A DISCRETE-TIME MULTIPLE EVENT PROCESS SURVIVAL MIXTURE (MEPSUM) MODEL FOR INVESTIGATING THE ORDER AND TIMING OF MULTIPLE NON-REPEATABLE EVENTS

Danielle O. Dean

A thesis submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Arts in the Department of Psychology (Quantitative)

Chapel Hill
2012

Approved by:

Daniel J. Bauer, Ph.D.

Robert C. MacCallum, Ph.D.

Michael J. Shanahan, Ph.D.

ABSTRACT

DANIELLE O. DEAN: A discrete-time multiple event process survival mixture (MEPSUM) model for investigating the order and timing of multiple non-repeatable events

(Under the direction of Daniel Bauer)

Traditional survival analysis was developed to investigate both the occurrence and the timing of an event, but researchers have recently begun to ask questions about the order and timing of multiple events. A multiple event process survival mixture model is developed here to analyze non-repeatable events measured in discrete-time that are not mutually exclusive. The model assumes the population is composed of a finite number of subpopulations of individuals who are homogeneous with respect to the risk of multiple events over time, in order to parsimoniously describe the underlying multivariate distribution of hazard functions. The model builds on both traditional univariate survival analysis and univariate survival mixture analysis. The model is applied to two empirical data sets, one concerning transitions to adulthood and another concerning age of first use of a number of substances. Promising opportunities, as well as possible limitations and future directions are discussed.

DEDICATION

To my father, Ed Dean.

ACKNOWLEDGEMENTS

I am grateful for the guidance and support of my advisor, Dr. Daniel Bauer, and my committee members, Dr. Michael Shanahan and Dr. Robert MacCallum, as well as my undergraduate advisor, Dr. Linda Isbell, who encouraged me to pursue the field of quantitative psychology. I am thankful for the members of the L. L. Thurstone Psychometric Laboratory, especially Ruth Baldasaro and Nisha Gottfredson, as well as Shawn Bauldry for their insight. I would also like to thank my friends and family – especially Adrian, Tara, and my parents – for their support.

This research uses data from Add Health, a program project directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill, and funded by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Information on how to obtain the Add Health data files is available on the Add Health website (http://www.cpc.unc.edu/addhealth). No direct support was received from grant P01-HD31921 for this analysis.

TABLE OF CONTENTS

LIST OF TABLES viii

LIST OF FIGURES x

LIST OF ABBREVIATIONS xii

LIST OF SYMBOLS xiii

Chapter

1. INTRODUCTION 1

1.1 Traditional Univariate Discrete-time Survival Analysis 6

1.2 Univariate Discrete-time Survival Mixture Analysis 15

1.3 Summary 20

2. A DISCRETE-TIME MULTIPLE PROCESS SURVIVAL MIXTURE MODEL 22

2.1 Model Definition 23

2.2 Software 26

2.3 Model Identification 27

2.4 Simulation Demonstration 31

2.4.1 Introduction 31

2.4.2 Methods 33

2.4.3 Results 36

2.4.4 Discussion 42

2.5 Utilizing the Model in Practice 43

2.5.1 Purpose of Model ... 43

2.5.2 Introduction to Model Building ... 46

2.5.3 Model Comparison ... 47

2.5.4 Suggested Model Building Strategy ... 50

3. EMPIRICAL EXAMPLE 1 – TRANSITIONS TO ADULTHOOD ... 56

3.1 Methods ... 58

3.1.1 Data ... 58

3.1.2 Measures ... 59

3.1.3 Analysis ... 63

3.2 Results ... 64

3.3 Discussion ... 88

4. EMPIRICAL EXAMPLE 2 – SUBSTANCE USE ONSET ... 92

4.1 Methods ... 93

4.1.1 Data ... 93

4.1.2 Measures ... 94

4.1.3 Analysis ... 97

4.2 Results ... 100

4.3 Discussion ... 126

5. DISCUSSION ... 130

5.1 Simulation Study ... 130

5.2 Empirical Examples ... 132

5.3 Other Limitations and Future Directions ... 133

APPENDIX ... 135

REFERENCES ... 137

LIST OF TABLES

Table 1: *Bias in average hazard* 37

Table 2: *Absolute bias in hazard* 38

Table 3: *Absolute bias in lifetime distribution functions* 39

Table 4: *Average bias in the probability of class membership* 41

Table 5: *Age at time of interview for individuals sampled in Wave IV Add Health* 59

Table 6: *Number of individuals with missing data in Add Health* 60

Table 7: *Number of event occurrences and sample estimated hazard probabilities in Add Health* 61

Table 8: *Model fit to Add Health data* 65

Table 9: *Add Health median event time within latent classes* 71

Table 10: *Add Health odds ratios for five class solution*………………………………..74

Table 11: *Substances included in each class of drugs in NSDUH* 94

Table 12: *Age at time of interview for individuals sampled in 2009 NSDUH* 95

Table 13: *Number of individuals with missing data - NSDUH* 96

Table 14: *NSDUH model fit with unstructured hazard functions* 100

Table 15: *NSDUH model fit with quadratic hazard functions* 101

Table 16: *NSDUH median event time within each latent class for four class solution* .. 104

Table 17: *NSDUH median event time within each latent class for five class solution* ... 108

Table 18: *NSDUH median event time within each latent class for six class solution* 112

Table 19: *NSDUH logit parameters from model with covariates compared to model without covariates* 117

Table 20: *NSDUH odds ratios for class membership in four class solution* 119

Table 21: *Predicted probabilities of class membership in NSDUH evaluation and validation sample depending on covariates* 123

Table 22: *NSDUH validation sample logit parameter estimates* 125

Table 23: *NSDUH multinomial logit parameters for class membership in evaluation and validation sample, with last class “early hard drug users” as a referent* 126

Table 24: *Average estimated standard deviation of all of the hazard parameters* 135

Table 25: *Empirical average standard deviation of all of the hazard parameters over the 100 replications* 135

Table 26: *Ratio of estimated average standard deviation of hazard parameters to empirical average standard deviation of all of the hazard parameters over 100 replications* 135

Table 27: *Bias in average logit* 136

Table 28: *Absolute bias in logit* 136

Table 29: *Proportion of 95% confidence interval coverage of logit parameters* 136

LIST OF FIGURES

Figure 1: *Estimated discrete-time survival analysis functions for Add Health data on age of college degree* ... 10

Figure 2: *Functions for the three different levels of (constant) risk* ... 33

Figure 3: *Bias in the average hazard, weighting over latent classes* ... 37

Figure 4: *Absolute bias in the hazard, weighting over latent classes* ... 38

Figure 5: *Absolute bias in lifetime distribution functions, weighting over latent classes* ... 40

Figure 6: *Absolute average bias in the probability of class membership, weighting over latent classes* ... 41

Figure 7: *Add Health sample observed functions* ... 62

Figure 8: *Add Health hazard functions for five class solution* ... 69

Figure 9: *Add Health lifetime distribution functions for five class solution* ... 70

Figure 10: *Add Health lifetime distribution functions for redundant classes found in six class solution* ... 72

Figure 11: *Add Health simple path diagram of model with covariates* ... 73

Figure 12: *Add health hazard indicators across all ages and events found in model without covariates compared to model with covariates* ... 74

Figure 13: *Add Health lifetime distribution functions depending on gender* ... 79

Figure 14: *Add Health lifetime distribution functions depending on race (Caucasian and African-American)* ... 83

Figure 15: *Add Health lifetime distribution functions depending on race (Hispanic and other race)* ... 84

Figure 16: *Add Health lifetime distribution functions depending on parental education* ... 87

Figure 17: *NSDUH sample observed hazard and lifetime distribution functions* ... 96

Figure 18: *NSDUH hazard functions for the four class solution* ... 105

Figure 19: *NSDUH lifetime distribution functions for the four class solution* ... 106

Figure 20: *NSDUH hazard functions for the five class solution* ... 109

Figure 21: *NSDUH lifetime distribution functions for the five class solution* 110

Figure 22: *NSDUH hazard functions for the six class solution* 113

Figure 23: *NSDUH lifetime distribution functions for the six class solution* 114

Figure 24: *NSDUH simple path diagram of model with covariates* 116

Figure 25: *NSDUH four class model implied lifetime distribution functions depending on gender* 120

Figure 26: *NSDUH four class model implied lifetime distribution functions depending on race* 121

LIST OF ABBREVIATIONS

Add Health	National Longitudinal Study of Adolescent Health
ARD	Average absolute Residual lifetime Distribution probability
ARH	Average absolute Residual Hazard probability
CF	"College then family" pathway
CW	"College and work" pathway
EM	Expectation-Maximum
EP	"Early parenthood" pathway
LL	Log-Likelihood
MAR	Missing At Random
MCAR	Missing Completely At Random
MEPSUM	Multiple Event Process Survival Mixture
NMUP	Non-medical use of prescription drugs
NSDUH	2009 National Survey on Drug Use and Health
W	"Work only" pathway
WF	"Work then family" pathway

LIST OF SYMBOLS

α	Intercept in model for prediction of the hazard
β	Effect of time-invariant predictor on the hazard
C	Categorical latent variable
D	Lifetime distribution function
δ	Dummy variable with a value of 1 for uncensored individuals, 0 for censored
f	Probability of experiencing an event at $T = j$
γ_0	Intercept in multinomial regression of class membership
γ	Effect of time-invariant predictor on class membership
h	Hazard function
i	Individual
I	Indicator
j	Discrete time point
J	Total number of time points
k	Specific latent class
K	Number of categories or classes of the latent categorical variable
κ	Effect of time-varying predictor on the hazard
L	Likelihood
m	Index for specific time period
M	Total number of parameters
N	Sample size
p	Event process
P	Total number of event processes

π	Proportion of individuals belonging to specific latent class
q	Index for specific time-invariant predictor
Q	Total number of time-invariant predictors
R	Total number of time-varying predictors
S	Survival function
t	Continuous time
T	Event time
θ	Latent variable representing unobserved heterogeneity
V	Total number of observed event history variables
w	Index for specific latent class
W	Sample weight
X	Value of time-invariant predictor
y	Event history response
Z	Value of time-varying predictor

CHAPTER 1

INTRODUCTION

Survival analysis is a useful tool for understanding both the occurrence and the timing of events. While survival analysis was originally developed to investigate the human lifetime, it is equally applicable to questions regarding the occurrence of any type of event, and there are numerous applications in the social and behavioral sciences. For example, clinical psychologists investigating the occurrence of affective illnesses or therapy termination benefit from the survival analysis framework (e.g. Corning & Malofeeva, 2004), as do developmental researchers who investigate the transition from one developmental stage to another (e.g. Ha, Kimpo, & Sackett, 1997), and researchers following students' entrance and exit from school (e.g. Bowers, 2010).

Event history data is rather unique in it aims to analyze both "if" and "when" an event occurs, yet there are often individuals who do not experience the event within the time frame of the study. Traditional linear and logistic regression techniques are not suited for this kind of missing data problem, termed censoring. For censored individuals, it is unknown when they will experience the event, or in some cases whether they will even experience the event at all. Survival analysis techniques were formulated to analyze this type of data (Singer & Willett, 2003; Lee & Wang, 2003). The basic statistical concepts of survival analysis depend on whether the time variable measuring the state of the event is assumed to be continuous or discrete. Continuous-time survival methods assume time can be measured exactly – thus there should be no “ties” in the dataset where two or more people have the same event time. While it may be logical to think of time as a continuous variable, this assumption is often unrealistic in practice. This is especially true for data collected in the social and behavioral sciences, as researchers frequently ask for the year or age of an event rather than the exact date. Also, events can sometimes only occur at discrete points in time (e.g. number of therapy sessions before dropout). In addition, discrete-time methods can be used to approximate the results of a continuous-time survival analysis (Vermunt, 1997), and are conceptually and computationally simpler. As such, the remainder of the paper assumes time is measured on a discrete scale.[1]

Moving beyond traditional survival analysis, researchers have recently begun to ask questions about the order and timing of multiple events. Multivariate survival models relax the standard requirement that all time variables are univariate and independent (see Hougaard, 2000). Recurrent event models, parallel data models, and competing risks models are three common multivariate survival tools. Recurrent event models are useful for examining the repeated occurrence of an event, such as the birth of a child, or the sequential occurrence of disparate events, such as children’s progress through stages of moral reasoning (e.g. Willett & Singer, 1995). Parallel data models have been proposed to examine the lifetimes of several individuals who are related in some way, such as a study of an event history of twins (e.g. Hougaard, Harvald, & Holm, 1992). Competing risk models account for the occurrence of mutually exclusive events; Ventura et al. (2000) used such a model to investigate the competing risk of psychotic exacerbation and

[1] For a comprehensive resource on continuous-time survival analysis methods, see Lee and Wang (2003).

depressive exacerbation following a major life event for individuals in the early course of schizophrenia.

While there has been great progress on the analysis of multivariate event history data through models such as those mentioned above, there is a demonstrated need for new analytic methods in investigating the order and timing of different non-repeatable events which are not mutually exclusive and do not necessarily occur in a sequential manner. Many researchers investigating several such events have resorted to completing a separate survival analysis for each event, and have not directly examined the interdependence of the events. For example, Schwartz et al. (2010) investigated how positive youth development influenced tobacco, alcohol, illicit drug, and sex initiation by conducting four separate survival analyses. Similarly, Scott et al. (2010) examined the influence of gender and marital status on the first onset of mood, anxiety, and substance use disorders by conducting several survival analyses. While analyzing each event separately can be useful, it gives no insight on how the events are related to each other.

In order to investigate the interdependencies between events, several dynamic survival analysis approaches have been proposed for a subset of survival methods concerning events measured in continuous-time; in these models, the rate of change in one process depends on the state of another process. One such approach, developed by Cox and Lewis (1972), cross classifies states on two or more events, yielding one process with multiple states, where the transition rates between the states are studied. Kalbfeisch and Prentice (1980) developed an approach similar to this, but the multiple events need not be cross classified, allowing for mutual dependencies to be more easily examined. In this method, one of the event processes is the dependent variable, and is predicted by other time-varying event processes.

Vermunt (1997, p.175) similarly suggested for multiple processes measured in discrete-time that researchers specify one of the events as the dependent variable and treat others as time-varying covariates.[2] However, researchers must rotate the dependent variable and run multiple models in order to investigate the reciprocal relationships. Malone et al. (2010) used a different approach for discrete-time data called dual-process discrete-time survival analysis, which expands on associative latent transition analysis (Bray, Lanza, & Collins, 2010). This approach models two time-to-event processes concurrently by linking the processes to each other, similar to a cross-lagged panel design. They used the model to test the gateway drug hypothesis by using a highly constrained latent transition matrix to model and test the cross-links between time to illicit drug use and time to licit drug use.

The discrete-time methods proposed above to investigate the interdependencies of multiple events require one of the event processes be specified as the dependent variable or are difficult to expand to more than two events. In addressing the need for a new model, this paper has two main objectives. The first objective of this paper is to introduce a discrete-time Multiple Event Process SUrvival Mixture (MEPSUM) model, a latent variable approach to analyzing the interdependencies between multiple non-repeatable events which are measured in discrete-time. The approach is mathematically similar to single-event discrete-time survival mixture analysis (Muthén, & Masyn, 2005), but is

[2] Often researchers use a time-lag for the independent variables to prevent reversed causation (Tuma & Hannan, 1984, p.268).

conceptually different in some ways and has several advantages in addition to incorporating multiple events.

The second objective of the paper is to demonstrate the usefulness of the model through two empirical analyses. The first empirical example concerns the timing and occurrence of four different markers of adulthood: parenthood, marriage, full-time work, and obtaining a college degree from individuals in Wave IV of the National Longitudinal Study of Adolescent Health (Add Health). The second empirical example, using data from the 2009 National Survey on Drug Use and Health (NSDUH), examines age of first use of numerous different substances such as alcohol, tobacco, marijuana, cocaine, and several other hard substances. Two examples are used both to demonstrate the applicability of the model to different domains and to examine the performance of the model when different numbers of events are examined, as well as when some of the events have a much lower probability of occurrence.

This paper is organized into five chapters. In the remaining part of Chapter 1, the basic concepts of traditional univariate discrete-time survival analysis will be introduced, as well as single-event survival mixture analysis. The fundamental concepts in these sections will be used in order to introduce the discrete-time multiple event process survival mixture model in Chapter 2. Chapter 2 also includes a small simulation study simply demonstrating the ability of the model to capture population parameters from data generated under the model for a small number of conditions. Chapter 3 regards the first empirical example – the analysis of transitions to adulthood, and Chapter 4 regards the second empirical example – the analysis of substance use onset. Chapter 5 has concluding remarks.

1.1 Traditional Univariate Discrete-time Survival Analysis

As a first step in discrete-time survival analysis, we must define several important concepts. A survival process under study encompasses different states, or categories of the event variable, and an event is defined as the transition from one state to another. Univariate survival analysis was developed for situations when the event under study can only occur once (e.g. death), or only the first event is examined (e.g. age of first marriage). This process is assumed to have only two states (event has not yet occurred; event has occurred). To survive past a certain time implies that the event under study has not occurred. The period that someone is at risk of an event is termed the risk period, and an individual is only at risk of an event if he or she has not yet experienced the event. Individuals who are able to experience the event at a certain point in time form the risk set. An event history analysis can then be defined as the analysis of the risk set in order to determine the probability of event occurrence during the risk period.

Another important concept in survival analysis is censoring, a general term referring to missing data in the analysis of event histories. An individual is censored if his or her event time is unknown, and a distinction can be made between whether this unknown event time is before or after the time period under study (left and right censoring, respectively). It is more common for an individual's unknown event time to be after the time period under study; this happens when the study concluded before the event occurred for the individual or the individual drops out from the study before the event occurred. In a retrospective study, this type of censoring occurs when an individual – who is younger at the time of interview than the last age examined in the study – has not yet experienced the event. For example, in studying the event of first marriage up to age 30, an individual who is 23 years old at the time of interview is censored for all time points representing ages 24 – 30. For the rest of this paper, an assumption will be made that all individuals are right-censored only, as right-censoring is the most common form of censoring in social and behavioral sciences.[3]

It will be assumed in this paper, as it generally is in survival analysis, that the censoring mechanism is noninformative. This corresponds to the assumption of ignorable missingness, including both missing-completely-at-random (MCAR) and missing-at-random (MAR) (Little & Rubin, 1987; Enders, 2010). If the censoring mechanism is independent of event times, the censored observations may be treated as MCAR. If the censoring mechanism is independent of event times, conditional on the set of observed covariates, the censored observations may be treated as MAR. The assumption of noninformative censoring is tenable if censoring is determined in advance by design. This is usually the case in event history studies, as the investigator determines the ending time of a study in a prospective study and the last age in a retrospective study. The assumption of noninformative censoring is important, for we can then assume all non-censored individuals at each time period are representative of all individuals who would have remained in the study if censoring had not occurred. This allows generalization to the entire data set and thus the original population.

To formalize univariate survival analysis, let T denote the event time, and j the discrete time point, with j =1, 2, …, J. There are many methods of characterizing the probability distribution of the event time. The simplest way is to define the probability of experiencing an event at a specific time period:

[3] See Yamaguichi (1991) and Vermunt (1997) among others for implications of left-censoring.

$$f_j = P(T = j) \tag{1}$$

Another option is the survival function, which is defined as the probability that an individual survives longer than j and is denoted S_j:

$$S_j = P(T > j) = 1 - \sum_{m=1}^{j} f_m \tag{2}$$

with $S_j = 1$ at $j = 0$. The survival function is often used to find descriptive measures of the event history, such as the median lifetime: an estimate of the time period when the event has occurred for fifty percent of the population. Such descriptive measures are important when there is censoring, as measures such as the sample mean will not be useful in describing the center of the distribution when the event time is not known for all individuals.

An equally useful function known as the lifetime distribution function defines the probability that an individual has experienced the event by time j:

$$D_j = P(T \le j) = 1 - S_j = \sum_{m=1}^{j} f_m \tag{3}$$

Importantly, the number of individuals who experienced the event at $T = j$ is unknown if there are censored individuals. Thus, neither the survival function nor lifetime distribution function can be directly estimated, as f_j is unknown.

The hazard probability h is the first function that can be estimated with both censored and uncensored individuals. It is the conditional probability that the event occurs at j given that it did not occur prior to j:

$$h_j = P(T = j \mid T \ge j) = P(T = j \mid T > j - 1) = \frac{P(T = j)}{P(T > j - 1)} \tag{4}$$

The hazard for time *j* is estimated as the number of events that occur at *j* over the number of individuals in the risk set. It thus tells us the unique risk of event occurrence for each time period among those eligible to experience the event, which is exactly what we want to know: whether and when events occur. It is estimable with censored individuals as it is a conditional probability computed only using individuals in the risk set, and can be computed for every time period when event occurrence is recorded.

It is important to note that the hazard function can be re-written in terms of f_j and S_j, under the assumption of noninformative censoring:

$$h_j = \frac{f_j}{S_{j-1}} = 1 - \frac{S_j}{S_{j-1}} \tag{5}$$

This relationship is useful in obtaining an estimate of the survival function when there are censored individuals, as Equation (4) can be rearranged to show:

$$S_j = \left[S_{j-1}\right]\left[1 - h_j\right] \tag{6}$$

Given this relationship and the fact the survival function is equal to one at $j = 0$ (no individual experienced an event before the beginning of the time variable) this leads to the idea that the survival probability at time period *j* is the product of the hazard probabilities for each of the earlier time points:

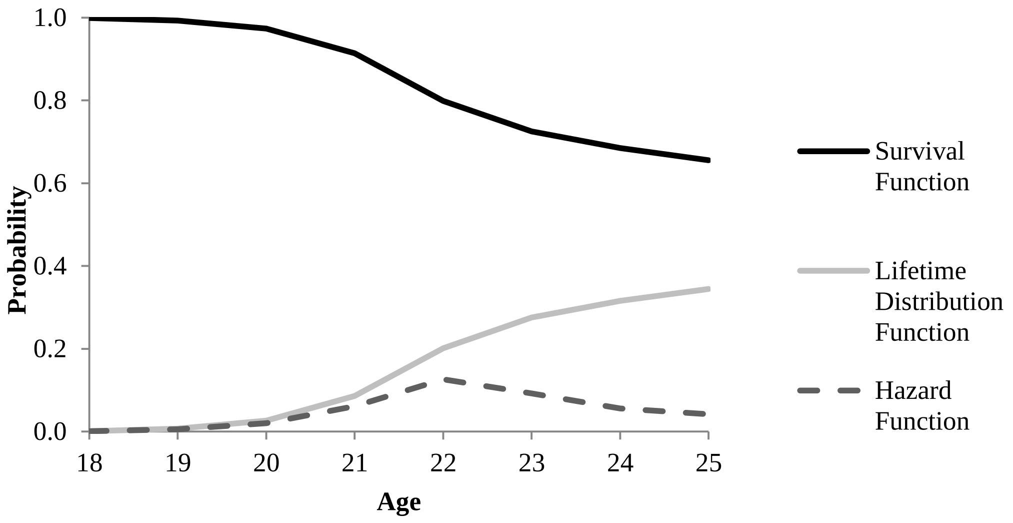
$$S_j = \prod_{m=1}^{j} \left(1 - h_m\right) \tag{7}$$

The lifetime distribution function can similarly be estimated indirectly from the hazard probabilities, or by the simple relationship between D_j and S_j given in Equation (3).

See Figure 1 for a graphic example of the relationship between the different survival analysis functions. It displays the survival function, lifetime distribution function, and the hazard function estimated from the National Longitudinal Study of

Adolescent Health Wave IV data on the age of college degree. In this case, the hazard represents the probability of obtaining a college degree at each age given a college degree had not yet been obtained. We can thus identify that the event is most likely to occur, given it had not occurred at an earlier age, at age 22 for the individuals in this sample. The survival function and lifetime distribution function were indirectly estimated using the hazard function estimates. The survival function estimates the proportion of individuals without a college degree at each age, and the lifetime distribution function estimates the proportion of individuals with a college degree at each age. Note that the survival and lifetime distribution function change more rapidly in periods when the hazard is high, and more slowly in periods when the hazard is low.

Figure 1: *Estimated discrete-time survival analysis functions for Add Health data on age of college degree*



Now that the probability distribution of the duration of an event occurrence or nonoccurrence has been defined, the next objective of a survival analysis is to investigate how covariates affect the event times. This is achieved by modeling the probability distribution and adding covariates to the model to examine their influence. As the hazard function is the most useful way to describe event history data, given it is estimable even with censored individuals and reveals the risk of event occurrence at each time period, it is used as the dependent variable in a survival analysis model. As hazards are conditional probabilities bounded by 0 and 1, the hazard is often transformed so it can be easily regressed on covariates and time variables; such a transformation prevents inadmissible predicted values. In line with Singer and Willett (1993), a logit link function will be used for the remainder of the paper, but other link functions such as the complementary log-log link are equally applicable to all of the survival methods discussed hereafter. The unstructured hazard function at time *j* without covariates is then given by:

$$logit(h_j) = ln\left(\frac{h_j}{1-h_j}\right) = \alpha_j \tag{8}$$

where α_j is the intercept parameter for time *j*. This model represents the log-odds of event occurrence as a function of the time period only.

There are almost countless ways to expand on the simple unstructured discrete-time hazard model discussed here (e.g. Singer and Willett, 2003). For example, instead of allowing an intercept for each time period which places no constraints on the shape of the hazard, it is possible to have a polynomial representation of time. When the number of time periods is large or some time periods have very small risk sets, it can be advantageous to fit a more parsimonious model. A structured hazard can also be advantageous for estimation purposes when the hazard is near 0 is some time periods, as this can result in convergence problems. A constant hazard function results from restricting the intercept in the link function to be constant over time. Without covariates, this is given by:

$$logit(h_j) = \alpha \qquad (9)$$

An expanded polynomial representation of the hazard function is also common. For example, a quadratic hazard function is given by:

$$logit(h_j) = \alpha_0 + \alpha_1 Time_j + \alpha_2 Time_j^2 \qquad (10)$$

For simplicity purposes, the remainder of the chapter will focus on the unstructured hazard with a logit link function, but the equations that follow can be easily generalized to alternative functions as mentioned above.

Time-invariant predictors (X_1, X_2, …, X_Q) for person i (i = 1, 2, …, n) are often added to the model in Equation (8) in such a way that each parameter β_Q represents a shift in the baseline logit hazard function for a one unit increase in the value of the predictor X_Q, controlling for the effects of all other predictors in the model. The model for the log-odds of event occurrence for person i in time period j as a function of the predictor values represented by the Q × 1 vector $\mathbf{X}_\mathrm{i}$ (X_1, X_2, …, X_Q)' is:

$$logit(h_{ij}) = ln\left(\frac{h_{ij}}{1-h_{ij}}\right) = \alpha_j + \boldsymbol{\beta}'\mathbf{X}_i \qquad (11)$$

The model can be rewritten using the exponential function to be in terms of the odds of event occurrence:

$$\frac{h_{ij}}{1-h_{ij}} = exp\left(\alpha_j + \boldsymbol{\beta}'\mathbf{X}_i\right) = exp\left(\alpha_j\right) exp\left(\boldsymbol{\beta}'\mathbf{X}_i\right) \qquad (12)$$

This reformulation reveals that the model invokes a proportional odds assumption, in that the effect of each predictor is postulated to be the same for each time period, and that a one unit increase in X_Q increases the odds of event occurrence exp(β_Q) times, compared to subjects in the baseline group (i.e. X_1, X_2, …, X_Q = 0). The proportional odds assumption can be relaxed by allowing the predictor to have time-varying effects:

$$logit(h_{ij}) = \alpha_j + \boldsymbol{\beta}'_j \mathbf{X}_i \tag{13}$$

Additionally, time-varying predictors represented by the R × 1 vector $\boldsymbol{Z}_{ij}$ may be added to the model:

$$logit(h_{ij}) = \alpha_j + \boldsymbol{\beta}'_j \mathbf{X}_i + \boldsymbol{\kappa}'_j \mathbf{Z}_{ij} \tag{14}$$

The model is a simple variant of logistic regression, varying over time period:

$$h_{ij} = \frac{exp(\alpha_j + \boldsymbol{\beta}'_j \mathbf{X}_i + \boldsymbol{\kappa}'_j \mathbf{Z}_{ij})}{1 + exp(\alpha_j + \boldsymbol{\beta}'_j \mathbf{X}_i + \boldsymbol{\kappa}'_j \mathbf{Z}_{ij})} \tag{15}$$

In aiming to understand how this model is fit to data, let y_{ij} represent the event history response for person i at time period j, where y_{ij} = 1 if the event occurred for person i at time period j and y_{ij} = 0 if the event had not yet occurred for person i at time period j. Due to the conditional nature of the hazard function, individuals only contribute data at time period j if they experienced the event at that time period or they had not yet experienced the event by that time period. Individuals also do not contribute data if they are censored, under the assumption of ignorable missingness as discussed earlier. Therefore, the number of time periods can vary across individuals (j = 1, 2, …, J_i where J_i is the time period with the last non-missing value for individual i). For individual i who is uncensored (thus y_{ij} = 1 at $j = J_i$), the probability of the entire event history response pattern represented by the vector $\mathbf{y}_\mathrm{i}$ (y_{i1}, y_{i2}, …, y_{iJi})' is given by:

$$P(\mathbf{y}_i \mid T = J_i) = h_{iJ_i} \prod_{j=1}^{J_i - 1} \left(1 - h_{ij}\right) \tag{16}$$

For individuals who are censored, the probability of the response pattern $\mathbf{y}_\mathrm{i}$ is given by:

$$P(\mathbf{y}_i \mid T > J_i) = \prod_{j=1}^{J_i} \left(1 - h_{ij}\right) \tag{17}$$

The likelihood may be written as:

$$L = \prod_{i=1}^{n} \left(P(\mathbf{y}_i \mid T = J_i)^{\delta_i} \, P(\mathbf{y}_i \mid T > J_i)^{1-\delta_i} \right) \tag{18}$$

where δ_i is a dummy variable with a value of one if the individual is uncensored and zero if censored, thus serving as a device for selecting the appropriate probability by which to multiply. Substituting Equation (16) and Equation (17) into Equation (18), the likelihood is:

$$L = \prod_{i=1}^{n} \left[\left(h_{iJ_i} \prod_{j=1}^{J_i - 1} \left(1 - h_{ij}\right) \right)^{\delta_i} \left(\prod_{j=1}^{J_i} \left(1 - h_{ij}\right) \right)^{1-\delta_i} \right] \tag{19}$$

which is used to find optimal parameter estimates.

Allison (1982) and Singer and Willett (1993) note that the probability of the event history response pattern can be rewritten using the event history response variable y_{ij}, which serves a similar function as the indicator variable δ_i in Equation (19) in it selects the appropriate probability by which to multiply:

$$P(\mathbf{y}_i) = \prod_{j=1}^{J_i} \left(h_{ij}{}^{y_{ij}} (1 - h_{ij})^{(1-y_{ij})} \right) \tag{20}$$

The likelihood is then:

$$L = \prod_{i=1}^{n} \prod_{j=1}^{J_i} \left(h_{ij}{}^{y_{ij}} (1 - h_{ij})^{(1-y_{ij})} \right) \tag{21}$$

For all time periods before event occurrence (y_{ij} = 0), the function multiplies by 1 - h_{ij}, and for the time period when the event occurs (y_{ij} = 1), the function multiplies by h_{ij}.

Censored individuals only contribute to the likelihood through the (1 - h_{ij}) terms, as they do not experience the event within the time frame under study.

As noted by Singer and Willett (1993) and Allison (1982), the likelihood function in Equation (21) is identical to the likelihood function for a sequence of V ($V = J_1 + J_2 + \ldots + J_n$) independent Bernoulli trials with parameters h_{ij}. As such, we can treat the V dichotomous observed variables y_{ij} as a collection of independent observations with a hypothesized logistic relation with covariates. In other words, the event history response for an individual at each discrete time period can be treated as a separate, independent observation. This allows estimation via standard logistic regression procedures (e.g. Allison, 1999).

1.2 Univariate Discrete-time Survival Mixture Analysis

All survival analysis models impose an assumption that there is no unobserved heterogeneity. Vaupel and Yaskin (1985) famously demonstrate the potential impact of unobserved heterogeneity: a hazard function may seem to follow a specific form when in fact it does not. The problem occurs when there are individuals with different levels of risk for an event – for individuals who are at highest risk of an event tend to experience the event first. This phenomenon can produce event patterns for a population that are very different than for subpopulations of that population, such as those at high risk of the event.

For example, suppose a researcher investigated the onset of depression and genetic risk factors were not introduced into the model. Suppose there were two subpopulations of individuals – one with a high genetic risk of depression and one without – and within each subpopulation the risk of depression was constant over time.

