveness of condom usage as a technique to prohibit transmission: condom is used perfectly $B = 1$, condom is used with 50 percent effectiveness $B = 0.5$, condom is used completely ineffectively $B = 0$.

ATTENUATION OF TRANSMISSION WITH CONDOMS

As alluded to above, the simulator incorporates the ability to model condom usage as a method to attenuate transmission. The simulator accounts for the fact that condom usage affects both fertility and the transmission of HIV. If condoms are used with an effectiveness of say 0.85, then the values of both e in the Fertility Module and B in the HIV module will be set to 0.85. If some other non-barrier method of contraception is used, then e will take a non-zero value and B will remain zero. This arrangement allows the simulator to examine the

different population and personal-level outcomes generated from fertility control programs that use either barrier or no-barrier forms of contraception[37].

ANTI RETROVIRAL TREATMENT

Antiretroviral therapies interfere with the replication of the HIV. The cumulative result of the many variations on this type of treatment is to lower viral load to nearly or actually immeasurable levels and keep it there. Once that is achieved it may rebound and at the very least stabilizes. The effect is to improve and stabilize the patient's overall health and effectively halt the progression of the disease. The effectiveness of the therapy depends on a large number of factors leading to a range of outcomes spanning virtually perfect suppression for unlimited periods of time to limited suppression for limited periods of time. The "average" result might be described as substantial suppression for a long time.

The current version of the simulator was not designed to implement *virtual* treatment programs, but it can crudely approximate a treatment program by altering the profile of the viral load disease progression parameter. For treated children the average time to death is extended to five to ten years, and treated adults live about twice as long as their untreated counterparts with negligible infectivity during the majority of that time.

RANDOM INFECTION

In one way or another a real population is composed of heterogeneous subpopulations that mix more freely an more consistently through time with themselves and only occasionally with

[37] It is this question that originally motivated the creation of the simulator to begin with: to examine the unintended negative results of implementing a non-barrier method of contraception as part of a fertility control program in a population with endemic HIV. Along the way it became obvious that a lot of other questions could also be examined with the simulator.

members of other subpopulations. It was suggested earlier that an important process in the spread of an HIV epidemic is the transportation of the HIV between these various subpopulations; in this case urban centers separated by long distances and urban and rural centers separated by distance and culture.

In its current form the simulator creates a closed, homogenous virtual population that might represent one of the subpopulations described above. As a result, the problem arises of how to initiate an HIV epidemic in the population and how to model the reality that there is continuous contact with "outside" populations that provides a high-risk route through which new infections are brought into the population. Because the current simulator does not model multiple populations, it cannot expose its virtual population to the kind of continuous inoculation that affects a real population.

To solve this problem a very small random fraction of uninfected adults between the ages of twenty and fifty is infected each month, in most simulations twelve per 100,000 which gives an uninfected person a 95 percent chance of surviving between age fifteen and 50 uninfected.

PARAMETER STRUCTURE

All of the parameters governing the HIV Module are constant through time: those that define the viral load disease-progression indicator, the transmission probabilities and the random infection rate. Consequently, they are stored in two simple tables.

DESIRABLE ADDITIONS

As discussed above, there must be a better disease progression submodel for future versions of the simulator. Substantial work has already been done to improve this aspect of the model, and a new submodel of disease progression that uses both the CD4 count and the viral load as primary indicators will be completed soon. This will allow a much more satisfactory modeling of HIV/AIDS in an individual and provide the correct duration-specific indicators for assessing an individual's health and infectivity. This is important because an individual's overall health affects their behavior and their susceptibility to other infections and hence their mortality. Obviously, an individual's infectivity is critical to the modeling of transmission.

Once a better disease progression model has been built, it will become possible and attractive to model opportunistic infections that often affect HIV-positive individuals. For Africa, both TB and Malaria are major opportunistic infections that should be immediately considered for modeling.

Along with a better and more comprehensive disease model, it will be necessary to differentiate and improve the mortality model. Specifically, it will be necessary to create a cause-specific model of mortality that is able to make use of the disease data produced by the modeling of HIV, TB and Malaria.

The fact that multiple populations, migration and prostitution are critical to the movement and transmission of HIV in Africa was discussed at some length. Although it will greatly increase the complexity of an already complex model, it is worth investigating the potential to model

multiple heterogeneous populations, movement between them, and some intrapopulation heterogeneity to account for prostitution.

Last, it is necessary and relatively easy to implement vertical transmission of HIV from mother to child during breastfeeding.

NUPTIALITY MODULE

The Nuptiality Module regulates the dynamics of marital union[38] formation and dissolution. This together with its sister module governing the dynamics of non-marital union formation and dissolution form the most critical components of the simulator. Anderson and Garnett (Anderson, Ng, Boily, and May 1989; Anderson 1991a; Anderson et al. 1991; Anderson 1992; Anderson and May 1992; Anderson, May, Ng, and Rowley 1992; Anderson 1996; Garnett and Anderson 1993a; Garnett and Anderson 1993b; Garnett and Anderson 1994a; Garnett, Swinton, and Parker 1994; Garnett and Anderson 1996a; Garnett, Hughes, Anderson, Stoner, Aral, Whittington, Handsfield, and Holmes 1996) have convincingly demonstrated that the nature of the pairing process largely determines the evolution of an HIV epidemic. Consequently, it is important that models of pairing processes are sufficiently flexible and precise to describe realistic behavior. This is in fact a potential shortcoming of Garnett and Anderson's model of a population with HIV. Their model (Garnett and Anderson 1993a; Garnett and Anderson 1994a) defines the mixing preferences (and flows) for males and females separately. A constraint is applied to force the number of males of a specified type (age, sexual activity level etc.) who will form new unions during each time step to equal the number of females of a specified type who will form pairs with those males. Because these two numbers are not necessarily equal it is necessary to force one of the other to conform to the opposite sex, and this is essentially what Garnett and Anderson do. The problem with this, as they themselves point out, is that the dynamics of the pairing process are dominated by

[38] A "union" in the context of the simulator is the pairing of a man and woman. There are two basic types, both having the potential to be fertile. *Marital* unions are similar in duration, parity, contact frequency and social status to a conjugal union of

the sex that is chosen to determine the number of pairs that will form, and consequently the partner acquisition dynamics of the dependent sex may be forced to be very different from what that sex actually prefers. Garnett and Anderson address this issue in one of the most recently published versions of their model by devising a scheme whereby the two sexes split the difference between their preferred mixing patterns and both compromise to the extend necessary to bring the overall flows of male partners and female partners into balance.

The Garnett and Anderson model is specified at the population level using differential equations that move proportions of groups from one state to another during each time step instead of moving individuals, and this is why they must explicitly address the problem of harmonizing the mixing preferences of males and females. A model specified at the individual-level, such as the one described here, is able to conceptualize and implement pairing in such a way as to avoid the need for *ad hoc* procedures to balance the preferences of males and females.

FORMATION OF UNIONS

LOGIC

The pairing process is fundamentally different from other processes in that it brings together two individuals with different preferences and produces a single conceptual entity that is composed of the two individuals. The difficulty in modeling the process arises from the fact that there are independent at-risk individuals with potentially different preferences who must *decide* whether or not they want to form a union. The Garnett and Anderson model defines

---

the type described in Nuptiality, under Demographic Characteristics above; and *extramarital* unions (affairs) are typically of much shorter duration and may not provide the same level of (or exposure to) contact between the lovers.

the individual men and women's preferences for forming a union separately and then calculates how many of each *want* to form a union during a time step. Because the preferences of the sexes do not necessarily match and there are different numbers of each at any point in time, the number of men who decide they want to form a union based on *their own* preferences does not match the number of women who decide that they want to form a union based on *their own* preferences. Defining the men's and women's preferences separately results in a fundamental inconsistency in which the number of men who want to form a union does not match the number of women who want to form a union.

Unit of Analysis and Union Formation Probabilities

To solve the inconsistency, the unit of analysis is shifted from individual people to couples, and the probabilities of union formation are specified with respect to potential couples of various types. Potential couples are formed by pairing individual men and women within cells defined by various characteristics of the men and women. For example, a cell may consist of all men aged 40 to 44 who have two wives and all unmarried women aged 30 to 34. The maximum number of unions that can occur in such a cell is equal to the lesser of the number of males and females in the cell – if there were 250 males and 300 females, the maximum number of unions is 250. This framework allows a probability of union formation to be defined that is equal to the number of unions that form in a cell divided by the maximum number of unions that could form in the cell. If empirical probabilities of this type are known, it is possible to assign individuals to new unions. Empirical probabilities like this are estimated and displayed in Part 2 and are used by the simulator to pair males and females into marital unions.

To accomplish the pairing, males and females within a cell are randomly paired until the number of potential unions is equal to the maximum possible. For each potential union a random number is drawn and compared to the hazard of forming a union in that cell, and if the random number is less than that hazard of union of union formation for that cell, the union is initiated and recorded by the simulator. Furthermore, during each time step all at-risk individuals of each sex are exposed to all at-risk individuals of the opposite sex once until they are paired with someone or until they have been exposed to everyone, and the order in which they are exposed to each other is random. The process through which this is accomplished is somewhat complicated and is described briefly below.

Males and females are first categorized by the attributes of interest into various sex-specific groups. When males are exposed to marital union formation, they are organized into five-year age groups, and within each five-year age group, into three marriage cardinality groups: men with zero, one, or two or more wives. Only unmarried females are exposed to marital union formation, and they are organized into five-year age groups. The result is a set of male groups and female groups that can be paired with each other; the pairing of a male group with a female group forms a cell of the type described above – for example, men aged 30 to 34 with no wives and women aged 50 to 54. A list of all possible groupwise pairings of this type is kept, and at the beginning of each time step, this list is randomized. Once all the male and female groups have been defined, all possible male group/female group pairs are created one at a time in the order defined by the randomized list of all pairs. In this way all at-risk males are exposed to all other at-risk females in a completely random fashion. If a pair is created, the

individuals are removed from the at-risk groups and not further exposed to union formation during the time step.

The important advantages of this method of pairing are: 1) it does not require an *ad hoc* procedure to equate the number of individual males and females who will form pairs during a time step, and 2) the process is free of the so called *scaling problem.* The scaling problem arises from the fact that the number of potential male/female pairs in the population is equal to the number of males multiplied by the number of females and consequently grows at a rate that is roughly proportional to $\left(\frac{N}{2}\right)^2$ where N is the size of the population. If the total number of potential pairs is the unit of analysis, then the probabilities that would have to be applied to the potential pairs to determine the initiations of unions would also have to scale to the power two. This rapidly becomes an unreasonable approach for several reasons, including the fact that something growing to the power two quickly produces a number that is too large to work with easily, and if the number of males and females is not very nearly equal, the exponent will not be two but something that could be quite different. An approach of this nature requires the kind of *ad hoc* fixing up that is not desirable. The method described above avoids the scaling problem by using the total possible unions as the basic unit of analysis instead of the total number of potential pairs – where *possible* refers the maximum number of unions that can be formed within each cell during a time, and this scales linearly with the total size of the population because it depends directly on the number of the sex with the fewest at-risk in each cell.

The hazards used to determine the initiation of unions can be obtained from an empirical estimate using good data describing the initiation of unions in a real population, or they may be generated by a function that reflects a theoretical understanding of how the pairing process should work, or the two may be combined to produce a synthetic set of hazards of union formation. This feature of the method makes it very flexible and also potentially very realistic and specific.

It also allows the hazards of union formation to change in response to the composition of the at-risk populations of males and females. It is likely that as the age structure of either or both of the male and female at-risk populations changes, or as the marital cardinality composition of the male population changes, the cell-specific hazards of union formation will also change to compensate for the relative lack of or abundance of possible pairs in different cells. For example, it is likely that the evacuation of the age groups between 25 and 40 by HIV/AIDS and the concomitant lack of potential spouses in those age groups will cause the likelihood of union formation between younger people, between older people, and between young and old people to increase above what it would be if there are reasonable numbers of middle aged people with whom to form pairs. This dynamic adjustment of the pairing probabilities is likely and must be important in determining the overall dynamics of the pairing. The simulator does not currently model dynamic hazards of union formation, but the potential to do so has been built into the method through which pairings are created. Again, dynamic hazards of union formation could be estimated from empirical data or could be synthesized from non-dynamic empirical hazards and some theoretical understanding of how dynamic hazards might behave.

The main drawback to this model of pairing is its relatively serious demand for empirical data. Few data sources in Africa are able to produce enough reliable data to fully describe pair formation, much less dynamic hazards of pair formation. This can be viewed as a problem or as something that can be solved by using a range of reasonable guesses that will elucidate the important attributes of the pairing process and the effects it may have on other population processes.

POLYGYNOUS MARRIAGE

The simulator is implemented to model polygynous marriage. In this system men may be married to more than one woman simultaneously while women may not be married to more than one man at the same time. Consequently, the at-risk populations for the formation of new marital unions are all men and unmarried women.

The simulator recognizes three different types of union formation:

1. unions formed between an unmarried man and an unmarried woman,
2. unions formed between a married man with one wife and an unmarried woman,
3. and unions formed between a married man with two or more wives and an unmarried woman.

This defines three types of at-risk pairs and consequently requires three sets of age/age-specific hazards of union formation, presented below. The simulator classifies men into one of the three categories and exposes them to the risk of forming a marital union with all unmarried women using the hazard of union formation associated with their existing marital cardinality (how many wives they already have).

EMPIRICAL HAZARDS OF UNION FORMATION

The empirical linearly scaling hazards of union formation presented in Part 2 are smoothed using two dimensional moving averages to produce the hazards of union formation used by the simulator, presented below in Figure 99 to Figure 101.

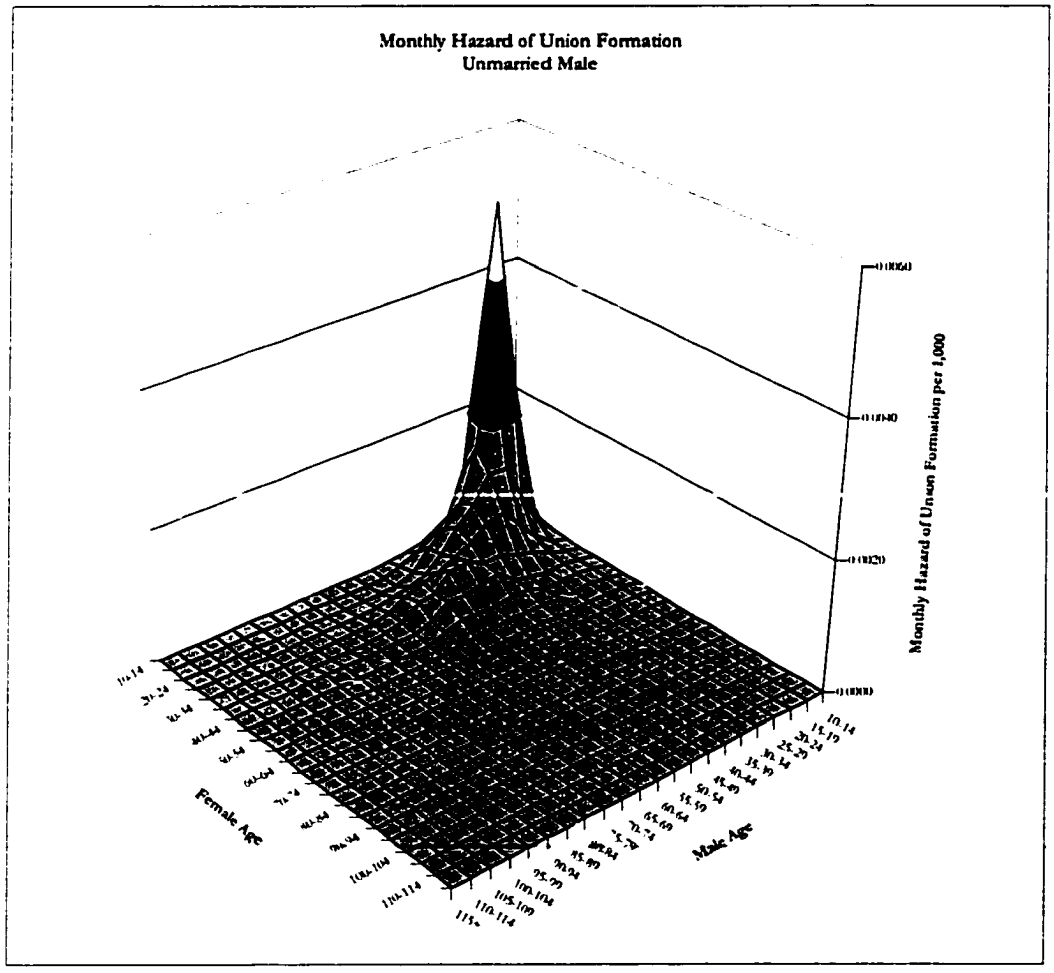

Figure 99: Monthly Hazard of Union Formation for Couples with an Unmarried Male

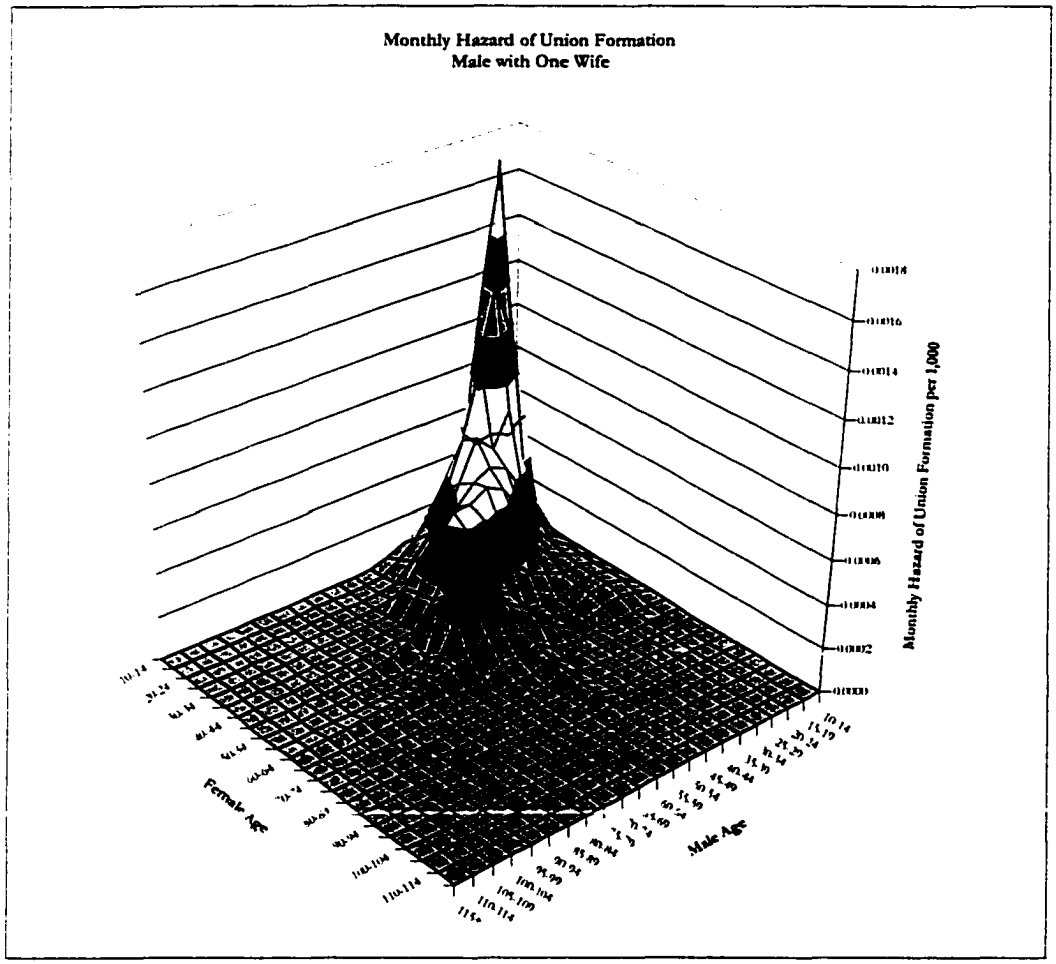

Figure 100: Monthly Hazard of Union Formation for Couples with Male with One Wife

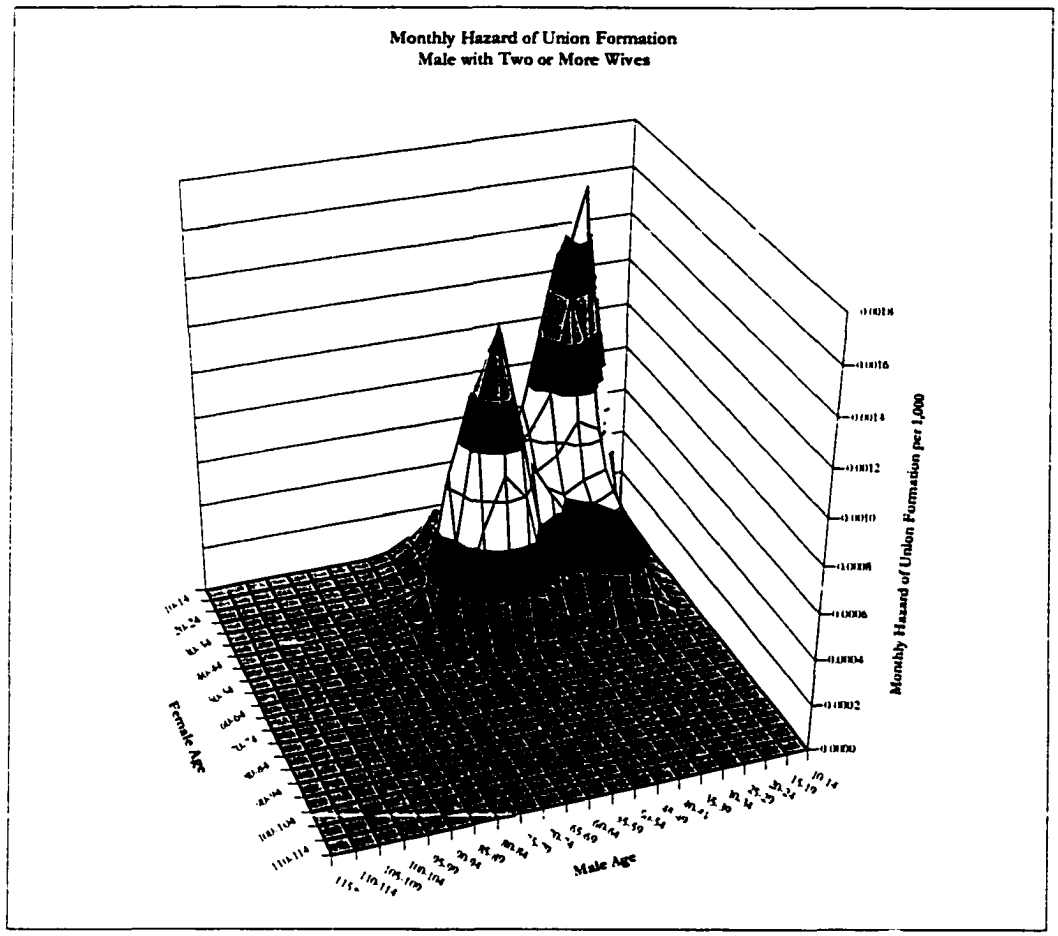

Figure 101: Monthly Hazard of Union Formation for Couples with Male with Two or More Wives

PARAMETER STRUCTURE

The monthly male age/female age-specific hazards of union formation are stored in three tables, and for each of those there is a set of 40 tables that store period-specific modifications to the base hazards. During each period, the base hazards are retrieved and modified according to the period-specific modifications stored in the modification table for the current period. All of the modifications are defined as odds ratios relating to the base period and the base age/age group, males ten to fourteen and females ten to fourteen. This allows all of the

base age/age hazards to be modified in the same way by changing only one parameter for each subsequent period. The age/age modifications are accomplished individually by modifying the corresponding age/age parameter in the period-specific table of odds ratios.

