Estimating Discrete-time Survival Models as Structural Equation Models

Shawn Bauldry Kenneth A. Bollen

August 13, 2008

Abstract

Survival analysis is a common tool in the analysis of the timing of events. Survival models have been applied to a vast number of areas of sociological research. In this paper we show how to embed survival models in structural equation models (SEMs) while taking advantage of additional features possible in a SEM approach. Using empirical examples of time to promotion and timing of intercourse, we show that we can replicate the more traditional survival model results, but more importantly, we show how to control for measurement error and permit mediating variables. By doing so we provide additional options useful to sociologists who use survival models.

1 Introduction

Following their introduction into the discipline in the late 1970s, survival models have become a common tool in the analysis of the timing and occurence of events of sociological significance. Sociologists have drawn on survival models, or event-history models, to further our understanding of social processes, including the diffusion of protests (e.g., Andrews and Biggs 2006), marriage formation (e.g., Sweeney 2002), the passage of laws (e.g., Behrens, Uggen, and Manza 2003), and organizational demography (e.g., Sørensen 2004). As the use of survival models has spread throughout sociology the models have been extended to address various departures from the standard model. For instance, Allison (1984) covers how to handle multiple types of events and repeated events, Guo (1993) demonstrates how to address left-truncated data, and Allison (2005) reviews fixed-effects survival models. Until recently, however, it has not been known how to accomodate measurement error in the covariates and how to simultaneously model systems of equations allowing for the effects of mediators. Accommodating measurement error is important since ignoring such error will bias our estimates of the impact of covariates and will mislead researchers. Simultaneous equations permits a fuller understanding of a variable's direct, indirect, and total effects. Survival analysis as currently practice reveals only the direct effect which can differ from its indirect effects.

Recent work in the area of educational statistics has demonstrated that survival models can be treated as a specific case in the more general class of structural equation models (Asparouhov, Masyn, and Muthén 2006; Muthén

and Masyn 2005; Masyn 2003). In this article, we provide an overview of this work with a focus on discrete-time survival models for a sociological audience. These prior works do not consider multiequation models and mediating variables. Given the advantages of doing so, we also include a discussion of such models in our presentation. Following a brief discussion of discrete-time survival models, we illustrate how these models fit into a structural equation modeling framework. Next, using Allison's (1995) data on the careers of biochemists, we demonstrate how to estimate several basic models and in the process document that one obtains the same results as the more conventional approach to estimating these models. We conclude with an extended empirical example in which we illustrate the advantages of this approach in accounting for measurement error and allowing for mediating and indirect effects.

2 Discrete-Time Survival Models

The two primary types of survival models are continuous-time and discrete-time. Continuous-time models are typically employed when the "exact" timing of an event is known (e.g., the day and year of death). Discrete-time models, in contrast, are generally used when only the interval in which an event occurs is known or the event itself occurs in discrete intervals (e.g., the year in which respondents marry). In practice, the distinction between whether an event is measured in continuous-time or discrete-time is arbitrary (in a technical sense, all events of sociological interest can be considered as measured in discrete-time)

and the choice between continuous-time and discrete-time models depends on other factors, such as the existence of ties in the data (i.e., individuals who experience the event of interest at the same time point), whether a specific parametric function of time is appropriate, and computational considerations. As many sociological applications involve the use of discrete-time models, we focus on this class of models in this work; however, it is also possible to estimate continuous-time survival models in a structural equation modeling framework (see Asparouhov et al. 2006).