The population hazard model in this case would not be constant; it would in fact decline over time. This decline in the hazard could be correctly interpreted as the population average trend (Xue & Brookmeyer, 1997). However, the decline in the hazard in this case would not reflect that the risk of depression for an individual decreases over time, as the decline in the hazard function only represents the changing composition of the risk set.

Many researchers have proposed adding a latent variable to the hazard model to account for unobserved heterogeneity in the continuous-time framework. Vaupel, Manton, and Stallard (1979) proposed including a continuous latent variable to account for the unobserved heterogeneity which they called "frailty." The continuous latent variable, or random effect, was assumed to have a multiplicative and proportional effect on the hazard rate. The hazard rate as a function of continuous time represented by *t* is given by:

$$h_{t|\theta} = h_t \theta \tag{22}$$

and θ was assumed to have a particular distributional form. A gamma distribution with mean of 1 and variance of $1/\gamma$ was proposed by some (Vauepl, Manton, & Stallard, 1979; Tuma & Hannan, 1984), but many other distributional forms have been proposed (see Hougaard, 2000).

Instead of a parametric characterization of θ, Heckman and Singer (1982, 1984) proposed a non-parametric heterogeneity model for continuous-time; their model is equivalent to a latent class model where the population is assumed to be composed of a finite number of mutually exclusive and exhaustive groups (Goodman, 1974). With a categorical latent variable *C* composed of *K* (*k* = 1, 2, …, *K*) categories, the marginal hazard rate at time point *t* can be defined as:

$$h_t = \sum_{k=1}^{K} \pi_{tk} h_{tk} \theta_k \tag{23}$$

where θ_k is the mean multiplicative effect on the hazard rate for latent class *k* and π_{tk} is the proportion of the population belonging to that class at time *t* (Vermunt, 1997). Heckman and Singer (1982) define the number of classes as the "number of mass points," signifying the use of latent classes as a mathematical device for capturing unobserved heterogeneity.

Non-parametric unobserved heterogeneity models have also been proposed for discrete-time. Land, Nagin, and McCall (2001) introduced a multilevel model which incorporates nonparametric specifications of unobserved heterogeneity through a piecewise-constant hazard function modeled with the Poisson distribution. Their model is the discrete-time equivalent to the model proposed by Heckman and Singer (1982, 1984) with the added ability to account for clustering of data. Another mixture model for discrete-time data was proposed by Steele (2003) which accounts for "long-term survivors": those who are known a priori to have a zero hazard throughout the study. The model is in a sense a two class model, long-term survivors or not, and uses information on covariates to determine the probability that censored individuals are members of the long-term survivor class.

More general frameworks for accounting for unobserved heterogeneity in discrete-time have been proposed. Vermunt (1997) presents a general model for discrete-time survival analysis with latent variables in a log-linear framework. The framework is also useful for many multivariate survival analyses, such as repeated measures or related observations, as it can incorporate multiple correlated latent variables. Muthén and Masyn (2005) also present a general framework for modeling a single event in discrete-time, where latent classes of individuals have different hazard and thus survival functions. As part of the general framework, they consider a generic multiple class model, a long-term-survivor model with two classes, and a multiple class model which combines the hazard model with a growth mixture model.

Let us now consider a simplified version of the general multiple class discrete-time survival model proposed by Muthén and Masyn (2005). For person *i* at discrete time period *j* belonging to class *k*, the probability of event occurrence is given by the hazard model:

$$logit(h_{ijk}) = \alpha_{jk} + \boldsymbol{\beta}'_{jk}\mathbf{X}_i + \boldsymbol{\kappa}'_{jk}\mathbf{Z}_{ij} \qquad (24)$$

where α_{jk} represents the intercept for time period *j* in class *k* or the log odds of event occurrence in class *k* for an individual with all predictor values equal to 0; $\boldsymbol{\kappa}_{jk}$ represents a $R \times 1$ logit parameter vector for the effects of the time-varying covariates $\mathbf{Z}_{ij}$; and $\boldsymbol{\beta}_{jk}$ represents a $Q \times 1$ logit parameter vector for the effects of the time-invariant covariates $\mathbf{X}_\mathrm{i}$ that may also vary across the time periods. If the *j* subscript for time is removed for the effects of the time-invariant predictors represented by the vector $\boldsymbol{\beta}_{jk}$, a proportional odds assumption is imposed, as discussed in the previous section.

The probability of the event history response pattern represented by the vector $\mathbf{y}_i$ for person *i* within latent class *k* is the same as in a traditional discrete-time model – see Equation (20) – only with a *k* subscript to note the hazard function is conditional on latent class:

$$P(\mathbf{y}_i \mid C_i = k) = \prod_{j=1}^{Ji} \left(h_{ijk}{}^{y_{ij}} (1 - h_{ijk})^{(1-y_{ij})} \right) \tag{25}$$

The prediction of class membership – by time-invariant covariates only – is added through a general multinomial logistic regression model; the probability of person *i* belonging to class *k* is given by:

$$\pi_{ik} = \frac{\exp(\gamma_{0k} + \boldsymbol{\gamma}'_k \mathbf{X}_i)}{\sum_{w=1}^{K} \exp(\gamma_{0w} + \boldsymbol{\gamma}'_w \mathbf{X}_i)} \tag{26}$$

where the last class is a reference class with $\gamma_{0k} = 0$ and $\boldsymbol{\gamma}'_k = 0$. As with all latent class models, the probability of a specific individual's response pattern is a weighted function of the probability of class membership given by Equation (26) and the probability of the specific event history response pattern given class membership (Equation (25)):

$$P(\mathbf{y}_i) = \sum_{k=1}^{K} \left(\pi_{ik} \prod_{j=1}^{Ji} \left(h_{ijk}{}^{y_{ij}} (1 - h_{ijk})^{(1-y_{ij})} \right) \right) \tag{27}$$

where $\sum_{k=1}^{K} \pi_{ik} = 1$. The likelihood is then:

$$L = \prod_{i=1}^{n} \left[\sum_{k=1}^{K} \left(\pi_{ik} \prod_{j=1}^{Ji} \left(h_{ijk}{}^{y_{ij}} (1 - h_{ijk})^{(1-y_{ij})} \right) \right) \right] \tag{28}$$

under the assumption of non-informative censoring. When there is a single latent class (K = 1 and π_{ik} = 1), the likelihood simplifies to the univariate survival analysis likelihood in Equation (21). As noted in the previous section, the dichotomous observed values y_{ij} can be treated as independent observations due to the equivalence of the likelihood function for a hazard model to independent Bernoulli trials with parameters h_{ijk}.

The model in Equation (28) is a special version of a latent class model – or what is sometimes referred to as latent class regression (Formann, 1992) – where the prevalence of the latent classes and the hazard probabilities are parameters to be estimated. Important to note, traditional latent class models aim to account for the dependence between the observed variables through the addition of latent classes. In this case, however, the observed variables y_{ij} are by definition independent. As the observed variables y_{ij} are independent, a multiple class discrete-time survival model with unstructured hazard probabilities is not identifiable without covariates. In other words, latent class models add classes in order to satisfy the condition of local independence – that conditional on latent class, the observed variables are independent. But, in the absence of covariates, more than one class is unnecessary as the y_{ij} for the survival model are already independent.

For example, in the two class long-term survivor model, a distinction between long-term survivors versus those who are at risk of the event can only be made based on covariate values (Land, Nagin, & McCall, 2001; Muthén & Masyn, 2005). As different covariates may produce nontrivial differences in the formation of the latent classes, results may be highly dependent upon the set of covariates that are included. This is clearly an undesirable feature of the univariate mixture survival model, but fortunately one that will not be shared by its multivariate extension, as will be discussed in Chapter 2.

1.3 Summary

As was discussed in this chapter, traditional univariate survival analysis provides an important conceptual and analytic framework from which to evaluate if and when events occur. Extensions to the basic model, which are not discussed at length in this paper, include accounting for competing events and recurrent events. One recent extension that provides an important foundation for the model that will be introduced in the next chapter is the finite mixture survival model that was discussed in Section 2 of this chapter. While the model was originally motivated from the desire to account for unobserved heterogeneity, in extending this model to multiple events the latent classes will also serve to capture the interdependencies across multiple event processes. This extension to the univariate survival mixture model will be termed a multiple event process survival mixture model and will be introduced in the next chapter.

CHAPTER 2

A DISCRETE-TIME MULTIPLE PROCESS SURVIVAL MIXTURE MODEL

The discrete-time Multiple Event Process SUrvival Mixture (MEPSUM) model is a finite mixture model, specifically a special type of latent class model designed to accommodate data on the occurrence of multiple non-repeatable events. The model assumes that the population is composed of a finite number of subpopulations of individuals who are homogeneous with respect to the risk of multiple events over time. The latent classes obtained through the MEPSUM model are a convenient statistical devise for parsimoniously describing the underlying multivariate distribution of hazard functions. In other words, the model is a non-parametric way to capture associations between events through identification of classes of individuals with similar risk, or hazard, for multiple events over time. The model is easily expanded beyond two events and enables researchers who aim to analyze multiple events to utilize all individuals in their dataset, including those with censored event times.

Substantively, the model allows researchers to understand both the order and timing of the events through examination of the hazard functions both within each latent class and across latent classes. Additionally, both the survival function and lifetime distribution function for each event can be compared between each class and across latent classes, as these functions may be estimated indirectly from the fitted hazard functions through Equation (3) and Equation (7). Predictors can be incorporated into the model in several different ways to investigate potential influences on the risk for multiple events over time.

In the sections that follow, the model is formally defined (Section 2.1) and software for fitting the model (Section 2.2) as well identification of the model is then discussed (Section 2.3). A small simulation study is used to investigate the performance of the model under different conditions (Section 2.4) and a proposed model building approach is outlined (Section 2.5).

2.1 Model Definition

To reiterate points made in Chapter 1, the model focuses on discrete-time survival data on non-repeatable events. Suppose the event history variable y_{ipj} for person i represents whether an event of type p (p = 1, 2, …, P) occurs at time period j (j = 1, 2, …, J_{ip}) and the response vector $\mathbf{y}_{\mathrm{i}}$ holds the event history variable across all time periods and processes $\left[\left(y_{i11},...,y_{i1J_{i1}}\right),\left(y_{i21},...,y_{i2J_{i2}}\right),...,\left(y_{iP1},...,y_{iPJ_{iP}}\right)\right]'$. The total number of time points under study for event process p is represented by J_p. Note the flexibility of the model in that the number of time periods studied can vary between processes, the width of the time periods can vary within processes, and the length of the vector can vary between individuals.

Let y_{ipj} = 0 if the event for process p did not occur for individual i at that time period or earlier and y_{ipj} = 1 if the event occurred at that time period. By framing the data in this way, individuals only contribute data at j for process p when they are in the risk set at j for process p, similar to a standard univariate survival analysis. For example, consider two event processes (e.g. onset of depression and onset of an anxiety disorder), which are both measured at each age from 10 years old to 14 years old. An individual who responds at age 15 with no history of either disorder would have the event history (0 0 0 0 0) for each process. In contrast, consider an individual who is measured at age 13 who was diagnosed with an anxiety disorder at age 11. The event history for depression would only include data from ages 10 to 13 (0 0 0 0), and the event history for anxiety would only include data from ages 10 to 11 (0 1). Censored data is ignored under the assumption of ignorable missingness, as discussed in Chapter 1.

The probability of event occurrence ($y_{ipj} = 1$) for event process p in time period j within latent class k is represented by h_{pjk}. Within latent class k, h_{pjk} is modeled using a simple unstructured discrete-time hazard function with time-specific intercept α_{pjk}:

$$logit(h_{pjk}) = \alpha_{pjk} \tag{29}$$

As in Equation (24), both time-invariant and time-varying covariates may be added to the model, such that covariates can have a direct effect on the hazard functions:

$$logit(h_{pijk}) = \alpha_{pjk} + \boldsymbol{\beta}'_{pjk}\mathbf{X}_i + \boldsymbol{\kappa}'_{pjk}\mathbf{Z}_{ij} \tag{30}$$

Restrictions may be placed on the influence of the covariates for parsimony. For example, by dropping the j subscript on $\boldsymbol{\beta}_{pjk}$, a proportional odds assumption is invoked for the time-invariant predictors, and by dropping the k subscript for $\boldsymbol{\kappa}_{pjk}$ and/or $\boldsymbol{\beta}_{pjk}$ the influence of the covariates can be restricted to be the same across classes. It is also possible to structure the hazard function. For example, a quadratic hazard function may be imposed:

$$logit(h_{pijk}) = \alpha_{0pk} + \alpha_{1pk}Time_j + \alpha_{2pk}Time_j^2 + \boldsymbol{\beta}'_{pjk}\mathbf{X}_i + \boldsymbol{\kappa}'_{pjk}\mathbf{Z}_{ij} \tag{31}$$

However, caution is needed before imposing such a structure – even after examining the shape of the sample estimated hazard function – as it is possible that the hazard function has a certain shape across latent classes but a different shape within latent classes. This issue will be discussed further in a recommended model building approach outlined in Section 2.5.

The model assumes that the hazard functions across event processes are associated because the population is comprised of a finite number of subpopulations, where individuals have common hazard functions within latent class. The model assumes that all associations between the hazard functions are captured between the latent classes, so that the observed hazard indicators are independent within latent class. This implies the probability of a specific response vector within a given latent class *k* can be obtained by simply multiplying the probability of all of the responses:

$$P(\mathbf{y}_i \mid C_i = k) = \prod_{p=1}^{P} \prod_{j=1}^{Ji} \left(h_{pijk}{}^{y_{ipj}} (1 - h_{pijk})^{(1-y_{ipj})} \right) \tag{32}$$

Similar to the y_{ij} variable in Equation (20) and Equation (25), the indicator variable y_{ipj} simply functions as a device for selecting the appropriate probability by which to multiply. When the event occurs (y_{ipj} = 1) for process *p* at time period *j*, the model multiplies by h_{pijk}, versus event nonoccurrence for process *p* at time period *j* when the model multiplies by (1 - h_{pijk}).

The overall probability of response pattern $\mathbf{y}_\mathrm{i}$ is a weighted average across all of the latent classes of the probability of being in latent class *k* given by π_{ik} and probability of $\mathbf{y}_\mathrm{i}$ given latent class *k* as defined in Equation (32):

$$P(\mathbf{y}_i) = \sum_{k=1}^{K} \pi_{ik} P(\mathbf{y}_i \mid C_i = k) \tag{33}$$

where π_{ik} is modeled using standard multinomial logistic regression. With time-invariant predictors $\mathbf{X}_\mathrm{i}$, this is given by:

$$\pi_{ik} = \frac{\exp(\gamma_{0k} + \boldsymbol{\gamma}'_k \mathbf{X}_i)}{\sum_{w=1}^{K} \exp(\gamma_{0w} + \boldsymbol{\gamma}'_w \mathbf{X}_i)} \tag{34}$$

where the last class is a reference class with $\gamma_{0k} = 0$ and $\boldsymbol{\gamma}'_k = 0$, and $\sum_{k=1}^{K} \pi_{ik} = 1$. This leaves us with the final equation for the probability of an event history response vector:

$$P(\mathbf{y}_i) = \sum_{k=1}^{K} \left(\pi_{ik} \prod_{p=1}^{P} \prod_{j=1}^{Ji} \left(h_{pijk}{}^{y_{ipj}} (1 - h_{pijk})^{(1-y_{ipj})} \right) \right) \tag{35}$$

and the likelihood function:

$$L = \prod_{i=1}^{n} \left[\sum_{k=1}^{K} \left(\pi_{ik} \prod_{p=1}^{P} \prod_{j=1}^{Ji} \left(h_{pijk}{}^{y_{ipj}} (1 - h_{pijk})^{(1-y_{ipj})} \right) \right) \right] \tag{36}$$

which is used to find optimal parameter estimates. In large sample surveys, individuals are often drawn with unequal selection probabilities and the contribution of individual *i* may be weighted by a sample weight *W*, which is often computed as the inverse probability of selection into the sample or through a function that also takes other features of the survey into account (Kish, 1965; Lohr, 2009). The likelihood in this case is given by:

$$L = \prod_{i=1}^{n} W_i \left[\sum_{k=1}^{K} \left(\pi_{ik} \prod_{p=1}^{P} \prod_{j=1}^{Ji} \left(h_{pijk}{}^{y_{ipj}} (1 - h_{pijk})^{(1-y_{ipj})} \right) \right) \right] \tag{37}$$

where the sample weight effectively serves as a frequency weight, representing the number of times that each person’s individual likelihood should be replicated.

2.2 Software

The model may be fit using latent variable modeling software such as Mplus (Muthén & Muthén, 1998-2010) or Latent Gold (Vermunt & Magidson, 2005), which obtain maximum-likelihood model parameter estimates using an Expectation-

Maximization (EM) algorithm. In this algorithm, class membership is considered missing, and individuals' posterior probabilities of class membership are computed in the E step, given the parameter estimates. Then estimates of model parameters are updated given the posterior probabilities of class membership in the M step. While the basic algorithm implemented in the programs is the same, these two programs differ in several ways, including how they address an issue that commonly arises with modeling the probability of a binary outcome with a logit link: the logit is undefined if the probability is exactly zero or one. This could occur in time periods where there is no risk of event occurrence. To address this issue, Mplus implements default bounds on the logits of ±15, while Latent Gold utilizes a Bayesian approach in including a Dirichlet prior for the latent and conditional response probabilities that serves to smooth parameter values away from the boundary solution.[4] No matter what software program is selected, researchers should remain cognizant of the methods employed by the program to address this issue. Related to fitting the model, it should be noted that mixture models in general are susceptible to converge at local rather than global maxima. Multiple starting values should be used, and the convergence pattern should be monitored (McLachlan & Peel, 2000; Hipp & Bauer, 2006).

2.3 Model Identification

Identification of latent class models rests on the fulfillment of two conditions (Abar & Loken, 2012). First, the data must provide more unique pieces of information than parameters in the model; in other words, it is necessary for the model to have positive degrees of freedom. Second, the probability distributions for the possible

[4] By implementing such a prior, the estimation method is not truly maximum-likelihood estimation but instead posterior mode estimation, which can be seen as a penalized form of maximum-likelihood.

response patterns must be linearly independent, so that the information matrix of the model parameters is positive definite (McHugh, 1956). In a traditional latent class analysis with K latent classes with I indicators, there are 2^I possible response patterns, and $(K - 1) + K \cdot I$ parameters. The degrees of freedom is then $2^I - ((K - 1) + K \cdot I) - 1$ which is the number of response patterns minus the parameters minus one for the restriction that the frequency counts across the response patterns must sum to the sample size.

Both conditions mentioned above are necessary for identification. For example, Goodman (1974) demonstrated that a three class model with four indicators has 1 degree of freedom, yet is not identifiable due to a non-positive definite information matrix. However, it is possible to identify such a model by imposing constraints on the parameters which limit the number of estimated parameters. Covariates included in the model to predict class membership can also influence identification of the model, and adding a continuous covariate has been shown to improve estimation and recovery of the parameters as long as the covariate has some degree of predictive validity (Abar & Loken, 2012).

While confirming the degrees of freedom of a model is relatively straightforward, establishing that the information matrix is invertible is more difficult. Unfortunately, researchers cannot necessary rely on warning messages from standard software packages. In fact, model estimation can proceed normally in standard software, with boundary estimates effectively serving as a priori constraints which would have been necessary in order to identify the model (Abar & Loken, 2012). This complicates the establishment of identification in practice. For example, a five class, five indicator latent class model has seven degrees of freedom, and some have claimed that this model is identified (Magidson & Vermunt, 2001) while others have said it is not (Formann, 2003). Abar and Loken (2012) note that it is possible that this model is only identified when the true population estimates are on the boundary of the parameter space (i.e. probability of 0 or 1). Yet when the boundary solutions do not represent true population parameters, the model will result in more classification errors than would otherwise be expected (Abar & Loken, 2012).

As the MEPSUM model is a special type of latent class model, the conditions described above which are necessary for identification of a latent class model are also necessary for the MEPSUM model. However, general rules of thumb – such that identification should be questioned when the number of classes is equal to or greater than the number of indicators – are not applicable, due to the structured missingness that results from the unique nature of the event history response variables, which will be discussed further below.

As mentioned in Chapter 1, the unstructured hazard MEPSUM model with multiple latent classes is not identified in the absence of covariates when simplified to the situation where only one event process is studied (Muthén & Masyn, 2005). Considering this in more detail, suppose there are J time periods under study. There are $J + 1$ possible response patterns: one for each of the time periods, plus one for individuals who did not experience the event within any of the time periods. Note that there are less possible response patterns than a standard latent class analysis (which would have 2^J possible response patterns) due to the conditional nature of the data; once an individual experiences the event, they are no longer eligible to experience the event again. When using unstructured hazard functions, the number of parameters is however the same as a standard latent class analysis, and is equal to (K - 1) + $K \cdot J$. The degrees of freedom are then equal to (J + 1) - ((K - 1) + $K \cdot J$) - 1. For a one class MEPSUM model for one event, there would be 0 degrees of freedom and the model is just identified. For multiple latent classes, the model is not identified for only one event process without covariates.

The above generalizes to the degrees of freedom for a MEPSUM model with multiple events with unstructured hazard functions. With J_P time periods for event process p, there are $\prod_{p=1}^{P}\left(J_p+1\right)$ possible response patterns and the number of parameters is equal to $(K-1)+K\sum_{p=1}^{P}J_p$. The degrees of freedom is then $\prod_{p=1}^{P}\left(J_p+1\right)-\left((K-1)+K\sum_{p=1}^{P}J_p\right)-1$. In the situation where each event process is studied for the same number of time periods, this simplifies to $\left(J+1\right)^P-\left(\left(K-1\right)+K\cdot P\cdot J\right)-1$. For example, a two class MEPSUM model for three events each measured over three time periods would result in 64 possible response patterns, 19 parameters, and 44 degrees of freedom. Thus, unlike the survival mixture model for one event, the MEPSUM model for multiple events can have positive degrees of freedom for multiple classes, even with unstructured hazard functions and in the absence of covariates. As can be seen above, this is due to the fact that with multiple event processes, the observed variables are still independent within event process, but are not independent across processes (resulting in more unique pieces of information than parameters). The latent variable is thus able to capture independencies between the hazard functions of the different process through the addition of latent classes.

Not all possible combinations of the MEPSUM model with multiple events will have positive degrees of freedom. For example, a two class model with unstructured hazard functions for two events, each measured over two time periods will have 9 possible response patterns, 9 parameters, and negative degrees of freedom (-1). Additionally, as discussed earlier, positive degrees of freedom does not ensure identification. It has been my experience that the MEPSUM model is not identified when only two events are studied with unstructured hazard functions, even when the implied degrees of freedom is positive. This could be the result of a non-positive definite information matrix, or due to very near zero correlations between event history indicators across events, resulting in less information than that which is implied through calculation of the number of possible response patterns. This could also occur because the number of actual observed response patterns is much smaller than the number of possible observed response patterns. Small correlations between event history indicators across processes can also result in an information matrix that is so empirically near non-positive definite that the software fails to reach a solution or results in boundary estimates. Researchers should carefully monitor the estimation process and parameter values that are output, and start values may assist in the convergence process.

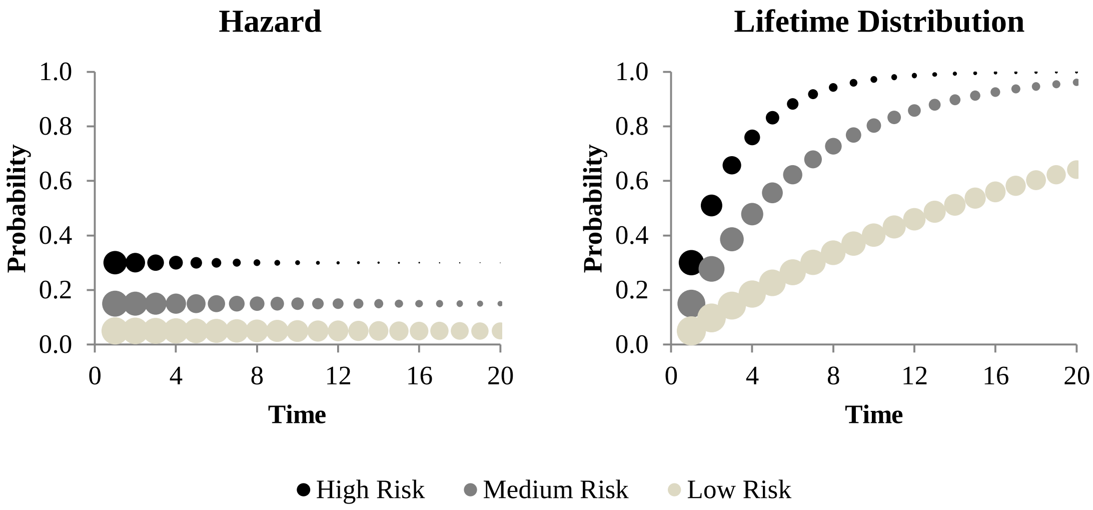
2.4 Simulation Demonstration

2.4.1 Introduction

The goal of this simulation is simply to demonstrate that the model can recover characteristics of data that are generated under the assumption that the population is truly comprised of a certain and finite number of latent classes. While latent classes are characterized by different hazard functions, individuals within a given class are assumed to have the same hazard functions. While the conditions of the simulation that follow are arguably overly simplistic, the purpose is to have a contained demonstration rather than a thorough investigation of all aspects of the model and model building approach. To this end, the number of latent classes in the population is held constant at three and only the three class model with unstructured hazard functions is fit in the simulation. The proportion of individuals within each latent class is also held constant, with each latent class size equal (π_k = 0.333). The shape and level of the hazards within latent class is held constant. Last, the sample size is held constant at 10,000. While this sample size is much larger than is typical in psychology, it is actually smaller than both sample sizes in the two empirical examples that follow.

The simulation is a 2 x 3 x 2 design, for a total of 12 conditions. First, the number of events is varied to be either 4 or 8; and second, the number of time periods is varied to be 5, 10, or 20. The variations in number of events and number of time periods are both similar to conditions in the two empirical examples that follow. Last, the class separation is varied. In the first class separation condition, which is labeled "good," the first class has a relatively high risk for all events over all time periods, defined as a constant risk of 0.30. The second class has a moderate risk for all events over all time periods, defined as a constant risk of 0.15, and the last class has a smaller risk for all events over all time periods, defined as a constant risk of 0.05. In the second "class separation" condition, which is labeled "poor," the first class and second classes are the same as above, with a constant risk of 0.30 and 0.15, respectively. The third class is defined to have half of the events with a high risk (0.30) and half of the events with a moderate risk (0.15), and again for simplicity this risk remains constant over time. Note that the hazard functions are held constant, and that the lifetime distribution functions are nonlinear as a result. Functions for the three risk levels are displayed in Figure 2, with the size of the bubble indicating the relative size of the risk set (i.e. number of individuals within a latent class eligible to experience the event).

Figure 2: *Functions for the three different levels of (constant) risk*



While the risk set for the three different levels of risk is of equal size at the first time period, the size of the risk set for high risk events diminishes faster than the risk set for the medium and low risk events, as individuals with high risk are more likely to experience the event at each time period.

2.4.2 Methods

Data were generated in SAS 9.2 and the simulation was run in Mplus 6.12 with 100 replications for each condition, totaling 1200 analyses. Boundary values on the logit of ± 15 were allowed per the Mplus default. Due to practical limitations on the amount of time necessary to run the model with random starting values and ensuring replication of the log-likelihood, the population generating values were given as starting values, so random starts were not necessary. This also assisted with the problem of label switching in latent class simulations, a phenomenon where the classes may be correctly captured but are in the incorrect order, making analysis of the results difficult. As another check on this problem, a label switching algorithm was developed in SAS which ordered the classes based on the median of all of the logit parameters within each latent class.[5]

The raw bias of a parameter was computed as the difference between the population generating value and the average value of the parameter found by the model across replications. As there are 60 hazard parameters even in the condition with the smallest number of parameters, two different summary values for bias of the hazard were calculated. The first summary value is the bias of the average of all of the parameters within latent class *k*, given by:

$$\text{Bias in Average Hazard, Class } k = \frac{1}{J \cdot P} \sum_{p=1}^{P} \sum_{j=1}^{J} \left(\hat{h}_{pjk} - h_{pk} \right)$$

with h_{pk} equal to the population hazard value (which was constant across time for process *p* within latent class *k*) and $\hat{h}_{pjk}$ equal to the average value across the 100 replications. The average absolute amount of bias in the average hazard across latent classes is then given by:

$$\text{Bias in Average Hazard} = \frac{1}{3 \cdot J \cdot P} \sum_{k=1}^{3} \left| \sum_{p=1}^{P} \sum_{j=1}^{J} \left(\hat{h}_{pjk} - h_{pk} \right) \right|$$

But the value above indicates the “bias in the average hazard”: only the bias in all of the hazard indicators averaged together. Within a latent class for instance, the model could underestimate the hazard for two events by 0.10 over all time periods and

[5] The median was taken rather than the mean due to concerns about just a few parameters – such as when a logit went to a boundary value when the risk of event occurrence was low – influencing a summary measure of the entire class.

overestimate the hazard for the other two events by 0.10 over all time periods, and this summary measure would indicate the bias in the average parameter is 0. As such, I also examined the average amount of bias for each parameter separately, and took the absolute value when averaging across time periods and events, termed the "absolute bias in hazards." With *J* time periods and *P* events, the absolute bias in hazards in class *k* is calculated as:

$$\text{Absolute Bias in Hazards, Class } k = \frac{1}{J \cdot P}\sum_{p=1}^{P}\sum_{j=1}^{J}\left|\hat{h}_{pjk} - h_{pk}\right|$$

The average across latent classes was also calculated:

$$\text{Absolute Bias in Hazards} = \frac{1}{3 \cdot J \cdot P}\sum_{k=1}^{3}\sum_{p=1}^{P}\sum_{j=1}^{J}\left|\hat{h}_{pjk} - h_{pk}\right|$$

The next measure of bias that was investigated was the absolute bias of the lifetime distribution functions, only examined through one summary measure:

$$\text{Absolute Lifetime Distribution Bias, Class } k = \frac{1}{J \cdot P}\sum_{p=1}^{P}\sum_{j=1}^{J}\left|\hat{D}_{pjk} - D_{pjk}\right|$$

where D_{pjk} is the population value for the lifetime distribution function for process *p* at time period *j* within latent class *k* and $\hat{D}_{pjk}$ is the average value found by the model across the 100 replications. Only one summary measure is needed because the lifetime distribution functions are not constant over time, and thus an overall "average" across time periods and events is not as interpretable as the average hazard is. The average amount of absolute bias across latent classes was again found by summing the bias over the latent classes and dividing by 3. The final measure was bias of the class size, which was investigated in the probability scale, and the average amount of class size bias across classes was also computed by summing the absolute value of the bias for each class and dividing by 3.

While other results – including recovery of the parameters in logit scale, standard deviation of the parameters, and 95% confidence interval coverage of the logit parameters – are tabled in the Appendix, bias of the hazard functions, lifetime distribution functions, and class size will be the focus of the results section that follows.

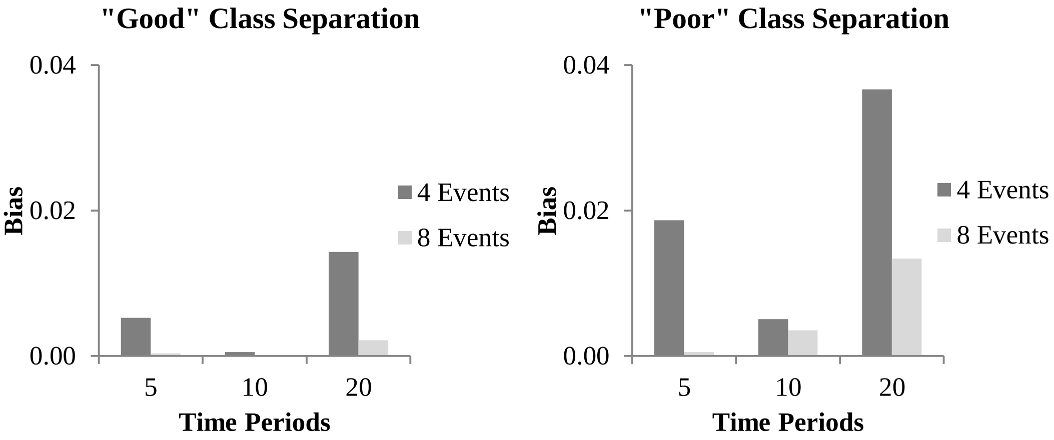
2.4.3 Results

Across all conditions, the model recovered the average of all of the hazard indicators well, with the difference between the population hazard and the average hazard less than 0.02 on average across the latent classes for the good separation condition, and less than 0.04 on average across the latent classes for the poor separation condition. Average bias was generally smallest for 10 time periods, and bias was largest when the number of time periods was 20, specifically because of poor recovery of the parameters on average in the high risk class (and similarly, the mixed class which had half of the events with high risk). Bias is smaller when the number of events is larger. See Table 1 for the bias in the average hazard within each class and across classes for each condition, and Figure 3 for the bias of the average across classes for each condition.