The final parameter set allows the basic age/age hazards of union formation to be varied in five-year periods over a total period of 200 years.

DISSOLUTION OF UNIONS

LOGIC

The dissolution of unions is modeled as a function of duration of union, the age of the male and female who participate in the union, and the number of surviving children produced by the union. During each time step after a union is formed it is exposed to the risk of dissolution based on the current state of the factors enumerated above. Death of either or both spouses automatically terminates a union.

SAMPLE, UNIT OF ANALYSIS, AND HAZARD OF SEPARATION

The at-risk unit is a valid marital union with both members alive. Each union that survives into the time step is classified with respect to the number of surviving children it has produced and the duration of time it has been valid. Based on those two attributes and the ages of the male and female involved, a hazard of separation is retrieved, and a random number is drawn to determine if the union will dissolve during the time step – if the random number is less than the corresponding hazard, the union dissolves and that fact is recorded in the database.

The base hazards of separation are defined with respect to *new* unions, those that have existed over a duration between zero and four years. During each subsequent five-year duration

group the base male age/female age-specific hazards of separation are modified by a duration-specific factor allowing the overall hazard of separation to be a function of the duration over which the union has been valid. The hazards and these factors are derived from the logistic regression event history model that is presented in Part 2. That model estimates the hazard of separation separately for unions with various numbers of surviving children. For each of those groups, the hazard of separation is modeled using dummy variables to be a function of male and female age and the duration of the union. Interaction terms are included for male/female age, but not for age and duration. The result is a fully independent specification of the hazards with respect to the ages of the man and woman in the union, and a separate effect that captures the duration effect. The structure of that model is reflected in the way in which the simulator handles the hazards of separation. The important result is that the whole male age/female age set of hazards is modified in the same way during each subsequent period, each male/female age-specific cell is increased or decreased to the same extent so that the male/female age-specific hazards are not a allowed to vary independently over the duration of a union.

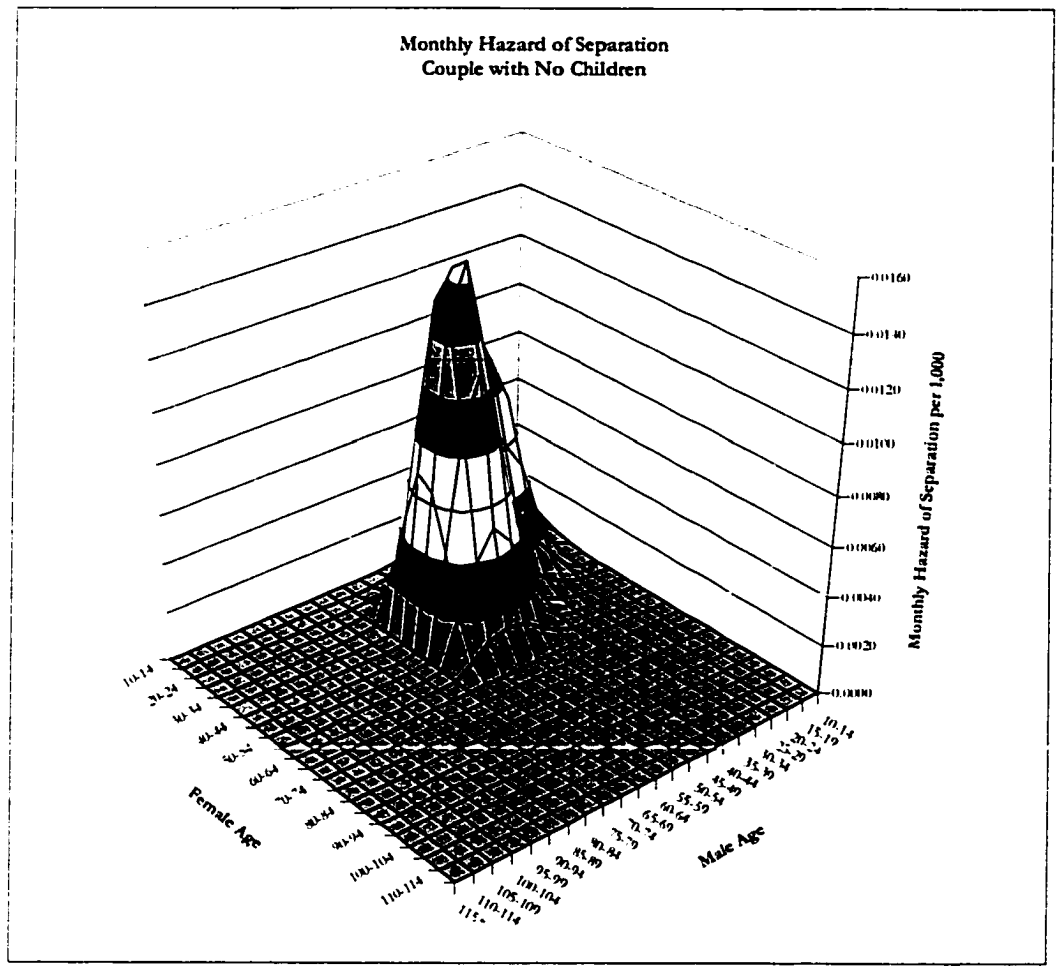

Figure 102: Age/Age-Dependent Monthly Hazard of Separation for Couples with No Children

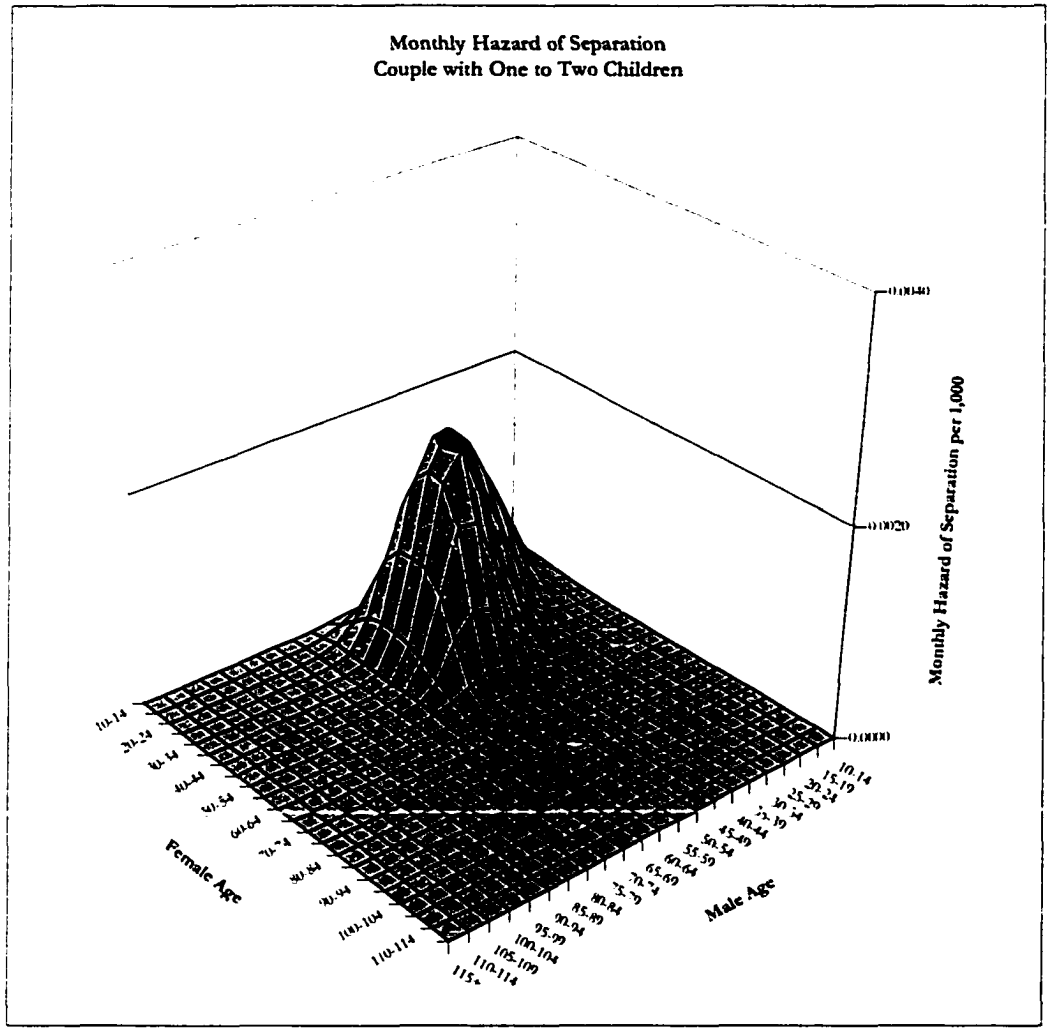

Figure 103: Age/Age-Dependent Monthly Hazard of Separation for Couples with One to Two Children

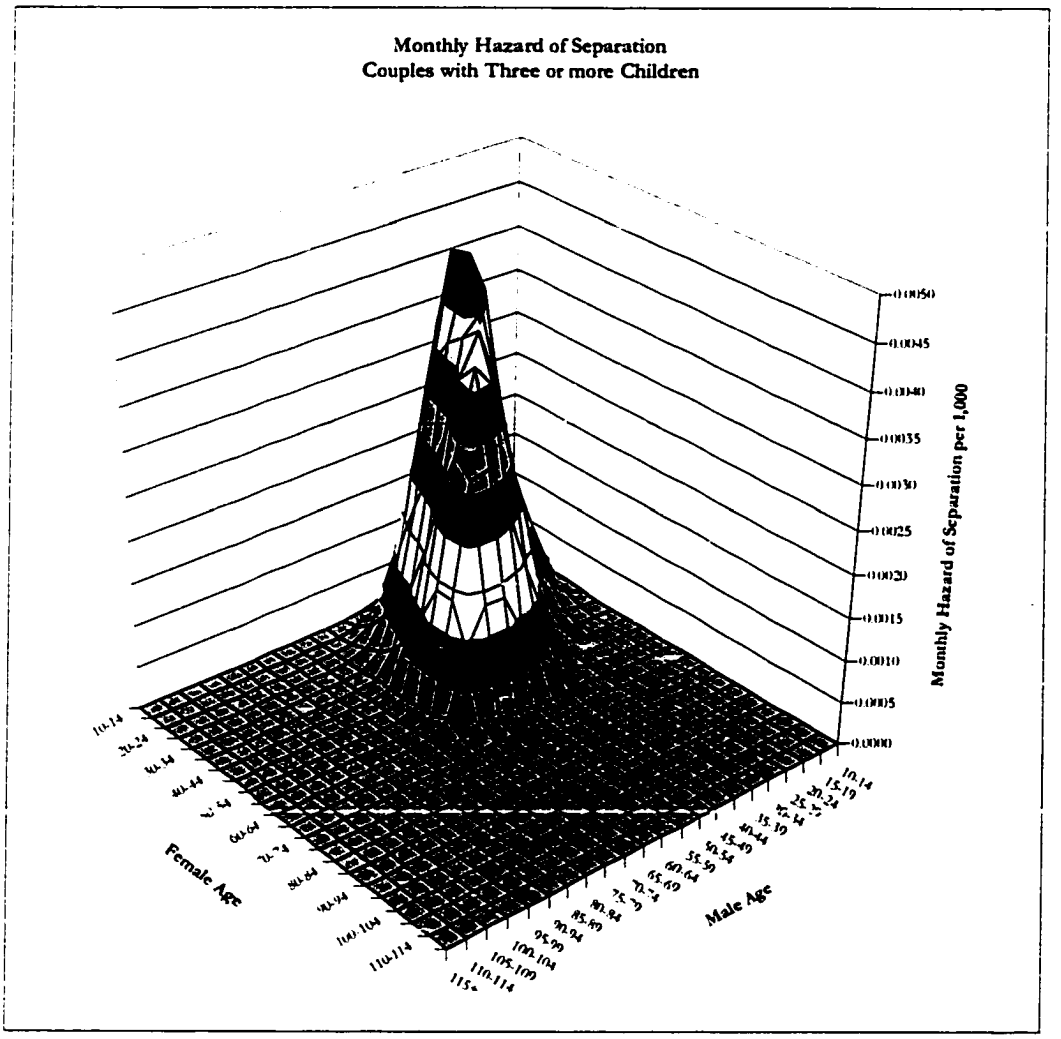

Figure 104: Age/Age-Dependent Monthly Hazard of Separation for Couples with Three or More Children

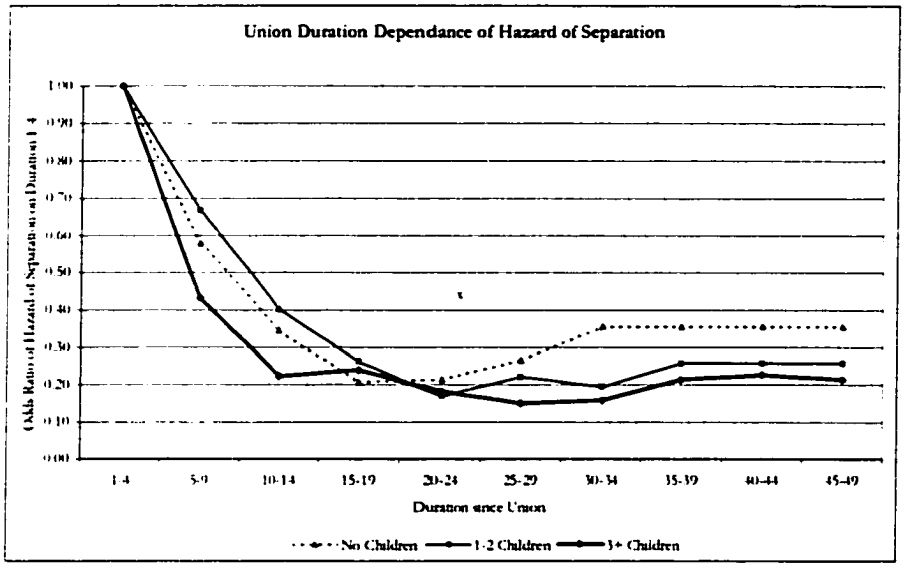

Figure 105: Union Duration Dependence of Monthly Hazard of Separation

PARAMETER STRUCTURE

Like the hazards governing the formation of unions, the base hazards governing the dissolution of unions are stored in three tables corresponding to the three fertility categories for which the hazards are specified. For each of those three tables, there are 40 additional tables specifying period and age/age-specific modifications to the base hazards. The result is a parameter set that can be varied in five-year periods over a total period of 200 years.

The duration-specific component does not support period-specific variation in this version of the simulator, but it is easy to add that capability to a future version.

AFFAIRS MODULE

The Affairs Module governs the formation and dissolution of extramarital unions between men and women. This module is critical to the simulator because it is able to define the extramarital pairing between men and women, and together with the Intercourse Module the degree to which men and women are engaging in sexual intercourse outside of marriage.

THEORETICAL DYNAMICS OF AFFAIRS

The heading for this section contains the word "theoretical" in acknowledgement of the fact that we know virtually nothing about the real extramarital sexual dynamics of any African population. It is ironic and noteworthy that the Module that is arguably most important in determining the course of an HIV/AIDS epidemic is not based on a solid empirical foundation, or in fact on any empirical foundation at all. This results from the fact that there is no empirical data describing this aspect of behavior for the Gwembe Tonga, and there appears to be very little generalizable data of this type for any reasonable sized population in Africa, or the rest of the world for that matter.

Before proceeding it is worth noting that the lack of this information is startling given the long period of time over which the need has been known and the dramatic effects of the HIV/AIDS epidemic in Africa. It appears to be a strong testament to the ever-mighty taboos protecting the secrets of human sexual behavior!

An understanding of sexual networking is critical to the understanding HIV/AIDS on both individual and population levels. The lack of this understanding is perhaps one of the key

inhibitors preventing the more rapid design and deployment of effective prevention strategies. It is also one of the reasons why models like the one described here are useful. Using a simulator that is able to reproduce a variety of sexual networks it is possible to determine which types of networks are necessary for the spread of a sexually transmitted disease, what types are particularly well-suited to a rapid transmission of the disease, and to what degree sexual networks of a given type need to be modified to slow or stop the transmission of the disease.

The model of sexual networking that is implemented in the current version of the simulator reflects the general approach taken by Garnett and Anderson, my own intuition and general lack of knowledge in this area. The likelihood that a man and woman will form an extramarital union (affair) is assumed to be a function of their ages and the degree to which each of them is sexually active. It is assumed that an individual's level of potential for sexual activity is largely a result of biological factors[39], and it is therefore a characteristic of the individual that is assigned at birth and maintained throughout life.

The process through which people have sex outside of marriage is assumed to have two components: 1) a process of finding and forming a union with someone of the opposite sex other than someone to whom you are married, and 2) the process of having sexual intercourse with someone with whom you share an extramarital union. This allows for different dynamics to be specified for the pairing and intercourse components of the overall process.

---

[39] This is certainly not completely accurate as there are many social and environmental factors that will also contribute to an individual's sexual activity potential. However, faced with a complete lack of information that might gude the modeling of those components of an individual's sexual activity potential, I have decided not to attempt to model them.

SEXUAL ACTIVITY INDICES

At birth each individual is assigned a value from zero to five reflecting their potential for sexual activity; a value of zero corresponding to a very low potential for sexual activity and a value of five corresponding to a very high potential for sexual activity. For convenience sake the sexual activity potential of the population is assumed to be distributed according to a normal distribution with mean 3.0 and standard deviation 1.2. This assigns roughly 10.6 percent of the population to the lowest and highest values, 23.3 percent of the population to the next lowest and highest values and 32.3 percent of the population to the middle value.

SEXUAL ACTIVITY ASSOCIATIVITY

It is assumed that the formation of extramarital unions and the potential for sexual activity within those unions is influenced by the sexual activity potential of the two partners. The form of this effect is thought to be strongly assortative reflecting the fact that men with a high potential for sexual activity are more likely to form liaisons with women who also have a high potential for sexual activity. Within a given liaison the potential for sexual intercourse to occur is related to the overall potential for sexual activity possessed by the couple, in other words the sum of their individual sexual activity potential values.

AGE ASSOCIATIVITY

Age is also assumed to affect the potential to form extramarital unions but not to affect the potential for intercourse within those unions once they have formed. As with the sexual activity potential values, it is thought that the age preferences of both the males and females is strongly assortative reflecting the fact that men prefer to engage in extramarital relationships

with women who are younger than themselves by a given number of years, and that women generally adhere to the same preference in reverse.

FORMATION OF AFFAIRS

LOGIC

Like the formation of marital unions, samples of at-risk males and females are identified, categorized by various characteristics, members of each category from the two sexes are paired, and within each of the cells formed by those pairings, individual males and females are exposed to the risk of forming an extramarital union. The strategy employed is exactly analogous to that employed to form marital unions; the only differences being the nature of the categorization of at-risk males and females and the manner in which the monthly hazard of union formation for each type of possible couple is obtained.

SAMPLE STRUCTURE AND UNIT OF ANALYSIS

All males and females between the ages of ten and 80 are eligible to form an extramarital union, and there is no limit to the number of simultaneous extramarital unions that an individual of either sex may have.

Corresponding to marital union formation, the unit of analysis is the couple, and the probability of forming an extramarital union is defined with respect to the attributes of a possible couple and the duration of the time step. At-risk males and females are categorized into five-year age groups, and the groups defined in that way are paired to yield age/age cells in which possible extramarital unions are formed by randomly pairing males with females until the number of unions is equal to the lesser of the number of males and females in the cell. All

of those randomly chosen possible couples are then exposed to the hazard of forming an extramarital union.

AFFAIR FORMATION PROBABILITIES

The monthly hazard of forming an extramarital affair is a function of the ages and sexual activity potential values of the man and woman who are part of a possible pair. An age associativity score from zero to one is calculated based on the preferred age differential between the sexes, and a sexual activity potential associativity score, also between zero and one, is calculated based on the preferred sexual activity potential differential between the sexes.

In both cases the preferred associativity is defined by specifying the location and width of a *normal ridge* running through either the male age/female age space or the male sexual activity potential/female sexual activity potential space, an example of such a ridge is displayed in Figure 106. Imagine either of those two spaces as a two-dimensional field though which is running a linear ridge whose cross section is a normal distribution. The course of the ridge is defined by the line that it follows in the two-dimensional space, and the width of the ridge is defined by the variance of a normal distribution centered on the line. This specification takes three parameters to define the normal ridge: 1) the slope of the line that it follows, 2) the vertical offset of the line that it follows, and 3) the variance of the normal distribution that determines the "width" of the ridge. The slope corresponds to a linearly varying difference in the preferred values of the male and female attributes, the vertical offset corresponds to a constant difference in preferred values of the male and female attributes, and the variance corresponds the precision of the preferences; a small variance means that the male and female

values must be very close to the preferred values to yield a high associativity score while a larger variance means that male and female values that are some distance from the preferred values still yield a relatively high associativity score. To determine a couples' associativity score in either age or sexual activity potential space, the value of the normal ridge is calculated at the point defined by the male and female values.

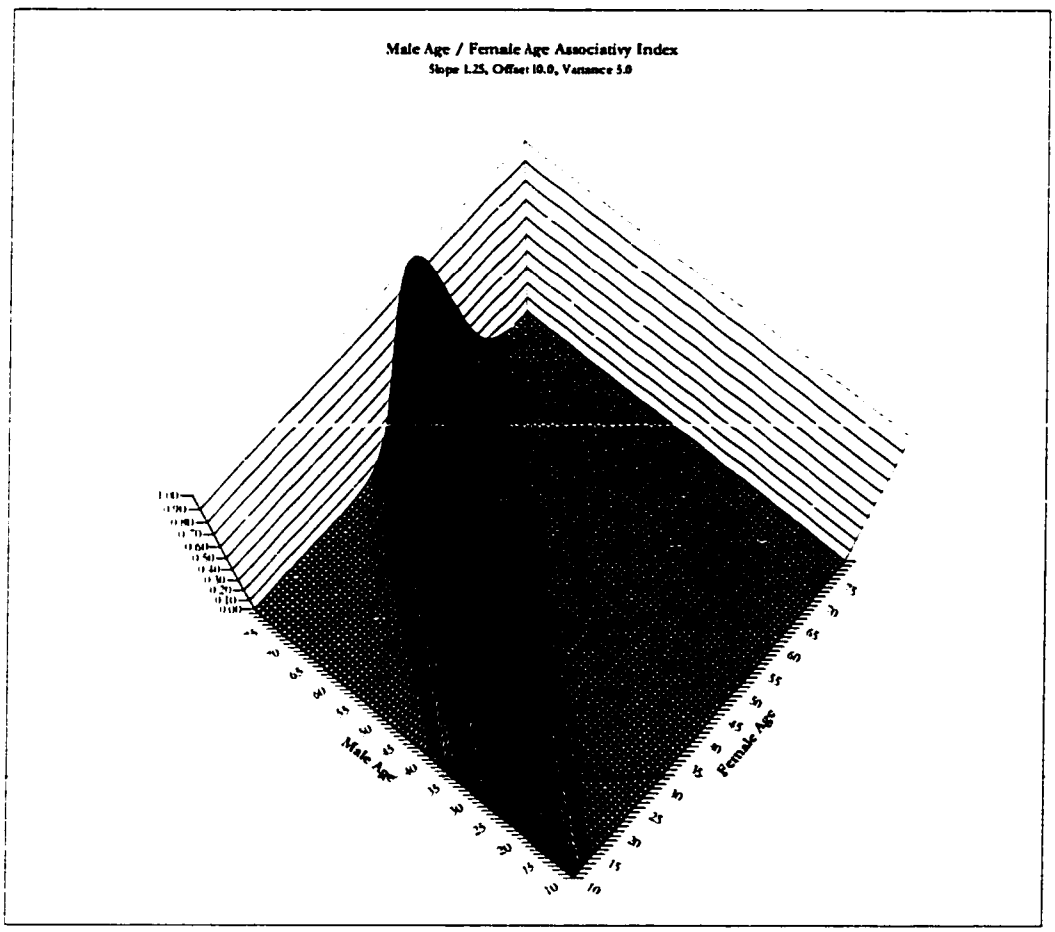

Figure 106: Male Age / Female Age Associativity Index

The normal ridge plotted in Figure 106 corresponds to a preferred couple whose male partner is about ten years older than the female partner, in which that preferred age difference increases by 0.25 years for every year of the female's age, and in which the preference is not too strong – as long as the couple is within five units or so to the maximally preferred difference, they will still have a reasonably high associativity score. The associativity parameters used by the simulator are displayed in Table 60 below.