Survival data consists of at least two pieces of information for each case regarding an event: (1) the time until an event occurs or the case is no longer observed and (2) an indicator of whether the case experienced the event. A case that does not experience the event while under observation is said to be censored as the time to event is unknown. Allison (1995) describes three types of censoring: Type I, Type II, and random (see also Singer and Willett (2003) for a helpful discussion). Type I censoring occurs when the censoring time is fixed by design and all observations share the same censoring time (e.g., the end of a study). Type II censoring occurs when cases are no longer observed after a predetermined number of events. Random censoring refers to the situation when a case is no longer observed for any reason not under the control of the investigator (e.g., sample attrition in a longitudinal study). These cases can be problematic for the analysis. If the reason for censoring, conditional on the covariates in the model, is related to the event of interest, then the censoring is considered informative and parameter estimates can be badly biased (Allison

1995).[1] For example, in an analysis of the duration of spells of unemployment, the people who are not able to be followed over time (i.e., censored) may be more likely to experience longer spells of unemployment. Unfortunately, it is not possible to test whether random censoring is informative or non-informative (but, see Allison (1995) for one approach to assessing the sensitivity of results to the potential presence of informative censoring). In the following development of the models we assume censoring is non-informative; if researchers have reason to believe otherwise, as usual, results should be interpreted with caution.

2.1 Survival and Hazard Functions

In the discrete-time framework with non-repeated events there are two probabilities of primary interest: the hazard probability and the survival probability. The hazard probability refers to the probability of a case experiencing an event in a time interval given that the case has not experienced the event in any previous time interval. Formally, let T be a discrete random variable, i an index for cases, and j an index for time periods. Then T takes on values T_i which indicate the time period j when case i experiences an event. The hazard probability expressed as a function of time is given as the conditional probability density function

$$h(t_{ij}) = \Pr[T_i = j | T_i \geq j]. \tag{1}$$

[1] As we will discuss below, the concepts of informative and non-informative censoring closely mirror different types of missing data.

The conditional nature of the hazard probability is important to keep in mind. A case can only experience an event in time period j if and only if the case has not already experienced the event. Such cases that are eligible to experience an event are said to be in the risk set.

The survival probability is defined as the probability of surviving, or not experiencing an event, beyond a given time period j. We can express the survival probability as

$$S(t_{ij}) = \Pr[T_i > j]. \tag{2}$$

As one would expect, the hazard and survival probability are closely related. In a discrete-time setting the survival probability can be considered the probability of not experiencing the event in any prior time period. For any single time period, the probability of not experiencing the event, given the case is in the risk set, is simply one minus the hazard probability. Taking the product, the survival probably can also be written as

$$S(t_{ij}) = \prod_{k=1}^{j} \left(1 - h(t_{ik})\right). \tag{3}$$

Using this formula, it is easy to calculate the estimated survival probability for any given time point using the estimates of the hazard probabilities (due to censoring, however, it is not possible to calculate the hazard probabilities from the survival probabilities).

2.2 Maximum Likelihood Estimation

The method of maximum likelihood is the most common approach to obtain estimates of the hazard probabilities for the population. For discrete-time survival models, the likelihood function expresses the probability of observing the pattern of the occurrences of the event in the data. The pattern of events arises from arraying the data in a case-period format such that each case contributes as many observations as time periods they are in the risk set. In this format an indicator variable (δ_{ij}) denotes for each time period whether a case experiences an event. As noted in (1), the hazard captures the probability that case i experiences an event in time period j conditional on being in the risk set. Therefore if case i experiences the event in time period j the case contributes $h(t_{ij})$ to the likelihood function. Conversely, if case i does not experience the event in time period j, the the case contributes $1 - h(t_{ij})$ to the likelihood function. Taking the product over all cases (indexed by N) and all time periods cases remain in the risk set (indexed by J_i), we can write the likelihood function as

$$L = \prod_{i=1}^{N} \prod_{j=1}^{J_i} \left[h(t_{ij})^{\delta_{ij}} (1 - h(t_{ij}))^{1-\delta_{ij}} \right]. \tag{4}$$

As we will demonstrate below, we derive an identical likelihood function estimating these models as structural equation models.