Table 1: *Bias in average hazard*

		Class Separation							
		Good				Poor			
Events	Time Periods	High Class	Medium Class	Low Class	Absolute Average	High Class	Medium Class	Mixed Class	Absolute Average
4	5	-0.009	-0.005	-0.001	0.005	-0.028	0.005	-0.023	0.019
	10	-0.001	-0.001	0.000	0.001	0.003	0.003	0.009	0.005
	20	0.042	0.001	0.000	0.014	0.055	0.003	0.052	0.037
8	5	-0.000	0.001	0.000	0.000	-0.000	-0.000	-0.001	0.001
	10	-0.000	-0.000	0.000	0.000	-0.004	-0.000	-0.006	0.004
	20	0.006	0.001	0.000	0.002	0.020	0.001	0.019	0.013

Figure 3: *Bias in the average hazard, weighting over latent classes*



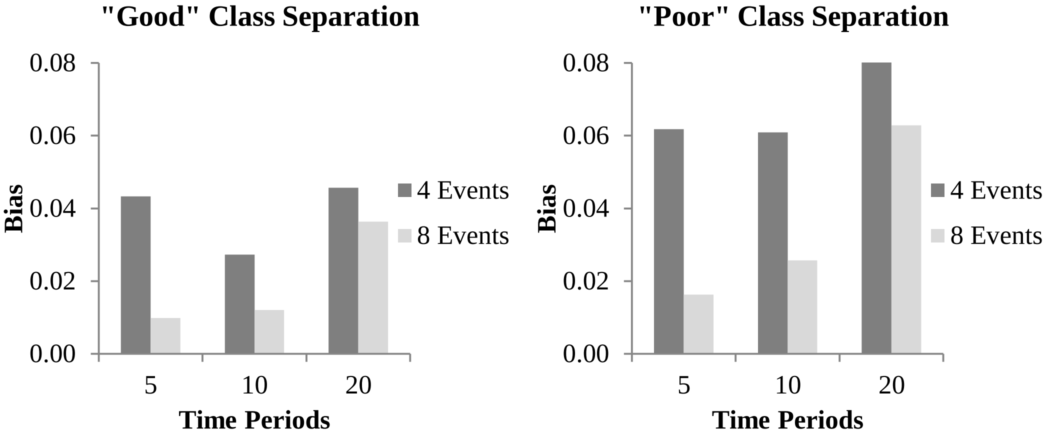
While more time periods provide more information, there is more sparseness in the data (i.e. the risk set diminishes in size), which makes the hazards more difficult to capture at later time periods. This relates to Figure 2, where we saw that the risk set grows especially small in the high risk set at later time periods. We see in Table 1 that while the total average bias is worse with twenty time periods, this is only due to the larger bias in the high risk class, which is to be expected due to the small number of individuals contributing data in that class at later time periods.

Next, examining the absolute bias of the hazards, the bias was again smaller for the good class separation condition and when the number of events was larger. The bias was worse when the number of time periods was 20, again due to poor recovery on average in the high risk class, as to be expected due to the diminishing risk set as discussed above. See Table 2 for the absolute bias in the hazard within each latent class and Figure 4 for the total average amount of absolute bias in the hazard across latent classes.

Table 2: *Absolute bias in hazard*

		Class Separation							
		Good				Poor			
Events	Time Periods	High Class	Medium Class	Low Class	Average	High Class	Medium Class	Mixed Class	Average
4	5	0.056	0.060	0.014	0.043	0.0738	0.0321	0.079	0.062
	10	0.050	0.026	0.006	0.027	0.0704	0.0312	0.081	0.061
	20	0.098	0.034	0.005	0.046	0.1078	0.0445	0.090	0.081
8	5	0.013	0.012	0.005	0.010	0.0184	0.0106	0.020	0.016
	10	0.022	0.010	0.004	0.012	0.0347	0.0112	0.031	0.026
	20	0.090	0.015	0.004	0.036	0.1002	0.0193	0.069	0.063

Figure 4: *Absolute bias in the hazard, weighting over latent classes*

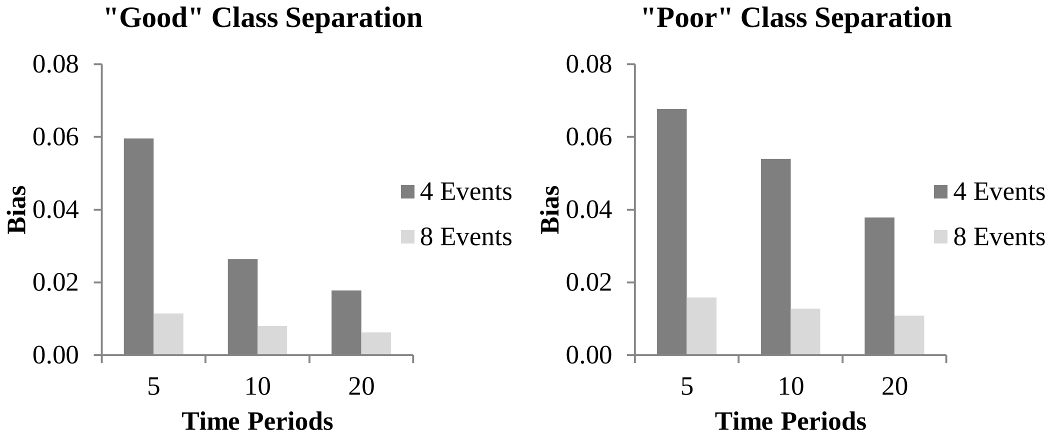


In the scale of the lifetime distribution functions, the trend was very clear: the bias was smaller with good class separation, more events, and more time periods. The best recovery was in the good class separation condition with 8 events and 20 time periods, when the absolute difference between the population lifetime distribution function and the absolute average found by the model was <0.01, and worst in the poor separation condition with only 4 events and 5 time periods, with the average absolute value of the difference between the population value and the model average across the 100 replications equal to 0.068. See Table 3 for the average bias in the lifetime distribution functions within latent classes depending on condition and Figure 5 for the average bias in the lifetime distribution functions averaging over latent classes.

Table 3: *Absolute bias in lifetime distribution functions*

		Class Separation							
		Good				Poor			
Events	Time Periods	High Class	Medium Class	Low Class	Average	High Class	Medium Class	Mixed Class	Average
4	5	0.046	0.104	0.029	0.060	0.056	0.061	0.087	0.068
	10	0.019	0.045	0.015	0.026	0.031	0.058	0.072	0.054
	20	0.014	0.027	0.012	0.018	0.024	0.039	0.051	0.038
8	5	0.010	0.017	0.007	0.011	0.014	0.013	0.020	0.016
	10	0.007	0.010	0.007	0.008	0.011	0.012	0.016	0.013
	20	0.004	0.008	0.007	0.006	0.008	0.011	0.013	0.011

Figure 5: *Absolute bias in lifetime distribution functions, weighting over latent classes*



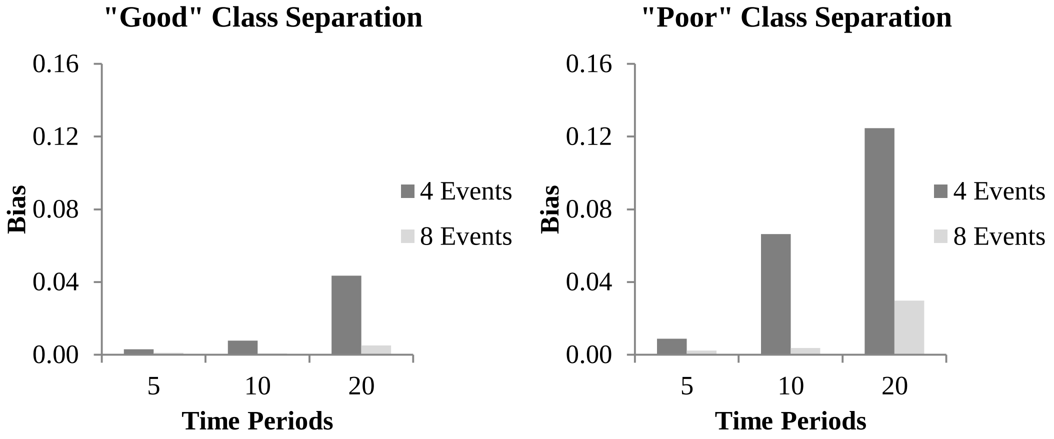
This finding is important to contrast with the recovery of the hazard functions. While the diminishing risk set negatively impacts recovery of the hazard (thus increasing bias with twenty time periods), the lifetime distribution function is a cumulative probability, and is not affected as much by the diminishing risk set. For example, once the lifetime distribution function reaches unity (i.e. cumulative probability of event occurrence is 1), the value of the hazard is irrelevant, as the lifetime distribution function will remain at unity. Thus, in the scale of the lifetime distribution function, the risk set diminishes in relation to the function growing closer to unity, which results in the influence of the hazard on the value of the lifetime distribution function diminishing over time, and thus more time periods results in better recovery on average.

The final result that is examined here is the bias in the probability of class membership (i.e. class size). The bias is smallest with more events and when the class separation is good. The bias grows larger with more time periods, possibly related to the fact with more time periods, the risk set diminishes in size and it becomes more likely that a boundary value is found for the logit. The model tends to overestimate the proportion of the population belonging to the high risk class, while tending to underestimate the size of the medium risk class. See Table 4 and Figure 6 for the average bias in the probability of class membership.

Table 4: *Average bias in the probability of class membership*

		Class Separation							
		Good				Poor			
Events	Time Periods	High Class	Medium Class	Low Class	Absolute Average	High Class	Medium Class	Mixed Class	Absolute Average
4	5	0.004	-0.004	-0.001	0.003	0.013	-0.007	-0.006	0.009
	10	0.010	-0.012	0.002	0.008	0.100	-0.060	-0.039	0.066
	20	0.064	-0.065	0.001	0.044	0.187	-0.122	-0.065	0.125
8	5	0.001	0.001	-0.002	0.001	0.001	0.002	-0.003	0.002
	10	0.001	-0.001	0.000	0.001	-0.002	0.006	-0.003	0.004
	20	0.008	-0.007	-0.001	0.005	0.045	-0.036	-0.009	0.030

Figure 6: *Absolute average bias in the probability of class membership, weighting over latent classes*



Thus, it is hypothesized that increasing the number of time periods makes recovery of the hazard more difficult because of the diminishing risk set, which possibly negatively affects class size estimates. This is consistent with the fact that the bias of the class size of the low risk class (i.e. “low class” in the good separation condition where the risk set remains large across all time periods) is small across all conditions.

2.4.4 Discussion

Two points are clear throughout all of the results; recovery is better 1) when the latent classes have non-overlapping hazard functions (i.e. good class separation condition), and 2) when there are more events. The effect of the number of time periods is not as straightforward. This is likely due to the fact with 20 time periods, the risk set (i.e. number of individuals in the population eligible to experience the event) grows extremely small for the events with high risk. This makes it more difficult for the model to correctly capture hazard parameters, and both the average bias of the hazard parameters and the absolute bias of the hazards is large as a result of poor recovery of the parameters in the high risk class. It is interesting that increasing the number of time periods also negatively affects recovery of the probability of class membership, and this is likely related to the difficulty encountered in estimating the hazard in the high risk class.

However, when examining the lifetime distribution function, the difficulty in capturing the hazard parameters in the high risk set is no longer an issue, as the lifetime distribution function is a cumulative probability and not affected as much when the cumulative probability is near one (when the risk set is small). Thus in this scale, the bias actually decreased with an increase in the number of time periods. In sum, while the number of time periods, class separation, and number of events affected recovery, the model was able overall to capture the population parameters with minimal bias, especially given only 100 replications were used for each condition in this study.

The simulation raises an interesting issue about how bias and other measures should be calculated when there is a large amount of structured missing data. The bias in the simulation was computed simply as the difference between the population generating value and the model estimated value, but it might be more appropriate to weight the bias based on the number of individuals able to experience the event. By doing so, the larger bias found with 20 time periods would likely be smaller. However, there would be many possible ways to do this – such as weighting based on the population known risk set or instead by the model implied risk set – and results would be influenced by the choice of weight, so the more straightforward calculation was used in this paper. Also, by investigating the bias in different scales (i.e. hazard and lifetime distribution), the impact this issue has on understanding recovery of the parameters is reduced.

2.5 Utilizing the Model in Practice

2.5.1 Purpose of Model

The simulation above investigated the recovery of discrete, true groups of individuals in the population, but mixture models are often also applied as an approximation of different forms of underlying heterogeneity (e.g. Heckman & Singer, 1982). Related to this, a distinction between indirect and direct applications of mixture modeling has been made in the literature (e.g. Titterington, Smith, & Makov, 1985; Dolan & van der Maas, 1998). In indirect applications, the purpose of the model is to approximate a distribution of unknown form, and within class estimates are often described as a heuristic examination of local conditions of the underlying distribution (Nagin, 2005). An analogy can be drawn to the “smoothing parameter” in fitting a loess curve to a scatter plot of data, where depending on the value of the smoothing parameter (similar to deciding on the number of latent classes), a different level of detail is revealed in examination of the distribution. It is common for indirect applications to reference the aggregate population rather than within class parameter estimates (Bauer & Curran, 2004; Bauer, 2007). Direct applications instead aim to determine the absolute “true” number of latent classes and focus on within class parameters and the assignment of individuals to latent classes (Titterington, Smith, & Makov, 1985).

This model is proposed to be an indirect application of mixture modeling, as it is employed as a mathematical device – a way to summarize the risk of multiple events. For example, without a mathematical model, describing the timing of four events measured over ten time periods each would be quite difficult, as there would be 14,641 possible combinations of response patterns. To my knowledge, there has been no other proposed model for summarizing this kind of data. Thus, rather than subjectively classifying individuals based on their response patterns and examining the resulting hazard functions within those groups, the model recognizes uncertainty in group membership and allows the examination of predictors on latent classes (Nagin, 1999). Using the model in this indirect way is similar conceptually to using a finite mixture distribution to approximate non-linear relationships between latent variables (Bauer, 2005).

By suggesting the model be used as an indirect application, I am implicitly stating that researchers should not use this method to propose or verify theories regarding the existence of “true” latent subgroups in the population or take the results to suggest that a specific individual will follow one of the pathways described by a latent class. It may not even be possible to determine whether there is truly a certain number of groups or test the assumption that there are a finite number of latent classes of individuals who have the same risk for multiple events over time. However, the model is useful in heuristically describing the heterogeneity in the multivariate distribution, and through its ability to investigate the influence of covariates, which will be discussed next.

When investigating the effects of the covariates, the MEPSUM model can be useful in two different ways. First, the effect of covariates on class membership can be examined, and second, model implied functions weighting over latent classes can be computed for different levels of a covariate. For example, if gender was entered as a covariate to predict class membership, the model would reveal both 1) odds of being in one latent class compared to another depending on gender and 2) model implied functions for the events weighted over the latent classes separately for females and for males. The second way of examining covariates can be thought of as providing an omnibus test of the effect of a covariate on all of the hazard functions simultaneously, controlling for all other covariates in the model.

Importantly, the MEPSUM model is a data-driven method, and the inclusion of auxiliary information is essential to understanding the utility of the latent classes which are derived from the model (Petras & Masyn, 2010). After all, as discussed in the literature on group-based growth mixture modeling, the number of subpopulations is not immutable within a sample and individuals do not belong to a single latent class where everyone in the latent class truly follows the same parameters (Sampson & Laub, 2005). The model may describe patterns of hazard functions which do not truly represent one “true” group of people, similar to how a growth mixture model can detect an additional latent class to account for non-normality in the distribution of repeated measures (Bauer & Curran, 2003a). Thus, examining how the heterogeneity in classes is influenced by covariates should be the end focus of the analysis.

To conclude, the MEPSUM model is likely most useful as a hypothesis-generating method about the influence of covariates on risk of multiple events over time rather than a hypothesis-testing one regarding the absolute number of latent classes of hazard functions. Including covariates and considering construct validation in this process are essential (Bauer & Curran, 2003b). As Nagin and Odgers (2010) argue for a related model, the purpose of latent groups in this model is to draw attention to differences in the causes and consequences of different pathways rather than to suggest the population is composed of literally distinct groups.

2.5.2 Introduction to Model Building

As mentioned in Section 2.5.1, the utility of the model rests in large part on the conclusions that can be drawn from the inclusion of covariates. The discrete-time MEPSUM model presented in this chapter allows for covariates to enter the model in several ways. For example, covariates may predict class membership through the multinomial logistic regression as in Equation (34), or they may have a direct effect on the hazard functions as in Equation (30). To further complicate matters, direct effects of the covariates on the hazard functions can vary not only over event process, but over time period as well as latent class. Another possibility would be to estimate a multiple group model, allowing the hazard functions within latent classes to differ depending on observed group membership.

While there are clear advantages to having such a flexible model, the number of possible specific models that could be explored is quite large when investigating the influence of a number of covariates. This is complicated by the fact that the optimal number of classes could differ depending on whether and how covariates are entered in the model. Additionally, these models may require substantial computational time to ensure a global maximum to the likelihood is found, limiting the practical number of models that a researcher could estimate. Fortunately, one can draw insight from related literature on latent class analysis and growth mixture modeling to formulate an appropriate model building strategy (e.g. Petras & Masyn, 2010; Bandeen-Roche, Miglioretti, Zeger, & Rathouz, 1997; Collins & Lanza, 2010). While the approach outlined below may serve as a guide, note that a different model building strategy may be warranted based on the substantive theory or purpose of the analysis.

2.5.3 Model Comparison

Before discussing the proposed model building strategy in more detail, it is first useful to note different strategies for comparing models. Models may be evaluated and compared using information criteria such as Akaike information criterion (AIC), Bayesian information criterion (BIC), and sample-size adjusted BIC (SABIC) as well as classification indices measuring the degree of uncertainty of classification or separation of the clusters (Akaike, 1974; Akaike, 1987; Schwarz, 1978; Bozdogan, 1987; Fraley, & Raftery, 1998; Celeux, Biernacki, & Govaert, 1997; Vermunt & Magidson, 2002). The Lo-Mendell-Rubin likelihood ratio test and parametric bootstrap likelihood ratio test are other common approaches to selecting the number of classes and evaluating model fit (Lo, Mendell, & Rubin, 2001; McLachlan, & Peel, 2000; Nylund, Asparouhov, & Muthén, 2007).   Researchers may also examine the results to determine whether a class is redundant or whether the probability of belonging to a class is very small, as parameter estimates in a low probability class may not be stable due to the small number of individuals contributing data to that class.

The traditional likelihood ratio chi-squared statistic, which assesses the extent to which the expected cell frequencies differ from the observed cell frequencies, will not be appropriate for most, if not all, applications of the proposed model. This is due to the nature of examining multiple events over time, as the multi-way frequency table will be large relative to sample size, and the statistic would not be well approximated. A likelihood ratio chi-square test for comparing a *k* class model to a *k* + 1 class model is also not an appropriate tool for deciding on the appropriate number of classes as class probabilities in the *k* + 1 class would have to be restricted to be 0, which is a boundary value, violating the regularity conditions necessary for a likelihood ratio chi-square test for nested models (Bishop, Fienberg, & Holland, 1975). However, the likelihood ratio chi-square test may be used to test different models with the same number of classes that differ based on the set of covariates entered into the model or based on restrictions placed on the parameters (e.g. proportional odds assumption).

One model selection step that may aid a researcher in this process – especially when structured hazard functions are used – is to compare the sample observed functions with the model implied functions weighting over latent classes. The aggregate model implied lifetime distribution function for process *p* is found by weighting the within-class function by the probability of class membership $\hat{\pi}_k$:

$$\hat{D}_{pj} = \sum_{k=1}^{K} \hat{\pi}_k \hat{D}_{pjk} \tag{38}$$

The average absolute residual lifetime distribution probability can be then be computed across all event processes. With *P* event process, each with J_P events, this is given by:

$$ARD = \frac{\sum_{p=1}^{P}\sum_{j=1}^{J_p}\left|D_{pj} - \hat{D}_{pj}\right|}{\sum_{p=1}^{P} J_p} \tag{39}$$

where D_{pj} is the sample observed lifetime distribution function for process *p*.

Computing the model implied hazard functions weighting over latent classes is not as straightforward, as the number of people eligible to experience the event in each class will decrease unevenly due to differential risk of event occurrence. Therefore, the population average hazard functions must be computed by weighting the within-class hazard functions not only by the probability of event occurrence, but also by the number eligible to experience the event at time *j* within a latent class *k*. The number eligible to experience the event is equal to the survival probability at time *j* – 1, and the model implied hazard function weighting over latent classes is then given by:

$$\hat{h}_{pj} = \frac{\sum_{k=1}^{K}\hat{\pi}_k \hat{S}_{p,j-1,k}\hat{h}_{pjk}}{\sum_{k=1}^{K}\hat{\pi}_k \hat{S}_{p,j-1,k}} \tag{40}$$

The average absolute residual hazard function across all event process is given by:

$$ARH = \frac{\sum_{p=1}^{P}\sum_{j=1}^{J_p}\left|h_{pj} - \hat{h}_{pj}\right|}{\sum_{p=1}^{P} J_p} \tag{41}$$

where h_{pj} is equal to the sample observed hazard function for process *p*. Ideally, the average residual hazard and lifetime distribution functions would be very close to 0, which is likely when the form of the hazard functions is left unstructured.

Researchers in practice should not rely on one measure or test to determine the number of classes, but rather should use a combination to determine the most appropriate number for their particular research goals. It is important to note that none of the indices or methods mentioned above has been studied in the context of the MEPSUM model, and they should thus be used as guides rather than rules in selecting the model. Also, simulation studies of these indices and methods are often completed under an assumption that there is a true number of latent classes rather than from the viewpoint of an indirect application of mixture modeling. The most important step a researcher can take is to carefully inspect each solution to ensure plausible parameter values as well as interpretability of the overall solution. The final number of classes or range of number of classes should be as small as possible while still allowing heterogeneity in the risk for multiple events over time to be effectively described.

2.5.4 Suggested Model Building Strategy

Step 1: Fit the model with unstructured hazard functions and no covariates

First, I suggest that a researcher fit the MEPSUM model with an increasing number of latent classes, without covariates in the model and with the shape of all hazard functions left unstructured. In selecting an appropriate number or range of number of classes, there must a balance between the need for a model that fits the data well with a desire for parsimony. I would recommend using the indices and methods discussed in Section 2.5.3 as a way to narrow the possible number of classes that will be examined more carefully, but then to examine the hazard and/or lifetime distribution functions from the model. In selecting the final number of classes, researchers must weigh two aspects of the model: 1) that without enough latent classes, the local independence assumption will be violated in that the model will not completely account for the association between the hazard indicators, and 2) with too many classes, the model will lose its function as a parsimonious description of the underlying process. One way to do this would be to determine whether the addition of a latent class is substantively meaningful in effectively describing the heterogeneity in the hazard functions.

Step 2: Determine if structured hazards are necessary

As the unstructured hazard function is the most flexible and general form, I have suggested that it should be used first for an increasing number of classes in order to examine the hazard functions with latent classes when no shape is imposed. However, if the number of time periods is large or some of the events have a very low risk of occurrence, a parametric form of the hazard function may be considered next. Caution is needed before imposing such a structure, as the shape of the hazard function may differ between latent classes, and each event should be considered separately as the hazard functions for the different events may have radically different shapes. Additionally, it is possible that the number of classes influences the shape of the hazard functions; for example, the shape of the hazard functions when examining a two class solution may be different than the shape of the hazard functions when a five class solution is examined. However, the solution found with unstructured hazard functions can serve as a guide to the form of the functions within a certain number of latent classes.

Step 3: Add covariates to predict class membership

Once the form of the hazard functions and an appropriate number or range of number of classes has been chosen, covariates can be incorporated directly into the model to predict class membership, as in Equation (34). The model built in this way implicitly assumes that the covariates are independent of the hazard functions conditional on the latent class variable. In other words, the model assumes that covariates only influence the probability of belonging to one latent class over another, but that the hazard functions within latent class are the same regardless of observed covariates. All effects of covariates on the risk of event occurrence are transmitted through the latent class variable and thus latent class membership is assumed to be sufficient to describe the risk of the events occurring over time:

$$E(h_{pijk} \mid C = k, \mathbf{X}_i) = E(h_{pijk} \mid C = k) \tag{42}$$

This assumption would be violated if the hazard functions within latent class were dependent upon an observed covariate.

For example, suppose a researcher was examining the risk of four events, one of which was the onset of depression, and suppose that men had a higher risk of depression across all time periods. If the hazard functions within latent classes for men and women were the same except that the hazard function for depression for men was higher across all time periods in each class (within-class differences), this would be a violation of the assumption. However, if there was a high risk of depression class, and the probability of belonging to that class was higher for men, it is possible that the differences between men and women in the risk of depression could be captured without violation of the assumption (between-class differences).

Step 4: Determine stability of the model with covariates

The solution obtained without covariates should then be compared to the solution obtained with covariates influencing class membership. As the covariates are assumed to affect only class membership, the substantive meaning and size of the clusters should remain unbiased by the inclusion or exclusion of the covariates (Petras & Masyn, 2010). If the substantive interpretation of the classes changes, this may indicate that the assumption that the covariates only influence class membership is violated (Marsh, Lüdtke, Trautwein, & Morin, 2009). In this case, direct effects of covariates on the hazard functions should be explored, or the number of latent classes should be reevaluated. Another possibility would be to estimate a multiple group model at this stage, if sample size warranted and there was reason to believe the hazard functions within latent classes would differ based on observed group membership. Thus, including covariates in the model after selecting an appropriate number of classes can serve as a verification tool of the stability of the model and to explore whether direct effects are necessary or a multiple group model should be considered. The ability to test the model in this way is an advantage over the single event version of the MEPSUM model, which is not identified in the absence of covariates.

Note that adding direct effects of covariates on the hazard functions substantially increases the complexity and interpretability of the model. If necessary, direct effects should initially be entered as class-invariant, as any parameter that varies over latent classes provides information to identify and discriminate the latent classes (Petras & Masyn, 2010). For example, including class-varying direct effects of a covariate on the hazard functions results in latent classes defined both by heterogeneity in the hazard functions and by heterogeneity in the effect of the covariate on the hazard functions. Also, if covariates are allowed to have both between-group effects in influencing class membership and within class effects through direct effects on the hazard functions, a multiple-group model would not be necessary as the model would already capture the between-group differences (Muthén, 2001).

However, a multiple-group model – where the grouping variable defines known subpopulations – might be preferred instead of including direct effects for substantive interpretation purposes. A multiple-group model can be used to relax the assumption that the hazard functions within latent class are the same for different known groups, and this model also allows the grouping variable to moderate the effect of each predictor on class membership (see Collins & Lanza, 2010). Another way to test whether the effect of a covariate differs among a known group would be to include an interaction term between the covariate and the group in the model.

The strategy outlined above assumes the researcher will include the predictors directly in the model rather than testing the effects of the covariates after classification has taken place. It has been shown that examining the effects of predictors in a two-step fashion, by first estimating the latent classes and then separately examining the effects of the covariates by the assigning individuals to a latent class by modal posterior probabilities results in significant biases in the estimation of the model parameters (Clogg, 1995; Hagenaars, 1993; Clark & Muthén, 2009). The simultaneous approach is also recommended over a pseudo class draw technique, which aims to account for the variability in the posterior probabilities (Wang, Brown, & Bandeen-Roche, 2005). However, this all assumes that the researcher is examining the effects of predictors on the latent classes; when distal outcomes are of interest, pseudo class draws may be the most appropriate strategy, or researchers may even include the outcome in the formation of the latent classes, depending on the purpose of the analysis (see Petras & Masyn, 2010).

Step 5: Explore influence of covariates

As the last step, the effects of covariates entered into the model to predict class membership can be explored in two ways: 1) through the odds of class membership in one latent class in relation to another latent class and 2) through the model implied hazard and lifetime distribution functions for the events, weighted over latent class. When covariates influence only group membership, the probability of class membership can be computed simply through Equation (34), and then Equation (38) and (40) above may be used to find aggregate model implied functions for a specific level of a covariate. These functions allow researchers to compare the effect of different covariates, controlling for other covariates in the model.

With ample sample size and a small number of categorical covariates, model implied hazard functions for different levels of a covariate can be compared to stratified sample observed hazard functions to obtain residual hazard functions, as will be demonstrated in Chapter 3. Ample sample size also permits the possibility of a split sample validation, where the researcher estimates the MEPSUM in one random half of the sample and then compares results found in the second half of the sample, as will be demonstrated in Chapter 4. The method and strategies for model building introduced in this chapter will now be demonstrated in the two chapters that follow.

CHAPTER 3

EMPIRICAL EXAMPLE 1 – TRANSITIONS TO ADULTHOOD

The aim of the first empirical example is to examine the order and timing of different transitions into adulthood. Researchers have long established that the events that occur over an individual's life are interdependent. For example, individuals may make decisions on whether they would like to continue their education based on their family status, such as whether they are married and have children (Marini, 1984). Early parenthood may lead an individual to postpone educational goals or start full-time work earlier than he or she would have otherwise. In contrast, a person's family behavior may depend on educational goals, such as an individual postponing parenthood based on whether he or she is currently in school or not (Hofferth & Moore, 1979).

Life course research is guided by the notion that an individual's development involves the order and timing of multiple social roles over time where the meaning of a given social role is dependent upon the presence or absence of other roles. Elder (1985) notes that the dynamics involved in the life course can be conceptualized through the notions of role trajectories and transitions, which are interdependent over time. Trajectories index the timing of social roles over time, such as an individual's path through schooling, employment, marriage, and parenthood; transitions mark changes in a role status, such as having a child (Macmillan & Copher, 2005). Transitions are given meaning and form depending on the trajectory in which they are embedded (Elder, Johnson, & Crosnoe, 2003). The interconnectedness of trajectories and transitions identify pathways through the life course that mark the general structures of the life course (Macmillan & Eliason, 2003). These pathways are greatly shaped by social institutions and historical forces (Shanahan, 2000; Shanahan, Miech, & Elder, 1998).

As Macmillan and Eliason (2003) note, the phenomenon of the life course as a whole, characterized as interlocked pathways of social roles over time, has seldom been the object of research. This is likely due to the fact that without an appropriate statistical model, a researcher who aims to examine the order and timing of numerous social roles would be confronted with hundreds or thousands of possible combinations of movement into social roles over time (Hogan, 1978). Instead of investigating the multidimensional nature of the life course, researchers typically focus on one aspect of the life course, such as timing of an individual's first child; then they examine this event in isolation from other life course events using traditional methods such as linear and logit regression and univariate event history models. However, as the significance of a role depends on the role configuration, dissecting the life course in such a way limits our understanding of the life course as a dynamic phenomenon (Macmillan and Eliason, 2003).

In aiming to understand the dynamic, multidimensional nature of the life course, the MEPSUM model proposed in Chapter 2 was applied to the timing of four transitions: marriage, parenthood, college degree, and the beginning of full-time work. The purpose of this analysis is both to demonstrate the model's applicability to life course theory and to build on prior research by examining the latent classes which reveal pathways to adulthood, or patterns of the events over time. The life course pathways found from this model are not expected to be the only pathways through the life course nor are they expected to reveal true groups of people, but they provide a glimpse at the underlying multivariate distribution of pathways, of which there are likely thousands of possibilities. Additionally, this example is useful in examining the ability of the model to detect differences in pathways taken by different social groups.

A key concern in life course theory is how membership in different social groups can influence the life course pathway of an individual, and to this end, several predictors were added to the model to examine their influence on the probability of belonging to a certain pathway, or latent class. In other words, the model is useful in understanding the mechanisms leading to different pathways through the life course. In particular, the influence of gender, race, and parent education was examined. Consistent with prior literature, it is hypothesized that all three predictors have a significant influence on heterogeneity in the hazard functions over time (e.g. Mahaffy, 2003). Only a small number of categorical covariates were examined so that model implied functions could be compared to sample observed functions of the sample stratified by the different levels of the covariates, in order to investigate the ability of the model to detect group differences.