TABLE 60: ASSOCIATIVITY PARAMETERS

	Associativity Dimension	
Parameter	Age	Sexual Activity Potential
Slope	1.0	1.0
Offset	7.5	0.0
Standard Deviation	3.0	1.0
Influence Exponent	$\alpha = 1.0$	$\sigma = 0.5$

The age and sexual activity potential associativity scores are combined to produce the hazard of extramarital union formation according to Equation 32 below.

$$H_{as} = A^{\alpha} \cdot S^{\sigma}$$

Equation 32: Monthly Hazard of Extramarital Union formation

Where H_{as} is the monthly hazard of extramarital union formation as a function of the age and sexual activity potential associativity scores, A is the age associativity score, S is the sexual potential associativity score, α is an exponent governing the influence of the age component, and σ is an exponent governing the influence of the sexual activity potential component.

The multiplicative combination of the indices satisfies the constraint that their combination must lie between zero and one and be equal to zero when either or both of the indices is equal to zero. By varying the exponents and their relative value the contribution of the two factors can be adjusted. Figure 107 displays all of the possible combinations of age associativity and sexual activity potential associativity scores and the monthly hazard of extramarital union formation associated with each combination, for the values of the exponents used by the simulator.

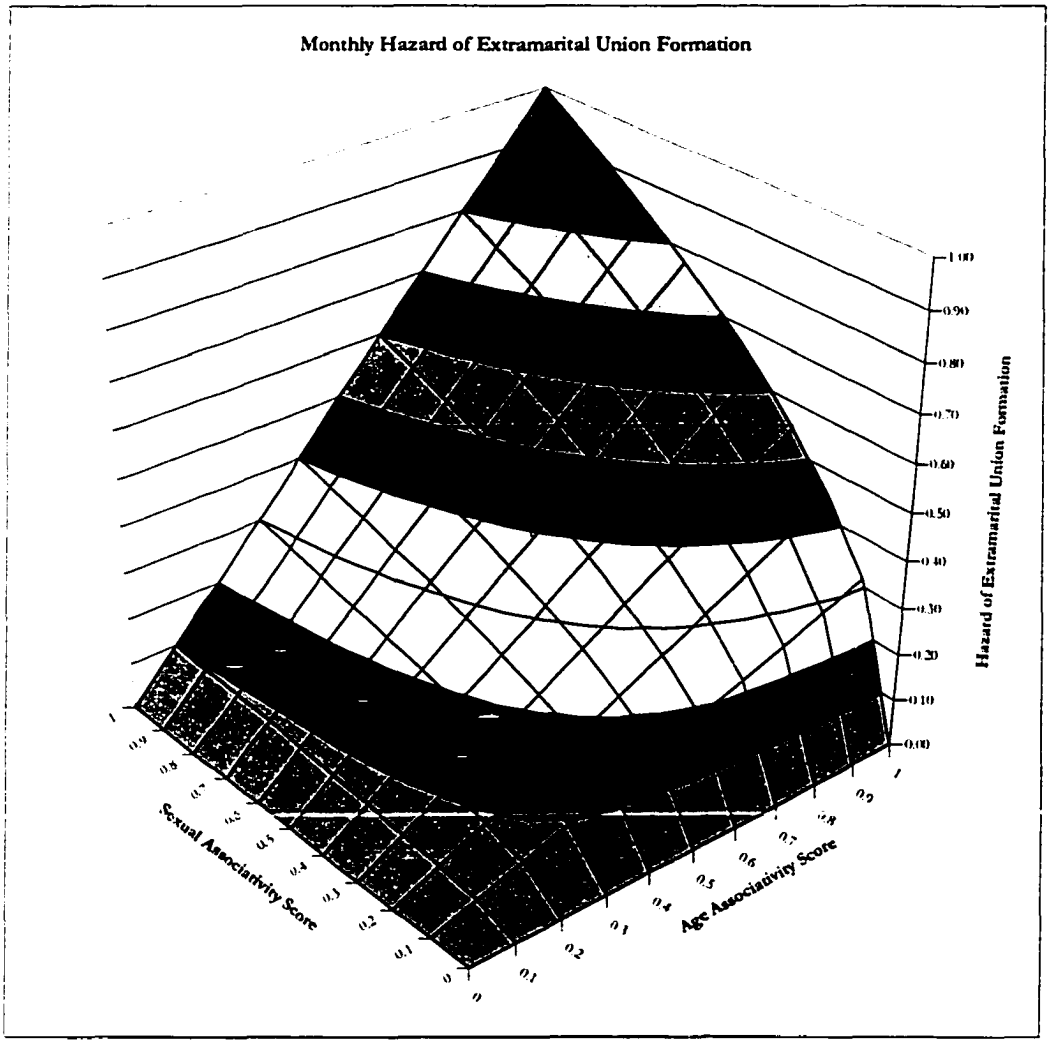

Figure 107: Monthly Hazard of Extramarital Union Formation

These values give the age associativity score a linear influence over the hazard, while the sexual activity potential score has an influence proportional to its square root – this means that as long as the sexual activity potential score is not too close to zero it does not have as strong an effect on the hazard.

SEXUAL INTERCOURSE WITHIN AFFAIRS

During the time when an affair is valid, the two people participating in the affair are exposed to the hazard of engaging in sexual intercourse. Again, since there is no empirical basis on which to model the hazard of experiencing intercourse within an affair, a simple intuitively reasonable model is chosen that assumes that the daily hazard of intercourse within an affair is proportional to the daily hazard of intercourse that a similar couple would experience if they were married *plus* an extra amount that is proportional to the sum of the sexual activity potential values of the two individuals participating in the affair. This extra amount that is added based on the combination of the sexual activity potential values of the partners is displayed in Figure 108. The simulator uses a maximum addition of 0.2.

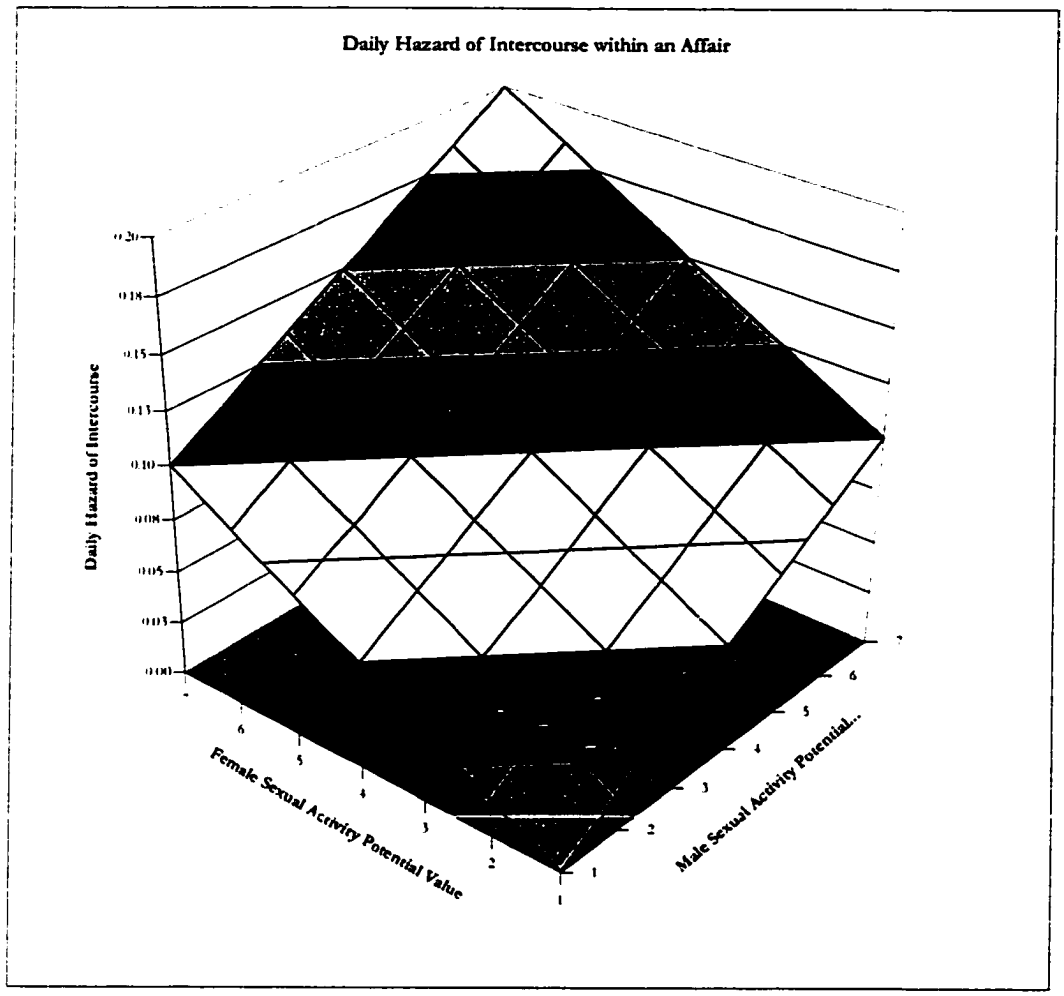

Figure 108: Daily Hazard of Intercourse within an Affair

The values displayed in Figure 108 are those that are used in the current implementation of the simulator.

DISSOLUTION OF AFFAIRS

Affairs are subject to a constant hazard of dissolution of 0.45 per month. At that rate the half life of a cohort of affairs is 1.45 months. A future implementation of the simulator might

make the hazard of dissolution a function the duration of the affair, the outcome of any fertility events occurring within the affair, other attributes of the affair itself, and the characteristics of the individuals involved in the affair. However, in light of the fact that virtually nothing is known about the likelihood that an affair will end, this version of the simulator keeps it simple.

SUMMARY OF AFFAIRS

The sexual intercourse that takes place outside of marriage is critically important in determining how a sexually transmitted disease spreads through a population, what proportion of the population it is able to infect, how fast it is able to spread, and how stable the epidemic that it creates is. It is unfortunate that there is a substantial lack of empirical data on which to build a model of this process

In the absence of that data, a model of this type may help in defining the important dimensions of the process and how they interact with each other. This can be accomplished by performing a range of simulations with parameters that span the plausible parameter space, or with different forms of the submodels that govern the dynamics of affairs and sexual intercourse within affairs. Comparing the results from those simulations will help to understand the underlying features of the process and highlight the features of the system that are most critical and thus most worth investigating. This may help in prioritizing the investigation into the different aspects of sexual behavior that must be undertaken.

The current version of the simulator assumes that people will form affairs based on the relative ages and sexual activity potential values of the men and women. The close a possible couple's

age difference is to the preferred age difference (7.5 years), and the closer their sexual activity values are to each other, the more likely they are to form an affair. This means that men and women whose age difference is near optimum and who both have similar sexual activity potential values are most likely to form affairs. All affairs are subject to a constant hazard of dissolution that gives a cohort of affairs a half life of about 1.45 months.

MODULE INTERACTIONS & ASSOCIATIONS

IMPLICIT INTERACTIONS

The simulator models all of the important demographic processes of a complete, closed population. Because it is a probabilistic state-transition machine, each transition depends on the state of the individual at-risk unit, and the total number of transitions from one state to another depend on both the states of the at-risk units and how many of them there are (the base population at risk). Demographic effects are able to affect the future composition of the population and how quickly it will change by feeding back through those two channels – namely a global influence over the size of the at-risk population and what distribution of states it occupies at any given point in time. It is the necessity to capture these kinds of interactions that provides the motivation for building a model of this type in the first place. These effects are not modeled explicitly but result from the combined functioning of the explicitly modeled processes.

One of the most important implicit interactions is between the transmission of HIV and fertility. Each act of sexual intercourse exposes a couple to the risk of conception and the risk of transmitting HIV if they are a discordant couple, and it seems natural that an explicit modeling of sexual intercourse is the easiest and best way to insure that fertility and the transmission of sexually transmitted diseases are properly linked. That is exactly what the model does with the result that the levels of fertility and transmission of HIV in the population are naturally in harmony because they are both mediated through the intercourse model.

EXPLICIT INTERACTIONS

Explicit interactions between processes are those that are identified and provided with a defined form, and all of the explicit interactions implemented in the simulator have to do with the state of being HIV positive.

HIV

In the current version of the simulator, being HIV positive affects:

- the probability that a conception will lead to a miscarriage,
- the fecundity of an infected female,
- the daily hazard of intercourse between a male and female one or both of whom are infected,
- the probability of transmitting the HIV from an infected to an uninfected person through sexual intercourse,
- the probability of transmitting the HIV to a child if the mother is infected,
- the probability that a possible couple with one or both possible partners infected will form a marital union,
- the probability that a marital union will dissolve if one or both of the partner is(are) infected, and
- the probability that an infected individual dies.

Being HIV positive should at the least also affect :

- the probability that an infected individual will engage in an extramarital affair,
- the probability that two eligible (willing) people will form an extramarital affair if one or both are infected, and
- the probability that an extramarital affair will end if one or both of the participants is(are) infected.

NUPTIALITY

It is assumed that an HIV-positive individual will experience a reduced likelihood of forming a marital union, and that the magnitude of the reduction is related to the degree to which the

individual is affected by the disease. The viral load parameter disease progression parameter discussed above is the overall indicator of how sick an infected individual is at any time after they have been infected. Consequently, the viral load parameter mediates the degree to which the hazard of union formation is reduced when one or both of the possible partners are HIV-positive. The maximum affect of being HIV-positive is governed by a parameter that specifies by what maximum amount the hazard of union formation may be reduced by an infected partner. Equation 33 describes the influence of HIV on the hazard of formation of a marital union.

$$H_i = H_b \cdot \left(1 - R^N \cdot V_m\right) \cdot \left(1 - R^N \cdot V_f\right)$$

Equation 33: Effect of HIV on Hazard of Marital Union Formation

Where H_i is the hazard of marital union that is applied to at-risk possible couples, H_b is the base hazard of marital union formation that prevails when neither partner is HIV-positive, R^N is the maximum amount of the base hazard that can be subtracted by an HIV-positive individual, and V_m and V_f are the values of the viral load disease progression parameters for the male and female respectively. Remember that the disease progression parameter takes values between zero and one corresponding to the degree to which an infected individual is both infective and ill.

As Equation 33 makes clear, if an individual is HIV negative with a viral load disease progression parameter value of zero, they exert no influence on the base hazard. As they become sicker and their viral load disease progression parameter increases, they subtract

progressively more from the base hazard up to the maximum specified by R^N. When both possible partners are infected, the affect that they exert is multiplicative with a maximum reduction of $(1-R^N)^2$.

The hazard of marital union dissolution is also affected if one or both of the partners in a union are HIV-positive. In this case, the hazard of union dissolution is increased by one or both of the partners being HIV-positive, and as with the hazard of union formation the effect is mediated by the viral load disease progression parameter. This insures that the strength of the effect is approximately keyed to the state of each individual's infection, but more importantly that the strength of the effect has the correct dependence on the duration since infection.

$$S_i = S_b + \left(\frac{1-S_b}{2}\right) \cdot R^S \cdot (V_m + V_f)$$

Equation 3-4: Effect of HIV on the Hazard of Marital Union Dissolution

Where S_i is the hazard of marital union separation that is applied to at-risk couples, S_b is the base hazard of marital union separation that would prevail when neither partner is HIV-positive, R^S is the maximum fraction of half of the hazard of not separating that can be added to the hazard of marital union separation by an infected partner, V_m is the viral load disease progression parameter value for the male, and V_f is the viral load disease progression parameter value for the female.

According to Equation 34, if neither partner is infected the hazard remains unchanged; if one is infected, a fraction of half of the hazard that the union will not dissolve $(1-S_b)$ is added that is proportional to the disease state of the infected individual; and if both are infected, they both contribute an addition up to a maximum of $(1-S_b)\cdot R^S$.

AFFAIRS

As alluded to above, the fact that an individual is HIV-positive should affect the likelihood that they will become eligible for, initiate, and terminate an extramarital affair. At this time the simulator does not model these affects because I have not had sufficient time to identify (or imagine) what those affects might be. The logical implementation of the effects is straightforward, and they will be included in future versions of the simulator.

MORTALITY & MORBIDITY

One of the primary outcomes of being HIV-positive is an early death. The simulator takes this into account by adding a fraction of the hazard of surviving to the hazard of dying based on the disease state of an infected individual. Like all of the HIV effects, the effect on mortality operates through the viral load disease progression parameter so that the effect is duration-dependent and keyed to the progression of the disease within an individual. Equation 35 describes how HIV affects the hazard of death.

$$M_i = M_b + (1-M_b)\cdot R^M \cdot V$$

Equation 35: Effect of HIV on the Hazard of Death

Where M_i is the hazard of death that is applied to at-risk individuals, M_b is the base hazard of death that would prevails when the individual is not infected, R^M is the maximum fraction of the hazard of surviving that can be added by the fact that the individual is infected, and V is the viral load disease progression parameter indicating at what stage in the disease the individual is.

If a person is not infected there is no affect of HIV since V is equal to zero. The maximum affect is exerted when V is equal to 1, when an infected person is at the height of the disease and experiencing full blown AIDS, and in that case the hazard of death is increased by adding a maximum of $(1 - M_b) \cdot R^M$.

FERTILITY

The overall reproductive potential of the population is affected by HIV in a number of ways. Infected men and women presumably face a diminished likelihood of forming marital and extramarital unions, to a degree that is proportional to how sick they are. This results in women experiencing a lower exposure to intercourse and hence a lower fecundability. Moreover, an infected woman's body is also affected, and there is some evidence that infected women are less fecund. Together with reduced exposure to intercourse, this definitely reduces fecundability.

After conception, there is substantial increase in the incidence of miscarriages in women who are infected resulting in fewer births. However, since a miscarriage quickly places a woman back at risk of conception, the affect on fertility is not as great as it may seem.

Finally although it does not affect the number of births, it is true that infected mothers have a substantial likelihood of infecting their offspring at birth and through breastfeeding thereafter. The fact that a large fraction of infants born to infected mothers end up dying shortly after birth significantly reduces the net fertility of infected women.

UNION FORMATION & DISSOLUTION

As discussed above, HIV-mediated reduction in the hazard of union formation and increase in the hazard of union dissolution has the net effect of reducing women's exposure to intercourse and hence their fecundability. The specific form of the relationship between an individual's HIV status and those two hazards is specified directly above in the section dealing with the effect of HIV on Nuptiality.

INTERCOURSE

An individual infected with HIV is less likely to engage in sexual intercourse. The form of this effect is captured in Equation 37 which is analogous to Equation 33 because it deals with a couple-specific phenomena – one that requires two individuals to occur.

$$I_i = I_b \cdot \left(1 - R^I \cdot V_m\right) \cdot \left(1 - R^I \cdot V_f\right)$$

Equation 36: Effect of HIV on Hazard of Intercourse

Where I_i is the daily hazard of intercourse experienced by couples, I_b is the base hazard of intercourse that prevails when neither partner is HIV-positive (different for married and non-married couples as described above), R^I is the maximum amount of the base hazard that can

be subtracted by an HIV-positive individual, and V_m and V_f are the values of the viral load disease progression parameters for the male and female respectively.

As with Equation 33, the minimum effect when neither partner is infected is for the applied hazard to equal the base hazard (no effect), and the maximum effect when both partners are infected and at a terminal stage in the disease is for the base hazard to be reduced by $(1-R^I)^2$.

FECUNDITY

The biological fecundity of women infected with HIV is reduced. The model accomplishes this by applying Equation 37.

$$F_t = F_b \cdot (1 - R^F \cdot V)$$

Equation 37: Effect of HIV on Fecundity

Where F_t is the fecundity (monthly hazard of conception) that is applied, F_b is the base fecundity that is experienced when the woman is not infected, R^F is the maximum fraction that can be subtracted from the base fecundity when an infected woman is at the terminal stages of the disease, and V is the viral load disease progression value of an infected women.

Uninfected women experience the base hazard of conception while an infected woman with terminal AIDS has her fecundity reduced by a factor of $(1-R^F)$.

FETAL WASTAGE

The proportion of conceptions that end in miscarriage is greater for women who are infected with HIV. The simulator models this using

$$K_i = K_b + (1-K_b)\cdot R^K \cdot V$$

Equation 38: Effect of HIV on Fetal Wastage

Where K_i is the ratio of conceptions that lead to a miscarriage to the number of conceptions that will lead to a birth that prevails, K_b is the base ratio that is experienced when the woman is not infected, R^K is the maximum fraction of the base ratio than can be added to the base ratio when a woman is infected and at the terminal stage of the disease, and V is the woman's viral load disease progression value.

Again, the minimum effect when a woman is not infected is to leave the base ratio unchanged, the maximum effect when a woman is at the terminal stages with AIDS is for the base ratio to be increased by adding a maximum of $(1-K_b)\cdot R^K$

VERTICAL TRANSMISSION

The probability that an infected mother will transmit the HIV to her newborn infant is composed of two variable components. The basic vertical transmission rate is the underlying probability shared by all infected mothers no matter at what stage of the disease they are. The

variable component is added to that and depends on the stage of the disease at which the mother is when she gives birth. The variable component is keyed to the mother's viral load disease progression parameter so that it takes into account the duration dependent aspect of a mother's infectivity. This relationship is modeled with the no familiar form of an additive effect, Equation 39

$$T_i = T_b + (1 - T_b) \cdot R^T \cdot V$$

Equation 39: Perinatal Vertical Transmission Probability

Where T_i is the probability of transmitting the HIV from an infected mother to her newborn infant at the time of birth, T_b is the base vertical transmission probability experienced by all infected mother's and their infants, R^T is the maximum of fraction of the probability of not transmitting the HIV that can be added to the base probability by a mother who is in the terminal stage of AIDS, and V is the mother's viral load disease progression value.

The probability that an uninfected mother transmits the HIV to her newborn is zero. For mothers who are infected, the probability ranges from T_b to $T_b + (1 - T_b) \cdot R^T$.

PARAMETER VALUES USED FOR HIV INTERACTIONS

The values for the various "R" parameters used to mediate the HIV effects in the simulator are displayed in Table 61.

TABLE 61: HIV INFLUENCE PARAMETER VALUES

"R" Parameter	Value
R^N	0.80
R^S	0.45
R^M	0.90
R^I	0.80
R^E	0.10
R^K	0.90
R^T	1.00
T_3	0.20

SIMULATION OUTPUT

Three types of simulations are run to test the simulator. First, a stable population must be generated so that any changes observed when HIV is introduced are due to the impact of HIV and not some naturally occurring change in an unstable population. The resulting stable population is then infected with HIV or infected with HIV and a simulated treatment program and allowed to run into the future. The three simulated populations are then compared at various points in time to ascertain to what degree infection with HIV and infection with HIV and a treatment program cause the populations to be "different" on a number of measures including:

- overall size,
- growth rates,
- sex composition,
- age structure,
- HIV prevalence,
- HIV incidence,
- raw numbers of HIV cases living in the population,
- mortality rates,
- fertility rates.

One of the aims of creating a stochastic individual-level simulator is to be able to repeat each simulation many times in order to estimate the variance in the group of simulated populations. Unfortunately, that is not possible within the time frame necessary to complete this work given the computational capacity at my disposal. Consequently, one must view what is presented here as a *proof of concept* that clearly demonstrates that the simulator works and has the potential to be useful, *NOT* as a polished product ready for immediate use. This limitation in capacity

was recognized early in the process of creating this model and was addressed in parallel to the completion of this phase of the work. A powerful computer has been acquired to move this work forward in a substantively useful way.