In most research sociologists are interested in estimating the relationship between a set of covariates and the hazard probabilities. One of the advantages of survival models is that covariates are not forced to be static, they may vary

over time. In the discrete-time setting the logit link is the most widely used and the resulting model can easily incorporate both time invariant and time-varying covariates. Let $\boldsymbol{X}$ and $\boldsymbol{X}_j$ respectively be matrices of time invariant and time-varying covariates, $\boldsymbol{\beta}$ and $\boldsymbol{\kappa}$ vectors of coefficients, and $\boldsymbol{\alpha}$ a vector of intercepts for each time period. Then we can write the logit of the hazard probability as

$$\log\left(\frac{h(t_{ij})}{1-h(t_{ij})}\right) = \boldsymbol{\alpha} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{X}_j\boldsymbol{\kappa}. \tag{5}$$

In this form of the model, the effects of both the time invariant and the time-varying variables are constrained to be the same for each time period. This is often referred to as the proportional hazard odds property. In many instances, it is not reasonable to assume that the effects of the covariates will remain constant over time. For example, one might imagine that the effect of self-esteem on the hazard of first intercourse could change as individuals grow older. Fortunately, it is easy to relax the proportional hazard odds property and test it; one simply adds interactions between the covariates and the time period indicators contained in the vector of intercepts.

2.3 Structural Equation Models

In this section, we provide a brief overview of structural equation models (SEMs) before turning to how we can estimate discrete-time survival models in a structural equation modeling framework. SEMs are a general class of statistical models consisting of multiequation systems that represent relationships between

latent and observed variables. Many models commonly employed by sociologists can be estimated as SEMs. For instance, general linear models, factor analysis, and simultaneous equations are all special cases of SEMs (Bollen 1989).

Structural equation models have two primary components: the latent variable model and the measurement model. The latent variable model is

$$\boldsymbol{\eta}_i = \boldsymbol{\alpha}_{\boldsymbol{\eta}} + \boldsymbol{B}\boldsymbol{\eta}_i + \boldsymbol{\Gamma}\boldsymbol{\xi}_i + \boldsymbol{\zeta}_i, \tag{6}$$

where $\boldsymbol{\eta}_i$ is a vector of latent endogenous variables, $\boldsymbol{\alpha}_{\boldsymbol{\eta}}$ is a vector of intercept terms, $\boldsymbol{B}$ is a matrix of coefficients containing the effects of the latent endogenous variables on each other, $\boldsymbol{\xi}_i$ is a vector of latent exogenous variables, $\boldsymbol{\Gamma}$ is a coefficient matrix containing the effects of the latent exogenous variables on the latent endogenous variables, and $\boldsymbol{\zeta}_i$ is a vector of disturbances. The i subscript indexes the ith case in the sample. In this model we assume that $E(\boldsymbol{\zeta}_i) = 0$, $COV(\boldsymbol{\xi}_i', \boldsymbol{\zeta}_i) = 0$, and that $(\boldsymbol{I} - \boldsymbol{B})$ is invertible. Exogenous variables are determined outside the system of equations, while endogenous variables are influenced by other variables in the system.

The measurement model links the latent to the observed variables (indicators). The following two equations capture this relationship:

$$\begin{aligned} \boldsymbol{y}_i &= \boldsymbol{\alpha}_{\boldsymbol{y}} + \boldsymbol{\Lambda}_{\boldsymbol{y}}\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i \\ \boldsymbol{x}_i &= \boldsymbol{\alpha}_{\boldsymbol{x}} + \boldsymbol{\Lambda}_{\boldsymbol{x}}\boldsymbol{\xi}_i + \boldsymbol{\delta}_i, \end{aligned} \tag{7}$$

where $\boldsymbol{y}_i$ and $\boldsymbol{x}_i$ are respective vectors of indicators of the latent variables in

respectively $\boldsymbol{\eta}_i$ and $\boldsymbol{\xi}_i$, $\boldsymbol{\alpha_y}$ and $\boldsymbol{\alpha_x}$ are respective vectors of intercept terms, $\boldsymbol{\Lambda_y}$ and $\boldsymbol{\Lambda_x}$ are respective matrices of factor loadings, and $\boldsymbol{\varepsilon}_i$ and $\boldsymbol{\delta}_i$ are vectors of disturbances. We assume that $E(\boldsymbol{\varepsilon}_i) = E(\boldsymbol{\delta}_i) = 0$ and that $\boldsymbol{\varepsilon}_i$, $\boldsymbol{\delta}_i$, $\boldsymbol{\xi}_i$, and $\boldsymbol{\zeta}_i$ are all uncorrelated.