3.1 Methods

3.1.1 Data

The data for this example come from Wave I and Wave IV of the National Longitudinal Study of Adolescent Health (Add Health; Harris et al., 2009). Add Health began in the 1994-1995 school year with a nationally representative sample of adolescents from 80 high schools and 52 middle schools in the United States selected with unequal probability of selection. The individuals were then followed from adolescence into adulthood through four in-home interviews. Parental interviews were also completed during the first wave. The last interview, Wave IV, was completed in 2008, when the majority of the sample was twenty-four to thirty-two years old (see Table 5). At each wave, information was gathered on respondents' social, economic, psychological, and physical well-being. Wave IV in-home interviews were completed for 15,701 individuals.

Table 5: *Age at time of interview for individuals sampled in Wave IV Add Health*

Age	Frequency	Cumulative Percent
24	30	0.19
25	665	4.43
26	1808	15.94
27	2273	30.42
28	2822	48.39
29	2959	67.24
30	2885	85.61
31	1857	97.44
32	347	99.65
33	50	99.97
34	5	100.00

3.1.2 Measures

Four role status variables were examined: marriage, college graduation, full-time work, and parenthood. For each age from 18-30, a binary variable for each status was created indicating whether the individual occupied the status for the first time at that age (coded 1), or had not occupied the status by that age (coded 0). Once the individual occupied one of the role statues, they no longer contribute data for the remaining ages (coded as missing). See Table 6 for the extent of missing data, including censoring. To account for the fact that a small percentage of individuals occupied one of the roles before they were eighteen years old, the binary variable for age 18 will represent whether the individual occupied the status for the first time at age 18 or younger. In essence, this is structuring the first time period to be wider (from birth to age 18) than any of the other time periods, which all represent one year.

Table 6: *Number of individuals with missing data in Add Health*

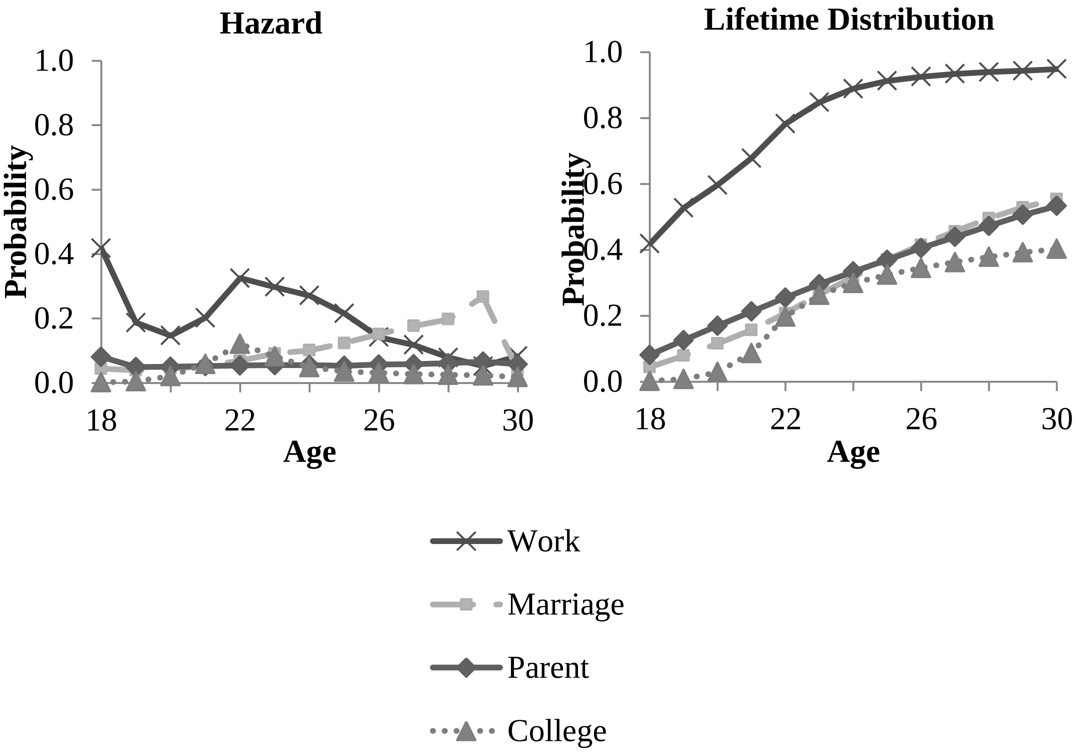
Event	Number Uncensored (%)	Number Censored (%)	Number Missing (%)
Parent	7664 (48.68%)	8000 (50.95%)	57 (0.34%)
Marriage	7648 (48.71%)	7912 (50.39%)	141 (0.90%)
College Graduation	6207 (39.53%)	9487 (60.42%)	7 (0.04%)
Full-time Work	14795 (94.23%)	860 (5.48%)	46 (0.29%)

The role status variables were taken from the Wave IV Add Health interview. The month and year of the individual's first marriage was used to find the age of the respondent when they first married. The year of the respondent's first degree (associate's degree, bachelor's degree, or graduate degree) after high school was used to determine the age at which the first post-high school degree was obtained, by using the age the respondent was for the majority of that year. The date of birth of the respondent's oldest child was used to determine the age at which the respondent first became a parent. The age when the person first began full-time work was directly measured in the Add Health interview. The sample observed hazard probabilities for each event process are listed in Table 7 and displayed in Figure 7. The sample observed lifetime distribution function for each event process is also displayed in Figure 7.

Table 7: *Number of event occurrences and sample estimated hazard probabilities in Add Health*

	Parent		Marriage		College Graduation		Full-time work	
Age	Event	Hazard	Event	Hazard	Event	Hazard	Event	Hazard
18	1227	0.08	536	0.03	12	0.00	6229	0.40
19	712	0.05	534	0.04	95	0.01	1809	0.19
20	723	0.06	597	0.04	313	0.02	1166	0.15
21	685	0.06	678	0.05	905	0.06	1362	0.21
22	660	0.06	766	0.06	1697	0.13	1692	0.33
23	641	0.06	858	0.07	1103	0.10	1033	0.30
24	614	0.06	816	0.08	605	0.06	655	0.28
25	578	0.06	810	0.08	433	0.04	417	0.24
26	567	0.06	677	0.08	351	0.04	208	0.17
27	444	0.06	538	0.08	275	0.03	128	0.14
28	375	0.07	415	0.08	191	0.03	67	0.10
29	254	0.07	254	0.07	125	0.02	28	0.06
30	135	0.06	131	0.06	67	0.02	14	0.06

Figure 7: *Add Health sample observed functions*



Three predictors were examined, each of which was assessed during Add Health Wave I: gender, race, and parental education. Gender was measured as a two-category item of male (46.83%) and female (53.17%). The measurement of race was simplified to a four category item of Caucasian (52.87%), African-American (20.62%), Hispanic (15.92%), and other (10.59%). Parent education was measured as the highest level of education achieved by either parent on a three point scale of less than high school (12.85%), high school degree (25.33%), or any schooling beyond high school (61.82%). Sampling weights given by Add Health accounting for the unequal probability of selection are used. Individuals with missing data on any of the covariates (<1.5%) or sample weights (<1 %) are excluded from the analysis, resulting in a final analysis sample of N = 14,557.

3.1.3 Analysis

The discrete-time MEPSUM model proposed in Chapter 2 was fit to the data using the robust maximum likelihood estimator (MLR) accounting for sample weights in Mplus 6.12. The first model was run on the four event processes across the thirteen time points, without covariates, from one to six latent classes with unstructured hazard functions. To ensure a global maximum likelihood solution, at least 1,000 random sets of starting values were used for each model, with the best 500 retained for final optimization, and the resulting solutions monitored to ensure the final loglikelihood was replicated. The number of classes was chosen based on a combination of information criteria, classification indices, interpretability (e.g. no clusters are redundant or small enough to warrant concern about the stability of parameter estimates), and parsimony. The resulting hazard functions were then used to indirectly estimate the lifetime distribution functions for each process within each latent class through Equations (3) and (7); these results were used to describe prototypical pathways of the events over time.

As discussed earlier, life course theory is built on the notion that there are interdependent trajectories over time, but is also concerned with how membership in different social groups can influence the likelihood that an individual follows one pathway over another. After the class enumeration process was complete, the next model included the covariates as predictors of class membership, as in Equation (34). Including covariates after selecting the number of classes allows for verification of the stability of the model (Petras & Masyn, 2010).

Finally, model implied lifetime distribution functions weighting over latent classes were computed to investigate the fit of the model and ability to detect group differences. I purposefully only examined a small number of categorical covariates so that model implied lifetime distribution functions could be compared to observed lifetime distribution functions from the sample stratified by the different levels of the covariates. This is only reasonable due to the extremely large sample size and small number of categorical covariates, allowing stratification and computation of hazard and lifetime distribution functions by gender, race, and parental education. In practice, it will likely only be possible to compute model implied functions for certain levels of the covariates rather than being able to empirically compare the model implied and sample observed functions.

3.2 Results

The MEPSUM model was fit with an increasing number of latent classes with unstructured hazard functions. Information criteria suggested solutions with at least six classes were optimal, while entropy suggested misclassifications were smallest for the two and four class solutions (Table 8). The shape of the hazard functions was different across latent classes and the unstructured hazards form of the MEPSUM model was deemed optimal. Examining the hazard and lifetime distribution functions more carefully for each of the solutions revealed a five class solution was optimal; the five class solution was able to more effectively describe heterogeneity in the risk of the events over time than the four class solution but the same was not true when increasing from a five class to a six class solution. The five class solution will first be described, and will then be compared to the six class solution to describe why the five class solution was chosen.

Table 8: *Model fit to Add Health data*

Latent Classes	-2LL	Number of Free Parameters	BIC	AIC	Smallest Class	Entropy
1	-102521.76	52	205541.99	205147.53	N/A	N/A
2	-98444.65	105	197895.81	197099.29	0.33	0.79
3	-97481.09	158	196476.75	195278.19	0.26	0.74
4	-96784.46	211	195591.54	193990.93	0.11	0.76
5	-96425.50	264	195381.66	193379.00	0.10	0.71
6	-96087.98	317	195214.68	192809.97	0.09	0.72

In the five class solution, the first class ($\hat{\pi}_1$ = 0.168) is characterized by high early risk of work ($\hat{h}_{18}$ = 0.63), followed by an increasing risk of transition into family roles. The risk of marriage starts low ($\hat{h}_{18}$ = 0.03) and increases rapidly to a high risk of 0.80 at age 29. The median event time for marriage is in between ages 21 and 22, with nearly a 1.00 cumulative probability of marriage by age 30. The risk of parenthood also starts low ($\hat{h}_{18}$< 0.01), and increases in a linear fashion, though the risk is never as high as that for marriage for any specific age (e.g. $\hat{h}_{28}$ = 0.24). By age 30, the model implied probability of being a parent is 0.86 for this class, with the median parenthood age between ages 24 and 25. The risk of college graduation is low throughout all of the time periods (maximum is $\hat{h}_{29}$ = 0.03), with a small cumulative probability of graduating college by age 30 ($\hat{D}_{30}$ = 0.17). This first class will be labeled a “work then family” pathway (WF).

The second class ($\hat{\pi}_2$ = 0.102) is characterized by a moderate risk of transitioning into both college and work roles in the mid-twenties, followed by an increasing risk of transitioning into parent and marriage roles in the later twenties. Specifically, the risk of college peaks around ages 22 ($\hat{h}_{22}$ = 0.42) and the risk of work also peaks around ages 22 to 24 ($\hat{h}_{22}$ = 0.43, $\hat{h}_{24}$ = 0.45). The median age for both beginning full-time work and for college graduation is between ages 21 and 22. The risk of transitioning into marriage is relatively low in the early twenties ($\hat{h}_{22}$ = 0.15) but increases into the late twenties ($\hat{h}_{27}$ = 0.58). Risk of parenthood similarly is low in the early twenties ($\hat{h}_{22}$ = 0.04), but steadily increases throughout the twenties ($\hat{h}_{30}$ = 0.41). The median age of marriage is between 23 and 24 with nearly a 1.00 probability of marriage by age 30, and the median age of parenthood is between 26 and 27, with high probability of parenthood by age 30 ($\hat{D}_{30}$= 0.88). This second class will be labeled a "college then family" pathway (CF).

The third latent class ($\hat{\pi}_3$ = 0.217) is characterized by moderate risk of college and work in the mid-twenties, similar to the CF pathway mentioned previously, only the risk of transitioning into any family role is low throughout the entire period under study. The risk of college is moderate, at least above 0.20, for all ages after 21. The risk is especially high at age 22 ($\hat{h}_{22}$ = 0.42) and age 30 ($\hat{h}_{30}$ = 0.61). The median college graduation age is between 21 and 22, with a 0.99 probability of graduating college by age 30. The risk of work is similarly moderate for all time periods after age 21 (e.g. $\hat{h}_{22}$ = 0.37, $\hat{h}_{30}$ = 0.36), with a 0.98 probability of transitioning into full-time work by age 30. The risk of transitioning into a parent role is less than 0.03 for all ages, and the risk of marriage is similarly low, peaking at 0.11 at age 28. By age 30, there is a 0.38 cumulative probability of transitioning into marriage and only a 0.09 cumulative probability of transitioning into parenthood. This will be labeled a "college and work" pathway (CW).

The hazard functions for the fourth latent class ($\hat{\pi}_4$ = 0.222) look remarkably different than the other classes, in the risk for all events decreases over time and the risk of transitioning into a parent role is especially high at early ages. At age 18, the risk of beginning full-time work is 0.59 and the risk of parenthood is 0.35. The median age for beginning full-time work is less than age 18, with a cumulative probability of beginning full-time work of 0.95 by age 30. While decreasing in magnitude, the risk of parenthood remains high in comparison to the other latent classes (e.g. $\hat{h}_{22}$ = 0.30 compared to $\hat{h}_{22}$ = 0.13 in the WF pathway). The cumulatively probability of becoming a parent is 0.70 as early as age 20 and reaches 0.90 by age 24. The risk of marriage is also the highest at age 18 ($\hat{h}_{18}$ = 0.15) and decreases throughout the time period under study ($\hat{h}_{30}$ = 0.05), with the median marriage time between ages 24 and 25. The risk of college graduation is very low throughout the entire time period (maximum $\hat{h}_{26}$ = 0.02), with a small cumulative probability of graduating college by age 30 ($\hat{D}_{30}$ = 0.13). This class will be labeled "early parenthood" pathway (EP).

In the fifth class ($\hat{\pi}_5$ = 0.291), the risk for transitioning into family roles as well as the risk of college is extremely low throughout all of the time periods, and the risk of work is highest at early ages and then decreases. The risk of work is 0.54 at age 18, and quickly and steadily decreases, with a risk of less than 0.10 of beginning full-time work for each age after 23. The median age for transitioning into full-time work is less than age 18, with a 0.90 cumulative probability by age 30. The risk of marriage is never higher than 0.05 for any age, nor is the risk of parenthood or college graduation. The cumulative probability of transitioning into marriage is 0.23 by age 30, and is 0.26 for parenthood. The cumulative probability of graduating college by age 30 is 0.13. As this class is characterized almost completely by the transition into a work role only, this class will be labeled “work” (W).

Hazard functions for the 5 class solution, representing the unique risk of event occurrence at a given age or the probability of event occurrence given the event had not yet occurred are displayed in Figure 8. The lifetime distribution functions, displaying the cumulative probability of event occurrence by a given age, are shown in Figure 9. The median event time for an event process within a latent class occurs when the lifetime distribution function is equal to 0.50 (Table 9).

Figure 8: *Add Health hazard functions for five class solution*

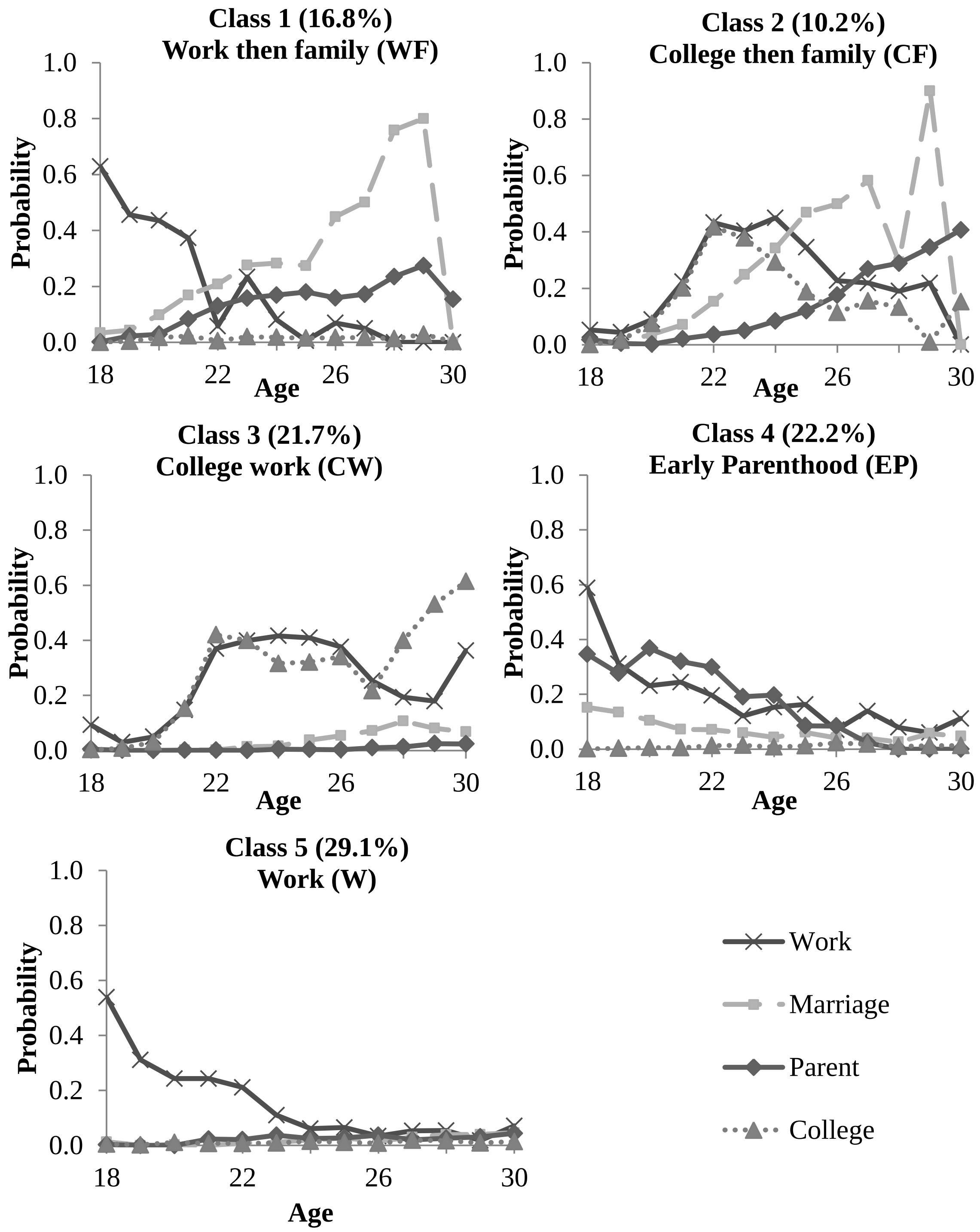
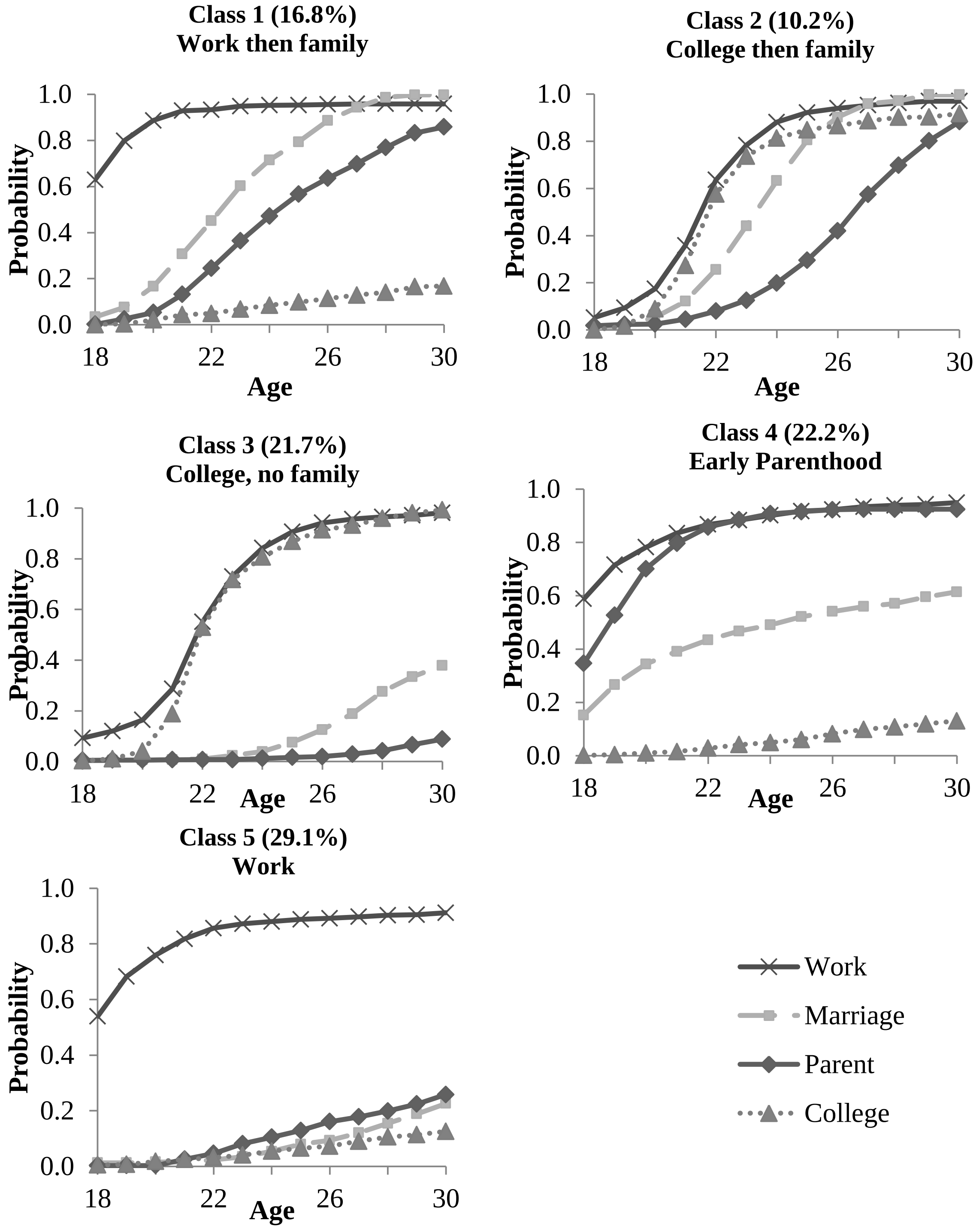


Figure 9: *Add Health lifetime distribution functions for five class solution*



Class 1 (16.8%)
Work then family

Class 2 (10.2%)
College then family

Class 3 (21.7%)
College, no family

Class 4 (22.2%)
Early Parenthood

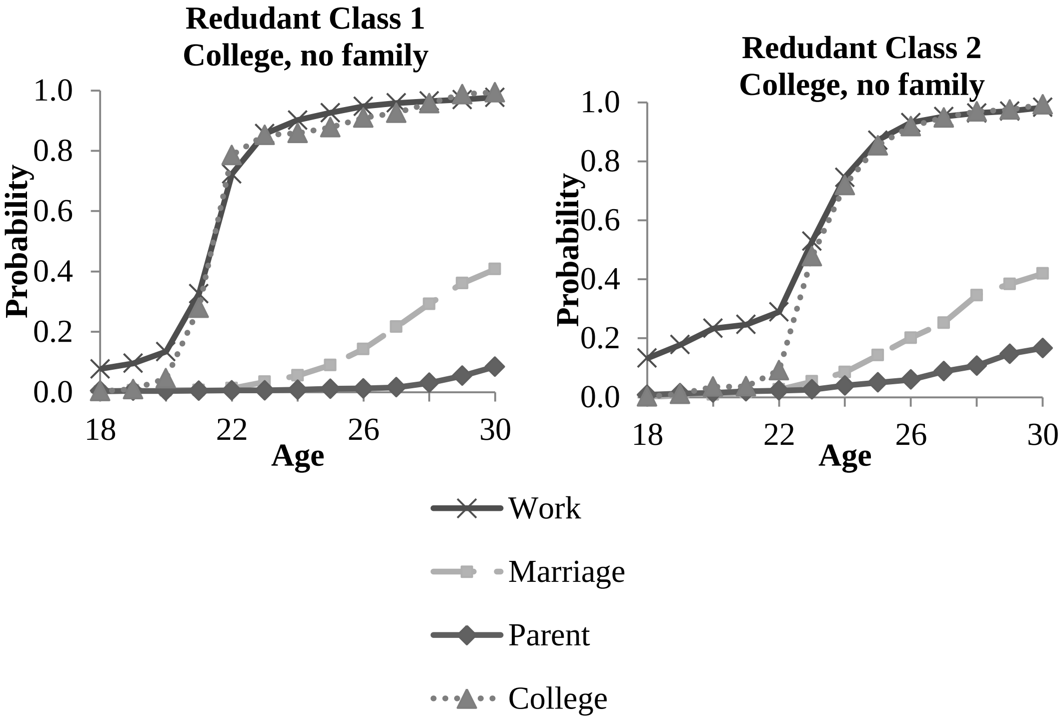
Class 5 (29.1%)
Work

Table 9: *Add Health median event time within latent classes*

Class	Label	Work	Marriage	Parent	College
1	WF	<18	22.5	24.5	-
2	CF	21.5	23.5	26.5	21.5
3	CW	21.5	-	-	21.5
4	EP	<18	24.5	18.5	-
5	W	<18	-	-	-

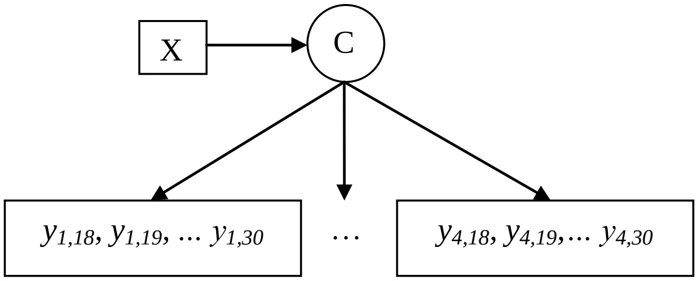
Examining results for the six class solution revealed a substantively redundant latent class, resulting in the five class solution being selected as the final solution. In the six class solution, the main difference is that the third class from the five class solution – the “college and work” pathway – split into two separate classes. The other classes remain virtually identical to the five class solution. The lifetime distribution function reveals the difference in the cumulative probability of marriage for any age for the two classes is smaller than 0.06. Similarly, the lifetime distribution functions reveal that the cumulative probability of parenthood by any age for the first redundant class is within 0.09 of the cumulative probability of parenthood by any age for the second redundant class. The main difference between the two classes is that the hazard or risk of college at age 22 is high for one class (0.78) while low for the other (0.09), yet this difference is only at that specific age. Both classes have a 0.99 cumulative probability of graduating college by age 30 and 0.98 cumulative probability of work by age 30. The lifetime distribution functions of these two classes are displayed in Figure 10.

Figure 10: *Add Health lifetime distribution functions for redundant classes found in six class solution*



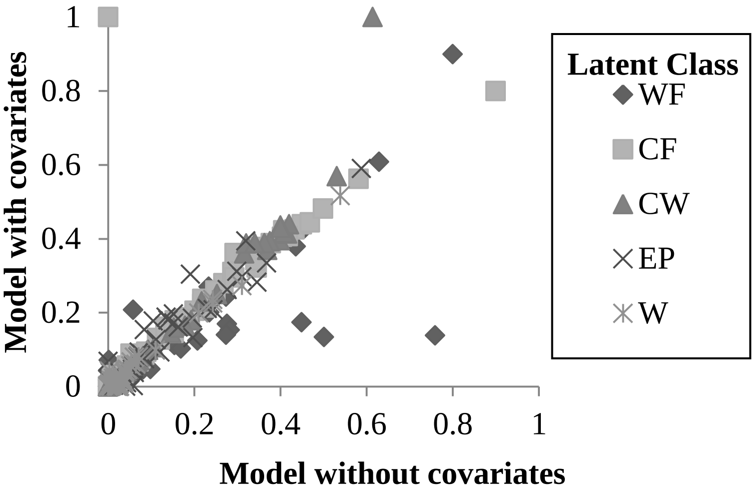
Thus, the increase in complexity from a five to a six class solution is not warranted in that it does not substantially increase our ability to describe heterogeneity in the hazard functions. The five class solution is selected at this stage as the optimal number of classes, and covariates are now entered into the model to predict class membership (Figure 11). By selecting the number of classes without covariates and then comparing the solution to that obtained with covariates predicting class membership, the stability of the model can be investigated.

Figure 11: *Add Health simple path diagram of model with covariates*


X
C
$y_{1,18}, y_{1,19}, \ldots y_{1,30}$
…
$y_{4,18}, y_{4,19}, \ldots y_{4,30}$

The size of the classes as well as the parameter estimates remained relatively stable even after the covariates were entered into the model. The correlation between the hazard for all of the events across all of the ages in all of the latent classes between the model estimated without covariates and the model estimated with covariates was 0.87. A plot of all individual hazards in the model estimated with covariates versus the model estimated without covariates is displayed in Figure 12. Note there are a few outliers in the plot. However, we must remember that the hazard function at later ages is less stable, as the number of individuals who remain eligible to experience the event grows smaller. In fact, many of the outliers that are found in this plot are at later ages and if we estimate the correlation between hazard indicators for all events between the model estimated without covariates and the model estimated with covariates excluding just age 30, the correlation increases to 0.94. Additionally, in the scale of the lifetime distribution function, the correlation between the cumulative probabilities for all events between the model estimated without covariates and the model estimated with covariates is 0.99. These results imply that the assumption of independence between the covariates and the hazard functions conditional on latent class has not been violated.

Figure 12: *Add health hazard indicators across all ages and events found in model without covariates compared to model with covariates*



As another check on the model, if we compare the aggregate model implied lifetime distribution functions and the sample observed lifetime distribution functions, we find that the average difference between the two sets of functions is small, *ARD* < 0.001. The difference between the aggregate model implied hazard functions and sample observed hazard functions is also small, *ARH* = 0.001. Thus, the model is capturing the observed overall risk of event occurrence well, as is expected with unstructured hazard functions.

Covariates in the final model were entered solely to predict class membership. As such, the model reveals the odds of being in one latent class compared to another depending on the level of a covariate. A complete list of all possible odds ratios is given in Table 10, with confidence intervals listed below the estimate, computed with a Bonferroni correction for multiple comparisons with $\alpha = 0.05$. This table reveals that gender, race, and parental education all significantly influence latent class membership, as several confidence intervals do not include 1 for each group of predictors.

Table 10: *Add Health odds ratios for five class solution*

		Gender	Race			Parental Education	
Class	Intercept	Female	Black	Hispanic	Other	No degree	College
WF v. W	0.63	1.68	**0.15**	0.58	0.72	1.05	1.00
	(0.35,1.15)	(0.71,4.00)	**(0.06,0.41)**	(0.30,1.11)	(0.34,1.53)	(0.54,2.06)	(0.64,1.55)
CF v. W	**0.10**	**4.74**	**0.30**	**0.28**	0.74	0.46	**3.26**
	(0.04,0.27)	**(2.46,9.15)**	**(0.15,0.57)**	**(0.11,0.72)**	(0.35,1.55)	(0.16,1.29)	**(1.60,6.63)**
CW v. W	**0.25**	**2.39**	**0.45**	0.54	1.24	0.40	**4.00**
	(0.14,0.44)	**(1.54,3.73)**	**(0.24,0.86)**	(0.26,1.13)	(0.68,2.26)	(0.15,1.08)	**(2.52,6.33)**
EP v. W	0.37	**4.94**	1.31	1.06	0.98	1.01	0.73
	(0.11,1.22)	**(2.95,8.26)**	(0.56,3.08)	(0.45,2.50)	(0.44,2.16)	(0.61,1.67)	(0.51,1.04)
WF v EP	1.71	**0.34**	**0.11**	0.55	0.74	1.04	1.37
	(0.59,4.90)	**(0.17,0.67)**	**(0.02,0.55)**	(0.22,1.38)	(0.24,2.25)	(0.62,1.76)	(0.89,2.12)
CF v EP	0.28	0.96	**0.23**	**0.27**	0.76	0.45	**4.49**
	(0.05,1.59)	(0.46,2.01)	**(0.09,0.60)**	**(0.12,0.60)**	(0.31,1.85)	(0.18,1.18)	**(2.40,8.42)**
CW v EP	0.67	**0.48**	0.34	0.51	1.27	**0.40**	**5.51**
	(0.26,1.69)	**(0.32,0.74)**	(0.12,1.03)	(0.23,1.15)	(0.52,3.10)	**(0.16,0.99)**	**(3.22,9.43)**
WF v CW	**2.55**	0.70	0.33	1.06	0.58	**2.60**	**0.25**
	(1.48,4.42)	(0.32,1.57)	(0.11,1.01)	(0.52,2.17)	(0.24,1.40)	**(1.01,6.79)**	**(0.15,0.42)**
CF v CW	0.41	**1.98**	0.66	0.52	0.60	1.14	0.81
	(0.12,1.46)	**(1.04,3.77)**	(0.33,1.30)	(0.23,1.17)	(0.28,1.28)	(0.28,4.60)	(0.35,1.88)
WF v CF	**6.16**	0.36	0.51	2.05	0.98	2.29	**0.31**
	(1.89,20.02)	(0.11,1.16)	(0.16,1.57)	(0.79,5.33)	(0.36,2.68)	(0.74,7.04)	**(0.14,0.67)**

Examining the influence of gender, we see that the odds of females being in either college pathway compared to the work pathway is over two times as great as the odds for males. Similarly, the odds for females being in the early parenthood pathway compared to the work pathway is 4.94 times as great as that for males. The odds of females being in the early parenthood pathway compared to the work then family pathway or the college then family pathway are also larger than the odds for males (2.94 and 2.08, respectively). Comparing the two college pathways, females are more likely to be in the college and family pathway than the college and work pathway compared to males. Generalizing over all of these findings, females are generally more likely to be in the early parenthood pathway, and males are generally more likely to be in the work pathway.