So, what is presented below must be viewed as a test of the simulator machinery, *not* as a realistic analysis of the impact of HIV on a population. The results presented are the very first to have been produced by the simulator, and because the prototype version of the simulator runs on a PC and takes many hours to generate the 25 years of simulation for each scenario, repeats, refinements and multiple trials have not been possible. Taken as a whole, the results of these initial runs are very encouraging as the clearly demonstrate that the simulator works and even the unrefined parameter set is able to generate a population with largely believable demographic indices. They also indicate clearly that the HIV module needs refinement, and more specifically, the manner in which HIV epidemics are initiated needs to be investigated in some detail.

Take as a whole, the initial results validate the conceptual and technical design of the simulator and clearly justify the additional work and resources necessary to transform the prototype into a fully functioning simulator with realistic parameter settings.

CALIBRATION – STABLE POPULATION

An initial seed population of 200 males "born" in month one and 200 females "born" in month 60 (so that they would be five years younger than the males) is created. This seed population is evolved into a stable population by simulating 1,500 months starting in month

300 and running to month 1,800. All of these values are chosen for convenience in order to make the population begin reproducing quickly.

STABLE POPULATION SIZE, GROWTH AND COMPOSITION

Figure 109 through Figure 112 demonstrate that the resulting population is very stable and growing at an average proportional rate of 4.97 percent during these five periods, see Table 62 below. The population counts are taken at the mid point of the five-year periods beginning in months 1,500; 1,560; 1,620; 1,680 and 1,740.

TABLE 62: STABLE POPULATION PROPORTIONAL GROWTH RATES

Period	Female	Male	Both
$1 \rightarrow 2$	5.02%	5.10%	5.06%
$2 \rightarrow 3$	4.96%	5.25%	5.10%
$3 \rightarrow 4$	5.04%	4.88%	4.96%
$4 \rightarrow 5$	4.76%	4.95%	4.86%
Weighted Average	4.92%	5.02%	4.97%

These growth rates exceed those of the Gwembe population from which the parameters are derived and are very high for a real population; however, they are within the realm of reason and suffice for the purposes of testing the simulator. Some investigation into the discrepancy will be carried out with an aim to improving future versions of the simulator.

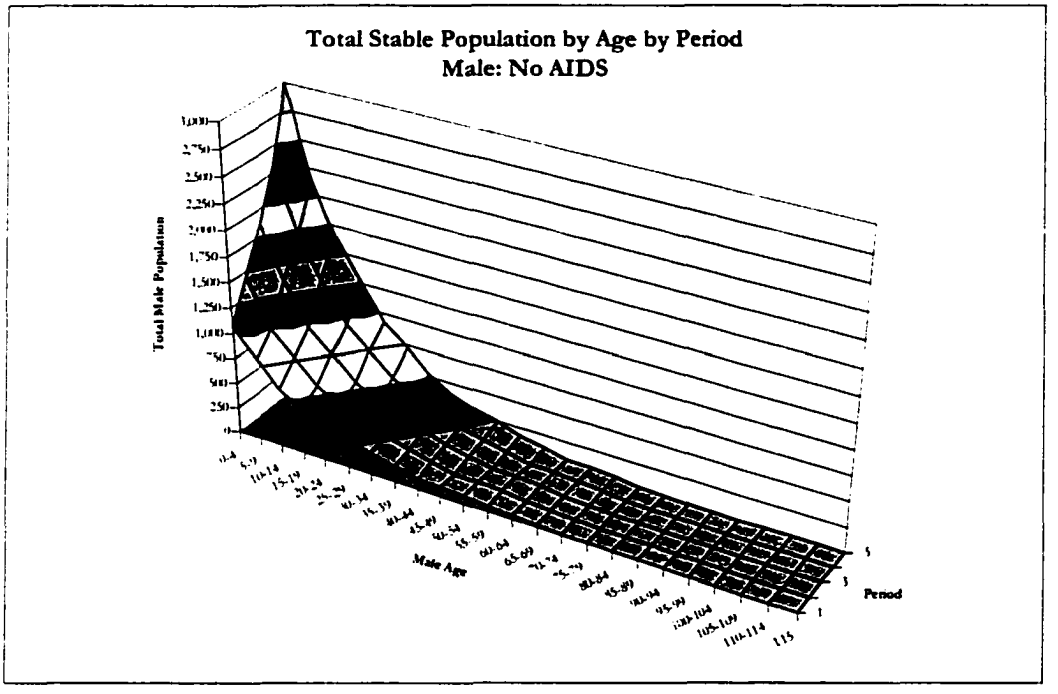

Figure 109: Total Stable Population by Age by Period – Male, No AIDS

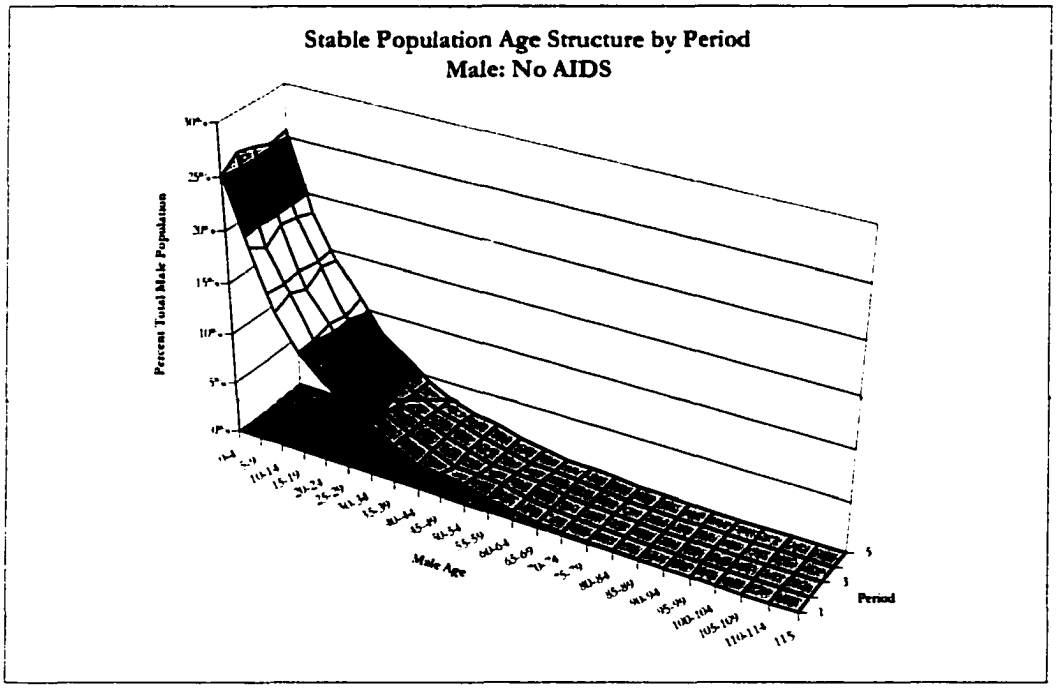

Figure 110: Stable Population Age Structure by Period – Male, No AIDS

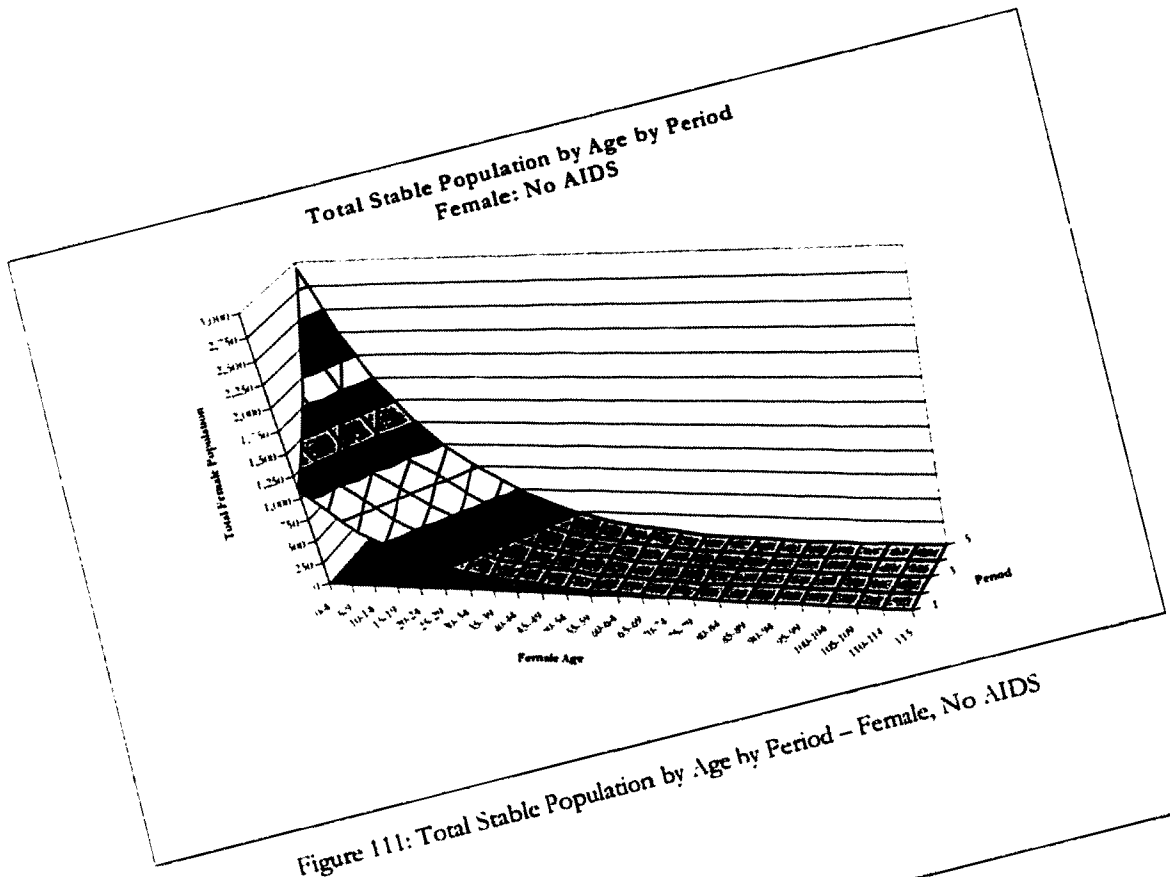

Figure 111: Total Stable Population by Age by Period – Female, No AIDS

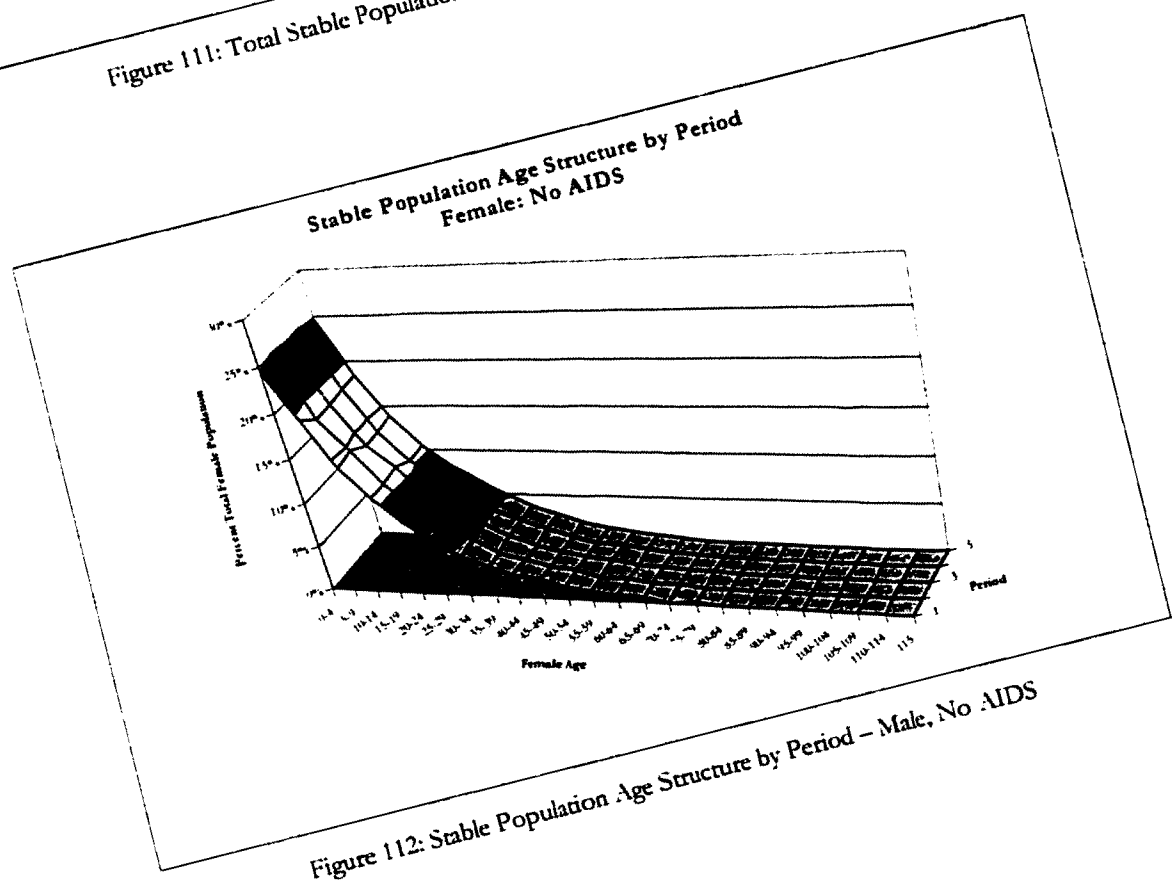

Figure 112: Stable Population Age Structure by Period – Male, No AIDS

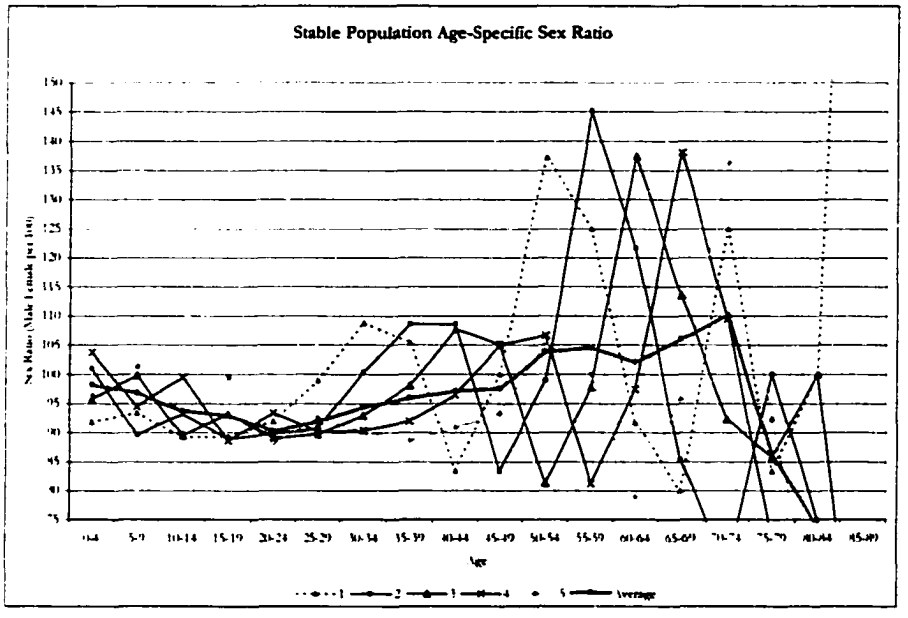

Figure 113: Stable Population Age-Specific Sex Ratio

Figure 113 displays the age-specific sex ratio for the stable population. The pronounced stochasticity at older ages results from small numbers: the fact that the cohorts aging through those ages were much smaller to begin with and by the time they reach those ages their numbers are substantially reduced anyway.

STABLE POPULATION MORTALITY RATES

Figure 114 and Figure 115 display the male and female age-specific mortality rates describing the stable simulated population during the five periods common to all of the simulations presented here: months 1,500 → 1,559; 1,560 → 1,619; 1,620 → 1,679; 1,680 → 1,739; and 1,740 → 1,799.

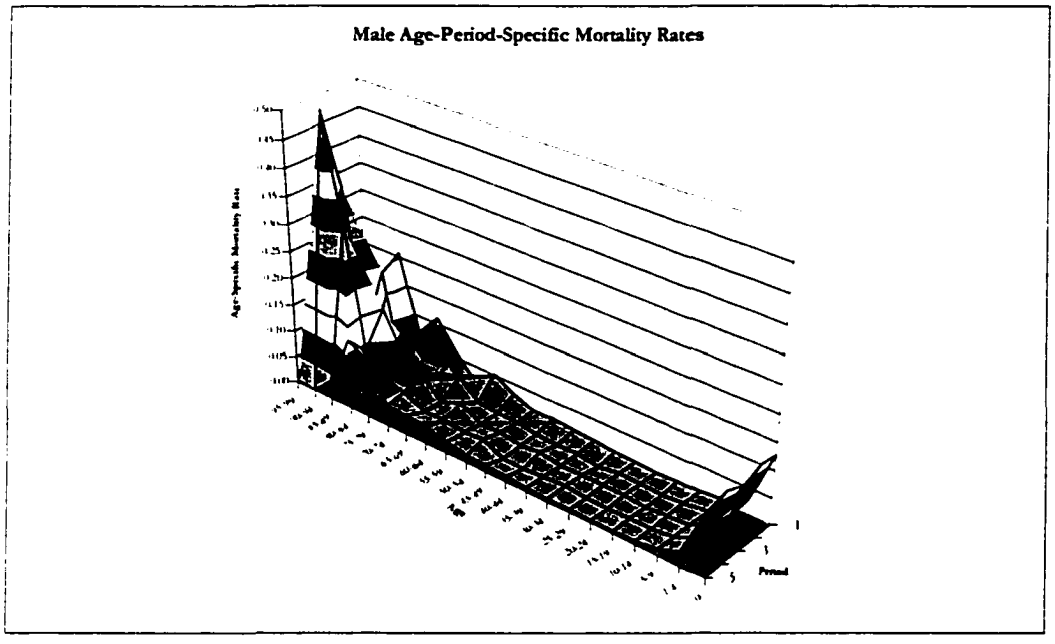

Figure 114: Male Age-Period-Specific Mortality Rates

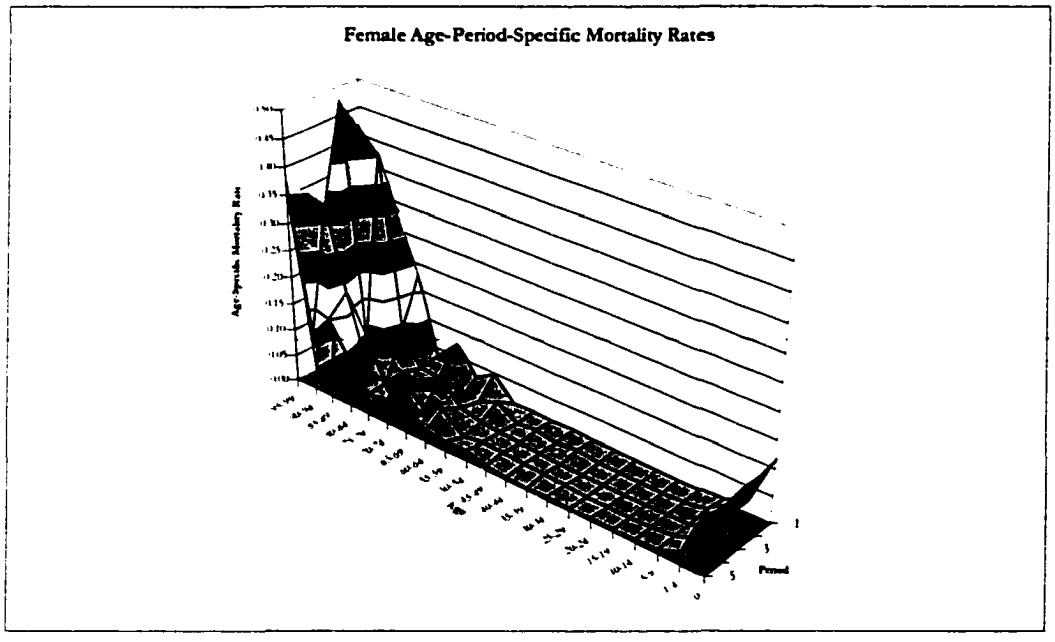

Figure 115: Female Age-Period-Specific Mortality Rates

In both cases the mortality rates appear reasonable and stable as time progresses. The marked stochasticity at older ages is due to the fact that there are increasingly fewer people alive at those ages at risk of dying, so that the law of large numbers prevails less and less as age increases and thereby the precision with which mortality is measured at those ages is progressively diminished – just as it with a "real" population. Confirming the realistic nature of the mortality rates are the expectations of life displayed in Table 63 which closely match the values of life expectancy calculated from the mortality rates used as parameters for the simulator – see Table 46.

TABLE 63: STABLE POPULATION PERIOD EXPECTATION OF LIFE AT BIRTH

Period	Male	Female
1	42.9	51.1
2	49.7	54.5
3	46.3	55.8
4	46.4	52.8
5	45.5	51.9
Weighted Average	46.2	53.2

STABLE POPULATION FERTILITY RATES

The stable population female age-specific fertility rates presented in Figure 116 are stable as time progresses, generally stochastic, and appear to have the typical age profile. However, they are noticeably higher than the age-specific fertility observed in the Gwembe, and they are especially high at relatively older ages. Moreover, the total fertility rates (TFR) presented in

Table 64 confirm that there is too much fertility in the simulated population. This fact explains the higher-than-desired growth rates prevailing in the simulated population.

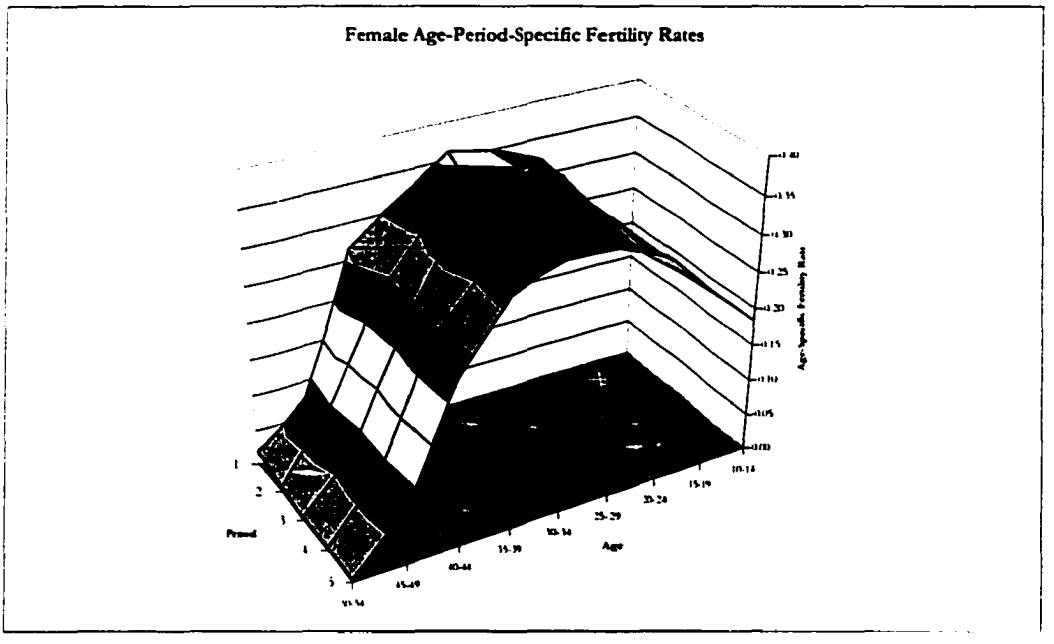

Figure 116: Stable Population Female Age-Specific Fertility Rates

TABLE 64: STABLE POPULATION FEMALE AGE-PERIOD-SPECIFIC FERTILITY RATES

	Period				
Age	1	2	3	4	5
10-14	0.176564	0.173632	0.185683	0.183855	0.184159
15-19	0.239297	0.246098	0.234338	0.236363	0.243158
20-24	0.319978	0.304974	0.300310	0.288433	0.307848
25-29	0.348385	0.349059	0.345336	0.334039	0.335044
30-34	0.361671	0.326582	0.334248	0.340686	0.337659
35-39	0.315687	0.307797	0.331415	0.306843	0.311706
40-44	0.265369	0.222446	0.217388	0.235247	0.226549
45-49	0.076967	0.093023	0.073620	0.078650	0.069888
50-54	0.019048	0.007187	0.003863	0.003034	0.003506
TFR (10-54)	10.6	10.2	10.1	10.0	10.1
TFR (15-49)	9.6	9.2	9.2	9.1	9.2

The fertility module does not currently model the decline in fecundability experienced by women as they age through their reproductive years, and it is likely due to this limitation that fertility is too high at older ages. Consequently, fecundity age-dependence will be added to the next version of the simulator.