SEMs have been generalized to account for categorical dependent variables. Following Long (1997), a limited-dependent variable model can be written as

$$y_i^* = \boldsymbol{X}_i\boldsymbol{\beta} + \varepsilon_i, \tag{8}$$

where y_i^* is an unobserved continuous variable related to the observed categorical variable y_i. The precise relationship depends on the nature of the categorical variable. For a dichotomous variable, the relationship can be written as

$$y_i = \begin{cases} 1 & \text{if } y_i^* > \tau \\ 0 & \text{if } y_i^* \leq \tau \end{cases}. \tag{9}$$

This equation represents a threshold model. When the underlying continuous variable exceeds some threshold τ, a 1 is observed, otherwise a 0 is observed. In order for the model to be identified, one must assume a distribution for the disturbance ε_i. In practice, the default assumption for SEMs is that the disturbance follows a standard normal distribution, in which case we have a probit regression model. For our purposes, however, it will be convenient to assume that the disturbance follows a standardized logistic distribution with a mean of 0 and a variance of $\pi^2/3$, in which case we have a logistic regression

model.

3 Discrete-time Survival Models as SEMs

To incorporate discrete-time survival models into a SEM framework we need a way to estimate the hazard probabilities for each time interval. Returning to our definition of a hazard probability as the conditional probability of a case experiencing an event in a time interval given the case is in the risk set suggests one approach. Imagine for each time period that each case in the risk set has an underlying latent propensity to experience an event and if a threshold is exceeded then the case experienced the event. In this set-up, if we assume a standardized logistic distribution for the disturbance, then the hazard probability for the jth time period is related to the threshold for the jth time period τ_j as

$$h(j) = \frac{1}{1 + \exp[-\tau_j]}. \tag{10}$$

We can then estimate the hazard probabilities for all of the time points simultaneously through a system of logistic models. We illustrate this in Figure 1, where the circles represent the underlying latent propensities to experience an event for each time period, the rectangles represent observed event indicators for each time period, and the jagged lines represent the nonlinear relationships between the latent propensities and the observed indicators.

– Figure 1 about here –

Estimating the hazard probabilities using this approach requires a wide format for the data rather than the long, or case-period, format needed for standard discrete-time survival models (see Figure 2). To prepare the data, the analyst needs to create event indicators for every time period in the data. For each case, the event indicators can take three values:

$$E_{ij} = \begin{cases} 0 & \text{did not experience event in time period } j \\ 1 & \text{experienced event in time period } j \\ . & \text{not in risk set in time period } j \end{cases}. \tag{11}$$

Once a case experiences an event or is censored, all of the subsequent event indicators should be set to missing. This ensures that only cases in the risk set are included in the estimation of the hazard probability for each time interval. Estimating the system of logistic models with missing data included relies on the assumption that the data are missing at random (Little and Rubin 2002), which in this setting corresponds to the assumption of noninformative censoring. With the data in wide format, time invariant covariates simply occupy a column for each, but time-varying covariates need to be prepared in a similar fashion to the event indicators (see Figure 2 for an illustration).

– Figure 2 about here –

3.1 Maximum Likelihood Estimation

Given the organization of the data, we take a different approach than used above in deriving the likelihood function in the SEM framework. Because the

data is in wide format and the cases not in the risk set for any time interval are treated as missing on the respective event indicators, we note that the conditional probability of experiencing an event in each time interval is given by

$$h(t_{ij}) = \Pr[E_{ij} = 1] \tag{12}$$

and, conversely, the conditional probability of not experiencing an event is given by

$$1 - h(t_{ij}) = \Pr[E_{ij} = 0]. \tag{13}$$

Letting δ_i be an indicator for whether a case experiences an event while under observation, we can write the individual likelihood function as

$$L_i = \Pr[E_{ij} = 1]^{\delta_i} \prod_{j=1}^{J_i - 1} \Pr[E_{ij} = 0]. \tag{14}$$