The odds for African Americans being in the work pathway compared to either college pathway or the work then family pathway are larger than the odds for Caucasians. The odds of African Americans being in the work then family pathway compared to the early parenthood pathway are smaller than the odds for Caucasians, as well as the odds of being in the college then family pathway compared to the early parenthood pathway (0.23). Overall, the model implies that African-Americans are generally more likely to be in the work pathway and the early parenthood pathway than Caucasians. Similarly, Hispanics are more likely to be in the work pathway and the early parenthood pathway than the college then family pathway than Caucasians. No differences between those of other races and Caucasians were found in terms of predicting class membership.

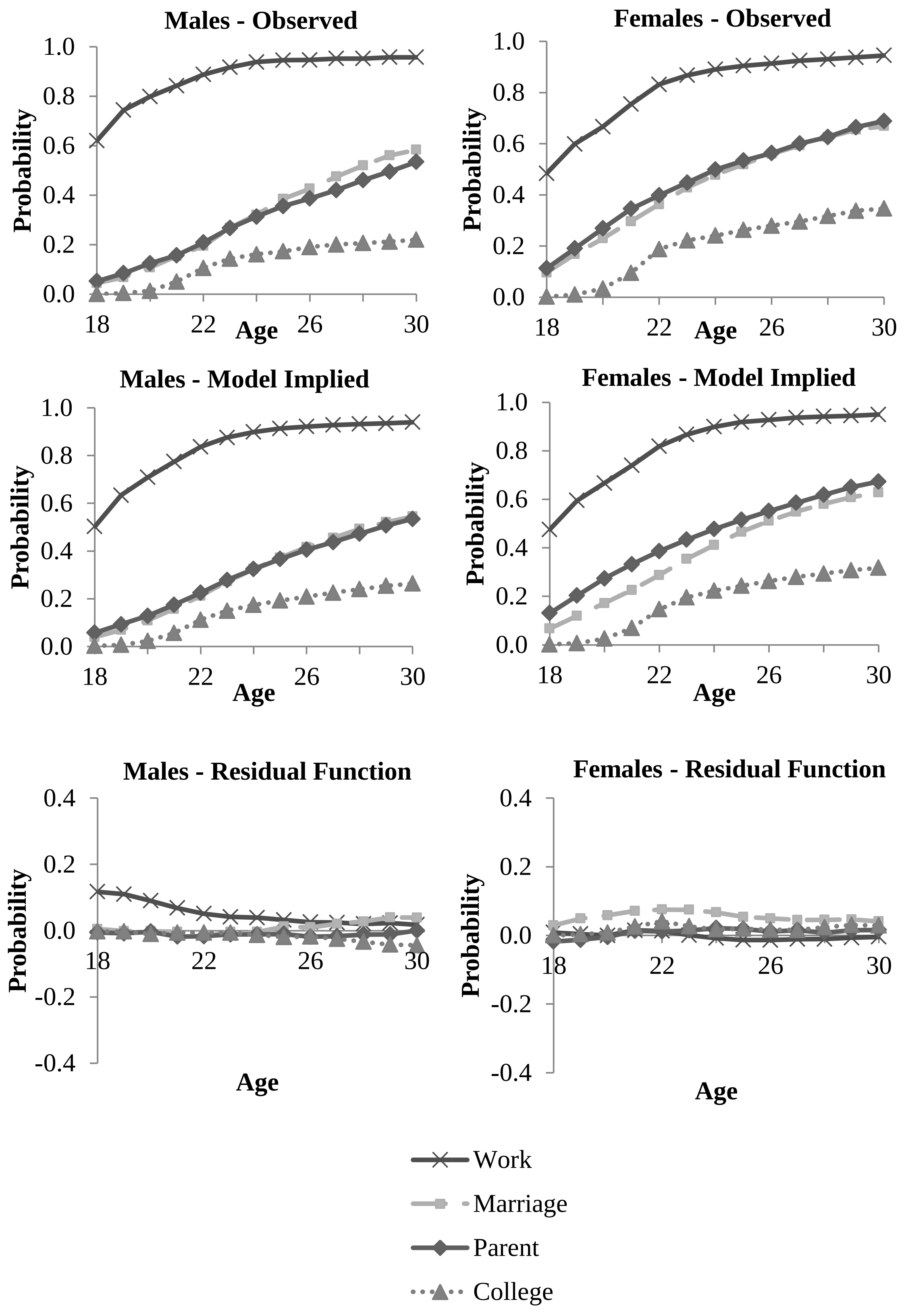
Parental education had an extremely consistent effect, in that the odds for individuals who had at least one parent with a college degree of being in a college pathway compared to any other pathway were significantly higher than for individuals who had a parent with a high school degree only. For example, the odds of being in the college then family pathway compared to the early parenthood pathway were 4.49 times as great for those individuals who had a parent with a college degree than those individuals who had a parent with a high school degree only. Significant differences were also found between those individuals who had neither parent graduate high school and those individuals who had at least one parent receive a high school degree only. Specifically, individuals who had neither parent graduate high school are more likely to be in the early parenthood and the work then family pathway versus the college and work pathway than individuals who had at least one parent receive a high school degree only.

In the spirit of an indirect application, the influence of covariates will also be examined by comparing aggregate model implied lifetime distribution functions weighting over latent classes for different levels of the covariates in the model. In order to compute the model implied hazard or lifetime distribution functions, the predicted probabilities of class membership can be found using only Equation (34) in this case, as covariates only affect the probability of class membership. For example, Caucasian females with a parent with a high school degree have a predicted probability of 0.21 of being in class 1 (“work then family” pathway), versus Caucasian males with a parent with a high school degree who have a predicted probability of 0.27 of being in that class. Once predicted probabilities have been computed for all of the classes, the model implied lifetime distribution functions can be found by weighting the within class lifetime distribution functions by the predicted probability of belonging to that class and then summing across latent classes, as in Equation (38).

Model implied lifetime distribution functions are computed for males and for females, holding race constant at Caucasian and parental education constant at high school degree (second row of Figure 13). The model implied lifetime distribution functions for work are consistent across gender for each age, with the cumulative probability of work by age 30 almost reaching unity ($\hat{D}_{30}$ = 0.94 for males and $\hat{D}_{30}$ = 0.95 for females). The model implies that females are more likely to be a parent by each age, such that by age 30, the cumulative probability of parenthood for males is 0.53 versus 0.67 for females. The model also implies a higher probability of becoming married by each age for females, with the median age of marriage between 25 and 26 for females versus between 28 and 29 for males. The cumulative probability of obtaining a college degree is slightly higher for females as implied by the model, with $\hat{D}_{26}$ = 0.26 for females and $\hat{D}_{26}$ = 0.21 for males.

The Add Health sample itself was then stratified by gender, dropping individuals who are not Caucasian and those whose parent has either no degree or a college degree in order to compute sample observed lifetime distribution functions with which to compare. This results in a sample size of 944 for males and 1,116 for females. Sample estimated lifetime distribution functions are displayed in the first row of Figure 13. Residual lifetime distribution functions are then calculated as the difference between the sample observed functions and the model implied functions and are displayed in the last row of Figure 13. The average difference between the sample observed and model implied lifetime distribution functions is quite small on average, *ARD* = 0.02, for both males and females.

Figure 13: *Add Health lifetime distribution functions depending on gender*



The model is worst at capturing the lifetime distribution function of full-time work for males, as it predicts males have a 0.50 probability of beginning work at age 18 or earlier while the sample observed function reveals the probability is 0.62. For females, the difference between the model implied lifetime distribution functions and sample observed functions is largest for marriage, with the model slightly underestimating the cumulative probability of marriage by each age (e.g. model implies $\hat{D}_{22}$ = 0.29 while sample estimated is D_{22} = 0.36).

However, the model is able to correctly capture many differences between the male and female lifetime distribution functions. For example, it captured that females are more likely at each age from eighteen to thirty to be a parent than males (holding race constant at Caucasian and parent education constant at high school degree). Specifically, for females, the model implied cumulative probability of parenthood by age 30 is $\hat{D}_{30}$ = 0.69, sample observed D_{30} = 0.67; for males, the model implied cumulative probability of parenthood is $\hat{D}_{30}$ = 0.53, sample observed D_{30} = 0.53. Similarly, the sample observed functions concur with the trend implied by the model that females are more likely to graduate college by each age than males (e.g. sample observed D_{26} = 0.28 for females and D_{26} = 0.19 for males) and that females are more likely to be married by age 30 than males (sample observed D_{30} = 0.67 for females and D_{30} = 0.58 for males).

Model implied lifetime distribution functions were also computed across the different races, keeping gender constant at male and parent education constant at high school degree (second row of Figure 14 and second row of Figure 15). The model implies that African-Americans and Hispanics have a higher probability of becoming a parent at early ages ($\hat{D}_{21}$ = 0.25 $\hat{D}_{21}$ = 0.20, respectively) compared to Caucasians and those of other races (both $\hat{D}_{21}$ = 0.17). However, the model implies the cumulative probability of becoming a parent by age 30 is relatively constant across races, all around 0.50. The model also implies that the lifetime distribution functions for full-time work are relatively constant across races. In contrast, the model implied cumulative probability of entering into marriage by age 30 is smaller for African-Americans ($\hat{D}_{30}$ = 0.40) than Caucasians ($\hat{D}_{30}$ = 0.54), Hispanics ($\hat{D}_{30}$ = 0.47), or those of other races ($\hat{D}_{30}$ = 0.50). The model also implied the cumulative probability of college by age 30 was higher for Caucasians ($\hat{D}_{30}$ = 0.26) and for other races ($\hat{D}_{30}$ = 0.28) than for African-Americans ($\hat{D}_{30}$ = 0.20) or Hispanics ($\hat{D}_{30}$ = 0.21).

The sample observed lifetime distribution functions were computed across races and compared to the model implied functions to assess fit (first row of Figure 14 and first row of Figure 15). Examining only males with a parent with a high school degree for comparison purposes resulted in sample size of 944 for Caucasians, 341 for African-Americans, 257 for Hispanics, and 143 for those of other races. The same trends for parenthood discussed above were found in the sample observed functions, in that African Americans and Hispanics were more likely to be a parent at an earlier age and that the cumulative probability of parenthood was relatively constant across races by age 30. The overall conclusions about differences between races on the cumulative probability of event occurrence across time were most different in the lifetime distribution function of beginning full-time work. As mentioned above and can be seen in the four different graphs, the model implied cumulative probability of beginning full-time work by any age was virtually identical across the different races; however, the sample observed functions revealed that African-Americans were much less likely to have begun full-time work at early ages (e.g. D_{18} = 0.42) than other races (Caucasians D_{18} = 0.62; Hispanics D_{18} = 0.60; other races D_{18} = 0.54), but that these differences decreased over time.

The sample observed functions for marriage were consistent with the model implied functions in they indicated African-Americans were less likely to be married by age 30 than other races; however, the model actually overestimated the rate of marriage for African-Americans (model implied $\hat{D}_{30}$ = 0.40; sample estimated D_{30} = 0.28). The sample-estimated functions for college were also consistent with the trend found in the model implied estimates in that African-Americans and Hispanics were less likely to have graduated college by age 30 than Caucasians or those of other races. Again, however, the model actually underestimated the differences, in that it overestimated the cumulative probability of college by age 30 for African-Americans and Hispanics (African Americans model implied $\hat{D}_{30}$ = 0.20 and sample observed D_{30} = 0.15; Hispanics model implied $\hat{D}_{30}$ = 0.21 and sample observed D_{30} = 0.16). Overall, the average amount of discrepancy between the sample observed lifetime distribution functions and model implied lifetime distribution functions across the four race categories was small, *ARD* =0.04 (Caucasian *ARD* = 0.02, African-American *ARD*=0.08, Hispanic *ARD* = 0.02, other race *ARD* =0.04).

Figure 14: *Add Health lifetime distribution functions depending on race (Caucasian and African-American)*

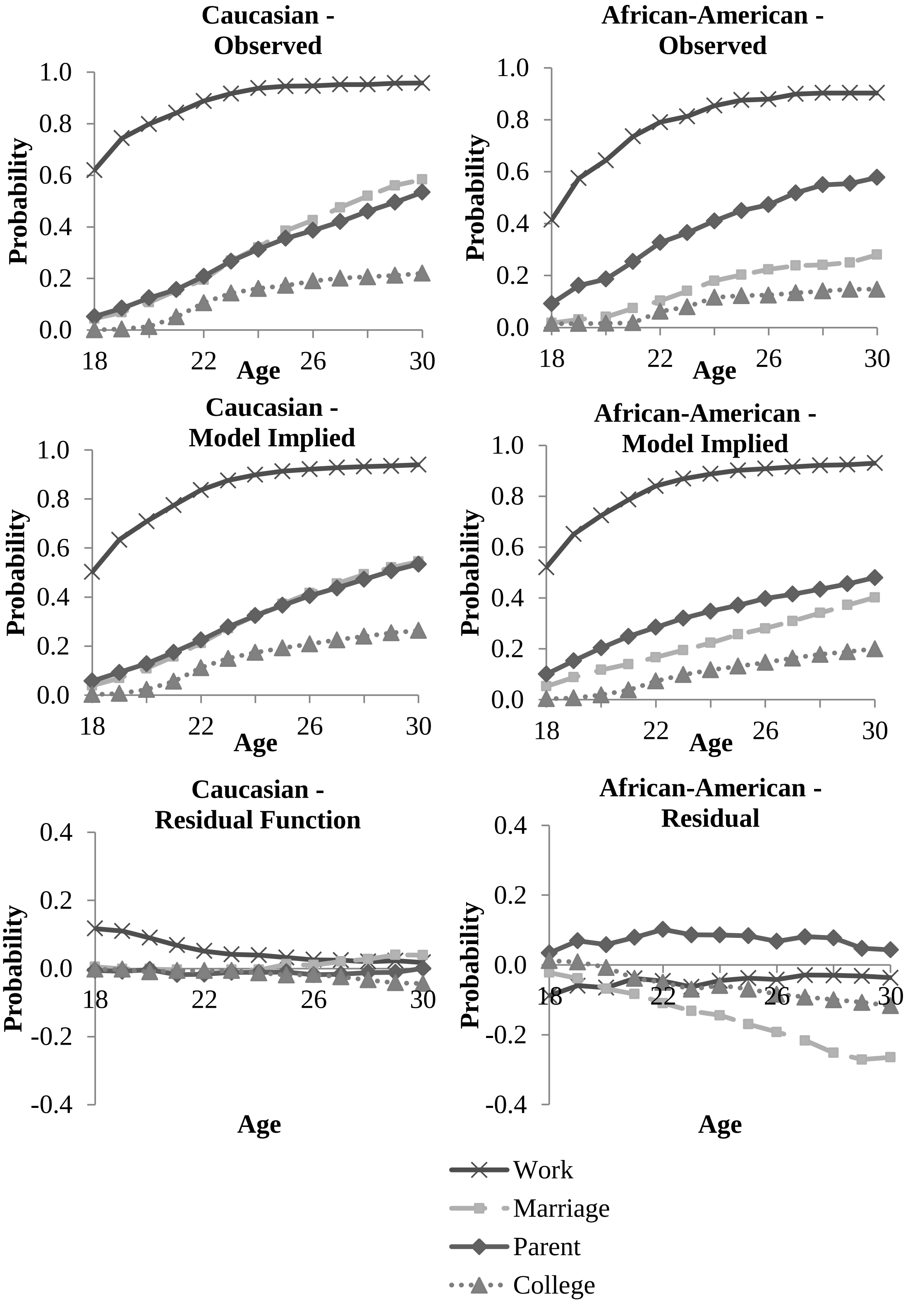
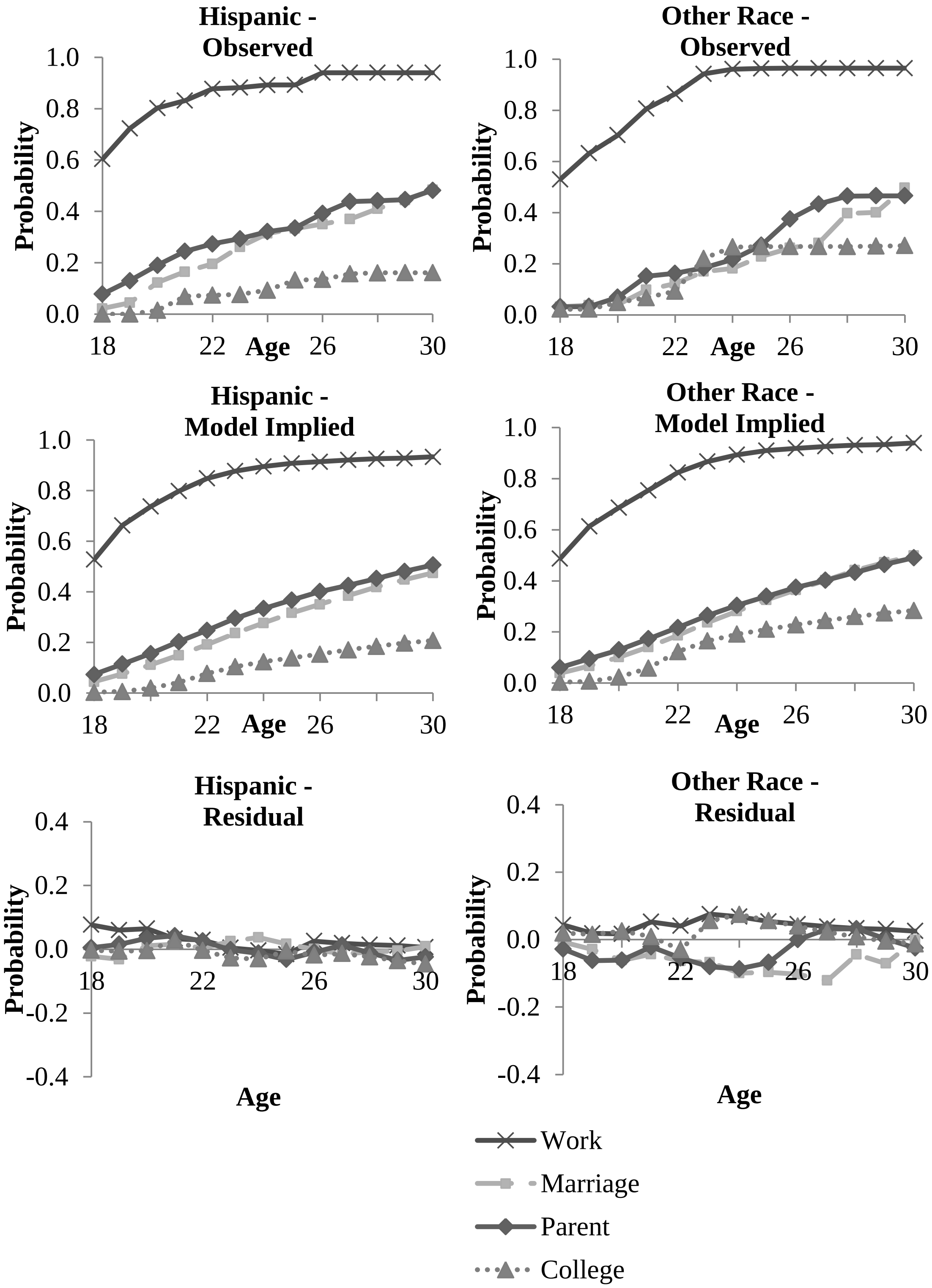


Figure 15: *Add Health lifetime distribution functions depending on race (Hispanic and other race)*

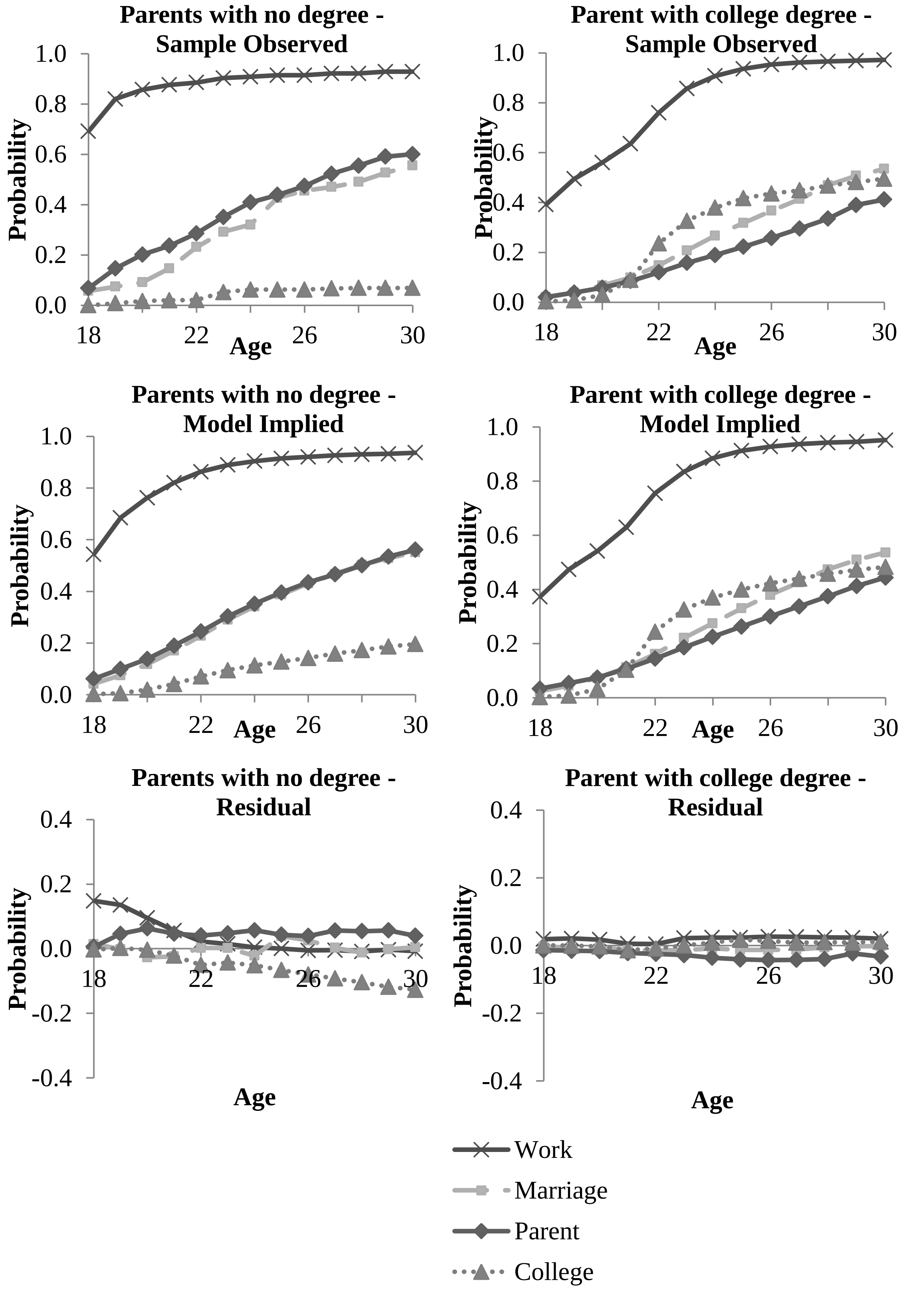


The last set of model implied lifetime distribution functions was computed across different levels of parental education, holding gender constant at male and race constant at Caucasian (Figure 16). The most dramatic difference between these functions is in terms of the cumulative probability of graduating college; individuals with a parent with a college degree have a much higher probability of graduating college by age 30 ($\hat{D}_{30}$ = 0.48) than individuals with a parent with a high school degree ($\hat{D}_{30}$ = 0.26) or no parent completing a high school degree ($\hat{D}_{30}$ = 0.20) as implied by the model. Related, the model predicts individuals who have a parent with a college degree have a smaller probability of beginning full-time work at earlier ages (e.g. $\hat{D}_{18}$ = 0.37) than individuals who have a parent with a high school degree ($\hat{D}_{18}$ = 0.50) or no degree ($\hat{D}_{18}$ = 0.54), but that there are no virtually no differences after age 24. The model implies that individuals who have a parent with a college degree also have a smaller risk of parenthood across all ages, and a smaller risk of marriage at earlier ages, but that the cumulative probability of marriage by age 30 is similar across parental education groups (range for $\hat{D}_{30}$ = 0.54-0.55).

Stratifying the Add Health sample by parent education and examining only Caucasian males for comparison purposes resulted in a sample size of 222 for neither parent with a high school degree, 944 for at least one parent with a high school degree only, and 2,536 for at least one parent with a college degree. The trends described by the model implied functions were found in the stratified sample observed functions in that those who had a parent with a college degree were much more likely to graduate college (D_{30} = 0.50) than for individuals who had neither parent graduate high school or at least one parent graduate high school but who had no further education (D_{30} = 0.07 and D_{30} = 0.22, respectively).

Note, however, that the model underestimated differences between these groups in that it overestimated the probability of graduating college for those with neither parent graduating high school (model implied $\hat{D}_{30}$ = 0.20; sample estimated D_{30} = 0.07). The trend was also consistent between the model implied and sample observed functions for work, with individuals with a parent with a college degree having a delay in the transition to full-time work (D_{18} = 0.39 versus parent with a high school degree D_{18} = 0.62). Also as implied by the model, individuals with a parent with a college degree had a smaller probability of parenthood across all ages as well as a smaller probability of marriage at early ages. Overall, the average difference between the model implied functions and the sample observed functions across the three parental educations categories was small, *ARD* = 0.03.

Figure 16: *Add Health lifetime distribution functions depending on parental education*



It is unclear whether the differences found between the sample observed lifetime distribution functions and the model implied functions weighting over latent classes are due to utilizing relatively few classes to capture the multivariate distribution of events, or due to possible minor misspecifications in the inclusion of covariates only in the multinomial model for class membership, as well as the fact interactions between covariates were not investigated. However, considering the small number of covariates included in the model, the mo del appears to be relatively stable and to be reproducing the observed patterns well.

3.3 Discussion

A five class solution was chosen for optimally describing heterogeneity in the hazard functions over time. The first class of the five class solution can be described as a work then family pathway, as it is characterized by transition into full-time work in the early twenties, followed by a high probability of transition into marriage and parenthood roles. Graduating college and transitioning into full-time work in the mid-twenties and then later transitioning into family roles characterize the second class, labeled a college then family pathway. The third class is also characterized by graduating college and transitioning into full-time work, but has a much lower probability of transitioning into marriage and parenthood roles by age 30, and is labeled a college and work pathway. The fourth class is characterized by a large probability of transitioning into parenthood by the early twenties, versus the last class, which is characterized by a large probability of transitioning into full-time work with a very low probability of transitioning into any other role.

These latent pathways capture heterogeneity in the risk for transitioning into multiple adulthood roles over time, and capture interdependence between the events through the delineation of latent classes. While useful in this way, we must be mindful that these pathways do not determine that there are truly only five transitions into adulthood. Rather, the pathways identified are prototypical pathways that heuristically and parsimoniously summarize the multivariate distribution of hazard functions for these measures.

The small number of covariates that were examined limits the substantive conclusions that can be drawn from this analysis, as there are certainly other variables that influence the probability of being assigned to one latent class over another. Additionally, interactions between gender, race, and parental education in predicting class membership may be of interest for future studies. However, the purpose of this empirical demonstration was to highlight the potential usefulness of the model for future research and also to investigate the ability of the model to detect group differences in the pathways over time. To this end, the limited number of covariates allowed empirical comparisons between the model implied lifetime distribution functions weighting over latent classes and sample observed functions in order to investigate the ability of the model to detect group differences in the risk of the events over time.

Considering the small number of covariates in the model, there is general consistency between the model implied and sample observed functions in the overall conclusions that were drawn, and the model does well at capturing overall differences in event occurrence across the ages examined. In terms of gender, the model detected that females are more likely to marry at earlier ages than males, as well as more likely to become a parent at earlier ages than males. This is consistent with previous literature on the transition into marriage and parenthood (e.g. Mahaffy, 2003; Martin et al. 2011). Additionally, the model implied that females are slightly more likely to earn a college degree consistent with other literature (e.g. Snyder & Dillow, 2011). The model also gives a larger picture into gender differences for all the events simultaneously by examining how females and males differ in the probability of taking one pathway over time versus another. The model generally predicted females are more likely to take an early parenthood pathway compared to most other pathways than males, and that females are more likely to belong to either of the identified college pathways versus the work-only pathway than males.

The model found African-Americans were less likely to enter into marriage and less likely to earn a college degree by age 30 than Caucasians. The model also implied African Americans were more likely to be in the early parenthood pathway as well as the work only pathway compared to Caucasians. Similarly, Hispanics were found to be more likely to be in the early parenthood pathway or work only pathway compared to the college with family pathway than Caucasians. It again should be noted that a very small number of covariates were entered into this model, and that racial differences found are not controlling for factors other than gender and parental education. For example, Ahituv, Tienda, and Hotz (2000) found that controlling for covariates such as income, test scores, parent education, and family structure, racial differences in school and work choices in the transition to adulthood largely disappear, and may even reverse direction. However, the results found are consistent with literature on racial differences in the transition to adulthood, without controlling for other factors (e.g. Snyder & Dillow, 2011; Ahituv et al., 2000).

The last predictor, parental education, had a consistent effect on the risk of class membership in that individuals with at least one parent with a college degree were more likely to take pathways involving college graduation and were more likely to have a delayed transition into full-time work. The cumulative probability of marriage by age 30 was similar across levels of parental education, but individuals with at least one parent with a college degree had a smaller probability of marriage at earlier ages, which could reflect postponement due to higher education.

While the influence of covariates here was found to be consistent with previous literature, examining predictors using the MEPSUM model is consistent with life course theory and the need to consider multiple life course roles simultaneously. The significance of a role depends on the role configuration, and the model avoids dissecting the life course in order to apply more traditional methods such as a univariate survival analysis. Also, by examining the multidimensional nature of the life course, the model gives insight into the possible mechanisms leading to differences in life course pathways. It is possible that a covariate influences the multivariate distribution of the risk of multiple events in a way that does not lend itself to be discovered by traditional methods that analyze events one at a time. For example, a covariate might increase the risk of transitioning into family roles for those who do not pursue college education but decrease the risk of transitioning into family roles for those pursuing a college education. Thus, the added complexity of the MEPSUM model has potential to increase our understanding of multiple transitions over time.

CHAPTER 4

EMPIRICAL EXAMPLE 2 – SUBSTANCE USE ONSET

Similar to the example in Chapter 4 where previous research has shown that the transition into different adulthood roles is interrelated, research on drug use is founded on the notion that the initiation of different substances are related to each other. This is conceptualized in the literature through the hypothesized existence of patterns of drug use, where the use of one drug is thought to be related to the subsequent use of another drug (Yamaguchi & Kandel, 1984). One popular theory, termed the gateway drug hypothesis, posits that the use of "softer" licit drugs leads to "harder" illicit drugs (Hamburg, Kraemer, & Jahnke, 1975). Alcohol, tobacco, and marijuana are commonly cited "gateway" drugs. For example, Wagner and Anthony (2002) found that tobacco and alcohol users were more likely to try marijuana than non-users, and likewise that marijuana users were more likely to try cocaine than non-marijuana users. They attributed the relationship in part to the "exposure opportunity" that occurs during the use of one drug, in that users of a drug are more likely to be offered a chance to try another drug.