HIV SIMULATIONS

Due to the long period of time necessary to complete an individual simulation on a PC, the HIV simulations presented here are the very first HIV simulations to be produced by this simulator. As such they are *initial* results that are on the forefront of this simulating effort. I want to remind the reader that these results do not necessarily reflect "realistic" situations as yet, and that they also contain some odd behavior that results from the specific choice of initial conditions and from the author's general lack of detailed experience with the working simulator. Again, I remind the reader to view these results as conclusive *proof of concept* rather than useful substantive exercises.

All of the HIV simulations start from using a stable population in month 1,500. Twenty percent of the initial population between ages fifteen and 49 is randomly infected, and thereafter each individual between ages fifteen and 49 is subject to a monthly hazard of infection from outside the simulated population of 0.0003 per month, discussed above under Random Infection. The untreated and treated AIDS simulations utilize a viral load indicator profiles displayed in Figure 117, as discussed above under Transmission.

The untreated viral load indicator for adults runs for 150 months corresponding to an average time from infection to death of about ten years. For children this period is 24 months. The

viral load indicator for treated adults runs for 300 months allowing a treated adult to live about twice as long; and additionally, the treated adult viral load indicator is of negligible value for the duration of the infection permitting a treated adult to be very *uninfective* during the bulk of their illness. The viral load indicator for treated children runs for 50 months also allowing them to live substantially longer, but not nearly as long as an infected adult.

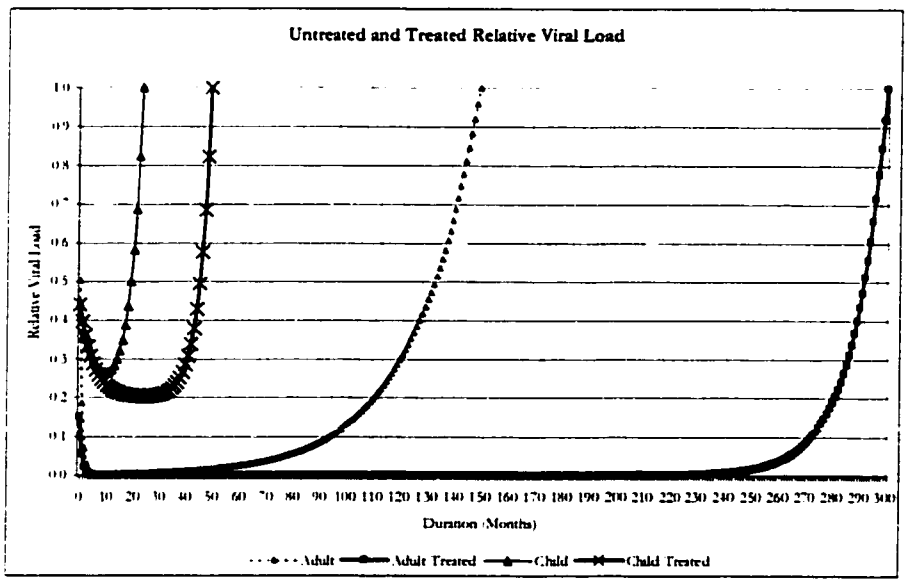

Figure 117: Untreated and Treated Relative Viral Load for AIDS Simulations

Several things were learned during the process of conducting these initial simulations. All of the first attempts to initiate and maintain an HIV epidemic just using random "external" infections failed, resulting in the very unrealistic starting condition of 20 percent infected adults. This was very surprising and must be investigated much more rigorously.

It is possible that an AIDS epidemic is very slow in the making and must reach some threshold before the pace of spread become significant. It is also possible that special sub-networks of

sexual contacts are necessary to allow a widespread HIV epidemic to develop, and that this simulator does not *yet* model those sub-networks. Those might include contact hubs such as commercial sex workers or a small number of very, very sexually active people who are able to spread the virus among otherwise well-separated cells within the population. Rather than viewing this as a defect in the current model, I think it clearly reveals that the creation of a rapidly growing, widely spread HIV epidemic relies on something other than normal, fertility-correlated levels of sexual contact between spouses and lovers, even with relatively high turnover of both. Rather, another element is necessary in the sexual networking, or the empirically observed transmission probabilities are far, far too low.

Adopting the initial conditions described does spark and maintain an epidemic, but it also produces undesirable and unrealistic transient effects including a massive short-term die-off of the initially infected adult population who are all simultaneously at the beginning of their infection period. Recently infected individuals should not die in such high numbers, indicating that the shape of the viral load indicator parameter needs to be altered to reduce the probability of death due to HIV infection immediately after infection ... or the HIV module needs to be modified in such a way as to more realistically model the progression of an individual from infection to AIDS. Both are likely necessary, and I have already begun work on an improved module to govern the natural history of HIV infection within an individual, and specifically to allow for two different indicators of an infection's status; one relating to the infected individual's health (a CD4 count indicator), and one relating to an infected individual's viral load and infectivity (a viral load count). These two indicators are related to each other in a relatively complex way as an infection progresses, both are substantially affected by

treatment, and both also interact with other opportunistic infections to create the full repertoire of infections that constitute AIDS.

POPULATION GROWTH

Figure 118 displays the annual proportional growth rates for the simulated populations infected with HIV compared to a simulated population not infected with HIV. Growth for the untreated population is initially much lower reflecting the initial die-off discussed above. However, it recovers quickly and appears to be still rising at the end of the 25 years of simulation. On the other hand, growth of the infected/treated population is similar to the uninfected population until the last period when it starts to decline. This appears to be the result of a delayed impact of the initial infection; the majority of those infected live for nearly twenty years following infection.

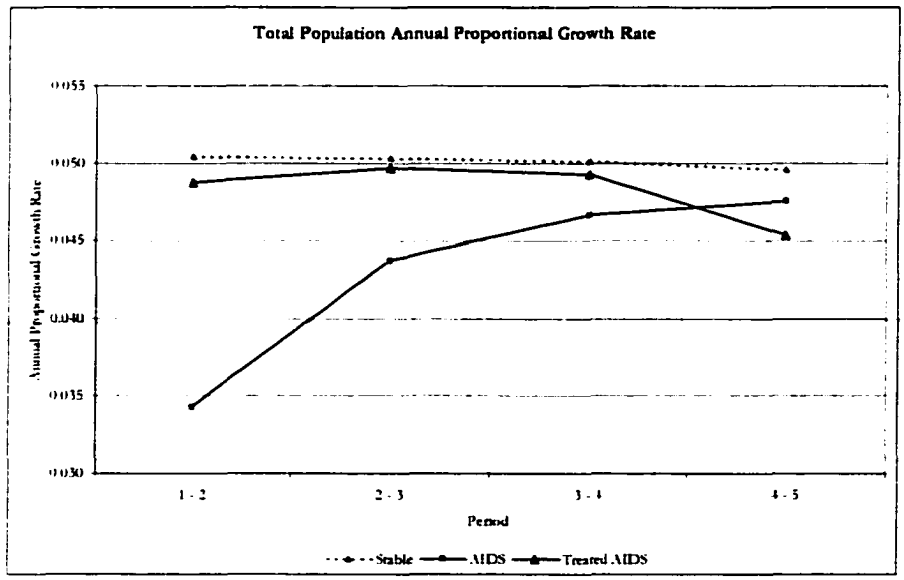

Figure 118: Total Population Annual Proportional Growth Rate for AIDS Simulations

MORTALITY

Figure 119 and Figure 120 display the period-averaged age-specific mortality rates for the populations infected with HIV. Very reasonable increases in mortality during the adult years between twenty and 50 are observed for the untreated population, while increases in the treated population are advanced in age and restricted to ages 35 to 50. There are no substantial differences observed between the male and female patterns for these relatively short durations of simulation; however, noticeable differences are observed in prevalence and incidence patterns (below) which probably require several more decades of simulation to have a full impact.

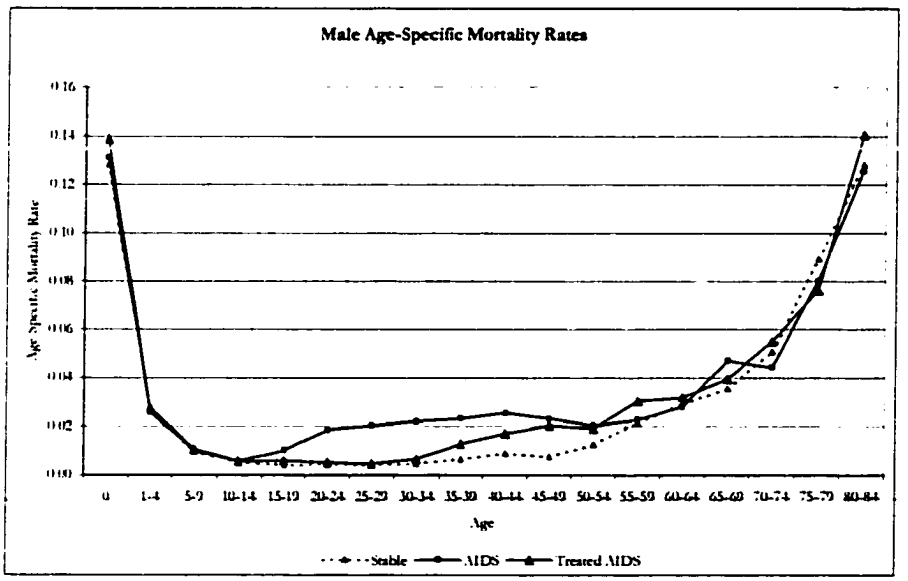

Figure 119: Male Age-Specific Mortality Rates for AIDS Simulations

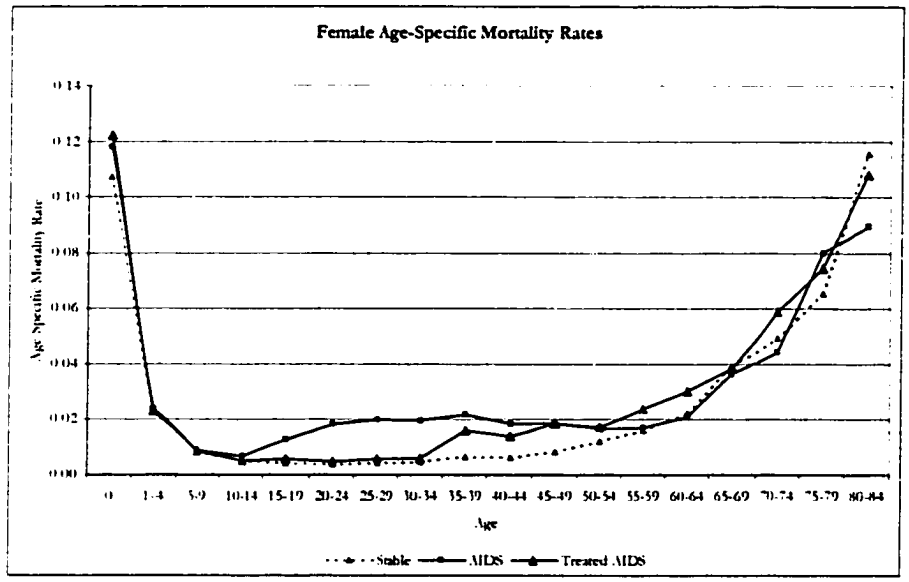

Figure 120: Female Age-Specific Mortality Rates for AIDS Simulations

FERTILITY

The impact of infection and treatment on fertility is evident in Figure 121. Infection with HIV reduces fertility at all ages substantially whereas treatment almost entirely negates this effect. This results from the fact that the viral load is suppressed for the bulk of the infected period for women who are infected and treated.

The level of fertility reduction appears to be about right given the empirically observed reduction of about twenty percent for HIV infected women. However, this is an aspect of the simulator which requires further tuning to bring the results into strict alignment with empirical observations.

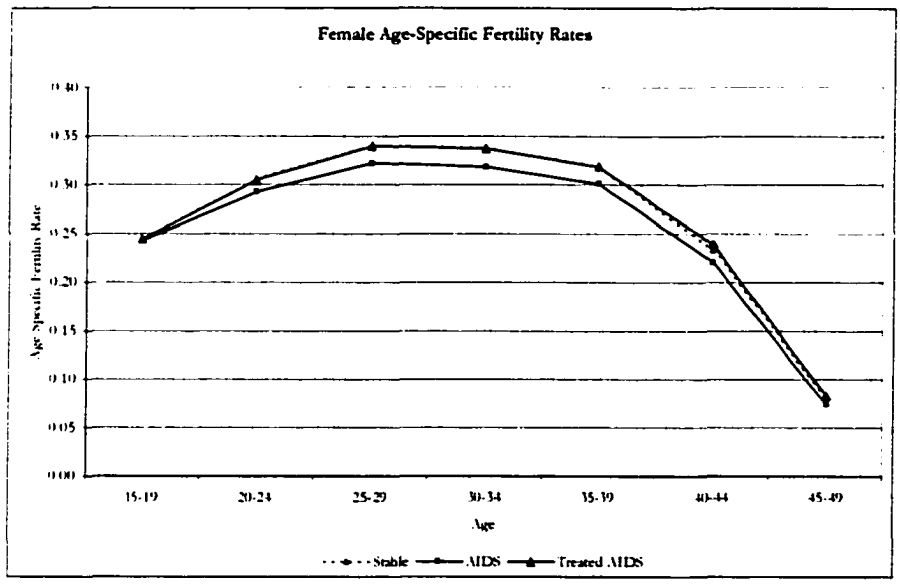

Figure 121: Female Period Total Fertility Rate for AIDS Simulations

AIDS

The following sections outline the levels of HIV-specific indices for the infected and treated populations. These are a first pass analysis of these simulations and do not constitute the sum total of all indices that can be calculated or the relationships between indices such as incidence and prevalence that it is possible to calculate.

INCIDENCE

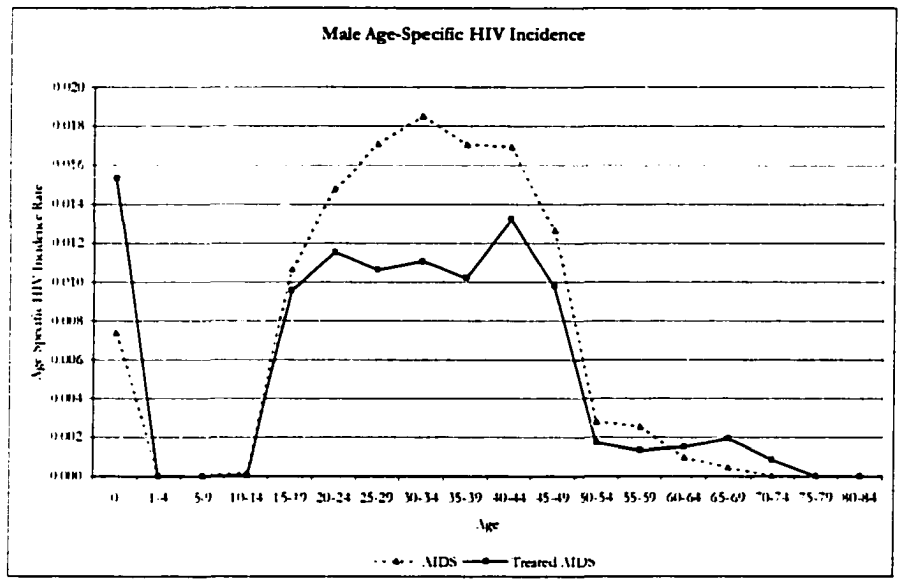

Figure 122: Male Age-Specific HIV Incidence Rate

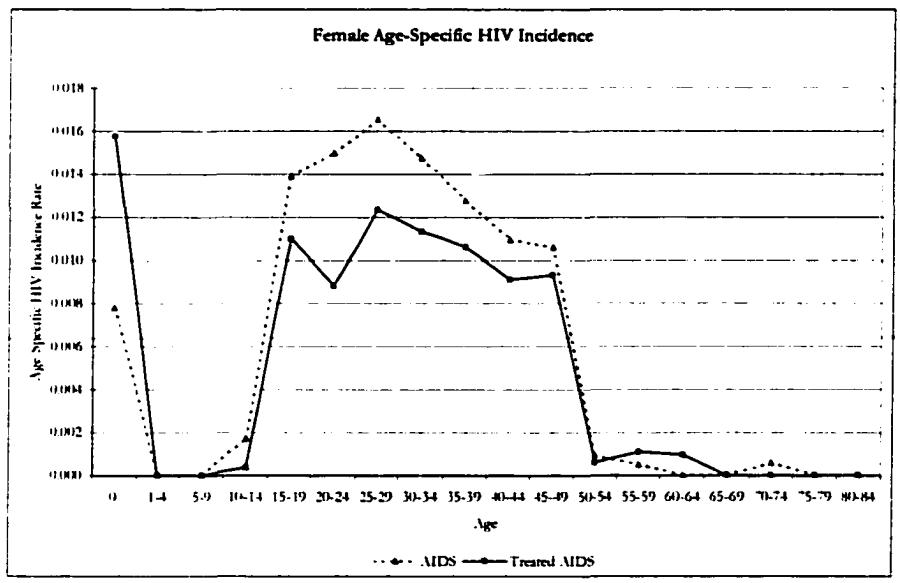

Figure 123: Female Age-Specific HIV Incidence Rate

It is clear that the HIV incidence is occurring at the correct ages; namely zero and the adult ages between fifteen and 50 for the most part. It is also clear that treatment significantly reduces incidence but does *not* bring it even close to zero. There is also a interesting but subtle difference in the age pattern of incidence for men and women with the male age profile being slightly more weighted toward older ages than the female profile. This makes sense given the fact that there is a substantial age difference between spouses and lovers. The low level of incidence observed at relatively older ages must result from the age-mixing inherent in the formation of affairs.

The most interesting finding is the fact that treatment actually *increases* the level of vertical transmission. It will have to be verified that this is not a "bug" in the simulator, and once that is confirmed, I can make statements regarding this finding.

PREVALENCE

Figure 124 and Figure 125 display the prevalence of HIV in the untreated and treated simulated populations. The prevalence is clearly greatest during the adult years reaching very realistic levels; levels that are actually low compared to the levels of twenty, 30 and even 40 percent observed in some populations in Southern Africa today. The sex differentials are again subtle but there; the male pattern is weighted slightly more toward older ages than the female pattern. The effect of treatment appears to have a slightly stronger sex bias with treatment substantially reducing prevalence, but more so for older females.

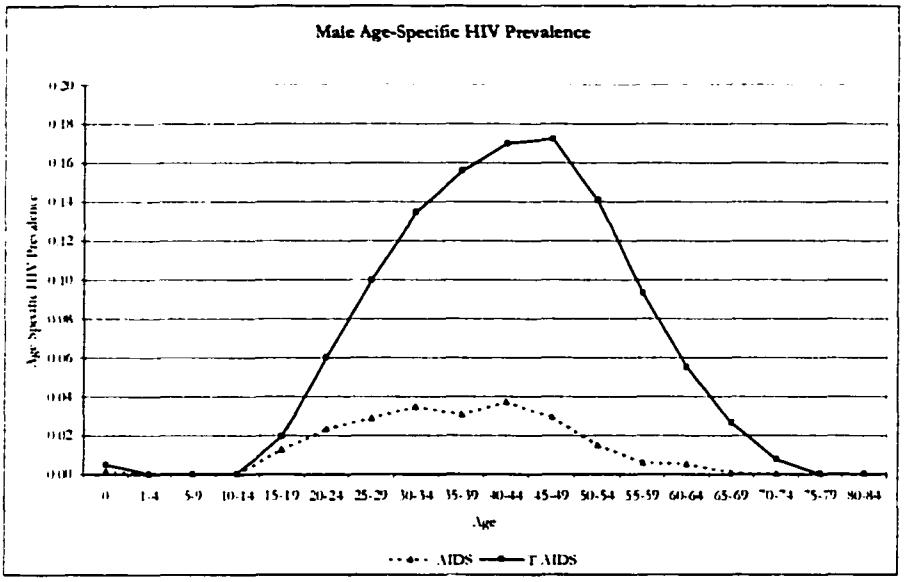

Figure 124: Male Age-Specific HIV Prevalence Rate

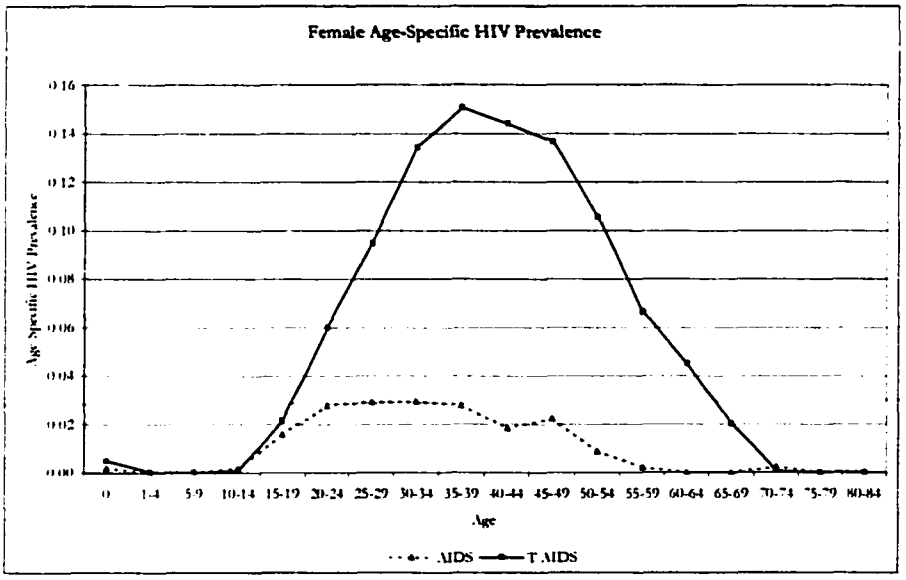

Figure 125: Female Age-Specific HIV Prevalence Rate

NEW CASES

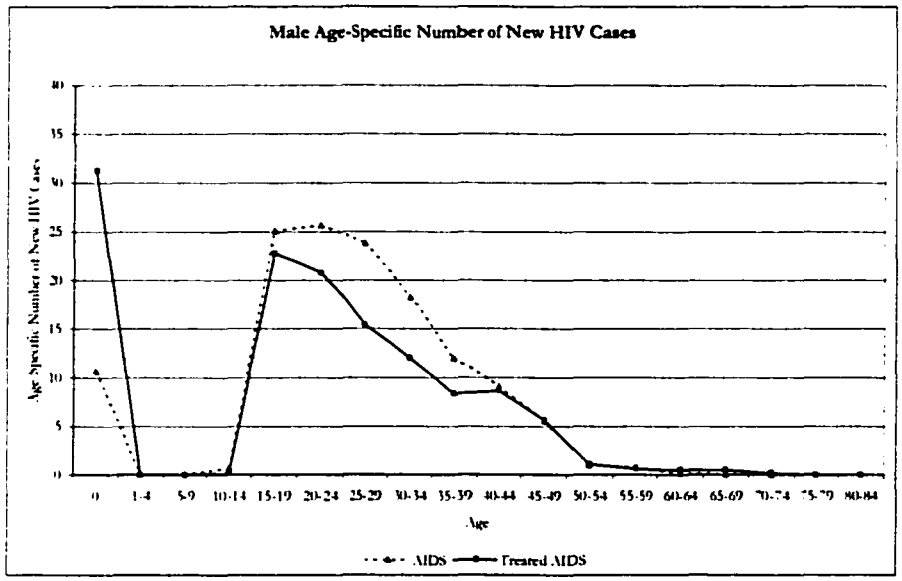

Figure 126: Male Age-Specific Number of New HIV Cases

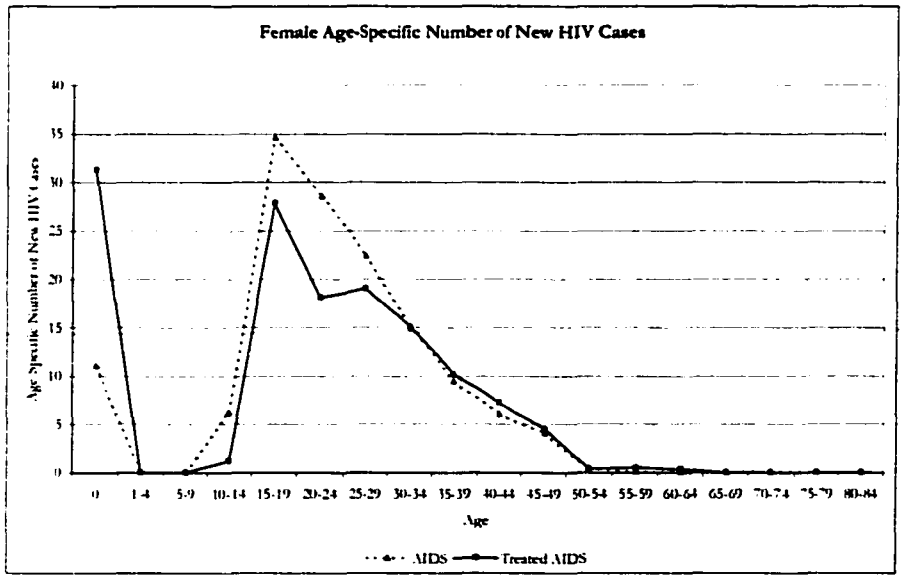

Figure 127: Female Age-Specific Number of New HIV Cases

The average raw number of new HIV cases by age for each five-year period of simulation is displayed in Figure 126 and Figure 127. Those figures most clearly demonstrate the sex-bias in the pattern of infection with the male distribution being shorter and broader extending more into older ages than the female distribution. In comparison, the female distribution peaks very early on and with substantially higher numbers.