Taking the product over all of the cases gives us the likelihood function for the sample, and with a little rearrangement and substiting in (12) and (13) we obtain the same likelihood function that we derived above (4);

$$\begin{aligned} L &= \prod_{i=1}^{N} \left[\Pr[E_{ij} = 1]^{\delta_i} \prod_{j=1}^{J_i - 1} \Pr[E_{ij} = 0] \right] \\ &= \prod_{n=1}^{N} \prod_{j=1}^{J_i} \left[h(t_{ij})^{\delta_{ij}} (1 - h(t_{ij}))^{1 - \delta_{ij}} \right]. \end{aligned}$$

The equality follows due to the fact that for the first $J_i - 1$ terms $\delta_{ij} = 0$ and for the final Jth term $1 - \delta_{ij} = 0$.

With an approach for estimating the hazard probabilities in place, we can extend our model to include time invariant and time-varying covariates. In Figure 3, we illustrate a model with a single time invariant covariate (x_1) and a single time-varying covariate (x_{2j}). In this model, we include a composite (η) that captures the time invariant influences on the hazard probability. In contrast to an endogenous latent variable, this composite does not have a disturbance associated with it, rather it is completely determined by the time invariant covariate. The composite does not vary over time, so the effects of the time-varying covariate are represented directly as paths from the covariate to the latent propensities for the respective time points. Note, however, that in this model the effects of the time-varying covariate are constrained to be equal across all of the time intervals. This corresponds with the proportional hazard odds property and, as noted above, can be relaxed and tested.

– Figure 3 about here –

More generally, the model illustrated in Figure 3 can be represented by the following system of equations:

$$\begin{aligned} \log\left[\frac{h(j)}{1-h(j)}\right] &= -\boldsymbol{\tau}_j + \eta + \boldsymbol{X}_j\boldsymbol{\kappa} \\ \eta &= \boldsymbol{X}\boldsymbol{\beta}. \end{aligned} \tag{15}$$

$\boldsymbol{X}$ and $\boldsymbol{X}_j$ are respectively matrices of time invariant and time-varying covariates, and $\boldsymbol{\beta}$ and $\boldsymbol{\kappa}$ are vectors of coefficients. As opposed to the vector of intercepts in (5), we have a vector of thresholds (τ_j). In addition, we see that

the log odds of the hazard is a function of η, which in turn is determined by $\boldsymbol{X\beta}$. In order to relax the proportional hazard odds property, we can simply free the constraint that the parameter estimates are equal across the time points. For time-varying covariates, this is easily reflected in our model by adding a subscript j to $\boldsymbol{\kappa}$. Since the composite η does not vary over time, in order to relax this constraint for time invariant covariates they needed to be added to the log hazard odds function. The system allowing for all coefficient estimates to vary over time can be written with the single matrix equation,

$$\log\left[\frac{h(j)}{1-h(j)}\right] = -\boldsymbol{\tau}_j + \boldsymbol{X\beta}_j + \boldsymbol{X}_j\boldsymbol{\kappa}_j. \tag{16}$$

4 Careers of Biochemists

To illustrate how to estimate a few basic discrete-time survival models in a structural equation modeling framework we draw on a dataset on the careers of biochemists provided as an example in Allison (1995). All of our structural equation models are estimated using Mplus (Version 5.1) (Muthén and Muthén 2007) and our comparison models are estimated using Stata (Version 10.0). We provide the Mplus code for all of our models in the Appendix.