The validity of the gateway drug hypothesis is still a point of contention (Golub and Johnson, 2001; Fergusson, Boden, & Horwood, 2006), but most researchers agree that the use of different drugs occur in clusters, and thus that the use of different drugs is related (Yamaguchi & Kandel, 1984; Hamburg, Kraemer, & Jahnke, 1975). Research has also focused on whether patterns of drug use vary over gender and racial groups. For example, Kandel and Logan (1984) found that overall patterns of drug use were similar for men and women, but that men were more likely to initiate all drugs. Vaughn, Wallace, Perron, Copeland, and Howard (2008) found that African-Americans were significantly more likely to initiate marijuana use before cigarettes compared to other ethnic groups, implying the patterns of drug use may differ depending on race.

The purpose of this example is to apply the discrete-time multiple event processes mixture model to drug use data to delineate patterns of drug use over time. The method is a novel way to examine a common research question in the drug use literature. While not a direct test of the gateway drug hypothesis, the MEPSUM model is useful substantively as a hypothesis generating method regarding the mechanisms leading to different patterns of drug use. This example will also be useful methodologically in investigating the utility of the model when numerous event processes are being studied and also when some of the event processes have a low hazard rate over all of the time periods. Other contrasts to the example in Chapter 3 are that these data necessitate the use of parametric hazard functions and that the larger sample size allows illustration of potential cross-validation procedures.

4.1 Methods

4.1.1 Data

The data for this example come from the 2009 National Survey on Drug Use and Health (NSDUH). The NSDUH is an annual survey providing national and state-level data on mental health and the use of both licit and illicit substances on randomly selected individuals twelve years of age or older. The data is available publicly through the Substance Abuse and Mental Health Data Archive (SAMHDA). The survey has four main objectives: 1) provide data on the patterns of substance use; 2) track trends in the use of various substances; 3) assess the consequences of substance use; and 4) identify groups at high risk for drug abuse (United States Department of Health and Human Services, 2010). The 2009 NSDUH recorded data from 55,772 individuals and includes information on the age of first use as well as lifetime, annual, and past-month usage for nine classes of substances: alcohol, cocaine, hallucinogens, heroin, inhalants, marijuana, non-medical use of prescription drugs (NMUP), stimulants, and tobacco. See Table 10 for a listing of the substances included in each class of drugs.

Table 10: *Substances included in each class of drugs in NSDUH*

Class	Included Substances
Alcohol	Beer, wine, and liquor
Cocaine/Crack	Cocaine powder, "crack," free base, and coca paste
Hallucinogens	LSD, PCP, peyote, mescaline, psilocybin, and ecstasy
Heroin	Heroin
Inhalants	Amyl nitrete, gasoline, glue, halothane, lighter gas, spray paints
Marijuana	Marijuana and hashish (also known as pot or grass)
NMUP	Nonmedical use of pain relievers, tranquilizers, sedatives
Stimulants	Methamphetamine, desoxyn, and methedrine
Tobacco	Cigarettes, chewing tobacco, snuff, cigars, and pipe tobacco

4.1.2 Measures

Age at time of interview was measured as a categorical variable, with categories representing each age from 12 years old to 21 years old, and categories of increasing width for 22 years of age and older. Individuals will be assigned the lowest age of the category to which they belong, and will be considered as censored for all ages after.[6]

[6] If any age other than the lowest age of the category was used, it would be implicitly assumed for anyone who was actually censored (had age been measured in integer values) that the event did not occur at all time periods, which could introduce a negative bias in the hazard probabilities. For example, if individuals in age category 24-25 were assigned the age of 25, the 25 year olds would have their data correctly measured, but the 24 year olds who had not experienced the event by age 24 would be assumed not to have

See Table 11 for the number of individuals in each age category. While there are many individuals who are older than thirty in the study, the events will only be examined to that age, as onset becomes increasingly less likely later in life.

Table 11: *Age at time of interview for individuals sampled in 2009 NSDUH*

Age	Frequency	Cumulative Percent
12	2561	4.59
13	2775	9.57
14	2930	14.82
15	3134	20.44
16	3128	26.05
17	3177	31.75
18	2716	36.62
19	2554	41.19
20	2344	45.40
21	2351	49.61
22-23	4591	57.84
24-25	4452	65.83
26-29	2702	70.67
30-34	2928	75.92
35-49	7863	90.02
50-64	3461	96.23
65+	2105	100.00

For each of the nine classes of drugs mentioned above, the variable that will be utilized is the age of first use of any of the substances included in the class. Thus there are nine event processes under study: time to first use of each class of substances. A summary of missing data, including the number of individuals with censored event times for each event process, is listed in Table 12. A binary variable was created for each of the nine event processes across ages 10 to 30 indicating whether the event had not yet occurred by that age (coded as 0), occurred at that age (coded as 1), and missing

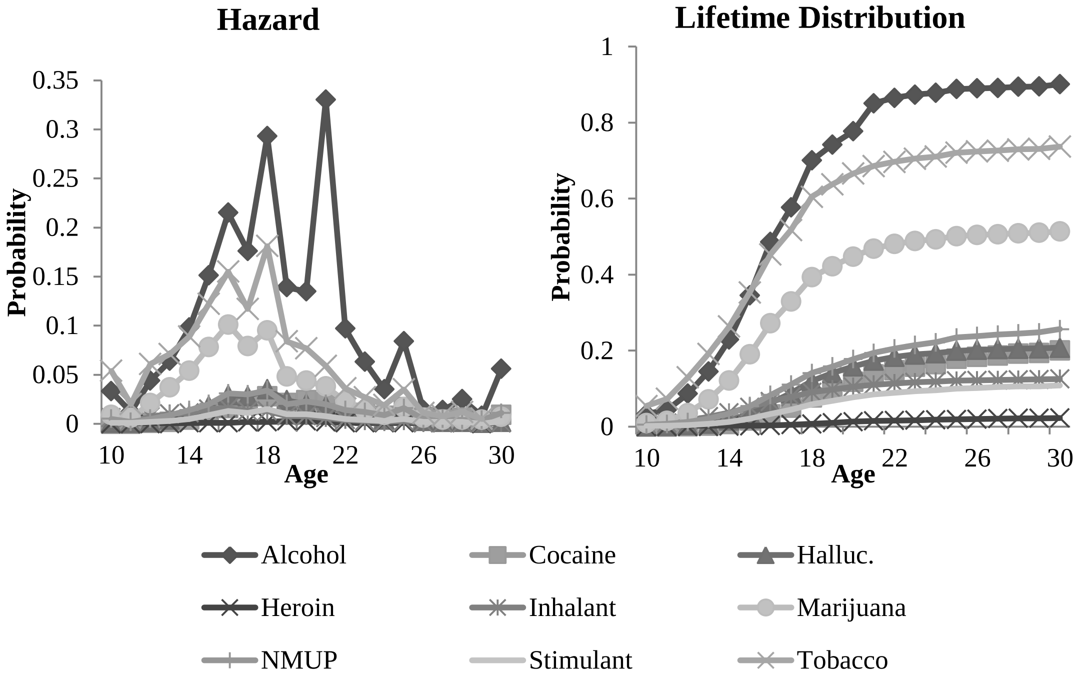
an event occurrence at age 25 when in reality they are just censored and thus should not have a value for that age.

otherwise. To account for the fact some individuals experience each event before the age of ten, the binary variable at age ten will represent whether the event occurred at that age, or any earlier age. The sample observed hazard and lifetime distribution functions are displayed in Figure 17.

Table 12: *Number of individuals with missing data - NSDUH*

Class	Number Uncensored (%)	Number Censored (%)	Number Missing (%)
Alcohol	39595 (70.99%)	15553 (27.89%)	624 (1.12%)
Cocaine/Crack	6620 (11.87%)	49120 (88.07%)	32 (0.06%)
Hallucinogens	7721 (13.84%)	48019 (86.10%)	32 (0.06%)
Heroin	783 (1.40%)	54955 (98.54%)	34 (0.06%)
Inhalants	5299 (9.50%)	50081 (89.80%)	392 (0.70%)
Marijuana	22009 (39.46%)	33560 (60.17%)	203 (0.36%)
NMUP	10229 (18.34%)	45523 (81.62%)	20 (0.04%)
Stimulants	3998 (7.17%)	51692 (92.68%)	82 (0.15%)
Tobacco	32114 (57.58%)	23648 (42.40%)	10 (0.02%)

Figure 17: *NSDUH sample observed hazard and lifetime distribution functions*



Race and gender are the only covariates included in the model.[7] The sample consists of 26,744 males (47.95%) and 29,028 females (52.05%). Race is measured as a seven category item but will be recoded for parsimony into a four category item of 1) White (61.93%), 2) African-American (12.80%), 3) Hispanic (16.24%), and 4) Other (9.03%), as the race categories Native American, Native Pacific Islands, Asian, or more than one race each compose less than five percent of the sample.

4.1.3 Analysis

The sample was split randomly in half into an evaluation sample and validation sample (N = 27,886 for each). The discrete-time MEPSUM model proposed in Chapter 2 was fit to the evaluation sample using the robust maximum likelihood estimator (MLR) in Mplus 6.12. As a first step, a model was fit to the data on the nine event processes, without covariates and with unstructured hazard functions. Yet even for a one class model, the number of parameters for an unstructured discrete-time MEPSUM model in this empirical example is quite large at 189, due to the large number of time periods and events under study. This is magnified for each increase in the number of latent classes; for example, a five class model would have 949 parameters. Additionally, several of the events have a low risk of occurrence, which can result in convergence issues as mentioned in Chapter 1. Therefore, a parametric form for the hazard functions was considered next, but only after examining results when the unstructured form of the

[7] Demographic variables including education, marital, and work status assessed at the time of interview are available as well as mental health status at the time of interview, but these variables will not be utilized in this analysis as the age range of individuals sampled in the survey make it difficult to compare across these categories, using them as a time-invariant predictor. Also, as these variables are assessed at the time of interview, they would not be true predictors of the latent classes, and also are not traditional time-varying covariates as they are only available at one point in time. In the future, a possible extension would be to address the prediction of distal outcomes by the latent classes.

hazard functions was used in order to determine whether the shape is constant between classes.

Models were then fit with quadratic parametric hazard functions with an increasing number of latent classes, up to six classes. Time was scaled in decades rather than years for estimation purposes. Without any constraints on the parameters, the model had trouble estimating for three or more classes, as the risk of at least one event was so low that the parameters could not be identified without Mplus imposing automatic constraints. Following the default Mplus places on the boundary value of logits, the intercept factors were constrained to be greater than -15.

This could be considered an empirical under-identification problem due to the fact the logit scale is unbounded. For example, the following two sets of parameters would both imply the cumulative risk of event occurrence over 20 time periods is less than 0.001: 1) intercept = -20, slope = 0, quadratic = 0 and 2) intercept = -9, slope = 2, quadratic = -5. This under-identification may also be an issue in identifying the parameters when the risk of an event is very large at early time periods, as the number of people able to experience the event at later time periods grows smaller.

However, while this is an issue in identifying the parameters of the model, as long as reasonable constraints are imposed, this will not influence the hazard functions in a probability scale. Due to the fact the hazard function is being modeled and the transformed parameters cumulate to calculate the lifetime distribution function, a conservatively low lower bound should be used. Note that the reasonableness of this value may depend on the number of time periods. For instance, constraining the intercept to be greater than -5 and with a slope and quadratic function of 0 results in a cumulative lifetime distribution of 0.05 for 7 time periods, which may be fine for some applications, but is also equal to a cumulative probability of 0.13 for 20 time periods, which may be too large for other applications. In the empirical example in Chapter 4, a much lower value such as -15 was needed due to the fact several of the events had a very low risk of occurrence. Constraining the intercept in this example to be at least -15 still allowed the possibility of a cumulative lifetime distribution of effectively 0 by age 30.

Each model was first run with at least 100 random sets of starting values. The solution found was then used as starting values for another set of replications with at least 500 sets of random starting values, constraining the slope and quadratic factors to 0 for each event within a latent class where the risk of the event was less than or equal to 0.001 across time periods. This greatly increased the estimation time of the model, allowing more replications to ensure a global solution to the likelihood. The final solution found, including the constraints on the slope and quadratic factors for low risk events, were used as starting values for the next analysis with an additional latent class.[8]

A model was then fit where race and gender were used as predictors of class membership. The size and parameters of the classes were monitored for change when the predictors were added to the model to investigate the stability of the model. The final model was then fit to the validation sample with all parameters constrained to the solution found in the evaluation sample. Measures of discrepancy – *ARD* and *ARH* – between the model imposed functions weighting over latent classes and the validation sample observed functions were computed to cross-validate the model. Finally, the MEPSUM

[8] Sensitivity analyses indicated no difference between constraining the low risk events from the beginning and starting with no constraints.

model was also fit independently in the validation sample and parameter estimates were then compared between the evaluation and validation sample.

4.2 Results

First, the unstructured hazard function form of the MEPSUM model was fit to the data with an increasing number of latent classes and model fit information is listed in Table 13. Model estimation began to break down at 5 classes, as the loglikehood was not replicated even after several thousand random sets of starting values, and Mplus gave a warning of a possible non-positive definite information matrix. This is likely due to the large number of time periods and event processes, as well as the fact the risk of event occurrence for several of the events is quite low.

Table 13: *NSDUH model fit with unstructured hazard functions*

Latent Classes	-2LL	Number of Free Parameters	BIC	AIC	Smallest Class	Entropy
1	-243380.78	189	488696.15	487139.57	N/A	N/A
2	-222525.20	379	448929.80	445808.40	0.29	0.80
3	-217680.64	569	441185.51	436499.29	0.12	0.75
4	-215563.32	759	438895.67	432644.64	0.09	0.69

The shape of the hazard functions in the one to four class solutions consistently indicated that all of the hazard functions followed a quadratic form even within classes. For parsimony and to determine whether more latent classes are necessary to effectively describe heterogeneity in the hazard functions, a MEPSUM model with quadratic form for all of the hazard functions was then fit with one to six latent classes (Table 14). Across all solutions, the model with quadratic hazard functions aggregated back to the sample observed functions well, *ARH* < 0.02, *ARD* < 0.01.

Table 14: *NSDUH model fit with quadratic hazard functions*

Latent Classes	"-2LL"	Number of Free Parameters	BIC	AIC	Smallest Class	Entropy
1	-219814.6	27	494859.2	494636.9	N/A	N/A
2	-226663.5	55	453890.0	453437.1	0.29	0.80
3	-221771.0	71	444268.7	443684.0	0.14	0.73
4	-219857.6	99	440728.5	439913.2	0.09	0.67
5	-218232.9	125	437745.3	436715.8	0.07	0.66
6	-217353.4	151	436252.4	435008.8	0.06	0.66

Information criteria suggested the model with least six classes was optimal, while entropy continued to decrease indicating classification errors increased with the addition of latent classes. Examining the hazard and lifetime distribution functions more carefully, there was not one clear optimal number of classes. Thus, in the spirit of an indirect application, the four to six latent classes solutions are now described, as it is useful both to examine the progression of including more latent classes as well as the impact of the selection of the number of classes on the substantive conclusions about the risk of multiple events over time.

The first class ("relative abstainers," $\pi_1 = 0.382$) in the four class solution is characterized by a small risk of initiating alcohol use peaking at age 21 ($\hat{h}_{21} = 0.14$) with a relatively moderate cumulative probability of initiating alcohol use ($\hat{D}_{30} = 0.74$) and tobacco ($\hat{D}_{30} = 0.40$) by age 30. The median event time for initiating alcohol use is between age 20 and 21. The cumulative probability of initiating any substance other than alcohol or tobacco is near 0 by age 30. For the remaining three latent classes, the cumulative risk of alcohol use by age 30 is 1.00.

The second class ("soft drug users," $\pi_2 = 0.375$) has a high cumulative probability of tobacco use ($\hat{D}_{30} = 0.94$) and moderate cumulative probability of marijuana use ($\hat{D}_{30} =$ 0.65). The risk of tobacco and marijuana use peaks around age 19, and the median event time is between age 15 and 16 for alcohol and tobacco, and between 19 and 20 for marijuana use. The risk of non-medical use of prescription drugs (NMUP) is larger than the first class but still relatively small ($\hat{D}_{30} = 0.17$) and the risk of all other drug use is likewise small across all the time periods, with the cumulative probability less than 0.10 for all other drugs.

The third latent class ("later hard drug users," $\pi_3 = 0.154$) is characterized by a peak in alcohol, tobacco, and marijuana use around 18, with a greater than 0.90 cumulative probability of initiating each of these substances by that age. The median event time for alcohol and tobacco is between 14 and 15, and the median event time for marijuana is between 15 and 16. The risk of the remaining substances all peak around age 21, with a 0.65-0.75 cumulative probability of having used cocaine, NMUP, and hallucinogens by age 30, and around a 0.35 cumulative probability of inhalant and stimulant use, and 0.05 cumulative probability of heroin use by age 30. The median event time for hallucinogens is between 19 and 20, for cocaine is between 20 and 21, and for NMUP is between 21 and 22.

The final class ("early hard drug users," $\pi_4 = 0.089$) similarly has a high risk of alcohol, tobacco, and marijuana use, beginning even earlier than the third latent class as the cumulative probability reaches above 0.95 by age 14 for each of these substances. The median event time for alcohol and tobacco is between age 11 and 12, and for marijuana between age 12 and 13. The risk of first use of the other substances also peaks earlier than the third latent class, generally around age 18, and the cumulative probability of using each substance is likewise larger than the third class, with around 0.80 cumulative probability of cocaine, NMUP, and hallucinogens, around 0.50 cumulative probability of inhalant and stimulant use, and 0.18 cumulative probability of heroin use. The median event time for hallucinogens and NMUP is between age 16 and 17, followed by cocaine between age 17 and 19, and inhalants between age 19 and 20.

The pattern of the fourth class is thus very similar to the third, only that the probability of beginning substance use is higher at earlier ages in the fourth class and the cumulative probability of initiating use of each of the substances is higher across all time periods in the fourth class. However, one interesting difference in pattern is NMUP has an earlier median event time in relation to other substances in the early hard drug users class, compared to the median event time of NMUP in the later hard drug users class, where the median event time for cocaine and hallucinogens preceded it. The median event time for each event process within each latent class is listed in Table 15. See Figure 18 for the hazard functions and Figure 19 for the lifetime distribution functions for the four class solution. In all hazard function graphs that follow, the hazard functions are only graphed during time periods when the cumulative probability is less than 1, because the hazard is irrelevant once the cumulative probability reaches 1 (no one is eligible to experience the event) and this would be solely an extrapolation of the model beyond what the data reveals.

Table 15: *NSDUH median event time within each latent class for four class solution*

Event	Relative Abstainer	Soft Drug Use	Later Hard Drug Use	Early Hard Drug Use
Alcohol	20.5	15.5	14.5	11.5
Cocaine	-	-	20.5	17.5
Hallu.	-	-	19.5	16.5
Heroin	-	-	-	-
Inhalant	-	-	-	19.5
Marijuana	-	19.5	15.5	12.5
NMUP	-	-	21.5	16.5
Stimulant	-	-	-	-
Tobacco	-	15.5	14.5	11.5

Figure 18: *NSDUH hazard functions for the four class solution*

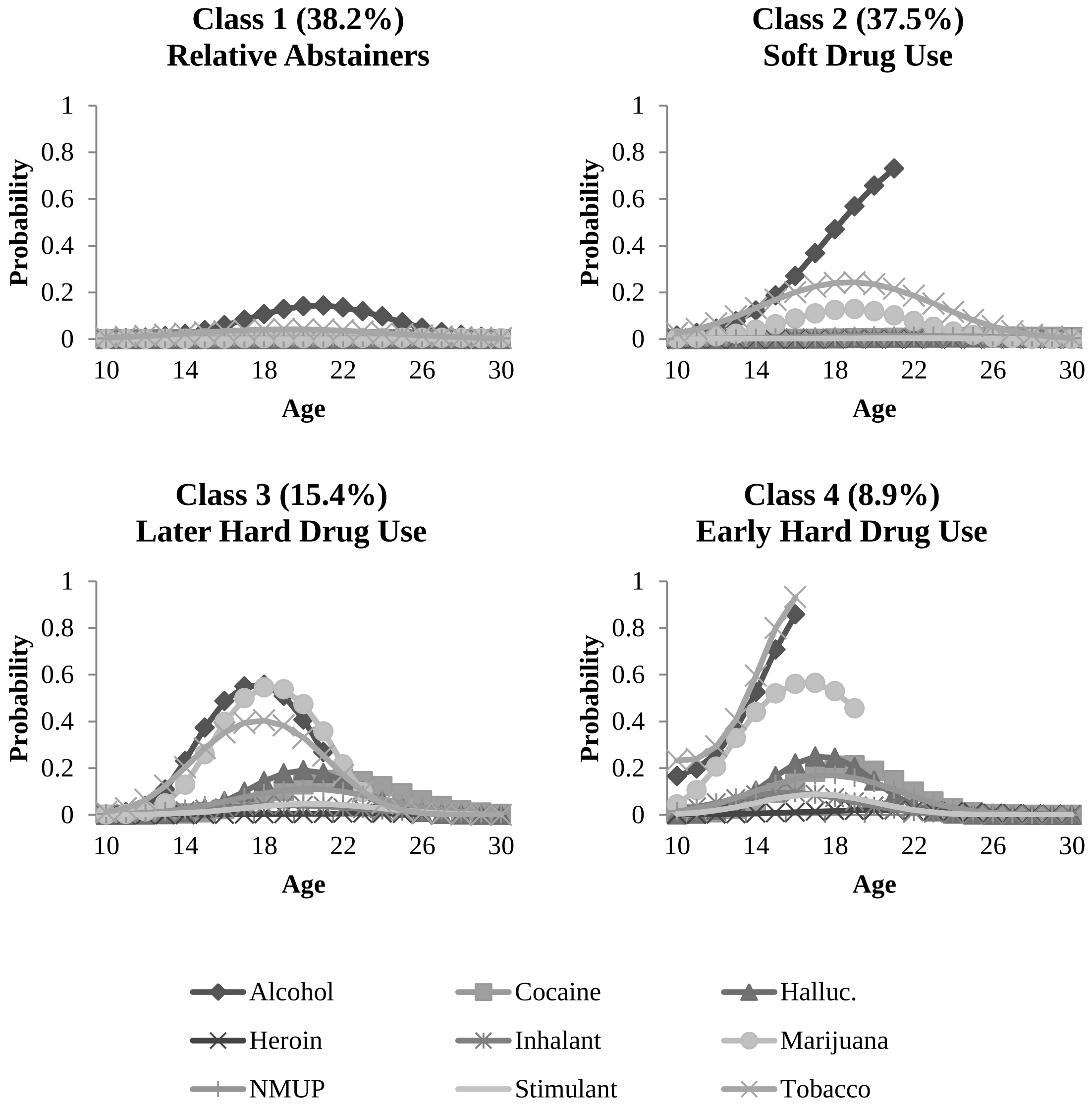
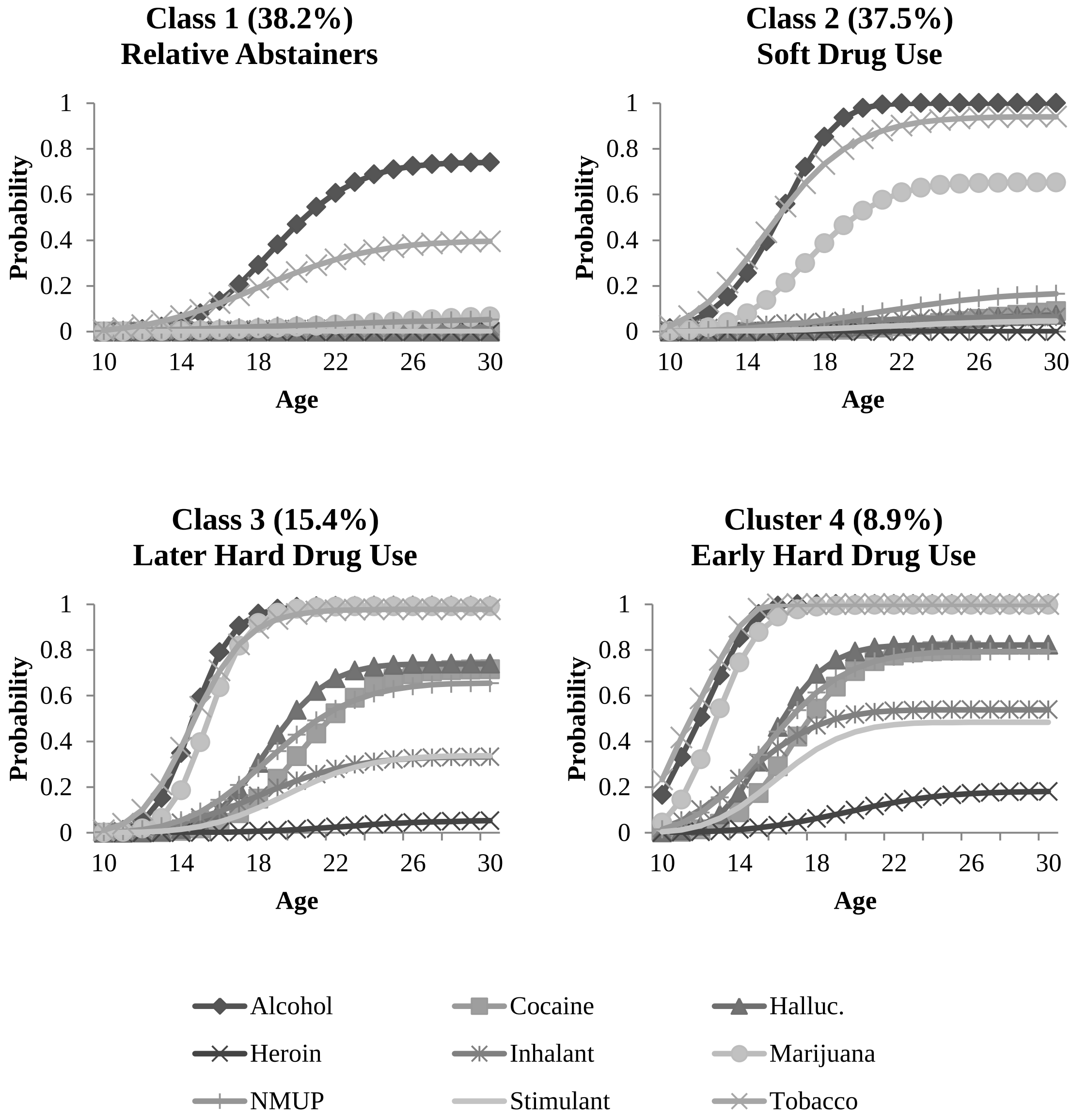


Figure 19: *NSDUH lifetime distribution functions for the four class solution*



In the five class solution, the first class (“relative abstainers,” $\pi_1 = 0.345$) again remained a class where the risk of ever trying alcohol by age 30 was relatively moderate ($\hat{D}_{30} = 0.72$), as was tobacco ($\hat{D}_{30} = 0.35$), with the risk of all other drug use less than 0.01 at any age. The second class of the five class solution (“soft drug users,” $\pi_2 = 0.308$) was also similar to the second class of the four class solution, with a small cumulative probability of trying any substance other than alcohol ($\hat{D}_{30}$ = 0.99), tobacco ($\hat{D}_{30}$ = 0.91), or marijuana ($\hat{D}_{30}$ = 0.55) by age 30. In comparison to the four class solution, the risk of marijuana use in this latent class was lower and the overall risk of all events had a more peaked shape rather than a more linear shape which would have indicated similar risk across all of the time periods. The median event time for alcohol (between age 16 and 17), tobacco (between age 16 and 17), and marijuana (between age 20 and 21) were all one year later than the "soft drug users" class in the four class solution.

The third class ("mainly soft drug users," π_3 = 0.138) is unique to the five class solution, with a more flat hazard function across the age ranges. While the risk is spread out over time, the cumulative probability of alcohol or tobacco use by age 30 is nearly 1, and the cumulative probability of marijuana use is also high ($\hat{D}_{30}$ = 0.82). The risk of using other substances is small at any individual age, but the cumulative probability for these other substances by age 30 is higher than the first two classes (e.g. cumulative probability of cocaine use is $\hat{D}_{30}$ = 0.28). The median event time for alcohol, tobacco, and marijuana is much earlier than the median event time for these substances in the "soft drug users" class (13.5, 12.5, and 16.5, respectively).

The fourth ("later hard drug users," π_4 = 0.142) and fifth class ("early hard drug users," π_5 = 0.067) from the five class solution are very similar to the third and fourth class from the four class solution. These classes start with a high risk of alcohol, tobacco, and marijuana use, and at a later age have a peak in the risk for first use of other substances. The fifth class has a higher risk of all substances at earlier ages than the fourth class and has especially high cumulative probabilities of ever trying other substances (e.g. probability of cocaine use is $\hat{D}_{30}$ = 0.90). See Figure 20 and Figure 21 for hazard and lifetime distribution functions for the five class solution, respectively and Table 16 for the median event time within each class.