In both cases treatment has a substantial impact in reducing the number of new cases during the adult ages, but anomalously (as mentioned above) increases the number of new cases due to vertical transmission. Again, it must first be determined whether or not there is a reversed sign somewhere in the model giving rise to this finding before it can be interpreted.

SUMMARY OF SIMULATION OUTPUT

The initial output from the simulator confirm that all of the demographic modules work well and require very little adjustment or augmentation to correctly and reliably model the complete demography of both monogamous and polygynous populations.

The results from the AIDS simulations are also positive but point to the fact that more needs to be done to understand and model individual sexual behavior and the natural history of the HIV within a single individual. Both of those avenues of investigation and development are ongoing as of this writing.

Perhaps most importantly, the successful running of the simulator confirmed that the underlying technology and methods used are appropriate and produce acceptable levels of performance, even on a comparatively underpowered personal computer. Even more

encouraging, being able to successfully run the simulator and produce results confirms that the structure of the simulator does offer the degree of flexibility and rich array of outputs that were envisioned from the outset.

In the final analysis, this endeavor is viewed as a completely successful proof of concept that fully justifies the additional time and funds that are required to move it from this stage to being a useful, generalized tool for population and health research.

Part 4

THE STRUCTURED POPULATION EVENT HISTORY REGISTER – S.P.E.H.R.

SPEHR CONCEPT

The Structured Population Event History Register (SPEHR) is a general relational data model designed to capture, store, verify, manipulate, extract and partially analyze longitudinal data describing a dynamic collection of entities that are related to each other in complex ways through their joint experience of salient events. Although the design is general, it was conceived specifically to manage the kind of data collected and manipulated by a long-lived, dynamic Demographic Surveillance System (DSS).

SPEHR is unique in two regards. First, it stores a complete description of the primary data stored in the database and the logic used by the database to manipulate those primary data; those *data that describe other data* are referred to as "meta data". Second, it uses the meta data to provide the meaning (or definition) associated with the primary data and the various connections between the primary data, thereby making the meta data an integral and indispensable part of the database. The meta data are themselves stored in the database as records in special meta data tables.

There are a number of useful consequences, the most important of which is that new *meaning* can be added to the database by simply adding new rows to the meta tables. In this fashion

new entities and their behavior can be described in the database without creating, deleting or amending the structure of any of the tables in the database. This allows the structure of the database to remain constant through time as new entities are added, perpetually maintains the descriptions (and meaning) of existing entities, and effectively isolates the users from both the physical and logical implementation of the database system. All that a user has to understand in order to maintain a SPEHR-based database is the logic through which meta data are interpreted by the database.

Another useful consequence is that because the meta data are necessary for the operation of the database, the meta data are always available to describe the data in the database, and it is *impossible to separate* and consequently *lose* the description of the primary data. This property of the model is critical to analysts who are usually not familiar with the day-to-day operation of the database but must have precise definitions of the data in order to conduct meaningful analysis.

Finally, the meta data provide a means through which the database can *understand itself.* This allows the development of general, automated processes to insure the integrity of the primary data stored in the database. Having access to the meta data allows the maintenance of *relationship integrity* in addition to standard relational and temporal integrity. By "relationship integrity", I refer to the validity of the stored representation of external relationships between modeled entities.

I have designed and successfully demonstrated the core SPEHR data model using a simple fictional model population. All of the primary design goals have been successfully realized

and tested in that demonstration. In addition to the core model I envision three more core components. The first to model primary data constraints, the second to model questionnaire (data capture) instruments, and the third to model database audits. Although these data models are still in the conceptual phase of development they are envisioned to incorporate the same *meta data-centric* design introduced in SPEHR, with the synergistic consequence that they can share meta data with the core SPEHR model and thereby increase the flexibility and utility of all components.

Continued development and testing of the SPEHR concept has the potential to be of great use. SPEHR is expressly designed to be general and thereby flexible enough to manage a wide array of different, even unique, longitudinal data sets. A successful test and subsequent full implementation of SPEHR (in the initial planning stages at the Agincourt DSS site in South Africa at the time of this writing) will produce a robust functioning software system that will have all of the capability necessary to manage data from any existing or new DSS site. In order to convert to the SPEHR-based system, a DSS site would have to go through a challenging conversion process, but after that was complete they would be able to avail themselves of all of the benefits of SPEHR. Moreover, a group of SPEHR-based systems would be able to easily share and compare *data, descriptions of data, and data collection instruments* amongst themselves in a technologically transparent fashion. This ability to share data without regard to the technicalities of the data system in which it is stored would greatly lower the barriers to collaboration between sites, including their ability to design comparable, compatible new multi-site studies.

DESIGN OBJECTIVES

SPEHR is a conceptual (logical) model of data that can manage continuously collected data describing the history of a human population. Data of this nature:

- are time-related or longitudinal,
- encode arbitrarily complex connections between individual members of the population and aggregates of individual members,
- must meet evolving substantive requirements, and
- must in parallel fulfill substantial operational requirements of the programs that collect them.

The overall objective of SPEHR is to provide a flexible, general standard conceptual and logical framework for capturing, validating, organizing, storing, manipulating and retrieving longitudinal population data. The ultimate desired result is to facilitate the sharing and comparing of data describing different populations through time.

STANDARDS AND COMPARABILITY

Successful **comparison** of data requires: 1) that the data have comparable *meaning*, or whose definitions are well enough understood to make comparison possible, and 2) that the data are stored and organized in ways that allow comparisons to be made. The first requirement is a substantive one and must be primarily addressed as such; however, the data storage logic can be helpful by tightly coupling primary data to its definition thereby making the formulation of a comparison feasible and as precise as possible. Not being able to meet the second requirement is surprisingly one of the most serious barriers prohibiting primary analysis of longitudinal population data, in addition to making comparisons very difficult. Because

longitudinal data are usually stored in complex relational databases, extracting them in a format that is amenable to analysis requires expert database skills not possessed by most analysts. Consequently, many of the data remain locked in their individual tables, never linked up, extracted, or analyzed – much less compared to other such data describing another population.

Improvements in the logic of data storage can be of great help regarding the second requirement. The logical design of the database can be standardized so that access to and manipulation of the data is performed using a unified logic that can to a large extent be automated.

SPEHR is the first step in the design of a database system of this sort. The SPEHR data model provides a standard conceptual and logical framework for storing and manipulating arbitrary longitudinal population data.

FLEXIBILITY AND EXTENSIBILITY

Flexibility refers to the ability of the data model to manage data that reflect a wide range of realities. In broad terms, a longitudinal data model needs to be flexible along two dimensions: time and structure.

On the time dimension, the data model needs to be able to gracefully manage data collected by various study designs including cross sections, panels, cohorts and continuous surveillance.

On the structure dimension, the data model must be able to represent and manipulate an almost unlimited variety of potential connections between modeled entities, and to be able to model any degree of connectedness from very little to extensive. For example, the data model

must be able to handle an unpredictable degree of relatedness brought about through male and female relationships that may take the form of elopements, non-marital conjugal unions, marital conjugal unions, short term affairs, long term affairs, intermittent affairs, repeated marital unions etc. Some studies may not care about all those or may consider them to be the same for their purposes. The data model must have the ability to model them all or to consider them the same or to collapse them in some manner. Where DSS systems are concerned, the model must at least be able to model all of the individual-based relationships between people in addition to be able to relate aggregates of people to each other and individuals; for example, a household to its members and the village to which it belongs – in addition to the *place* where it exists.

Extensibility refers to the model's ability to gracefully evolve and grow with the changing substantive requirements of the enterprise it is serving. Most importantly, the data model must be able to easily expand to model additional entities, relationships and constraints as new studies join the enterprise. Ideally, this expansion should be easy, transparent, not require logical modifications to the existing data model, *and* not disturb the ability of the model to continue modeling the existing entities, relationships and constraints. This last point is important in that it insures that the database is always able to retrieve valid information describing the past, even if some aspects of reality were recorded then, but are not recorded now – in other words, the system never forgets.

SPEHR incorporates a great deal of flexibility and extensibility by providing a data model logic instead of a data model *per se*. The logic is used to interpret meta data that describe the

modeled reality, and those meta data and thus the modeled reality can be augmented at any time.

SELF-DOCUMENTATION

Data documentation is very important but *very* boring, and hence is usually not done very well, or sometimes not at all. This results in the existence of data with no meaning. I have personally experienced this with census information recorded in ASCII data sets on magnetic tape. All of the values for each record are there stored in a long string of ASCII characters, but there is no dictionary to inform us of the length of each field or the definitions of the codes stored within each field; the result is completely meaningless garbage!

Self-documenting refers to the ability of a data management system to store all of the definitions of the primary data *with* the primary data in the database. A stronger version of this is to require the definitions in order for the database to operate.

SPEHR is self-documenting in the strong sense that it contains a full set of meta data describing the primary data in the database and the relationships between them, *and* those data are an integral and indispensable component of the database.

The meta data at least:

- provide a full description of the primary data that are stored in the database,
- provide a full description of the logic that is used to organize and manage the stored data,
- allow the system to automatically:
    - understand what is stored and how it is organized,
    - enforce constraint logic,
    - automate the manipulation and retrieval of primary data,

    - add new data structures,
- provide system users with full up-to-date data definitions at all times.

LOW MAINTENANCE

Any practical system must be as low maintenance as possible. Low maintenance comes in several forms for databases. The first is to minimize the amount of day-to-day tinkering that goes on to insure that a database runs properly and is useful, and the second is a reduction in the skill level required to access and successfully retrieve information from the database.

SPEHR aims to keep both as low as possible. Because the SPEHR data model does not require logical redesign of the database each time additions to the system are required, because the data organization logic is standardized, and because the meta data are always available, non-technical users should be fully empowered to access and manipulate data in a SPEHR-based database once they have learned the logic of SPEHR – with no other special skills required.

All of the day-to-day maintenance tasks can be automated, including:

- adding, editing, or deleting data,
- adding, or editing data structures,
- verifying data integrity,
- extracting operational reports, and
- extracting analysis data extracts.

Within the SPEHR data model, there is no need to ever add or delete tables from the database or add or delete relationships from the database. At most, new queries may need to be defined from time to time on SPEHR's base tables.

SECURITY

The primary security concern in most DSS databases is the requirement to guarantee the privacy of the respondents. This requires at least:

- controlling access to individual datum,
- controlling access to connections between individual datum,
- potentially preventing access to the data through subpoena, and
- providing a mechanism for the timely irreversible destruction of the data should they be in danger of being used to facilitate massive crimes against the respondents.

The last three points are not usually considered necessary, but I argue that they are potentially important, especially in the context of a longitudinal project that continuously collects high resolution, high quality data from a defined population.

Just the connections between modeled entities can be used to identify individual items based on the unique network of connections that exist for most modeled entities. For example, without knowing a person's identity, it is likely that they can be identified by drawing connections between them and their family members, the places they have lived and a few other pieces of otherwise "anonymous" information. So, it is important to control access to anonymous relationships that are stored in the database as well as protecting the entities themselves.

Long term projects that develop a high degree of trust with their study community may begin to record information that is of use to the criminal justice system as they investigate and prosecute criminal activity. It may also be that part of the study is to specifically study behavior that may be of a criminal nature, such as drug use. In these cases, protecting the privacy of the respondents means protecting their information from the authorities in a *legal* fashion. This can be easily accomplished by distributing the database over several legal jurisdictions so that the data stored in each jurisdiction is incomplete and not able to produce a valid view of any one individual. In so far as jurisdictions (like different countries) do not cooperate with the investigation and prosecution of crimes, the respondents are protected, and so are those responsible for the study – all information in a given jurisdiction can be provided to the authorities while legally withholding the information physically stored elsewhere. This security threat is in stark contrast to the much more troublesome situation in which the authorities themselves become the criminals.

The Nazis used the Dutch population register during World War II to quickly identify and apprehend the Jewish population in Holland, and the more recent conflict in Rwanda also involved explicit use of the census to more quickly identify victims. Population based information can be catastrophically misused when there is the will to do so, and in cases such as those mentioned above, the only way to prevent massive disaster is to quickly and permanently render the data useless. I assert that any responsible population-based project should have a well-developed plan to prevent their data from being used to coordinate genocide or the targeting of specific subpopulations. The easiest way to do this is to store one or more key tables on a single physical medium that can be irrevocably damaged or destroyed

on short notice. Regular offshore (overseas) backup insure that the data are not irrevocably lost.

To address these security concerns requires:

- thoughtful logical partitioning of the data into easily secured units,
- thoughtful physical partitioning of the data into legally inaccessible units,
- centralized control over access to the data, a
- at least one fusible relationship or table in the database.

SPEHR can accommodate all of these security concerns because it is based on a highly normalized representation of data that requires information to be broken down into its most basic unitary components, each of which is stored in a different table. These tables can be distributed across physical devices and those devices across physical space in ways that address each of the concerns described above. Access control to individual tables is already well-supported in all enterprise-level commercial DBMSs.

VALIDITY

Validity refers to the property of data having to do with the degree to which the data correctly and precisely reflect the reality that they describe. Consequently, validity has to do with both the meaning of the data and the fidelity with which they are recorded and organized. Insuring that the meaning of the data faithfully reflect reality is primarily the responsibility of the data collection system, whereas maintaining the fidelity of stored data is the responsibility of the database.

One of the most important principles in maintaining the fidelity of stored data is to minimize, or optimally disallow completely, the duplication of stored facts in the database. A fully *normalized* database does not duplicate anything whereas partially or not normalized databases store identical information with identical meaning in two or more locations in the database. Obviously, duplication provides the potential for the duplicate information to become inconsistent through inconsistent storage, editing or deletion. The fidelity with which information is stored is inversely related to the degree of duplication in a database. SPEHR has been *very carefully* designed to be **fully** normalized with **no** duplication of stored facts at all.

Another critical element to insuring the integrity of data is the specification and enforcement of *constraints*. Constraints act as filters preventing invalid data from being stored, preventing invalid links from being stored, and preventing invalid temporal facts from being stored. Constraints can be specified in many ways relating to all three of those dimensions. Many commercial DBMSs have built-in sophisticated machinery for providing referential integrity, but very few have sufficient support for temporal integrity, and relationship integrity cannot be supported in a general form by the DBMS. SPEHR provides good referential integrity, and has been designed with the potential to provide good temporal and relationship integrity. The provision of these last two is the subject of one of the additional components of SPEHR that is still in the conceptual design stage.

ANALYSIS-FRIENDLY

One of the primary limitations of most existing longitudinal population data management systems is the fact that they are very analysis-*unfriendly*. By this I mean that it is prohibitively

difficult for the average analyst to extract data from the database organized and formatted in a way that is amenable to analysis. Because relational databases are typically complicated collections of many related tables, it is necessary to completely understand the exact relationship between the tables and to have a firm command of SQL or some other relational query language to actually join the tables and extract useful information from them. Well-developed skills of this type are not widely available and usually reside only in the database manager, who is already fully occupied making the database run. Consequently, analysts are not able to have sufficient access to the data stored in the database.

To rectify this, the database needs to be organized around a consistent logic through which any of the data can be accessed and extracted using standard (automated) tools. Given this, the analyst has to have a firm grasp on the data model and the facility to use the standard extraction tools, *and that is all.* SPEHR provides the consistent logic necessary to implement this and begins to define the standard data manipulation and extraction tools. The further refinement of these extraction tools is an active area in the further development of SPEHR. As of now, the user can view the data using common "real world" views such as a "person" or a "marriage" with all of their attendant attributes. A generalized, automated system for defining "real world" views is under development.

Additionally, an accessible database should be able to produce user defined analysis extracts in standard formats such as: 1) fully rectangularized flat files based on an arbitrary unit of analysis, and 2) item-(time) granule files defined over arbitrary units of analysis and arbitrary granules of time (such as person-months or household-years). In its current version, SPEHR is able to

produce both formats easily, but not yet in a user-defined fashion. However, that potential is built into the SPEHR data model and simply requires some additional work to fully develop.

GENEALOGIES

The ability to model genealogies is important to all types of scientists who might utilize data pertaining to a whole population. Social scientists, and anthropologists in particular, regularly use genealogies to map families and categorize households and individuals. Epidemiologists may need them for the study of genetic disorders, and biologists and health care workers may require them to develop treatments for genetic disorders or to manage family-based health care programs.

One of the very few population-specific design goals is the requirement that SPEHR be able to completely model genealogies of any sort; generated by either monogamous or polygynous marriage systems. In fact, because genealogies are difficult to encode and manipulate in a general way, SPEHR's success in modeling genealogies would be very reassuring. Consequently, I am happy to observe that a SPEHR-based database can fully model and manipulate an arbitrary genealogy, for a demonstration of this please see www.samclark.net/SPEHR.

SPEHR DATA MODEL

This section describes the primarily components of the SPEHR relational data model. A relational data model specifies the structure of all of the tables and relationships between the tables that constitute a relational database. In so doing, it may also specify most or all of the referential integrity that is required of the data stored in those tables. Please refer to Appendix A and Appendix B for basic information on relational databases.

The SPEHR data model is highly abstracted, general, flexible and extensible. It is designed to capture, validate, store, manipulate, and retrieve longitudinal data describing any collection of items that:

- **experience meaningful events at definable points in time, and**
- **whose members are related to each other through the occurrence of meaningful events that affect more than one member.**

Integral to its functioning, SPEHR stores and maintains a complete set of meta data describing both the primary data stored in the database and the relationships between modeled entities.

The core SPEHR data model consists of modules and/or individual tables that model:

- events,
- item-episodes,
- experiences,
- attributes,
- shared experiences,
- multi-event processes, and
- meta data.

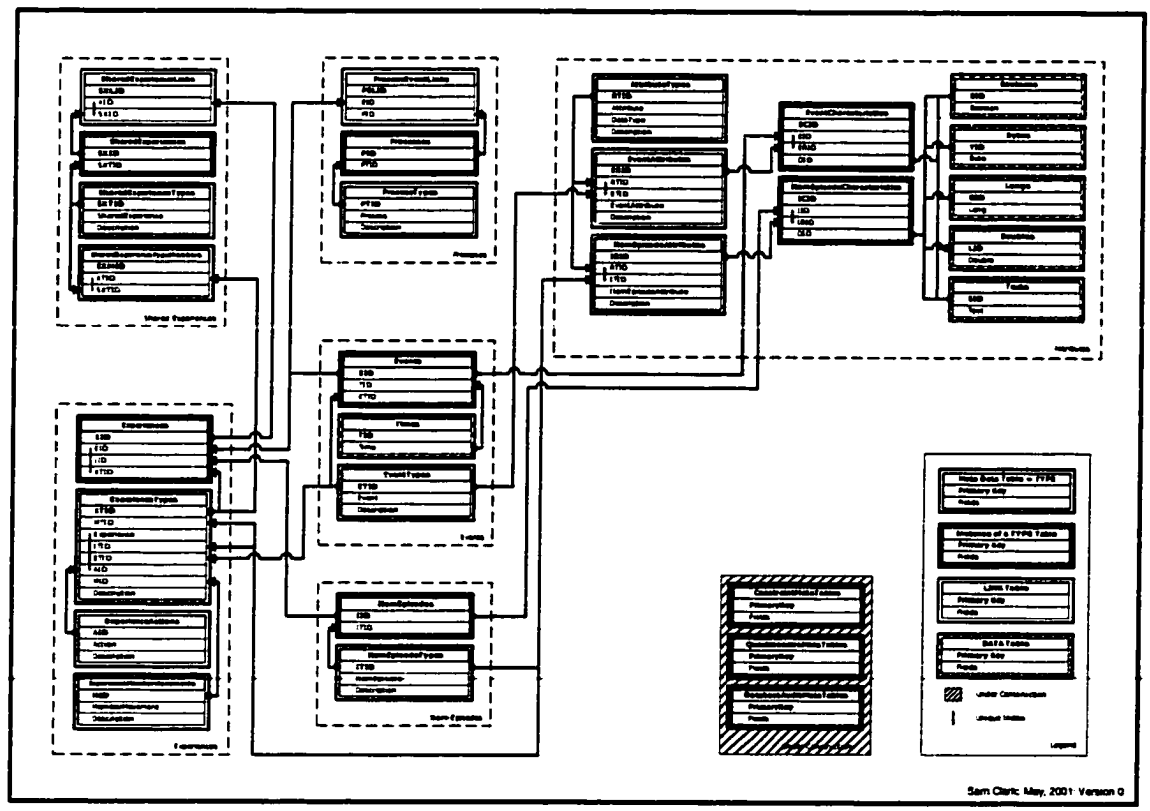

Figure 128: Structured Population Event History Register DATA MODEL

Figure 128 displays the SPEHR data model as a pseudo entity relationship diagram. It is included here for completeness although it is not necessary for the reader to fully digest it

because the concept behind each major component is described below in English. However, a few guidelines are appropriate.

The diagram is divided into modules that correspond to the major components of SPEHR: item-episodes, experiences, events, shared experiences, processes, and attributes. Within each block, the tables that model the associated functionality are diagramed with their attendant connections. Each block contains at least one meta table containing the list of "types" that can be instanced to create the individual entities that are at the core of the block. The connections between tables relate the unique IDs in each table to corresponding fields in other tables, and hence represent the connections and the cardinality of the connections between individual tables. Where a connector terminates in a sphere, the connection is of cardinality one, whereas where a connection terminates in an arrowhead, the connection is of cardinality greater than or equal to one. For example, the connection between the events table and the experiences table is one-to-many and is encoded by a sphere-to-arrowhead connection between the event ID field in the events table to the corresponding event ID field in the experiences table. Unique indices are represented by bold text or vertical lines encompassing the fields over which the index operates. All table-level IDs are unique and indexed.