The data for our examples consists of the years to promotion for 301 assistant professors of biochemistry (see Table 1 for descriptive statistics). The biochemists' careers were followed for up to 10 years after they were hired. Of the 301 professors, 72 percent were promoted during the 10 year period of ob-

servation. With this data we are able to examine the relationship between three time-invariant covariates and one time-vary covariate and the hazard of promotion. For time-invariant covariates, we have a measure of the selectivity of the undergraduate institution the individuals attended, whether the individual earned his or her Ph.D. from a medical school, and the prestige of the Ph.D. granting institution. For our time-varying covariate, we have a cumulative count of the number of article published by each individual for each year.

– Table 1 about here –

We first estimate a baseline model, illustrated in Figure 4 panel A, that just includes the hazard rate for each time period (see Table 2). Because this model does not include any covariates, the estimates for the hazard rates should be equal to the proportion of the sample promoted for that time period. This is, in fact, the case, which can be seen using the following formula relating the threshold estimates to the hazard rate and comparing the results with the descriptive statistics in Table 1:

$$h(j) = \frac{1}{1 + \exp(\tau_j)}. \tag{17}$$

For example, the estimated hazard probability for promotion in year 5 based on the model is $\frac{1}{1+\exp(1.09)} = 0.25$, which matches the proportion of the sample promoted in year 5 (see Table 1). We also note that the parameter estimates obtained from estimating the survival model as a SEM are within a hundredth of a decimal point of those obtained from estimating the model using the standard

logistic regression approach. The tau estimates are related to the intercept and parameter estimates for the time period indicators by

$$-\tau_j = \alpha + \beta_j, \tag{18}$$

so, for example, we see that the negative of the threshold estimate for year 3, 2.78, equals the intercept plus the estimate of the coefficient for the year 3 indicator variable, $-5.70 + 2.92 = -2.78$. Finally, we also note that we obtain the same log likelihood with both approaches.

– Figure 4 about here –

– Table 2 about here –

Adding covariates to the model (see Figure 4 panel B), we find that both undergraduate selectivity and the cumulative number of articles published are significantly positively associated with the hazard of promotion. As with our baseline model, we find that there are essentially no differences in the parameter estimates and the log-likelihoods between the two approaches. We also note that the standard errors for the estimates of the covariates are identical.

With this example we have little theoretical reason to expect the effects of the covariates may vary over time; however, it is generally a good idea to test whether the proportional hazard odds property should be relaxed. We illustrate how to do this by relaxing the constraint of equal parameter estimates for the effect of undergraduate selectivity (see Figure 4 panel C). We see some indication that the effect of undergraduate selectivity on the hazard of promotion varies

over time (see Table 3), but it does not seem to follow a meaningful pattern and for most years the effect is not statistically significant. Furthermore, a likelihood ratio test (see Table 4) indicates that the model relaxing the proportional hazard odds property does fit the data significantly better than a model maintaining the property. We also note that once again the parameter estimates we obtain using a SEM framework match those using a standard estimation procedure.

5 Extended Empirical Example

The last section illustrated that SEMs can incorporate the traditional discrete-time survival model and replicate results obtained with more traditional approaches. For our final empirical example we consider a model that illustrates some of the advantages of estimating discrete-time survival models as structural equation models. Research on the transition to first intercourse among adolescents has found religiosity and attitudes regarding sex are important predictors of initiation (Meier 2003). In theoretical models of this process, religiosity is posited to have both a direct effect on the probability of having sex and an indirect effect through its relationship to attitudes about sex. There are two advantages to estimating such a model as a structural equation model. First, religiosity and attitudes about sex are concepts for which we have several indicators. Treating these concepts as latent variables allows us to account for measurement error and to get a sense of how well the indicators capture the theoretical concepts. Ignoring the measurement error would create biased coef-

ficient estimates. Second, given that our theoretical model includes both direct and indirect effects, SEMs provide a convenient way of exploring these relationships.