Table 16: *NSDUH median event time within each latent class for five class solution*

Event	Relative Abstainer	Soft Drug Use	Mainly Soft Drug Use	Later Hard Drug Use	Early Hard Drug Use
Alcohol	20.5	16.5	13.5	14.5	12.5
Cocaine	-	-	-	21.5	16.5
Hallu.	-	-	-	19.5	15.5
Heroin	-	-	-	-	-
Inhalant	-	-	-	-	17.5
Marijuana	-	20.5	16.5	15.5	12.5
NMUP	-	-	-	20.5	15.5
Stimulant	-	-	-	-	19.5
Tobacco	-	16.5	12.5	14.5	11.5

Figure 20: *NSDUH hazard functions for the five class solution*

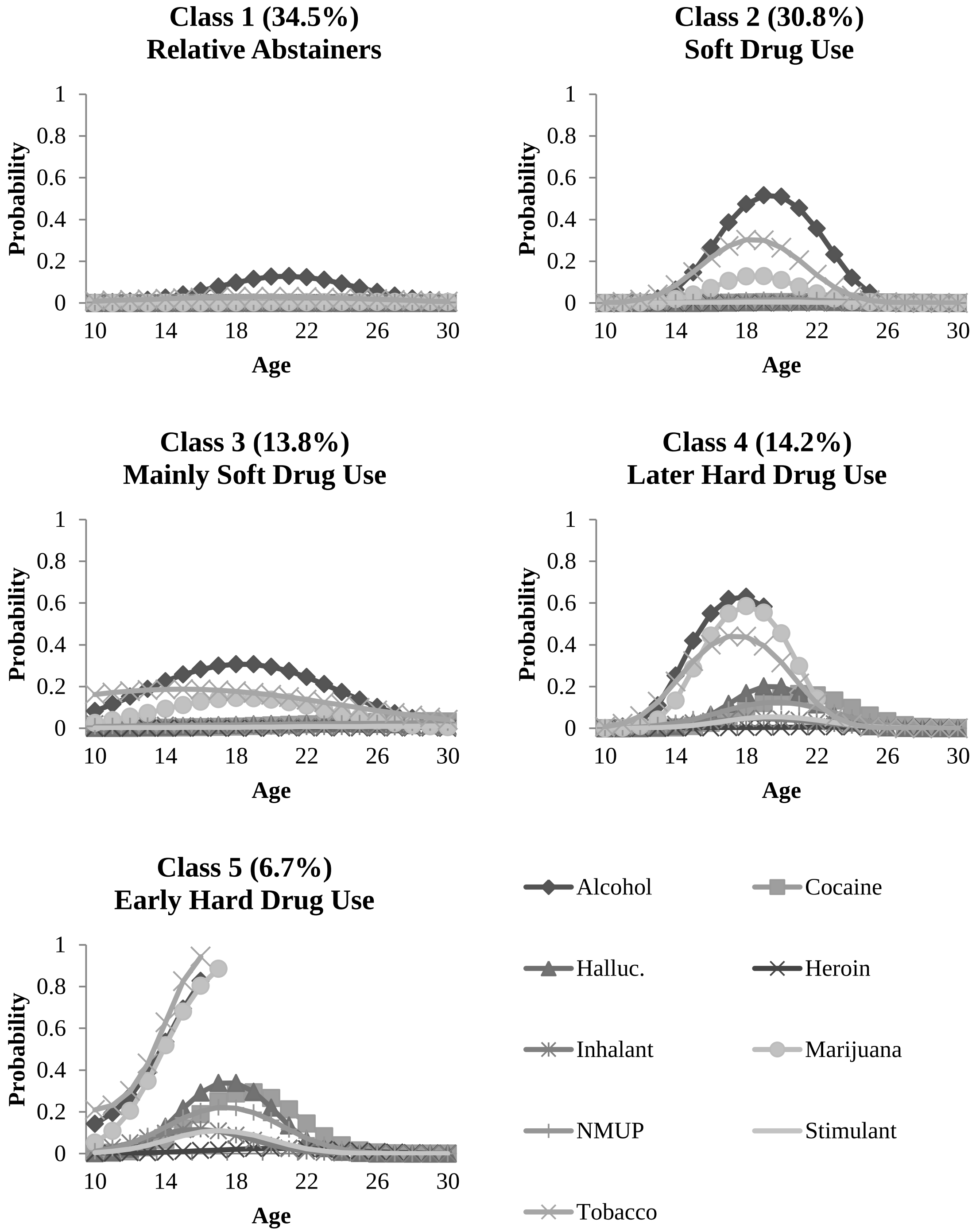
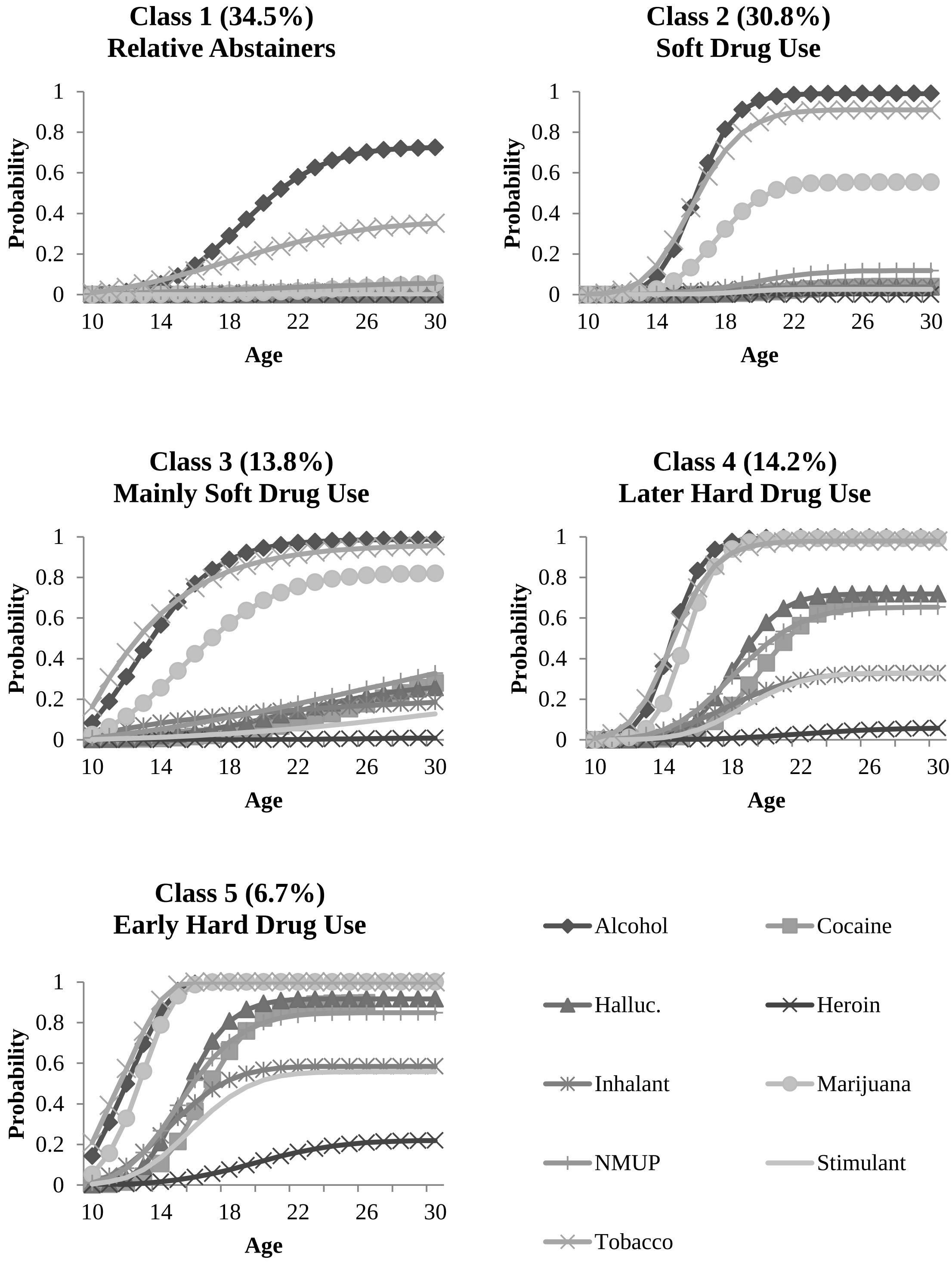


Figure 21: *NSDUH lifetime distribution functions for the five class solution*



Class 1 (34.5%)
Relative Abstainers

Class 2 (30.8%)
Soft Drug Use

Class 3 (13.8%)
Mainly Soft Drug Use

Class 4 (14.2%)
Later Hard Drug Use

Class 5 (6.7%)
Early Hard Drug Use

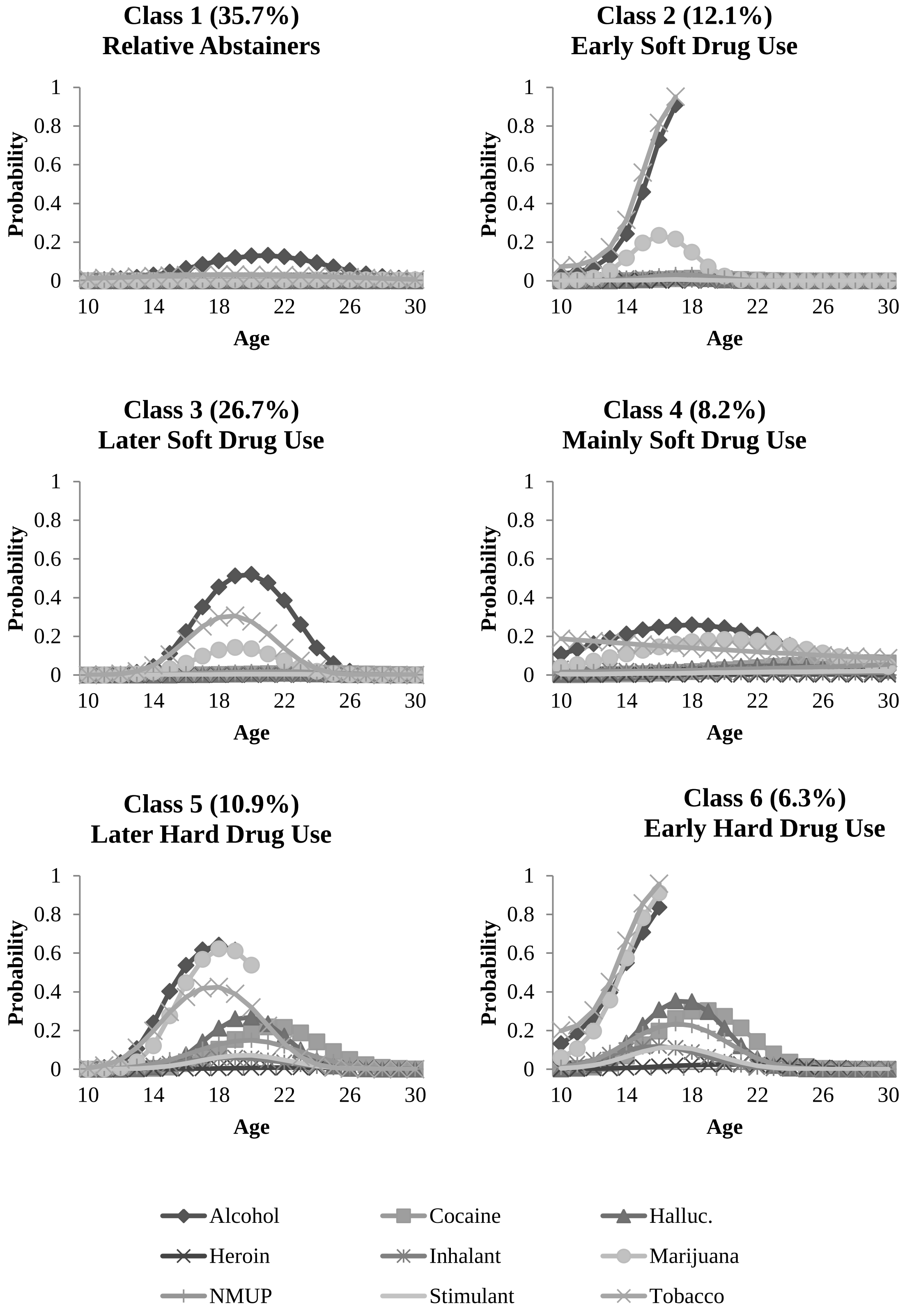
The main difference in the six class solution was that there were now two classes (class 2 and 3) which could be characterized as "soft drug users." Class 2 ("early soft drug users," π_2 = 0.121) has earlier median event times for alcohol (14.5), tobacco (13.5), and marijuana (16.5) than class 3 ("later soft drug users," π_3 = 0.267) which had median event times of 16.5, 17.5, and 20.5, respectively. The "early soft drug users" also had overall higher cumulative probability of marijuana use $\hat{D}_{30}$ = 0.69 versus $\hat{D}_{30}$ = 0.60. Note that while the differences were mainly in age and level, there was a slight difference in pattern in that the "early soft drug users" were more likely to initiate tobacco use at an earlier age than alcohol, versus the "later soft drug users" which had about the same risk for alcohol and tobacco initiation at early ages.

Class 4 ("mainly soft drug users," π_4 = 0.082) was also slightly different from the five class solution in that the cumulative probability of initiating hard drug use by age 30 was higher (e.g. cocaine $\hat{D}_{30}$ =0.48 versus four class solution where $\hat{D}_{30}$ = 0.28). The characteristic more linear risk for the events in this class remained the same as the five class solution. Class 1 ("relative abstainer," π_1 = 357), class 5 ("later hard drug users," π_5 = 0.109), and class 6 ("early hard drug users," π_6 = 0.063) remain virtually identical to the classes 1, 4 and 5 in the five class solution. See Table 17 for the median event times, Figure 22 for the hazard functions, and Figure 23 for the lifetime distribution functions within each latent class.

Table 17: *NSDUH median event time within each latent class for six class solution*

Event	Relative Abstainer	Early Soft Drug Use	Later Soft Drug Use	Mainly Soft Drug Use	Later Hard Drug Use	Early Hard Drug Use
Alcohol	20.5	14.5	16.5	13.5	14.5	12.5
Cocaine	-	-	-	-	20.5	16.5
Hallu.	-	-	-	-	18.5	15.5
Heroin	-	-	-	-	-	-
Inhalant	-	-	-	-	-	17.5
Marijuana	-	16.5	20.5	16.5	15.5	12.5
NMUP	-	-	-	-	19.5	15.5
Stimulant	-	-	-	-	-	19.5
Tobacco	-	13.5	17.5	12.5	14.5	11.5

Figure 22: *NSDUH hazard functions for the six class solution*



Class 1 (35.7%)
Relative Abstainers

Class 2 (12.1%)
Early Soft Drug Use

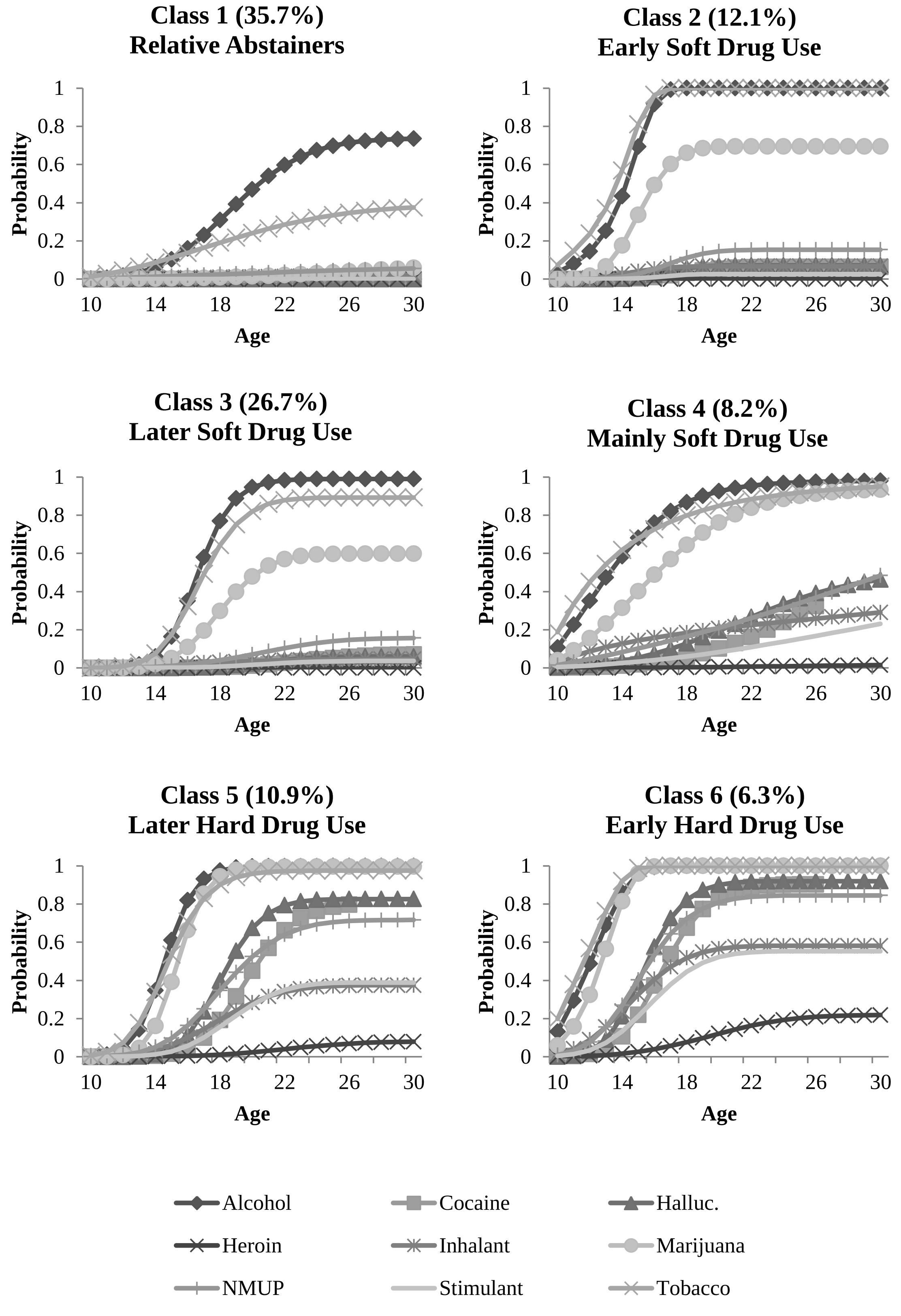
Class 3 (26.7%)
Later Soft Drug Use

Class 4 (8.2%)
Mainly Soft Drug Use

Class 5 (10.9%)
Later Hard Drug Use

Class 6 (6.3%)
Early Hard Drug Use

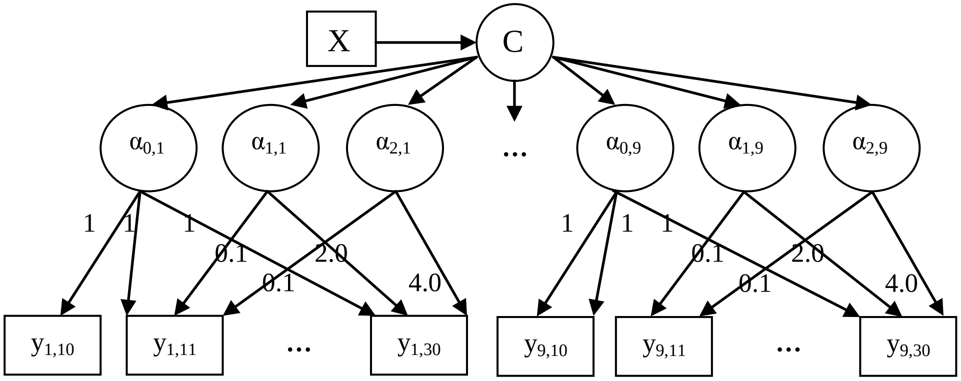
Figure 23: *NSDUH lifetime distribution functions for the six class solution*



Examining the four to six class solutions, we mainly find differences in the age at which risk is highest as well as the overall level of risk (e.g. large similarities between “early soft drug use” class and “later soft drug use class”), rather than large differences in the pattern of drug use. Depending on the end goal of the analysis, different solutions discussed above could easily be justified. It is argued here that the four class solution is the most clear and interpretable from a policy and intervention viewpoint; for example, being able to establish differences between early and later soft drug use might not warrant the added complexity of additional latent classes, as the pattern is similar between the two classes and the risk of hard drug use is small across all time periods. Instead, it might be useful to establish differences between the two classes of hard drug use as well as the differences between these classes and classes with a lower risk over time, and the four class solution allows us to parsimoniously investigate this.

Thus, covariates will be investigated with four classes only in order to keep this empirical example tractable. Additionally, with the small number of covariates in this analysis, the general conclusions drawn in the four class solution were similar for the five to six class solutions. The influence of covariates was again investigated both in terms of how covariates influence the probability of class membership and through the model implied population functions weighted over latent classes. Before doing so, however, the parameter estimates were compared from the solution with covariates influencing class membership (for a path diagram, see Figure 24) to the solution without covariates to check the stability of the hazard functions within latent class and the size of the latent classes.

Figure 24: *NSDUH simple path diagram of model with covariates*



The number of parameters is greatly reduced in comparison to the Add Health example in Chapter 3 due to the structured hazard functions. Therefore, in comparing the hazard functions from the model without covariates to the model found with covariates, it is reasonable in this case to directly compare the logit parameters (Table 18).

Table 18: *NSDUH logit parameters from model with covariates compared to model without covariates*

		Latent classes with covariates				Latent classes without covariates			
Event	Param.	1	2	3	4	1	2	3	4
Alcohol	α_0	-6.80	-4.29	-6.17	-1.64	-6.83	-4.24	-6.15	-1.62
	α_1	9.12	6.17	16.76	1.66	9.38	6.19	16.74	1.55
	α_2	-4.19	-1.30	-10.95	6.58	-4.35	-1.31	-10.97	6.91
Cocaine	α_0	-8.86	-10.41	-9.88	-6.96	-8.95	-10.30	-9.97	-6.95
	α_1	0*	6.69	13.64	13.19	0*	6.80	13.95	13.19
	α_2	0*	-1.98	-5.72	-7.67	0*	-2.04	-5.88	-7.70
Hall.	α_0	-8.79	-8.88	-9.61	-6.29	-8.84	-8.77	-9.78	-6.27
	α_1	0*	5.56	16.07	14.11	0*	5.69	16.62	14.09
	α_2	0*	-2.07	-7.97	-9.56	0*	-2.14	-8.29	-9.59
Heroin	α_0	-10.33	-10.83	-11.09	-7.41	-10.37	-10.72	-11.04	-7.39
	α_1	0*	0*	9.13	6.84	0*	0*	9.20	6.80
	α_2	0*	0*	-3.54	-3.32	0*	0*	-3.58	-3.30
Inhalants	α_0	-7.38	-5.25	-5.99	-3.91	-7.40	-5.21	-5.99	-3.89
	α_1	0*	0.12	6.15	5.89	0*	0.21	6.27	5.88
	α_2	0*	-0.69	-3.37	-5.22	0*	-0.71	-3.45	-5.24
Marijuana	α_0	-7.60	-6.10	-7.58	-3.09	-7.76	-6.14	-7.56	-3.07
	α_1	3.13	9.28	18.52	10.20	3.41	9.61	18.60	10.15
	α_2	-1.06	-5.29	-11.07	-7.74	-1.15	-5.47	-11.14	-7.71
NMUP	α_0	-6.39	-6.55	-6.28	-4.23	-6.49	-6.54	-6.24	-4.21
	α_1	1.45	3.85	8.08	6.95	1.56	3.99	8.10	6.91
	α_2	-0.74	-1.67	-3.91	-4.56	-0.77	-1.73	-3.94	-4.56
Stimulant	α_0	-8.31	-8.29	-7.93	-5.48	-8.33	-8.30	-7.93	-5.47
	α_1	0*	4.33	9.36	8.84	0*	4.57	9.48	8.86
	α_2	0*	-1.79	-4.47	-6.21	0*	-1.89	-4.56	-6.26
Tobacco	α_0	-5.20	-3.68	-4.48	-1.21	-5.08	-3.66	-4.52	-1.19
	α_1	4.14	5.72	10.44	-0.75	4.24	5.76	10.59	-0.86
	α_2	-2.18	-3.23	-6.66	11.52	-2.28	-3.29	-6.78	12.00

* = Parameter constrained

As can be seen, the parameters remain stable with the inclusion of covariates. Indeed, the correlation between the two sets of parameter values rounds to 1.00, implying the assumption that covariates only influence the probability of class membership had not been violated.

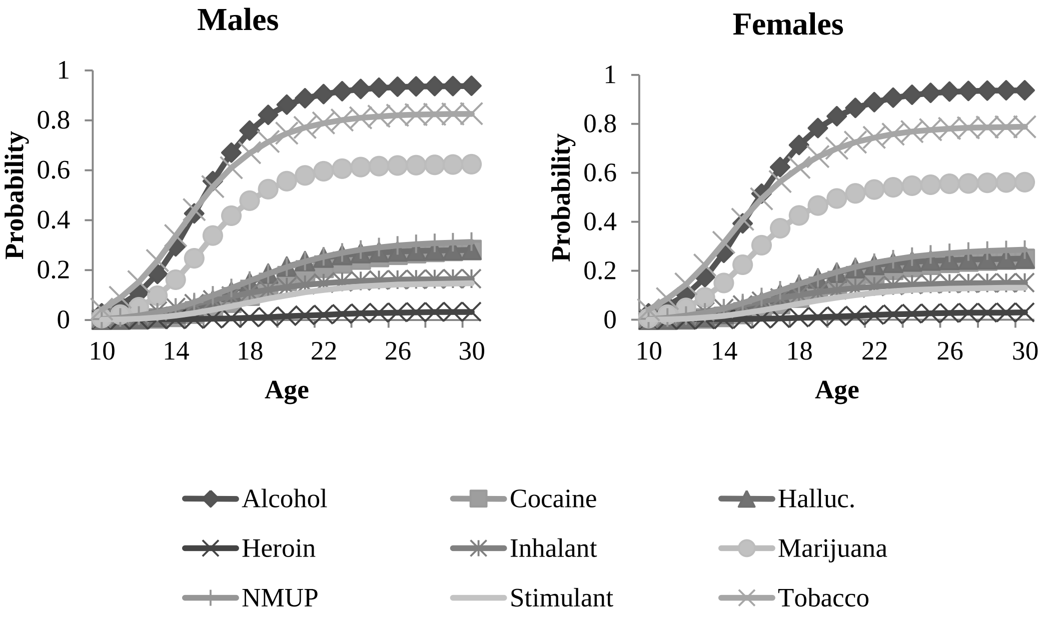
The odds ratios for class membership are displayed in Table 19, computed with a Bonferroni correction for multiple comparisons with $\alpha = 0.05$. The odds of females being assigned to the any of the soft or hard drug use classes in comparison to the relative abstainers class is smaller than the odds for males. The odds of females being in the early hard drug use compared to the soft drug use are likewise smaller than the odds for males. We find that the odds for Caucasians to be in any of the drug use classes in comparison to the relative abstainers class is larger than the odds for African-Americans, Hispanics, or those of other races. Likewise, we find that the odds for Caucasians to be in either of the hard drug classes in comparison to the soft drug class is higher than the odds for African-Americans. However, the odds for those of other races to be in the early hard drug class in comparison to the soft drug class is higher for those of other races than Caucasians. Interestingly, neither gender nor race significantly influence the probability of assignment to early hard drug use compared to late hard drug use.

Table 19: *NSDUH odds ratios for class membership in four class solution*

		Gender	Race		
Class	Intercept	Female	Black	Hispanic	Other
2 v. 1	**1.85**	**0.59**	**0.63**	**0.49**	**0.49**
	(1.51,2.27)	**(0.52,0.67)**	**(0.52,0.76)**	**(0.41,0.59)**	**(0.39,0.62)**
3 v. 1	0.87	**0.55**	**0.23**	**0.43**	**0.5**
	(0.67,1.12)	**(0.47,0.64)**	**(0.17,0.33)**	**(0.35,0.54)**	**(0.39,0.65)**
4 v. 1	**0.49**	**0.50**	**0.24**	**0.48**	**0.72**
	(0.37,0.64)	**(0.42,0.59)**	**(0.17,0.33)**	**(0.38,0.61)**	**(0.55,0.93)**
3 v. 2	**0.47**	0.94	**0.37**	0.88	1.02
	(0.40,0.54)	(0.79,1.11)	**(0.25,0.56)**	(0.69,1.13)	(0.75,1.37)
4 v. 2	**0.26**	**0.84**	**0.38**	0.98	**1.46**
	(0.22,0.32)	**(0.72,0.99)**	**(0.26,0.54)**	(0.77,1.23)	**(1.12,1.90)**
4 v. 3	**0.57**	0.90	1.01	1.10	1.43
	(0.45,0.70)	(0.73,1.11)	(0.60,1.69)	(0.82,1.49)	(1.03,1.99)

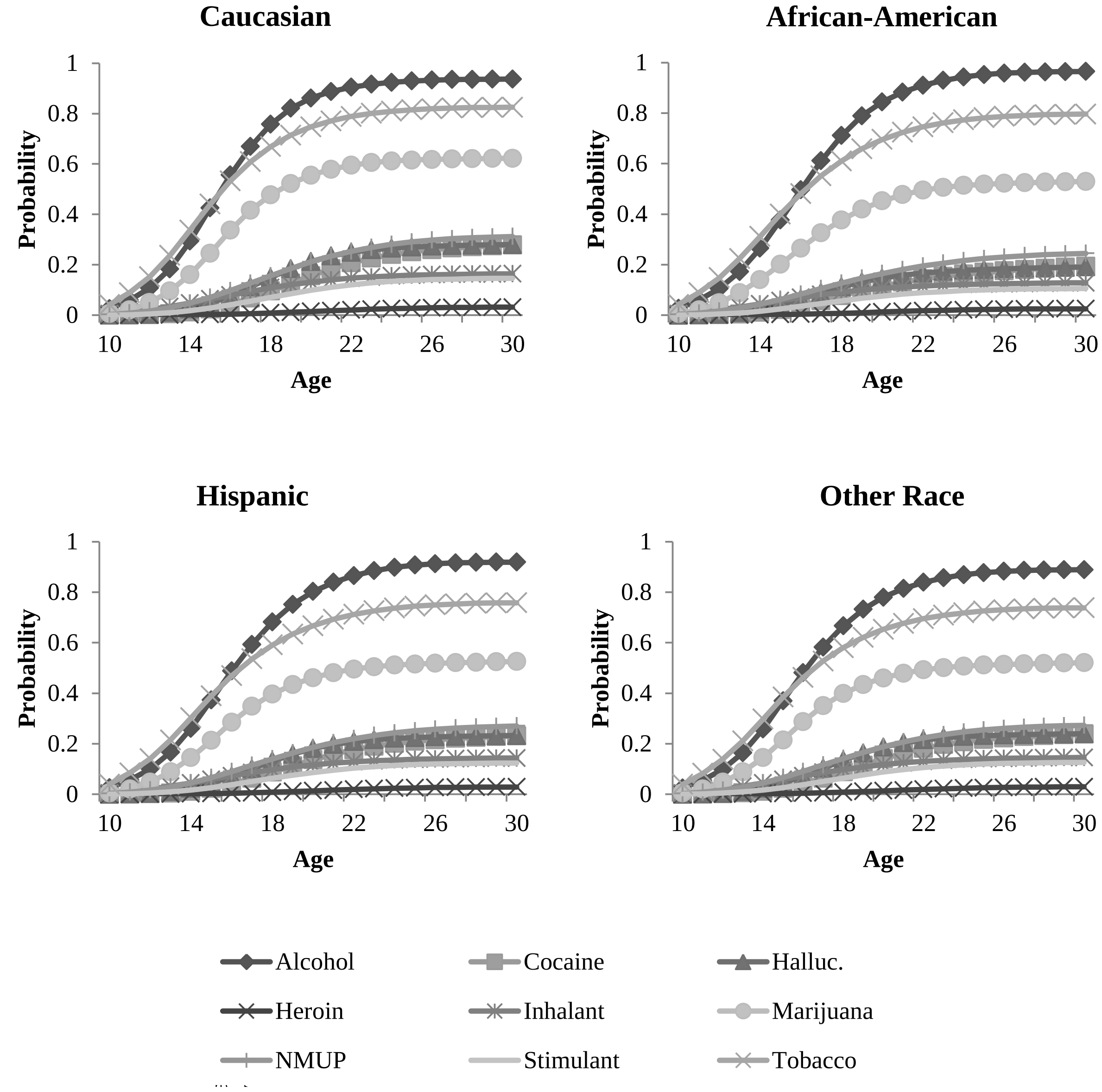
Next, the model implied lifetime distribution functions weighting over latent classes were computed depending on gender, with race kept constant at Caucasian (Figure 25). There seems to be almost no difference between the aggregate functions between males and females, other than males overall have a slightly higher risk for all of the events. This is consistent with the odds ratios that found the odds for males being in the soft drug use class as well as hard drug use classes in comparison to the relative abstainers class was higher than the odds for females. The average residual lifetime distribution probability between the functions for males and the functions for females is 0.02. The largest difference is in the lifetime distribution function for marijuana, with the cumulative probability at age 30 equal to 0.62 for males and 0.56 for females.

Figure 25: *NSDUH four class model implied lifetime distribution functions depending on gender*



Finally, model implied lifetime distribution functions were computed depending on race, with gender kept constant at male. The overall pattern again seemed to be similar across races, as implied by the model, but there were differences in level, with Caucasians having a higher lifetime distribution functions for all the events across all ages. In the scale of the lifetime distribution function, Caucasians were on average 0.03 higher than all other races.

Figure 26: *NSDUH four class model implied lifetime distribution functions depending on race*



The last step was to attempt to cross-validate the model, specifically looking to examine whether this final solution would fit the validation sample well and whether the same conclusions would be drawn in terms of the effects of covariates on the onset of different substances. First, the four class model without covariates was fit to the validation sample with all parameters constrained to the solution found in the evaluation sample. Measures of discrepancy – *ARD* and *ARH* – between the model imposed functions weighting over latent classes and the validation sample observed functions were computed to cross-validate the model in terms of recovery of the aggregate population functions (Cudeck & Browne, 1983). The model fit the validation sample nearly as well (compare evaluation sample *ARH* = 0.019, *ARD* = 0.005 with validation sample *ARH* = 0.019, *ARD* = 0.007). This speaks to the general consistency in the hazard and lifetime distribution functions between the two samples, but only references the ability of the latent classes to aggregate back to the population.

The next validation step that was taken was to again fit the four class model to the validation sample with all parameters constrained to the solution found in the evaluation sample, only allowing the effects of the covariates to be freely estimated on the probability of class membership. This was done in order to determine whether the covariates would have similar effects in the validation sample if the hazard functions within latent classes were equal to those found in the evaluation sample. Overall, the effects of the covariates were found to be equal in the validation sample. Predicted probabilities of class membership based on gender (holding race constant at Caucasian) and race (holding gender constant at male) are displayed in Table 20. The correlation between the predicted probabilities between the evaluation and validation sample is 0.99.

Table 20: *Predicted probabilities of class membership in NSDUH evaluation and validation sample depending on covariates*

	Evaluation Sample				
	Gender		Race		
Latent Class	Males	Females	African-American	Hispanic	Other
Relative Abstainers	0.24	0.36	0.40	0.40	0.37
Soft Drug Use	0.44	0.39	0.47	0.36	0.34
Later Hard Drug Use	0.21	0.17	0.08	0.15	0.16
Early Hard Drug Use	0.12	0.09	0.05	0.09	0.13

	Validation Sample				
	Gender		Race		
Latent Class	Males	Females	African-American	Hispanic	Other
Relative Abstainers	0.25	0.37	0.41	0.42	0.40
Soft Drug Use	0.45	0.38	0.47	0.36	0.34
Later Hard Drug Use	0.19	0.16	0.08	0.14	0.15
Early Hard Drug Use	0.11	0.09	0.04	0.09	0.11

In order to evaluate the within-class estimates, the MEPSUM model with covariates affecting class membership was also fit independently in the validation sample. The only constraints that were placed in this final validation analysis were the ones that were placed in the evaluation sample (i.e. constraining the slope and quadratic function to be 0 for low risk events within a class), rather than fixing the measurement of hazard functions within latent classes. See Table 21 for the validation sample logit parameter estimates. The validation sample parameter estimates correlated 0.97 with the evaluation sample estimates, which were listed above in Table 18.