This document is not intended as a technical treatise, so beyond displaying the diagram I will not provide a detailed technical description of the data model; rather, I will attempt to explicate the concepts governing the model.

EVENTS

An **event** is a meaningful instant in time corresponding to a fact associated with that instant and not with the instants immediately before or after. Typical events associated with longitudinal population studies include: **births, deaths, marriages, divorces, migrations, and movements between households.**

ITEM-EPISODES

An **item-episode** is a meaningful interval of time corresponding to a fact that is true during the interval and false immediately before and after. The "item" in the name calls attention to the fact that the episode can, and often does, correspond to the lifespan of a physical thing – an *item*. Typical item-episodes of interest to longitudinal population studies include:

- a **person** – alive and "existing" from their data of birth to their date of death,
- a **household** – recognized and valid from its date of incorporation to its date of disbanding,
- a **residency** – a person *at* a place from their data of arrival to their data of departure,
- a **marital union** – a valid contract pairing a man and woman from the date of union to the date of separation, and
- a **place** – a geographical entity that is located and named from its date of mapping to it becomes irrelevant.

EXPERIENCES

An **experience** is the manner in which an item-episode is affected by an event; or alternatively, the impact that an event has on an item-episode. Experiences *associate* item-episodes with events, and two *or more* item-episodes associated with an event are *related* to one another through their individual associations with the common event. A single item-episode may, and

most likely will, experience many events, and a single event may, and almost certainly does, affect many item-episodes. As a result, complex *many-to-many* relationships are possible.

The notion of experiences is the key concept behind SPEHR. Three aspects of this method of encoding relationships between things are of particular importance.

1. **The entire notion of relationships defined by experiencing events together is a natural and intuitive one that fits nicely with our day-to-day lives. As such it is not an arcane logical construct that is difficult for people to understand and manipulate.**

2. **The definition of relationships through the joint experience of events automatically associates every relationship with its duration and the circumstances that initiated and terminated it; and in fact, it is not possible to omit any of these attributes in this coding scheme. Rather than being an artificially constructed and enforced requirement of the storage scheme, the time dimension is a natural and necessary part of this coding scheme.**

3. **The system puts no limit on the complexity or cardinality of relationships that can be stored and manipulated. By this I mean that any number of items can share an experience and hence initiate or terminate some form of relationship between themselves, and each of their individual experiences can be very different, associating them in different ways with each other. The beauty of this is that it is possible to encode all of this complexity without creating new constructs within the database.**

Table 65 displays a simple demonstration of the principle of an experience. In this example two events Justine's birth and Adam's death affect a number of item-episodes in different ways. The birth is experienced by the newborn (Justine) as the *beginning* of her life, by her mother as *giving birth*, by her father as *having a birth* of one of his children, and by Durban, the place where she is born, as the *addition* of another person. Similarly, Adam's death is experienced by Adam as the *end* of his life, by Rachel (his wife) as the *loss* of her spouse, by

Justine as the *loss* of her father, by Durban, where he died, as the *loss* of a person, and finally by his marriage with Rachel as its *end*; and, all of these experiences (and item-episodes) are tied together by their link to the common event.

TABLE 65: EXAMPLE OF "EXPERIENCES"

Item-Episode	Experience	Event
Rachel	Gives Birth	*Justin's Birth*
Adam	Has Birth	*Justin's Birth*
Justine (newborn)	Begins Life	*Justin's Birth*
Durban (place of birth)	Adds Person	*Justin's Birth*
Rachel	Loses Spouse	**Adam's Death**
Adam	Ends Life	**Adam's Death**
Justine	Loses Father	**Adam's Death**
Durban (place of death)	Loses Person	**Adam's Death**
Adam and Rachel's Marriage	Ends	**Adam's Death**

Each column in Table 65 corresponds to a table in the SPEHR data model. Even though the Item-Episode and Event columns in Table 65 displays multiple instances of the same things, each "thing" is stored only once in the database. Each item in the Experiences column is stored separately because the experiences are what maps, or links, the item-episodes and events together, and through their joint linkages to the same events, the item-episodes to each other. The database tables corresponding to this example look similar to Table 66 through Table 68:

TABLE 66: RECORDS IN ITEM-EPISODES TABLE

Item-Episode ID	Item-Episode
1	Rachel
2	Adam
3	Justine
4	Durban
5	Adam and Rachel's Marriage

TABLE 67: RECORDS IN EXPERIENCES TABLE

Experience ID	Associated Item-Episode ID	Associated Event ID	Experience (Type)
1	1	1	Gives Birth
2	2	1	Has Birth
3	3	1	Begins Life
4	4	1	Adds Person
5	1	2	Loses Spouse
6	2	2	Ends Life
7	3	2	Loses Father
8	4	2	Loses Person
9	5	2	Ends Marriage

TABLE 68: RECORDS EVENTS TABLE

Item-Episode ID	Item-Episode
1	Justine's Birth
2	Adam's Death

Additionally, the Experience field in Table 67 does not actually contain data in the SPEHR model, but rather the ID number of a record in a meta table that specifies the various types of experiences that are possible.

ATTRIBUTES

The attentive reader will have noticed that there is not yet a facility for storing the individual attributes of specific items or events. In most databases, attributes are stored together with the things they describe using additional fields in a given table. However, because SPEHR stores things of many types in the same table (lots of different types of item-episodes in the item-episodes table, and lots of different types of events in the Events table), it is not possible to store all of the different sets of attributes corresponding to those different types of things in the same table. Instead it is necessary to store the attributes elsewhere and link them to the

things that they describe. This is done with a separate set of tables, one to store each type of data and several to link the attributes to the things that they describe. This organization degrades performance a little because many tables must be accessed and joined to attach attributes to things, but it is a necessary and very flexible mechanism for supporting the core SPEHR item-experience-event data structure.

SHARED EXPERIENCES

Experiences are often reciprocal or inextricably related. For example, a migration involves a person leaving one place and arriving at another, and there is no way that the migration can occur without the "leaving" and the "arriving". Those two experiences are reciprocal and inextricably linked. It turns out for various performance-related reasons that it helps to link those kinds of experiences explicitly, and SPEHR offers a facility to do that by grouping experiences into "shared experiences". This is also occasionally necessary to differentiate experiences that can be interpreted in ambiguous ways, but a discussion of this is beyond the scope of this document.

MULTI-EVENT PROCESSES

Many happenings are composed of a number of events that occur over a comparatively long period of time, and SPEHR provides a mechanism for grouping related but distinct events into "multi-event" processes. Initiating a marriage is a good example. That process often involves some sort of engagement event, the migration of one or both of the potential spouses, the movement of the potential spouses from their previous households to new ones, a wedding, and usually the formation of a new household. All of these events are part of the *process* of the new couple becoming married and likely are stretched over the course of several months or

years. Not having the ability to encode the fact that all of these events are related might (probably often) lead to a loss of information and potential misinterpretation of the individual events. Coding the multi-event process allows the database to represent the larger (causal) context in which the individual events take place.

META DATA

Meta data are at the heart of SPEHR. All item-episodes, events, experiences, and attributes are ***instances*** of types predefined for each. The number and description of the types can be augmented at any time simply by adding new rows to the various meta data tables that contain them. In this way the "types" that can be stored an manipulated by the database and the ways that they are connected (through "types" or experiences and "types" of attributes ...) can be updated and expanded at any time.

Both the database and the users of the database have access to this meta data, and both must use it to interpret the data stored in the database.

SPEHR DEMONSTRATION

A fully functioning demonstration of the SPEHR data model can be found at www.samclark.net/SPEHR implemented in Microsoft Access 2000. That demonstration adequately proves that a SPEHR-based database can store and manipulate data describing realistic population dynamics. Again, it is well beyond the scope of this document to describe in detail the construction and operation of the demonstration, and the reader is directed to the URL.

The SPEHR demonstration database contains SQL queries that calculate generations, genealogies, counts of events, counts of person-years, mortality rates, fertility rates, person-year files for event history analysis, item-history files, and much more using the SPEHR data model and the test population data that it contains. The SPEHR data model is easily able to manage a very wide range of complex population linkages over time and to efficiently retrieve the data in useful, user-recognizable forms. In short, SPEHR is the basis for a useful longitudinal data management system.

Appendix A

THE RELATIONAL MODEL OF DATA

The *relational model of data* is a theory describing a consistent framework in which to structure, manipulate and maintain the consistency of data. Data are stored and manipulated in *relations* (or tables), a set of *operators* manipulates relations to produce other relations, and relations are subject to certain *constraints* that insure the integrity of the data. For a lucid discussion of the relational model and relational database systems, refer to An Introduction to Database Systems by C.J. Date (Date 2000).

RELATIONS

A relation is composed of a heading and a body. The heading is a *predicate* composed of a number of attributes, each with its own domain of possible values. The body is a collection of *tuples* (or rows) that represent true propositions formed by choosing a value from the domain of each of the attributes specified in the predicate. Each tuple, therefore, represents a set of attribute values that when substituted into the predicate form a true proposition. This is a sufficient definition of a relation. As a consequence:

- There is no order specified (or necessary) for the attributes;
- There is no order specified (or necessary) for the tuples;
- Each tuple contains a unique value for each attribute; and,
- There are no duplicate tuples (each true proposition is recorded once only).

OPERATORS

The operators used to manipulate relations are similar, and in large part analogous, to the familiar set operators in mathematics. In fact, relations are carefully defined sets. All of the operators take relations as arguments and produce a relation as their result – this is the *relational closure property*. Closure is critically important because it insures that **I** can write and evaluate nested relational expressions; those in which an expression can be substituted in place of a relation. The most fundamental operators used to manipulate relations include:

- Restrict;
- Project;
- Union; and,
- Join.

Restrict produces a relation whose tuples satisfy a condition. In other words, **Restrict** allows one to select a subset of the tuples in a relation. **Project** produces a relation with only the attributes specified in the **Project** expression. The result contains all of the tuples of the original relation with only those attributes specified in the **Project** expression. **Restrict** produces a row-wise subset, and **Project** produces an attribute-wise subset.

Union produces a relation containing all of the tuples from two specified relations of the same type. Because all tuples in a relation must conform to the same predicate, the two relations contributing to the union must be based on the same predicate, or in other words, of the same type. This insures that the result is a valid relation, and closure is maintained.

Join produces a relation that consists of all tuples produced by forming all possible pairs of tuples from two specified relations with at least one common attribute subject to the condition that the values of the common attribute are equal. Join does precisely what its name implies - it takes two relations and 'glues' them together into 'wider' relation. If you imagine the input relations as sets of tuples, **Join** takes the Cartesian product of the two sets of tuples, and then selects only those new tuples for which the values of the common attributes of the input tuples are equal. There are many different varieties of the **Join** operator, the one I have described being the most general.

In addition to the manipulation operators, there is a set of operators used to create and modify relations and update the values of their attributes. These will be discussed briefly in the following appendix describing SQL.

CONSTRAINTS

The two crucial Data Integrity Constraints have to do with unique values that identify tuples. A *primary key* is an attribute whose value is unique for all the tuples in a relation. In general it is useful for all relations to have a primary key. The primary key can consist of a calculated attribute that combines values from other attributes to yield a unique value. However, for reasons I will not discuss here, it is cleaner and more efficient to have a single-attribute unique key in each relation.

A relations associated with another relation may store a value of the associated relation's primary key in one of their attributes, to allow joining of the associated tuples in each relation. The attribute that stores the values of the associated table's primary key is called a *foreign key*, for obvious reasons. Foreign keys are constrained in the sense that a value stored in a foreign key attribute must exist as a value of the primary key in one of the tuples of the associated relation.

In circumstances where tuples in associated relations are nonsensical in the absence of corresponding tuples in the associated relation, the use of primary and foreign key constraints insures that no such orphan (and meaningless) tuples will ever exist.

Again, the foregoing discussion of constraints is only the tip of a very large iceberg! Constraints come in many other forms, including the careful construction of the domains or types from which attribute values may be constructed and manipulated.

Appendix B

STRUCTURED QUERY LANGUAGE - SQL

SQL or Standard Query Language was invented and first implemented by IBM in the 1970s as SEQUEL (Structured English Query Language). SQL is the practical specification of the relational model, and working implementations of SQL are offered by many database systems vendors. Most of the widely available implementations of SQL are based on the **International Standard Database Language SQL (1992)** or SQL/92. Recently, an updated version of the specification has been produced and is referred to as SQL/99. However, at the time of this writing SQL/99 has not yet been implemented by any database systems vendor.

In very broad terms, SQL has three components:

- Data Definition Langauge or DDL components;
- Data Manipulation Language or DML components; and,
- Management components.

DDL - DATA DEFINITION LANGUAGE

The data definition statements allow the user to create, modify and drop (or delete) tables (what we discussed as 'relations') and statements to manage various other aspects of the data architecture of the database. For our purposes the important statements are:

- CREATE TABLE - CREATE AND DEFINE A NEW TABLE
- Modify table - MODIFY THE DEFINITION OF A TABLE
- Drop table - DELETE A TABLE FROM THE DATABASE

Following are examples of these commands and the output they produce using IBM's implementation of SQL, available in their DB2 Universal Database product. I create three tables that will be used in the next example.

CREATE TABLE EXAMPLE

Example B-1

```
CREATE TABLE people
(id Integer NOT NULL PRIMARY KEY,
name Varchar(20) NOT NULL,
sex Char(1) NOT NULL);
```

Example B-1 creates a table named 'people' with fields (attributes) id, name, and sex. id is a primary key, and none of the fields may contain a null value. For each field, the data type is defined after the field name and is self-explanatory.

Example B-2

```
CREATE TABLE locations
(id Integer NOT NULL PRIMARY KEY,
name VarChar(20) NOT NULL,
Type Char(1) NOT NULL);
```

Example B-2 creates a table named 'locations' with fields id, name, and type. Again, id is the primary key, and none of the fields can contain a null value.

```
CREATE TABLE lives_at                                    Example B-3
(id Integer NOT NULL PRIMARY Key,
pid Integer NOT NULL,
lid Integer NOT NULL,
FOREIGN KEY (pid) REFERENCES people (id),
FOREIGN KEY (lid) REFERENCES locations (id));
```

Example B-3 creates a table named 'lives_at' with fields id, pid, and lid. This is a *relationship* table that stores information linking people to the places where they live; hence, the name 'lives_at'. The id field is a primary key, no nulls are allowed in any fields, and the two fields pid and lid store values from the id fields of the people and locations tables and are, therefore, foreign keys.

MODIFY TABLE EXAMPLE

```
ALTER TABLE people                                       Example B-4
ADD COLUMN dob Date;
```

Example B-4 adds a date of birth (dob) field of type 'Date' to the people table.

DROP TABLE EXAMPLE

```
DROP TABLE people;                                       Example B-5
```

Example B-5 deletes the people table from the database.

DML - DATA MANIPULATION LANGUAGE

The data manipulation statements allow the user to manipulate and modify the data stored in the tables. These include:

- Insert
- Update
- Delete
- Select
- UNION

INSERT EXAMPLE

Example B-6

```
INSERT INTO people (id,name,sex)
VALUES(1,'Mary','F'),
      (2,'Paul','M'),
      (3,'Simon','M'),
      (4,'Jane','F'),
      (5,'Albert','M'),
      (6,'Ruth','F');
```

Example B-1 populates the people table with the listed values, shown below:

```
ID          NAME                 SEX
----------- -------------------- ---
          1 Mary                 F
          2 Paul                 M
          3 Simon                M
          4 Jane                 F
          5 Albert               M
          6 Ruth                 F
```

Example B-7

```
INSERT INTO locations (id,name,type)
VALUES(1,'Nairobi','U'),
      (2,'Accra','U'),
      (3,'Mtubatuba','R'),
      (4,'Sinafala','R'),
      (5,'Cairo','U');
```

Example B-7 populates the locations table with the listed values, shown below:

```
ID          NAME                 TYPE
----------- -------------------- ----
          1 Nairobi              U
          2 Accra                U
          3 Mtubatuba            R
          4 Sinafala             R
          5 Cairo                U
```

```
INSERT INTO lives_at (id,pid,lid)
  VALUES(1,1,4),
        (2,2,5),
        (3,3,2),
        (4,4,2),
        (5,5,1),
        (6,6,3);
```

Example B-8

Example B-8 populates the lives_at table with the listed values, shown below:

```
ID          PID         LID
----------- ----------- -----------
          1           1           4
          2           2           5
          3           3           2
          4           4           2
          5           5           1
          6           6           3
          7           3           4
          8           4           1
```

UPDATE EXAMPLE

```
UPDATE people
SET dob = '1950-6-4'
WHERE id = 1;
```

Example B-9

Example B-9 updates the dob field in the people table with the listed value, shown below after an update for all six records (tuples) in the people table:

ID	NAME	SEX	DOB
1	Mary	F	06/04/1950
2	Paul	M	10/12/1966
3	Simon	M	04/01/1980
4	Jane	F	09/12/1981
5	Albert	M	12/02/1991
6	Ruth	F	07/11/1991

DELETE EXAMPLE

```
DELETE FROM lives_at
  WHERE id = 20;
```

Example B-10

Example B-10 deletes the record(s) in table lives_at for which id is equal to 20.

SELECT and UNION EXAMPLES

```
SELECT p.name, l.name
FROM people as p, lives_at as la, locations as l
WHERE p.id = la.pid
AND la.lid = l.id
ORDER BY p.name;
```

Example B-11

Example B-11 produces the Cartesian product of the people, lives_at and locations tables and then selects the resulting records for which the id field from the people table is equal to the pid field from the lives_at table and the lid field from the lives_at table is equal to the id field from the locations table. Once the records fulfilling those conditions are identified, the name fields from the people and locations tables are selected, and the records are listed. This effectively *joins* the people, lives_at, and locations tables *on* the field pairs ([people].id ,pid) and (lid,[locations].id) subject to the condition that the resulting records must have equal values for those pairs of fields. In terms of the relational model, this select statement performs a join and a projection. The join is accomplished through the Cartesian product and the conditions placed on the id fields. This statement makes the join explicit and equivalent to a Cartesian product followed by a restriction; there is a more compact syntax that accomplishes the same

thing without specifying the equalities explicitly, but for pedantic purposes, I chose the most explicit form. The projection is accomplished through the selection of the name fields for retrieval from the result table. The order by clause sorts the resulting table on the name field from the people table. The statement is run on the tables created in the previous examples, and the result is shown below:

```
NAME                 NAME
-------------------- --------------------
Albert               Nairobi
Jane                 Nairobi
Jane                 Accra
Mary                 Sinafala
Paul                 Cairo
Ruth                 Mtubatuba
Simon                Accra
Simon                Sinafala
```

```
SELECT p1.name, p2.name
   FROM people as p1, people as p2
ORDER BY p1.name, p2.name;
```

Example B-12

Example B-12 produces a full Cartesian product of the people table with itself. In this case no restriction is put on the result of the Cartesian product resulting in a table consisting of records that pair each record in the people table with every other record in the people table, including itself. Again, a projection is performed to retrieve just the name fields from the resulting table. Last, the order by clause at the end of this statement sorts the resulting table first on the name field from the first instance of the people table and then on the name field of the second instance of the people table. The final result is shown below:

```
NAME                 NAME
-------------------- --------------------
Albert               Albert
Albert               Jane
Albert               Mary
Albert               Paul
Albert               Ruth
Albert               Simon
Jane                 Albert
Jane                 Jane
Jane                 Mary
Jane                 Paul
Jane                 Ruth
Jane                 Simon
Mary                 Albert
Mary                 Jane
Mary                 Mary
Mary                 Paul
Mary                 Ruth
Mary                 Simon
Paul                 Albert
Paul                 Jane
Paul                 Mary
Paul                 Paul
Paul                 Ruth
Paul                 Simon
Ruth                 Albert
Ruth                 Jane
Ruth                 Mary
Ruth                 Paul
Ruth                 Ruth
Ruth                 Simon
Simon                Albert
Simon                Jane
Simon                Mary
Simon                Paul
Simon                Ruth
Simon                Simon
```

Example B-13

```
SELECT p.name, l.name
FROM people as p, lives_at as la, locations as l
WHERE p.id = la.pid
AND la.lid = l.id
AND p.dob < '1980-01-01'
ORDER BY p.name;
```

Example B-13 is the same as Example B-11 with the addition of an additional restriction, p.dob < '1980-01-01', that retrieves only records for which the date of birth is prior to January 1, 1980, below:

```
NAME                NAME
------------------- --------------------
Mary                Sinafala
Paul                Cairo
```

Example B-14

```
   SELECT *
   FROM people
UNION ALL
   SELECT *
   FROM people;
```

Example B-14 selects all of the records from the people table and performs a union of those records to create the result below in which all of the records in the people table appear twice:

```
ID          NAME                SEX DOB
----------- ------------------- --- ----------
         1 Mary                F   06/04/1950
         2 Paul                M   10/12/1966
         3 Simon               M   04/01/1980
         4 Jane                F   09/12/1981
         5 Albert              M   12/02/1991
         6 Ruth                F   07/11/1991
         1 Mary                F   06/04/1950
         2 Paul                M   10/12/1966
         3 Simon               M   04/01/1980
         4 Jane                F   09/12/1981
         5 Albert              M   12/02/1991
         6 Ruth                F   07/11/1991
```

TRANSACTIONS

The archetypal example of a transaction is the transfer of money from one account to another in a bank's accounting database. The process involves two steps: 1) debit the transfer amount from customer A's account and 2) credit the transfer amount to customer B's account. Both of these updates must occur for the transfer to be complete, and under no circumstance can the first be performed without performing the second. A transaction is used to encapsulate these two updates into an atomic (see below) process that cannot save modifications to the database unless both updates are successful.

A transaction is a logical unit of work that must be completed in its entirety or not at all. Transactions are sometimes referred to as 'atomic' – in the sense that an atom is (in colloquial terms) the most fundamental unit of nature and cannot be subdivided. A process performed as a transaction only modifies the database when the entire process has been completed successfully. Consequently, it is possible to cancel a process performed as a transaction, even a complicated one, midway through without leaving the database in an inconsistent state. The cancellation could be user-initiated, the result of a system failure or any other intended or unintended action. Properly structured transactions are absolutely critical to maintaining the integrity of the database, and the SQL standard specifies the basic implementation of transactions.

SQL IN PRACTICAL USE

In practice, SQL is used in three ways. SQL can by used in a dynamic sense in either of two ways: 1) a user can type SQL into a command line interpreter (CLP) that parses the statements and passes them to the database management system for execution in real time, or 2) a program can construct SQL in real time based on parameter values or input from a user and then pass the SQL to the interpreter and onward to the database management system. In both of these situations, the SQL is interpreted and optimized in real time – hence the term 'dynamic' SQL. Static SQL consists of predefined, pre-optimized SQL statements that are usually invoked by a program and cannot be altered in real time. Static SQL usually runs faster because the interpretation, optimization and compile steps need not be performed, and for this reason, operations that are run many times without change are best implemented as static SQL.

Although SQL is a computational complete language, it is usually used as a data manipulation sublanguage that in combination with a fully-featured host language provides all of the functionality necessary to manage and manipulate the data, to perform the logic necessary for the application to run, and to interact with the user. Common host languages include C, C++, Java, Visual Basic and Cobol.

BIBLIOGRAPHY

Allison, Paul. 1982. "Discrete-Time Methods for the Analysis of Event Histories." Pp. 61-98 in Sociological Methodology, edited by S. Leinhardt. San Francisco: Jossey-Bass.

—. 1984. Event History Analysis. Regression for Longitudinal Event Data. Beverly Hills: Sage.

Anderson, R. 2001. Personal Communication. Washington DC.

Anderson, R. M., R. M. May, and A. R. McLean. 1988. "Possible Demographic Consequences of AIDS in Developing Countries." *Nature*, 332(6161):228-34.

Anderson, R. M., T. W. Ng, M. C. Boily, and R. M. May. 1989. "The Influence of Different Sexual-Contact Patterns between Age Classes on the Predicted Demographic Impact of AIDS in Developing Countries." *Annals of the New York Academy of Sciences*, 569:240-74.