5.1 Time to First Intercourse Data

For our analysis we use data from the Waves I and III of the National Longitudinal Study of Adolescent Health (Add Health). Add Health began as nationally representative sample of adolescents in grades 7 through 12 in 1994/95 (Harris et al. 2003). Respondents were followed for two additional waves of in-home interviews, with Wave III collected in 2002. We determine the age of first intercourse based on respondents' self-reports at Wave III. There is some concern about the accuracy of retrospective reports of age of first sex; however, Upchurch et al. (2002) found in an analysis of Add Health data through Wave 2, inconsistencies in reporting appeared to be random. In order to ensure that our measures of religiosity and attitudes regarding sex are temporally prior to initiation of sexual activity, we drop any cases in which the respondent reported having sex for the first time prior to the Wave I interview.[2] Finally, we drop a little less than 10 percent of the cases missing data on mother's education. These leaves us with an analysis sample of 9,914 respondents.

– Table 5 about here –

[2] Dropping these cases raises the possiblity of selection bias, which we do not address in this example (see Meier (2003) and Bearman and Brückner (2001) for approaches to accounting for selection bias or testing the robustness of the results).

Of 9,914 respondents, 18 percent did not initiate sexual activity between Wave 1 and Wave 3 of the study and so are treated as censored observations (see Table 5). The age of first intercourse in our sample ranges from 12 to 25. We only observe individuals for roughly seven years (the time between Wave 1 and Wave 3), so in creating our event indicators we have missing data at both ends of the age range. This should not be confused with left-censoring, but rather simply reflects the fact that respondents enter the data at Wave 1 at different ages.

Following Meier (2003), we consider four Likert-scale measures of religiosity: (1) "In the past 12 months, how often did you attend religious services?", (2) "In the past 12 months, how often did you attend such youth activities?", (3) "How often do you pray?", and (4) "How important is religion to you?". For the purpose of comparison, we construct a religiosity scale as the average of the non-missing items ($\alpha = 0.85$). Also following Meier (2003), we consider six attitudes about sex.[3] These items were only asked of respondents at least 15 years old and not married. Three of the attitudes express negative sentiments: If you had sexual intercourse, (1) "your partner would lose respect for you", (2) "afterward, you would feel guilty", and (3) "it would upset [name of mother]". Three others express positive sentiments: if you had sexual intercourse, (1) "your friends would respect you more", (2) "it would make you more attractive to women/men", and (3) "you would feel less lonely." Based on these items, we

[3] Meier also uses a seventh item: "if you had sexual intercourse, it would give you a great deal of physical pleasure." This item stood out in our measurement models and so we chose not to include it. [**Probably need a better/more precise justification for excluding this item.**]

created two scales: (1) negative sentiments about sex ($\alpha = 0.67$) and (2) positive sentiments about sex ($\alpha = 0.70$).

Research on the transition to intercourse has found that the probability of initiation differs by gender, race/ethnicity, and mother's education (Meier 2003, Bearman and Brückner (2001), Day (1992)). We include these background characteristics to account for the known differences. As our main purpose in this example is to explore the relationship between religiosity, attitudes regarding sex, and the initiation of sexual activity, this is not intended to be an exhaustive list of background characteristics that may be related to the transition to first intercourse.

In our first set of models we only use cases with complete data on the covariates. This involves dropping roughly 4,000 cases missing data on the measures of attitudes regarding sex due to the legitimate skip pattern noted above. Given the nature of the skip pattern and that we include the age of the respondent at Wave 1 as a covariate, one can make the case that these data are missing at random (MAR). Therefore, as a comparison model, we take advantage of the same method of handling missing data that we are using for the hazard component of the model to estimate the parameters using the full analysis sample.

5.2 Analysis of Time to First Intercourse

5.2.1 Measurement Models

Our first task is to develop the measurement models for our latent covariates religiosity and attitudes regarding sex. For religiosity we consider two mea-

surement models. Our first model involves a single latent dimension with four indicators, our second is a model with two latent dimensions each with two indicators (see Figure 5, Panels A and B). A model with a single latent dimension allowing for no measurement error and constraining the loadings to equal one is equivalent to the standard practice of creating a single scale from the four measures. We find that the model with a single latent dimension has a poor fit with the data according to a number of measures (see Table 6). In particular, the model χ^2 is significant, the RMSEA is well over 0.05, and the BIC is large and positive. As a justification for our second model, we felt that the concept of religiosity may consist of two dimensions: a dimension related to involvement in the religious community and a more personal dimension. For each of these dimensions we have two indicators. To identify the model, we scale the latent variable associated with involvement to our measure of attending services and we scale the latent variable associated with personal beliefs to the measure of the importance of religion. This model fits the data quite well. We find a non-signficant model χ^2, an RMSEA of 0.01, and a negative BIC.