Table 21: *NSDUH validation sample logit parameter estimates*

		Latent classes with covariates			
Event	Parameter	1	2	3	4
Alcohol	α_0	-7.03	-4.07	-6.70	-1.60
	α_1	10.09	6.28	18.30	-0.44
	α_2	-4.78	-1.76	-11.99	12.35
Cocaine	α_0	-9.48	-9.75	-10.72	-6.43
	α_1	0*	6.63	16.28	11.11
	α_2	0*	-2.05	-7.36	-6.13
Hallu.	α_0	-8.50	-8.95	-10.38	-6.05
	α_1	0*	7.04	18.60	13.21
	α_2	0*	-2.80	-9.62	-8.88
Heroin	α_0	-10.78	-9.27	-10.19	-7.03
	α_1	0*	0*	7.59	6.60
	α_2	0*	0*	-2.75	-3.44
Inhalants	α_0	-7.42	-4.99	-6.30	-3.84
	α_1	0*	-0.16	7.62	5.41
	α_2	0*	-0.51	-4.40	-4.75
Marijuana	α_0	-8.23	-5.87	-8.02	-2.65
	α_1	4.45	9.43	19.69	5.03
	α_2	-1.54	-5.40	-11.69	3.67
NMUP	α_0	-6.29	-6.01	-7.09	-4.19
	α_1	0.94	2.91	10.75	6.80
	α_2	-0.44	-1.15	-5.54	-4.48
Stimulant	α_0	-8.54	-8.25	-8.50	-5.17
	α_1	0*	4.33	11.33	7.42
	α_2	0*	-1.65	-5.74	-5.11
Tobacco	α_0	-5.15	-3.45	-4.94	-1.28
	α_1	4.77	5.23	12.39	0.02
	α_2	-2.63	-2.98	-8.16	8.18

* = Parameter constraint

Additionally, the effects of the covariates on class membership were highly correlated between the evaluation sample and validation sample, even when the only constraints were fixing the slope and quadratic factor to 0 for low risk events rather than constraining the hazard functions within class to be equal to the evaluation sample. With the last class as a referent, the multinomial logit parameters are given below in Table 22.

Table 22: *NSDUH multinomial logit parameters for class membership in evaluation and validation sample, with last class "early hard drug users" as a referent*

Sample	Latent Class	Intercept	Female	Black	Hispanic	Other Race
Evaluation	Relative Abstainers	0.71	0.70	1.44	0.74	0.33
	Soft Drug Use	1.33	0.17	0.98	0.02	-0.38
	Later Hard Drug	0.57	0.11	-0.01	-0.10	-0.36
Validation	Relative Abstainers	0.95	0.58	1.47	0.64	0.36
	Soft Drug Use	1.40	0.03	1.13	-0.04	-0.29
	Later Hard Drug	0.47	0.06	-0.01	-0.17	-0.39

The only different substantive conclusion drawn was that the odds for females being in the early hard drug class in comparison to the soft drug class was not significantly different than the odds for males, while the same odds for females in the evaluation sample was significantly smaller than the odds for males. However, all other conclusions remain the same, and the correlation between the parameters in the evaluation sample and validation sample with the last class as a referent is 0.98.

Thus, the model cross-validated well in the second half of the NSDUH sample, as the hazard functions within latent class were found to be very similar between the two subsamples, as was the effects of covariates on membership in the latent classes. It is important to note, however, that the sample size was very large in both the evaluation and the validation sample. The large sample size likely influenced the validation procedure, and the cross-validation results may not have been as strong at a more modest sample size.

4.3 Discussion

The risk for first use of the different substances was modeled using quadratic hazard functions within class, and there was not one clear optimal solution for the number of classes. This analysis provides an example of how – in an indirect application of mixture modeling – the end goal is description of the underlying multivariate distribution rather than deciding on the "true" number of subpopulations of individuals in terms of risk for initiating different substances. While covariates were investigated in the four class solution only to keep the scope of the example tractable, it would also be possible to compare model implied functions found using different number of classes and to compare substantive conclusions depending on the number of classes. Indeed, these analysis steps could even be used to determine the most appropriate number of classes to effectively describe the underlying multivariate distribution and the influence of covariates on this distribution.

In the four class solution, the first class represented relative abstainers with a relatively low cumulative probability of initiating alcohol and tobacco use, and near zero risk for all other substances. The second class of hazard functions could be described as soft drug users, with a high risk of initiating alcohol and tobacco, and a moderate risk of marijuana use. The third and fourth classes had a higher cumulative probability of initiating all drugs. In the third and fourth class, the hazard functions for alcohol, tobacco, and marijuana were found to be higher at earlier ages than the hazard functions for other harder substances, consistent with the gateway drug hypothesis. The five class solution was similar, only with a class emerging with a more linear risk of initiating substance use across the age range, which also had a small – rather than near zero like the other soft drug use class – cumulative probability of hard drug use by age 30. In the six class solution, the previous class of "soft drug users" was divided, with one class having a delayed risk of initiating substance use in comparison to the other. Gender and race were found to significantly predict class membership, and males and Caucasians were generally more likely to be in the soft and hard drug use classes in comparison to the relative abstainers class.

One interesting finding was that across all solutions, when comparing classes with similar patterns which differed in the age of peak risk and overall cumulative probability, the classes with earlier risk had higher overall cumulative probabilities of initiating drug use. For example, "later hard drug users" in the four class solution had smaller overall cumulative probabilities of initiating all substances than the "early hard drug users." Similarly, the "later soft drug users" class in the six class solution had a smaller cumulative probability of using the different substances than the "early soft drug users." This suggests that the age of initiating drug use is related to the probability of initiating subsequent drugs.

The differences found between the latent classes were mainly differences in the age of peak risk and the overall level, rather than large differences in the pattern. For example, no classes emerged that had higher earlier risk of a hard drug and later high risk of tobacco, alcohol, or marijuana. However, there were subtle differences, such as higher risk of tobacco at earlier ages in comparison to alcohol in the "early soft drug users" class in comparison to the "later soft drug users" class where the risk of alcohol and tobacco were similar at early ages. The latent classes were cross-validated in a split sample analysis. The hazard functions within latent class, as well as the effect of covariates on class membership, were found to be extremely similar in the evaluation and validation sample. However, this is limited somewhat in that the data were drawn from the same NSDUH sample; future work should look to cross-validate in an independent sample and to investigate the influence of other covariates on patterns found.

Related to cross-validation of the effects of covariates and on patterns of risk for substance use over time, it is interesting to note that the patterns of first use of substances are highly dependent on the time measure for the age of first use. In this example, data were collected retrospectively and the timing was measured to the nearest age. In almost all of the latent classes, the risk for alcohol, tobacco, and marijuana peaked at or close to the same age within a latent class. Similarly, the risk for other harder substances tended to peak around the same age within a latent class. Another study might find different patterns if the age range was narrowed, allowing a more detailed level of analysis of the risk for the events over time. This again relates to why the MEPSUM model should be used as a parsimonious description of the underlying distribution rather than a tool to discover the "true" number of subpopulations with similar risk for the events over time.

CHAPTER 5

DISCUSSION

A discrete-time multiple event process survival mixture (MEPSUM) model was introduced in this paper, which allows researchers to investigate the order and timing of multiple non-repeatable events that can occur at the same point in time. Both to be consistent with psychological and sociological theories, as well as to understand how the events are related to each other, it is important to consider the relationship between the hazard functions rather than to dissect the events in order to apply more traditional methods. A small simulation study was conducted and the MEPSUM model was applied in two empirical examples, and general conclusions, limitations, and areas for future research will now be discussed.

5.1 Simulation Study

A small simulation study was used in Chapter 2 to demonstrate the ability of the model to recover parameters from data generated under the assumption there are a finite number of subpopulations with the same risk for multiple events over time. The simulation found minimal bias in recovering the overall average of the hazard parameters, and recovery was best when the class separation was better and there were more events under study. The number of time periods had an interesting effect, in that bias was actually worse in the scale of the hazard, due to poor recovery in the high risk class, due to the diminishing risk set. However, in the scale of the lifetime distribution function, the bias on average decreased as the number of time periods increased.

The simulation was purposely small in scope, and many future directions are possible. First, the number of latent classes could be varied and models with different numbers of latent classes could be fit to investigate both optimal methods of model selection and influence of the selection on recovery. Additionally, the shape of the hazards could be varied and other forms of hazard functions could be investigated. The role of covariates and sample size are two other important aspects of the model that should be investigated further.

One important issue with the simulation is that the population was generated under the assumption there is truly a small number of latent groups with the same risk for multiple events over time, as is common in mixture modeling simulations (e.g. Lubke, & Neale, 2006; Lubke & Muthén, 2007). This simplified the process of analyzing recovery of the population parameters, and the goal of the simulation was just to demonstrate it could recover these parameters from data generated under the model. Yet as was discussed in Chapter 2, the purpose of the model is to approximate a complicated, but likely continuous, underlying multivariate distribution of hazard functions. The results from the simulation are thus limited in that they do not address a fundamental question of how the model could recover the population structure when there is not truly a small and finite number of latent groups. Thinking about the model in this way creates many new and interesting questions for future simulation research.

5.2 Empirical Examples

In Chapter 3, the MEPSUM model was used to capture heterogeneity in the hazard functions for multiple life course events and it found that gender, race, and parental education all significantly influenced latent class membership. A small number of categorical covariates was investigated, and a large sample size allowed stratification of the sample by different levels of covariates and comparison of model implied functions to sample estimated functions. Overall, there was general consistency in the functions implied by the model and the sample observed functions, such as females having a larger probability of parenthood at earlier ages than males.

While the model captured many trends exhibited by the sample-estimated hazard functions, the group differences were actually more exaggerated in the sample than in the model implied lifetime distribution functions. For example, the model implied those with neither parent earning a high school degree were less likely to obtain a college degree than others, yet this difference was larger in the sample than implied in the model. The statistical power of the model to detect differences in the simultaneous risk of multiple events over time should be investigated in future research, especially how model selection could possibly influence substantive conclusions and the power to detect the influence of covariates.

In Chapter 4, the MEPSUM model was used to investigate the risk for first use of nine substances, and a quadratic form of the hazard functions was determined to be optimal. This example was useful in examining model performance when there are many events, especially when some of the events have a very low risk of occurrence. It was found to be necessary to constrain the intercepts of the quadratic functions to be greater than a certain value in order to identify the parameters of the model. The model found that gender and race both significantly influenced heterogeneity in the risk for the events over time, and the model cross-validated well in a split sample validation. Results suggested that the age of first use is related to subsequent use, and that males and Caucasians are at particularly high risk for initiating hard drug use.

One important assumption that was made in this analysis was that the latent classes were equivalent across different cohorts of individuals. However, it is possible, even likely, that the risk of first use of different substances and the patterns of risk for multiple events over time has evolved over time. The range of ages in NSDUH is wide, and one interesting future research project would be to conduct a cohort analysis to examine the influence this might have on measuring the multivariate distribution of risk for multiple events, especially given that the large sample size in this data set can likely support such analysis.

5.3 Other Limitations and Future Directions

One limitation of this research project was that both empirical examples that were considered had a very large sample size. It is likely that a fairly large sample size is necessary for this model, as the model aims to determine patterns of risk for the events over time simply using binary variables on the timing of each event, but what exactly constitutes a “fairly large sample size” is unclear. This is an area for future research, but it is likely that the sample size necessary will depend in large part on the specifics of the data, such as the number of events, sample risk for the events, and number of time periods considered. A final limitation is that in both examples only a small number of covariates were examined, and neither a multiple group model nor direct effects were deemed necessary, as the model appeared stable after entering covariates. How the model performs with numerous covariates and with more complicated inclusions of covariates is yet to be seen.

This research should be extended to other situations common in the social and behavioral sciences, such as repeatable events. Combining the model with other frameworks might also be an interesting future direction, for situations when some of the variables in questions are appropriate for the MEPSUM model while others have more information (such as a continuous outcome measured over time). Additionally, how to model outcomes of patterns found, and how to consider mediation in this context is an area for future research. While there are many possible future directions, the model proposed in this paper provides an important framework from which to evaluate the interdependencies of multiple events which may occur at the same point in discrete-time.

APPENDIX

Table 23: *Average estimated standard deviation of all of the hazard parameters*

		Class Separation							
		Good				Poor			
	Time	High	Medium	Low		High	Medium	Mixed	
Events	Periods	Class	Class	Class	Average	Class	Class	Class	Average
4	5	0.451	0.583	0.950	0.661	0.217	0.975	1.376	0.856
	10	0.506	0.621	0.205	0.444	0.414	0.852	1.672	0.980
	20	1.229	0.443	0.155	0.609	1.985	0.689	1.562	1.412
8	5	0.080	0.119	0.127	0.109	0.098	0.130	0.164	0.131
	10	0.133	0.103	0.107	0.114	0.174	0.125	0.387	0.229
	20	0.898	0.154	0.112	0.388	0.875	0.208	0.893	0.659

Table 24: *Empirical average standard deviation of all of the hazard parameters over the 100 replications*

		Class Separation							
		Good				Poor			
	Time	High	Medium	Low		High	Medium	Mixed	
Events	Periods	Class	Class	Class	Average	Class	Class	Class	Average
4	5	1.039	1.915	1.834	1.596	0.507	2.199	2.197	1.634
	10	0.652	0.354	0.181	0.396	0.632	1.586	2.758	1.659
	20	2.258	0.678	0.144	1.027	2.903	2.415	2.686	2.668
8	5	0.075	0.119	0.119	0.105	0.092	0.122	0.151	0.122
	10	0.134	0.100	0.104	0.113	0.176	0.115	0.336	0.209
	20	2.134	0.154	0.112	0.800	2.458	0.379	1.746	1.528

Table 25: *Ratio of estimated average standard deviation of hazard parameters to empirical average standard deviation of all of the hazard parameters over 100 replications*

		Class Separation							
		Good				Poor			
	Time	High	Medium	Low		High	Medium	Mixed	
Events	Periods	Class	Class	Class	Average	Class	Class	Class	Average
4	5	0.435	0.305	0.518	0.419	0.428	0.443	0.626	0.499
	10	0.776	1.755	1.133	1.221	0.655	0.537	0.606	0.600
	20	0.544	0.653	1.071	0.756	0.684	0.285	0.582	0.517
8	5	1.067	0.999	1.069	1.045	1.057	1.064	1.086	1.069
	10	0.991	1.029	1.025	1.015	0.988	1.082	1.152	1.074
	20	0.421	1.002	0.999	0.807	0.356	0.549	0.512	0.472

Table 26: *Bias in average logit*

		Class Separation							
		Good				Poor			
	Time	High	Medium	Low	Absolute	High	Medium	Mixed	Absolute
Events	Periods	Class	Class	Class	Average	Class	Class	Class	Average
4	5	-0.047	0.264	0.322	0.211	-0.301	0.327	0.037	0.222
	10	0.037	0.031	0.010	0.026	0.040	0.248	0.355	0.214
	20	0.933	0.098	0.006	0.346	1.232	0.467	1.203	0.967
8	5	0.003	0.009	0.011	0.008	0.001	0.007	0.003	0.004
	10	0.004	0.006	0.006	0.005	-0.014	0.005	-0.023	0.014
	20	0.657	0.015	0.008	0.227	0.915	0.032	0.728	0.559

Table 27: *Absolute bias in logit*

		Class Separation							
		Good				Poor			
	Time	High	Medium	Low		High	Medium	Mixed	
Events	Periods	Class	Class	Class	Average	Class	Class	Class	Average
4	5	0.409	0.721	0.589	0.573	0.636	0.522	0.985	0.714
	10	0.326	0.221	0.132	0.226	0.618	0.442	1.059	0.706
	20	1.483	0.329	0.113	0.642	1.794	0.752	1.551	1.365
8	5	0.064	0.093	0.097	0.084	0.088	0.084	0.118	0.097
	10	0.105	0.081	0.084	0.090	0.184	0.089	0.201	0.158
	20	1.420	0.121	0.089	0.543	1.685	0.165	1.205	1.019

Table 28: *Proportion of 95% confidence interval coverage of logit parameters*

		Class Separation							
		Good				Poor			
	Time	High	Medium	Low		High	Medium	Mixed	
Events	Periods	Class	Class	Class	Average	Class	Class	Class	Average
4	5	0.766	0.647	0.818	0.744	0.812	0.881	0.786	0.826
	10	0.938	0.951	0.959	0.949	0.811	0.918	0.830	0.853
	20	0.870	0.982	0.963	0.938	0.769	0.902	0.831	0.834
8	5	0.958	0.958	0.959	0.958	0.961	0.960	0.969	0.963
	10	0.953	0.957	0.954	0.954	0.969	0.961	0.972	0.968
	20	0.897	0.959	0.954	0.937	0.886	0.967	0.921	0.925

REFERENCES

Abar, B., & Loken, E. (2012). Consequences of fitting nonidentified latent class models. *Structural Equation Modeling: A Multidisciplinary Journal*, *19,* 1-15.

Ahituv, A., Tienda, M, & Hotz, V. J. (2000) “Transition from School to Work: Black, Hispanic and White Men in the 1980s.” In R. Marshall (Ed.), *Back to Shared Prosperity: The Growing Inequality of Wealth and Income in America* (pp. 250-258). New York: M. E. Sharpe.

Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on automatic Control, AU-19*, 719–722.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52*, 317–332.

Allison, P. D. (1982). Discrete-time methods for the analysis of event histories. In S. Leinhardt (Ed.), *Sociological Methodology* (pp. 61-98). San Francisco: Jossey-Bass.

Allison, P. D. (1999). *Logistic regression using the SAS system: Theory and application*. Cary, N.C: SAS Institute.

Bandeen-Roche, K., Miglioretti, D. L., Zeger, S. L., & Rathouz, P. J. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association*, *92,* 1375-1386.

Bauer, D. J. (2005). A semiparametric approach to modeling nonlinear relations among variables. *Structural Equation Modeling: A Multidisciplinary Journal*, *12,* 513-535.

Bauer, D. J. (2007). Observations on the use of growth mixture models in psychological research. *Multivariate Behavioral Research*, *42*, 757-786.

Bauer, D. J., & Curran, P. J. (2003a). Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. *Psychological Methods*, *8,* 338-363.

Bauer, D. J., & Curran, P. J. (2003b). Overextraction of latent trajectory classes: Much ado about nothing? Reply to Rindskopf (2003), Muthén (2003), and Cudeck and Henly (2003). *Psychological Methods*, *8*, 384-393.

Bauer, D. J., & Curran, P. J. (2004). The integration of continuous and discrete latent variable models: Potential problems and promising opportunities. *Psychological Methods*, *9,* 3-29.

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. M.I.T. Press, Cambridge, MA.

Bowers, A. J. (2010). Grades and Graduation: A Longitudinal Risk Perspective to Identify Student Dropouts. *Journal of Educational Research, 103(3),* 191-207.

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52,* 345–370.

Bray, B. C., Lanza, S. T., & Collins, L. M. (2010). Modeling Relations among Discrete Developmental Processes: A General Approach to Associative Latent Transition Analysis. *Structural Equation Modeling, 17(4),* 541-569.

Celeux, G., Biernacki, C., & Govaert, G. (1997). *Choosing models in model-based clustering and discriminant analysis.* Technical report. Rhone-Alpes: INRIA.

Clark, S. L., & Muthén, B. (2009). Relating latent class analysis results to variables not included in the analysis. Manuscript submitted for publication and downloaded. Retrieved 17 Feb, 2012, from http://www.statmodel.com/download/relatinglca.pdf.

Clogg, C.C. (1995). Latent class models: Recent developments and prospects for the future. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences*, 311-352. New York: Plenum.

Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Corning, A. F., & Malofeeva, E. V. (2004). The application of survival analysis to the study of psychotherapy termination. *Journal of Counseling Psychology, 51(3),* 354-367.

Cox, D. R., & Lewis, P. A. W. (1972). Multivariate point processes. In L. M. Le Cam, J. Neyman, and E. Scott (Ed.), *Proceedings of the Sixth Berkeley Symposiom on Mathematical Statistics and Probability, vol. 3, Probability Theory* (pp. 401-48). Berkeley: University of California Press.

Cudeck, R. A. & Browne, M. W. (1983). Crossvalidation of covariance structures. *Multivariate Behavioral Research, 18*, 147- 167.

Dolan, C. V., & van der Maas, Han L. J. (1998). Fitting multivariate normal finite mixtures subject to structural equation modeling. *Psychometrika, 63(3),* 227-53.

Elder, G. H., Jr. (1985). *Life Course Dynamics*. Ithaca, NY: Cornell University Press.

Elder, G. H., Jr., Johnson, M. K., & Crosnoe, R. (2003). The emergence and development of life course theory. In: Mortimer, J. T., and M. J. Shanahan (eds.), *Handbook of the life course*, 3-22. Hingham, MA: Kluwer Academic Publishers.

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY US: Guilford Press.

Fergusson, D. M., Boden, J. M., & Horwood, L. (2006). Cannabis use and other illicit drug use: testing the cannabis gateway hypothesis. *Addiction, 101(4),* 556-569.

Formann, A. K. (1992). Linear Logistic Latent Class Analysis for Polytomous Data. *Journal of the American Statistical Association, 87(418),* 476-486.

Formann, A. K. (2003). Latent class model diagnostics from a frequentist point of view. *Biometrics, 59,* 189-196.

Fraley, C., & Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. Technical report No. 329, Department of Statistics, University of Washington.

Golub, A., & Johnson, B. D. (2001). Variation in Youthful Risks of Progression From Alcohol and Tobacco to Marijuana and to Hard Drugs Across Generations. *American Journal of Public Health, 91(2),* 225-232.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika, 61,* 215-231.

Ha, J. C., Kimpo, C. L., & Sackett, G. P. (1997). Multiple-spell, discrete-time survival analysis of developmental data: Object concept in pigtailed macaques. *Developmental Psychology, 33(6),* 1054-1059.

Hagenaars, J.A. (1993). *Loglinear models with latent variables*. London: Sage.

Hamburg, B. A., Kraemer, H. C., & Jahnke, W. (1975). A hierarchy of drug use in adolescence: Behavioral and attitudinal correlates of substantial drug use. *The American Journal of Psychiatry, 132(11),* 1155-1163.

Harris, K. M., Halpern, C. T. Whitsel, E. Hussey, J. Tabor, J. Entzel, P. & Udry, J. R. (2009). The National Longitudinal Study of Adolescent Health: Research Design. http://www.cpc.unc.edu/projects/addhealth/design.

Heckman, J. J., & Singer, B. (1982). Population heterogeneity in demographic models. In K. Land and A. Roders (eds.) *Multidimensional mathematical demography*. New York: Academic Press.

Heckman, J. J. & Singer, B. (1984). The identifiability of the proportional hazard model. *Review of Economic Studies, 51*, 231-241.

Hipp, J. R., & Bauer, D. J. (2006). Local solutions in the estimation of growth mixture models. *Psychological Methods, 11(1),* 36-53.

Hofferth, S. L., & Moore, K. A. (1979). Early childbearing and later economic well-being. *American Sociological Review, 44,* 784-815.

Hogan, D. P. (1978). The variable order of events in the life course. *American Sociological Review, 43,* 573-586.

Hougaard, P. (2000). *Analysis of multivariate survival data*. New York: Springer.

Hougaard, P., Harvald, B. & Holm, N.V. (1992). Measuring the similarities between the lifetimes of adult Danish twins born between 1881-1930. *Journal of the American Statistical Association, 87(417),* 17-24.

Kalbfeisch, J. D. & Prentice, R. L. (1980). *The statistical analysis of failure time data*. New York: Wiley.

Kandel, D. B., & Logan, J. A. (1984). Patterns of drug use from adolescence to young adulthood: I. Periods of risk for initiation, continued use, and discontinuation. *American Journal of Public Health, 74(7),* 660-666.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons.

Land, K. C., Nagin, D. S., & McCall, P. L. (2001). Discrete-time hazard regression models with hidden heterogeneity: The semiparametric mixed Poisson regression approach. *Sociological Methods & Research, 29(3),* 342-373.

Lee, E. T., & Wang, J. W. (2003). *Statistical Methods for Survival Data Analysis*. New York: J. Wiley.

Little, R. J. A., & Rubin, D. (1987). *Statistical analysis with missing data*. New York: Wiley.

Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika, 88(3),* 767-778.

Lohr, S. (2009). *Sampling: Design and Analysis*. Duxbury Press: Pacific Grove.

Lubke, G., & Neale, M. C. (2006). Distinguishing between latent classes and continuous factors: Resolution by maximum likelihood? *Multivariate Behavioral Research, 41(4),* 499-532.

Lubke, G., & Muthén, B. (2007). Performance of factor mixture models as a function of model size, covariate effects, and class-specific parameters. *Structural Equation Modeling: A Multidisciplinary Journal, 14:1,* 26-47.

Macmillan, R., & Copher, R. (2005). Families in the Life Course: Interdependency of Roles, Role Configurations, and Pathways. *Journal of Marriage & Family, 67(4),* 858-879.

Macmillan, R., & Eliason, S. R. (2003). Characterizing the life course as role configurations and pathways: A latent structure approach. In: Mortimer, J. T., and

M. J. Shanahan (eds.), *Handbook of the life course*. Hingham, MA: Kluwer Academic Publishers.

Magidson, J., & Vermunt, J. K. (2001). Latent class factor and cluster models, bi-plots and tri-plots and related graphical displays. *Sociological Methodology, 31,* 223-264.

Mahaffy, K. A. (2003). Gender, race, class, and the transition to adulthood: A critical review of the literature. *Sociological Studies of Children and Youth, 9,* 15-47.

Malone, P. S., Lamis, D. A., Masyn, K. E., & Northrup, T. F. (2010). A dual-process discrete-time survival analysis model: Application to the gateway drug hypothesis. *Multivariate Behavioral Research, 45(5),* 790-805.

Marini, M. (1984). Women's educational attainment and the timing of entry into parenthood. *American Sociological Review, 49(4),* 491-511.

Marsh, H. W. Lüdtke, O. Trautwein, U. & Morin, A. (2009). Classical latent profile analysis of academic self-concept dimensions: Synergy of person- and variable-centered approaches to theoretical models of self-concept. *Structural Equation Modeling: A Multidisciplinary Journal, 16,* 191-225.

Martin, J. A., Hamilton, B. E., Ventura, S. J., Osterman, M. J. K., & Kirmeyer, S. (2011). Births: Final data for 2009. *National Vital Statistics Reports, 60(1).* Hyattsville, MD: National Center for Health Statistics.

McHugh, R. B. (1956). Efficient estimation and local identification in latent class analysis. *Psychometrika, 21,* 331-347.

McLachlan, G. & Peel, D. (2000). *Finite Mixture Models*. New York: Wiley.

Muthén, B., & Masyn, K. (2005). Discrete-Time Survival Mixture Analysis. *Journal of Educational and Behavioral Statistics, 30(1),* 27-58.

Muthén, B. & Muthén, L. (1998-2010). *Mplus User's Guide. Sixth Edition*. Los Angeles, CA: Muthén & Muthén.

Muthén, B. (2001). Latent variable mixture modeling. In G. A. Marcoulides & R. E. Schumacker (eds.), *New Developments and Techniques in Structural Equation Modeling* (pp. 1-33). Lawrence Erlbaum Associates.

Nagin, D. S. (1999). Analyzing developmental trajectories: A semiparametric group-based approach. *Psychological Methods, 2,* 139-157.

Nagin, D. S. (2005). *Group-based modeling of development*. Cambridge, MA: Harvard University Press.

Nagin, D. S., & Odgers, C. L. (2010). Group-based trajectory modeling in clinical research. *Annual Review of Clinical Psychology, 6,* 109-138.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal, 14(4),* 535-569.

Petras, H. & Masyn, K. (2010). General growth mixture analysis with antecedents and consequences of change. In A. Piquero & D. Weisburd (Eds.), *Handbook of Quantitative Criminology* (pp. 69-100). New York: Springer.

Sampson, R. J., & Laub, J. H. (2005). Seductions of method: Rejoinder to Nagin and Tremblay's "Developmental trajectory groups: Fact or fiction?" *Criminology, 43,* 905-913.

Schwartz, S. J., Phelps, E., Lerner, J. V., Shi, H., Brown, C., Lewin-Bizan, S., Li, Y., & Lerner, R. M. (2010). Promotion as prevention: Positive youth development as protective against tobacco, alcohol, illicit drug, and sex initiation. *Applied Developmental Science, 14(4),* 197-211.

Schwarz, G. (1978). Estimating the dimensions of a model. *Annals of Statistics, 6,* 461–464.

Scott, K. M., Wells, J. E., Angermeyer, M. M., Brugha, T. S., Bromet, E. E., Demyttenaere, K. K., & Kessler, R. C. (2010). Gender and the relationship between marital status and first onset of mood, anxiety and substance use disorders. *Psychological Medicine: A Journal of Research in Psychiatry and the Allied Sciences, 40(9),* 1495-1505.

Shanahan, M. (2000). Pathways to adulthood in changing societies: Variability and mechanisms in life course perspective. *Annual Review of Sociology, 26,* 667-692.

Shanahan, M., Miech, R. & Elder, G. (1998). Changing pathways to attainment in men's lives, historical patterns of school, work, and social class. *Social Forces, 77,* 231-256.

Singer, J. D., & Willett, J. B. (1993). It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational Statistics, 18(2),* 155-195.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* New York: Oxford University Press.

Snyder, T., & Dillow, S. A. (2011). *Digest of Education Statistics, 2010*. Washington, DC: National Center for Education Statistics.

Steele, F. (2003). A discrete-time multilevel mixture model for event history data with long term survivors, with an application to an analysis of contraceptive sterilization in Bangladesh. *Lifetime Data Analysis, 9(2),* 155-174.

Titterington, D. M., Smith, A. F. M., & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. Chicester: John Wiley & Sons.

Tuma, N. B. & Hannan, M. T. (1984). *Social dynamics. Models and methods*. Orlando, FL: Academic Press.

United States Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Office of Applied Studies (2010). National Survey on Drug Use and Health, 2009. Ann Arbor, MI: Inter-university Consortium for Political and Social Research.

Vaughn, M., Wallace, J., Perron, B., Copeland, V., & Howard, M. (2008). Does Marijuana Use Serve as a Gateway to Cigarette Use for High-Risk African-American Youth?. *American Journal of Drug & Alcohol Abuse, 34(6),* 782-791.

Vaupel, J. W., Manton, K. G., & Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography, 16*, 439-454.

Vaupel, J. W., & Yashin, A. I. (1985). Heterogeneity's Ruses: Some Surprising Effects of Selection on Population Dynamics. *American Statistician, 39(3),* 176.

Ventura, J., Nuechterlein, K. H., Subotnik, K. L., Hardesty, J., & Mintz, J. (2000). Life events can trigger depressive exacerbation in the early course of schizophrenia. *Journal of Abnormal Psychology, 109(1),* 139-144.

Vermunt, J. K. (1997). *Log-linear models for event histories*. Thousand Oaks, CA: Sage.

Vermunt, J.K., and Magidson, J. (2002). Latent class cluster analysis. In: J.Hagenaars and A. McCutcheon (eds.), *Applied latent class analysis*, 89-106. Cambridge, UK: Cambridge University Press.

Vermunt, J. K. and Magidson, J. (2005). *Latent GOLD 4.0 Choice User's Guide*. Belmont Massachussetts: Statistical Innovations Inc.

Wagner, F. A., & Anthony, J. C. (2002). Into the world of illegal drug use: Exposure opportunity and other mechanisms linking the use of alcohol, tobacco, marijuana, and cocaine. *American Journal of Epidemiology, 165,* 918-925.

Wang, C., Brown, C. H., & Bandeen-Roche, K. (2005). Residual diagnostics for growth mixture models: Examining the impact of a preventive intervention on multiple trajectories of aggressive behavior. *Journal of the American Statistical Association, 100,* 1054-1076.

Willett, J. B., & Singer, J. D. (1995). It's déjà vu all over again: Using multiple-spell discrete time survival analysis. *Journal of Educational and Behavioral Statistics, 20(1)*, 41-67.

Xue, X., & Brookmeyer, R. (1997). Regression analysis of discrete time survival data under heterogeneity. *Statistics in Medicine, 16*, 1983-1993.

Yamaguichi, K. (1991). *Event history analysis*. Newbury Park, CA: Sage.

Yamaguchi, K., & Kandel, D. B. (1984). Patterns of Drug Use from Adolescence to Young Adulthood: II. Sequences of Progression. *American Journal of Public Health, 74(7)*, 668-672.