Anderson, R. M. 1991a. "Mathematical Models of the Potential Demographic Impact of AIDS in Africa." *AIDS*, 5(Suppl 1):S37-44.

—. 1991b. "AIDS and Its Demographic Impact." Pp. 214-20 in Disease and Mortality in Sub Saharan Africa, edited by R. G. Feachem and D. T. Jamison. Oxford: Oxford University Press.

Anderson, R. M., R. M. May, M. C. Boily, G. P. Garnett, and J. T. Rowley. 1991. "The Spread of HIV-1 in Africa: Sexual Contact Patterns and the Predicted Demographic Impact of AIDS." *Nature*, 352(6336):581-9.

Anderson, R. M., W. Ng, J. Rowley, and A. R. McLean. 1991. "The Anderson Approach: Potential Demographic Impact of AIDS – Uncertainties in Prediction." Pp. 63-76 in The AIDS Epidemic and Its Demographic Consequences. Proceedings of the United Nations/World Health Organization Workshop on Modelling the Demographic Impact of the AIDS Epidemic in Pattern II Countries: Progress to Date and Policies for the Future, New York, 13-15 December 1989. New York, New York/Geneva, Switzerland: United Nations. Department of International Economic and Social Affairs.

Anderson, R. M. 1992. "Some Aspects of Sexual Behaviour and the Potential Demographic Impact of AIDS in Developing Countries." *Social Science & Medicine*, 34(3):271-80.

Anderson, R. M. and R. M. May. 1992. "Understanding the AIDS Pandemic." *Scientific American*,:58-66.

Anderson, R. M., R. M. May, T. W. Ng, and J. T. Rowley. 1992. "Age-Dependent Choice of Sexual Partners and the Transmission Dynamics of HIV in Sub-Saharan Africa." *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 336(1277):135-55.

Anderson, R. M. 1996. "The Spread of HIV and Sexual Mixing Patterns." Pp. 71-86 in AIDS in the World II: Global Dimensions, Social Roots, and Responses. The Global AIDS Policy Coalition, edited by J. M. Mann and D. J. M. Tarantola. New York, New York: Oxford University Press.

Anonymous SOCSIM. 2001. "SOCSIM Documentation." http://www.demog.berkeley.edu/~marcia/c_doc.html. September 11, 2001.

—. 2001. "The SOCSIM Microsimulation Program: Selected Bibliography 1976-1998." http://www.demog.berkeley.edu/~wachter/socrefs.html. October 15, 2001.

Auvert, B., G. Buonamico, E. Lagarde, and B. Williams. 2000. "Sexual Behavior, Heterosexual Transmission, and the Spread of HIV in Sub-Saharan Africa: A Simulation Study." *Computers & Biomedical Research*, 33(1):84-96.

Auvert, Bertran. 1991. "The Auvert Approach: A Stochastic Model for the Heterosexual Spread of the Human Immunodeficiency Virus." Pp. 77-83 in The AIDS Epidemic and Its Demographic Consequences. Proceedings of the United Nations/World Health Organization Workshop on Modelling the Demographic Impact of the AIDS Epidemic in Pattern II Countries: Progress to Date and Policies for the Future, New York, 13-15 December 1989. New York, New York/Geneva, Switzerland: United Nations. Department of International Economic and Social Affairs.

Bongaarts, J. 1989. "A Model of the Spread of HIV Infection and the Demographic Impact of AIDS." *Statistics in Medicine*, 8(1):103-20.

Bongaarts, John and Robert G. Potter. 1983. Fertility, Biology, and Behavior: An Analysis of the Proximate Determinants. San Diego: Academic Press.

Brouard, Nicolas. 1991. "The Brouard Appraoch: Forecasting the AIDS Epidemic in an African Population." Pp. 84-89 in The AIDS Epidemic and Its Demographic Consequences. Proceedings of the United Nations/World Health Organization Workshop on Modelling the Demographic Impact of the AIDS Epidemic in Pattern II Countries: Progress to Date and Policies for the Future, New York, 13-15 December 1989. New York, New York/Geneva, Switzerland: United Nations. Department of International Economic and Social Affairs.

Chick, S. E., A. L. Adams, and J. S. Koopman. 2000. "Analysis and Simulation of a Stochastic, Discrete-Individual Model of Std Transmission with Partnership Concurrency." *Mathematical Biosciences*, 166(1):45-68.

Citro, Constance F. and Eric A. Hanushek, Editors. 1991a. Improving Information for Social Policy Decisions: The Use of Microsimulation Modeling, Technical Papers, vol. II. Washington, D.C.: National Academy Press.

—, Editors. 1991b. Improving Information for Social Policy Decisions: The Use of Microsimulation Modeling, Review and Recommendations, vol. I. Washington, D.C.: National Academy Press.

Coale, A. and J. Trussell. 1996. "The Development and Use of Demographic Models." *Population Studies*, 50(3):469-84.

Coale, Ansley J. and Paul Demeny. 1966. Regional Model Life Tables and Stable Populations. Princeton, New Jersey: Princeton University Press.

Colson, Elizabeth. 1960a. Social Organization of the Gwembe Tonga. Manchester: Manchester University Press.

—. 1960b. "Homestead, Household and the Family." Pp. 94-121 in Social Organization of the Gwembe Tonga. Manchester: Manchester University Press.

—. 1971a. The Social Consequences of Resettlement: The Impact of the Kariba Resettlement Upon the Gwembe Tonga. Manchester: University of Manchester Press.

—. 1971b. "The Family in Crisis." in The Social Consequences of Resettlement: The Impact of the Kariba Resettlement Upon the Gwembe Tonga. Manchester: University of Manchester Press.

Colson, Elizabeth and Thayer Scudder. 1987. For Prayer and Profit: The Ritual, Economic and Social Importance of Beer in Gwembe District, Zambia, 1950-1982. Stanford: Stanford University Press.

Date, C.J. 2000. An Introduction to Database Systems, 7th. Reading Massachusetts: Addison-Wesley.

Garnett, G. P. and R. M. Anderson. 1993a. "Factors Controlling the Spread of HIV in Heterosexual Communities in Developing Countries: Patterns of Mixing between Different Age and Sexual Activity Classes." *Philosophical Transactions of the Royal Society of London - Series B: Biological Sciences*, 342(1300):137-59.

—. 1993b. "No Reason for Complacency About the Potential Demographic Impact of AIDS in Africa." *Transactions of the Royal Society of Tropical Medicine & Hygiene*, 87(Suppl 1):S19-22.

—. 1994a. "Balancing Sexual Partnerships in an Age and Activity Stratified Model of HIV Transmission in Heterosexual Populations." *IMA Journal of Mathematics Applied in Medicine & Biology*, 11(3):161-92.

Garnett, G. P., J. Swinton, and G. Parker. 1994. "Sex Acts and Sex Partners." *Journal of Acquired Immune Deficiency Syndromes*, 7(9):989-90; discussion 990-2.

Garnett, G. P. and R. M. Anderson. 1995. "Strategies for Limiting the Spread of HIV in Developing Countries: Conclusions Based on Studies of the Transmission Dynamics of the Virus." *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*, 9(5):500-13.

—. 1996a. "Sexually Transmitted Diseases and Sexual Behavior: Insights from Mathematical Models." *Journal of Infectious Diseases*, 174(Suppl 2):S150-61.

Garnett, G. P., J. P. Hughes, R. M. Anderson, B. P. Stoner, S. O. Aral, W. L. Whittington, H. H. Handsfield, and K. K. Holmes. 1996. "Sexual Mixing Patterns of Patients Attending Sexually Transmitted Diseases Clinics." *Sexually Transmitted Diseases*, 23(3):248-57.

Garnett, Geoffrey P. and Roy M. Anderson. 1994b. "Balancing Sexual Partnerships in an Age and Activity Stratified Model of HIV Transmission in Heterosexual Populations." *IMA Journal of Mathematics Applied in Medicine & Biology*, 11:161-192.

—. 1996b. "Sexually Transmitted Diseases and Sexual Behavior: Insights from Mathematical Models." *The Journal of Infectious Diseases*, 174(Supplement 2):S150-S161.

Gray, R.H. 2000. Personal Communication. Washington, D.C.

Gregson, S., G. P. Garnett, and R. M. Anderson. 1994a. "Assessing the Potential Impact of the HIV-1 Epidemic on Orphanhood and the Demographic Structure of Populations in Sub-Saharan Africa." *Population Studies*, 48(3):435-58.

—. 1994b. "Is HIV-1 Likely to Become a Leading Cause of Adult Mortality in Sub-Saharan Africa?" *Journal of Acquired Immune Deficiency Syndromes*, 7(8):839-52.

Gregson, S., G. P. Garnett, R. Shakespeare, G. Foster, and R. M. Anderson. 1994. "Determinants of the Demographic Impact of HIV-1 in Sub-Saharan Africa: The Effect of a Shorter Mean Adult Incubation Period on Trends in Orphanhood." *Health Transition Review*, 4(Suppl):65-92.

Heuveline, Patrick. 2001. "The Demographic Impact of the HIV Dynamics for High Prevalence Populations in Southern and Eastern Africa." Presentation at 24th International Union for the Scientific Study of Population (IUSSP) General Conference sponsored by International Union for the Scientific Study of Population (IUSSP), Salvador, Brazil, August 22, 2001.

Himes, C.L., S.H. Preston, and G.A. Condran. 1994. "A Relational Model of Mortality at Older Ages in Low Mortality Countries." *Population Studies*, 48:269-291.

INDEPTH. 2000. "*INDEPTH* Announcement." http://www.indepth-network.org. September 18, 2000.

—. Forthcoming 2001. Population and Health in Developing Countries, vol. 1, Population, Health and Survival at *INDEPTH* Sites. Ottawa, Canada: International Development Research Centre.

John, Meredith A. 1991. "A Model of HIV-1 Transmission for Urban Areas of Africa." *Theoretical Population Biology*, 39(2):148-69.

Kaufman, L. and P.J. Rousseeuw. 1990. Finding Groups in Data: An Introduction to Cluster Analysis. New York: Wiley.

Lin, J. 1995. "Changing Kinship Structure and Its Implications for Old-Age Support in Urban and Rural China." *Population Studies*, 49(1):127-47.

Lotka, A. J. 1939. Théorie Analytique Des Associations Biologiques, Part. II. Analyse Démographique Avec Application Particulière À L'espèce Humaine, Actualités Scientifiques Et Industrielles, 780. Paris: Hermann et Cie.

Mathsoft Inc. 1999a. "Chapter 4: Cluster Analysis." Pp. 67-114 in S-Plus 2000: Guide to Statistics, vol. 2. Seattle Washington: Mathsoft Inc.

—. 1999b. "Chapter 4: Cluster Analysis." Pp. 102 in S-Plus 2000: Guide to Statistics, vol. 2. Seattle Washington: Mathsoft Inc.

Palloni, Alberto and Martha Glicklich. 1991. "Review of Approaches to Modelling the Demographic Impact of the AIDS Epidemic." Pp. 20-50 in The AIDS Epidemic and Its Demographic Consequences. Proceedings of the United Nations/World Health Organization Workshop on Modelling the Demographic Impact of the AIDS Epidemic in Pattern II Countries: Progress to Date and Policies for the Future, New York, 13-15 December 1989. St/Esa/Ser.A/119, edited by United Nations Department of International Economic and Social Affairs and World Health Organization Global Programme on AIDS. New York, New York/Geneva, Switzerland: United Nations.

Palloni, Alberto and Luis Lamas. 1991. "The Palloni Approach: A Duration-Dependent Model of the Spread of HIV/AIDS in Africa." Pp. 109-118 in The AIDS Epidemic and Its Demographic Consequences. Proceedings of the United Nations/World Health Organization Workshop on Modelling the Demographic Impact of the AIDS Epidemic in Pattern II Countries: Progress to Date and Policies for the Future, New York, 13-15 December 1989. New York, New York/Geneva, Switzerland: United Nations. Department of International Economic and Social Affairs.

Petersen, Trond. 1995. "Analysis of Event Histories." Pp. 453-517 in Handbook of Statistical Modeling for the Social and Behavioral Sciences, edited by G. Arminger, C. C. Clogg, and M. E. Sobel. New York: Plenum Press.

Preston, Samuel H., Patrick Heuveline, and Michel Guillot. 2001. Demography: Measuring and Modeling Population Processes. Oxford, UK: Blackwell.

Scudder, Thayer. 1962. Ecology of the Gwembe Tonga. Manchester: Manchester University Press.

—. 1967. "Lake Kariba Fishermen: Preliminary Report." Report, Food and Agriculture Organization of the United Nations, Rome.

Scudder, Thayer and Elizabeth Colson. 1977. "Long-Term Field Research in Gwembe Valley, Zambia." Pp. 227-254 in Long-Term Field Research in Social Anthropology, edited by F. G. New York: Academic Press.

—. 1980. Secondary Education and the Formation of an Elite: The Impact of Education on Gwembe District, Zambia. New York: Academic Press.

Siamwiza, Bennett Siamwiinde. 1993. "Hunger in the Gwembe Valley: A Case Study of Mweemba Chieftancy, 1905-1987.", History, University of Zambia, Lusaka.

Srikantan, K.S. 1982. "Population Models." Pp. 526-9 in International Encyclopedia of Population, vol. 2, edited by J. A. Ross. New York: Free Press.

Stanley, E. A., S. T. Seitz, P. O. Way, P. D. Johnson, and T. F. Curry. 1991. "The United States Interacgency Working Group Approach: The Iwg Model for the Heterosexual Spread of HIV and the Demographic Impact of the AIDS Epidemic." Pp. 119-136 in The AIDS Epidemic and Its Demographic Consequences. Proceedings of the United Nations/World Health Organization Workshop on Modelling the Demographic Impact of the AIDS Epidemic in Pattern II Countries: Progress to Date and Policies for the Future, New York, 13-15 December 1989. New York, New York/Geneva, Switzerland: United Nations. Department of International Economic and Social Affairs.

STATA Corporation. 1997. Stata Reference Manual a-F, vol. 1, Release 5. College Station Texas: STATA Press.

Struyf, A. and M. Hubert. 1997. "Integrating Robust Clustering Techniques in S-Plus." *Computational Statistics and Data Analysis*, 26:17-37.

United Nations, Department of International Economic and Social Affairs. 1982. Model Life Tables for Developing Countries, Population Studies, No. 77. New York: United Nations.

—. 1983. "Chapter III: Estimation of Child Mortality from Information on Children Ever Born and Children Surviving." Pp. 73-96 in Manual X: Indirect Techniques for Demographic Estimation, Population Studies, No. 81. New York: United Nations.

—. 1988. Computer Program: *Mortpak-Lite* 3.0 (IBM PC Compatible): New York: United Nations.

United Nations Department of International Economic and Social Affairs and World Health Organization Global Programme on AIDS, Editors. 1991. The AIDS Epidemic and Its Demographic Consequences. Proceedings of the United Nations/World Health Organization Workshop on Modelling the Demographic Impact of the AIDS Epidemic in Pattern II Countries: Progress to Date and Policies for the Future, New York, 13-15 December 1989. New York, New York/Geneva, Switzerland: United Nations.

Van der Ploeg, C. P., C. Van Vliet, S. J. De Vlas, J. O. Ndinya-Achola, L. Fransen, G. J. Van Oortmarssen, and J. D. Habbema. 1998. "Stdsim: A Microsimulation Model for Decision Support in Std Control." *Interfaces*, 28(3):84-100.

Van Vliet, C. 1995. "Use of the Response Surface Method for the Quantification of a Sexual Behavior Model." IIASA working paper: WP-95-110, Laxenburg, Austria: IIASA.

Wachter, Kenneth W. 1995. "2030's Seniors: Kin and Step-Kin." at Annual Meeting of the Population Association of America sponsored by Population Association of America, San Francisco, California, April 6-8, 1995.

—. 1997. "Kinship Resourced for the Elderly." *Philosophical Transactions of the Royal Society of London - Series B: Biological Sciences*, 352(29).

Wachter, Kenneth W., Debra Blackwell, and Eugene E. Hammel. 1997. "Testing the Validity of Kinship Microsimulation." *Journal of Mathematical and Computer Modeling*, 26:89-104.

Wachter, Kenneth W., John E. Knodel, and Mark VanLandingham. 2000. "AIDS and the Elderly of Thailand: Projecting Familial Impacts." Report 00-446, Population Studies Center at the Institute for Social Research, University of Michigan, Ann Arbor.

Wachter, Kenneth W. 2001. "SOCSIM: Description of the Program." http://www.demog.berkeley.edu/~wachter/socstory.html. September 11, 2001.

INDEX

Affair, 36, 205, 218, 223, 313, 316, 317, 322, 323, 324, 330
Anthropology, 52
Architecture, 240
Associativity, 315, 318, 319
Basic reproduction number, 231
Brass method, 172, 238
Calibration, 338
CD4, 264, 285, 286, 295, 348
Cluster, 23
Coale & Demeny, 7, 11, 18, 31, 34, 35, 36, 51
Colson, Elizabeth, 52, 53, 56, 57, 58, 62, 63, 65, 66, 67, 68, 70, 72, 84, 105, 147, 198
Conception, 254, 267, 269, 276
Contraception, 279
Cox, 97
Database, 241, 383, 388, 389
DB2 Universal Database, 241, 389
Demography, 52, 201, 208
DHS Survey, 136, 271
Dummy variable, 119, 131
Ecology, 55
Empirical, 260, 265, 290, 299, 304, 309
Epidemic, 205
Event history analysis, 96, 97
Events, 256, 270, 271, 291, 376, 379
Evidence of HIV, 154, 192
Experiences, 376, 378, 379, 380
Factor analysis, 21
Family, 71, 76, 78, 81
Fecundability, 270, 276
Fecundity, 333
Fertility, 124, 126, 130, 133, 134, 136, 139, 140, 141, 142, 150, 151, 152, 153, 275, 278, 292, 331, 344, 345, 351, 352
Genealogies, 81, 372
Genetic, 284
Granule, 254
GTRP, 23, 24, 52, 54, 62, 66, 67, 70
Gwembe Tonga, 23, 24, 52, 53, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 73, 85, 86, 88, 98, 99, 100, 104, 105, 109, 122, 133, 137, 147, 148, 149, 153, 154, 170, 198, 242, 260, 265, 313, 339, 344
Gwembe Valley, 52, 53, 55, 57, 59, 60, 61, 62, 64, 66, 147
Hazard, 110, 111, 113, 114, 115, 116, 117, 118, 123, 130, 131, 134, 136, 143, 144, 150, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 171, 184, 186, 187, 188, 189, 190, 191, 192, 193, 194, 195, 259, 262, 263, 273, 304, 305, 306, 307, 309, 310, 311, 312, 319, 321, 323, 328, 329, 330, 332
Hazard rate, 110, 111, 113, 114, 115, 116, 117, 118, 123, 130, 131, 134, 136, 143, 144, 150, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 171, 184, 186, 187, 188, 189, 190, 191, 192, 193, 194, 195, 259, 262, 263, 273, 304, 305, 306, 307, 309, 310, 311, 312, 319, 321, 323, 328, 329, 330, 332
History, 59, 62, 96, 246, 359, 374
HIV models, 232
HIV/AIDS, 8, 12, 14, 31, 32, 33, 37, 39, 40, 42, 43, 47, 48, 49, 50, 59, 61, 133, 148, 154, 155, 170, 171, 176, 192, 197, 198, 199, 200, 205, 206, 207, 208, 209, 210, 211, 212, 213, 214, 215, 216, 217, 218, 219, 221, 222, 223, 225, 226, 227, 228, 229, 231, 232, 233, 235, 236, 237, 238, 239, 240, 243, 247, 249, 250, 253, 254, 256, 257, 258, 259, 260, 262, 264, 265, 280, 281, 284, 285, 286, 287, 288, 289, 290, 291, 292, 293, 294, 295, 296, 297, 302, 313, 326, 327, 328, 329, 330, 331, 332, 333, 334, 335, 336, 337, 338, 340, 341, 346, 347, 348, 349, 350, 351, 352, 353, 354, 355, 356, 357
Implement, 241
Incidence, 353
Incubation, 288
INDEPTH Network, 1, 7, 11, 12, 13, 14, 15, 16, 17, 18, 20, 21, 22, 23, 25, 26, 27, 28, 29, 30, 31, 32, 35, 36, 39, 42, 43, 47, 48, 49, 50, 51
Infection, 288, 292, 293, 346, 351
Institute for African Studies, University of Zambia, 66
Interactions, 326, 327, 336
Intercourse, 256, 265, 269, 270, 271, 273, 276, 291, 292, 313, 322, 323, 332
Lake Kariba, 53, 55, 58, 59, 62
Levels, 106, 122, 134, 156, 164, 170
Logistic, 261
Macro model, 202, 203, 204, 205, 219, 227, 232
Marriage, 105, 112, 113, 115, 130, 297, 298, 303, 327, 332, 378, 379
Meta data, 365, 381
Micro model, 203, 220, 221, 223, 232
Microsoft, 81, 83, 241, 255, 381

Migration, 253
Miscarriage, 77, 82, 83, 279
Model, 7, 9, 11, 20, 31, 35, 39, 40, 44, 47, 48, 49, 50, 51, 109, 111, 119, 123, 131, 134, 136, 143, 150, 171, 184, 192, 246, 261, 373
Modular, 252
Module, 253, 258, 265, 275, 276, 281, 284, 285, 291, 292, 294, 297, 313, 326
Morbidity, 33, 330
Mortality, 7, 9, 11, 13, 14, 15, 16, 17, 18, 20, 21, 23, 25, 26, 27, 28, 29, 30, 31, 33, 36, 42, 43, 47, 48, 49, 50, 170, 171, 175, 184, 192, 258, 260, 330, 342, 343, 350, 351
Mortality pattern, 25, 35
Multinomial, 136
Network, 1
Nuptiality, 105, 297, 298, 327, 332
Oracle, 242
Output, 337, 357
Parameter, 253, 255, 263, 270, 271, 274, 282, 294, 306, 312, 319, 336
Person-year, 99, 100, 108, 109, 150, 171, 184, 195, 260
Polygyny, 121, 303
Population model, 201, 232
Population simulator, 242, 246, 292
Prevalence, 354, 355
Principal component model, 21
Regress, 9
Relocation, 59
Reproduction, 256, 257
Resettlement, 56, 105
Sample, 69, 71, 81, 85, 88, 89, 90, 91, 92, 93, 94, 95, 98, 99, 100, 107, 108, 307, 316
Scudder, Thayer, 52, 53, 54, 59, 62, 63, 65, 66, 67, 68, 70, 72, 84, 122, 147, 198
Security, 367
Separation, 101, 124, 125, 126, 127, 128, 129, 130, 131, 132, 307, 309, 310, 311, 312
Sex, 76, 84, 86, 88, 95, 100, 101, 111, 143, 146, 199, 256, 342
Sexual activity, 229
SPEHR – Structured Population Event History Register, 359, 360, 361, 362, 363, 364, 365, 366, 367, 369, 370, 371, 372, 373, 374, 375, 377, 378, 379, 380, 381, 382
S-Plus, 22
SQL, 74, 81, 83, 99, 100, 102, 242, 245, 371, 382, 386, 388, 389, 397, 398
STATA, 12, 23, 108, 109, 111, 123, 192, 195, 261
Stochastic, 199, 220
Total fertility rate, 141, 142, 146, 344, 345
Transmission, 211, 257, 281, 289, 290, 291, 292, 334, 335, 346
Trend, 119, 120, 121, 131, 132, 150, 151, 152, 156, 157, 158, 164, 165, 175, 186, 193, 194, 195
Trends, 59, 118, 131, 147, 175
Twin, 84
United Nations, 7, 18, 31, 32, 33, 35, 36, 51, 205, 209, 218, 223
University of Zambia, 62, 66
UNIX, 241
Validity, 369
Vertical transmission, 227
Zambezi River, 53, 58, 62, 64
Zambia, 35, 52, 53, 54, 55, 59, 60, 61, 62, 66, 67, 136, 192, 197, 198, 210, 271