– Figures 5 here –

Turning to the parameter estimates, we find both of the free factor loadings are significant in the expected direction of effect. We also find that three out of four of the measures have reliabilities between 0.7 and 0.85, but one, youth group attendance, has a more modest 0.42 reliability. Finally, we note that the estimated correlation between religious involvement and personal beliefs is 0.83. As we would expect, this correlation is high, but these do appear to be

two distinct dimensions. Given the strong overall fit and the sensible parameter estimates, we adopt this model with two latent dimensions of religiosity in the following analyses.

– Table 6 about here –

We consider a similar set of models for attitudes regarding sex (see Figure 6, Panels A, B, and C). We find that a one dimensional model does not fit the data. We find that a two dimensional model including latent variables for positive and negative attitudes related to sex has a substantially better fit with the data, but there are still some indications that the model does not adequately fit the data (see Table 6). The χ^2 remains significant, the RMSEA is greater than 0.05, and the BIC is positive. We examine a third model that introduces a correlation among the error terms for the two measures involving respect (one's partner and one's friends). We find that this model has a significantly better fit than the previous model, but the fit is still not ideal for the same reasons as the second model.

– Figure 6 about here –

Looking at the parameter estimates, we find that the free factor loadings are all significant and in the expected direction.[4] In terms of the reliabilities of the items, we find more variation than we did for the religiosity items. Three of the items have quite low reliabilities around 0.25: partner would lose respect, it would upset your mother, and friends would respect you more. Two items

[4] For ease of interpretation, we coded all of the attitude items such that high values would be theoretically associated with an increased likelihood of initiating sexual activity.

have reliabilities around 0.55: it would make you more attractive and you would feel less lonely. Finally, one item, you would feel guilty, has a high reliability of 0.82. The two dimensions of sexual attitudes do appear to capture fairly distinct dimensions as they have a relatively low correlation of 0.28. We also find that a significant negative association between the error terms for the two items measuring respect. This appears to be due to the fact that the original wording of the items had opposite valences, which were rendered the same by reverse coding the second item. Although the overall fit and the reliabilities of some of the items indicate that we do not have as strong a model for attitudes regarding sex as we do for religiosity, we do not see any other theoretically-motivated improvements to the model. So, in the following analyses, we use our third model.

5.2.2 Structural Models

As a basis for comparison, we first estimate a model in which we use only observed variables – that is, we use scales for religiosity, positive, and negative attitudes regarding sex (see Figure 7, Panel A). This corresponds to the typical sociological application where measurement error in covariates is ignored. In our second model, we replace the scales with the measurement models we developed for the two dimensions of religiosity and positive and negative attitudes related to sex (Figure 7, Panel B). Finally, in our third model, we build in the theoretical relationships among religiosity and attitudes related to sex (Figure 7, Panel C).

– Figure 7 about here –

In our first model, consistent with past research we find that religiosity is significantly negatively associated with the hazard of first intercourse (see Table 7). We also find that positive and negative attitudes regarding sex are significantly related to the hazard of first intercourse in the expected directions (we reverse coded the negative attitudes such that the expected effect is positive). When we account for measurement error in religiosity and attitudes related to sex a different pattern emerges. Although both dimensions of sexual attitudes remain significantly associated with the hazard of first intercourse, neither dimension of religiosity remains significant. This suggests that failing to account for measurement error may lead one to erroneously conclude that religiosity has a direct effect on the initiation of sexual activity among adolescents. This is not to say that religiosity has no effect. In model three we examine the effects of religiosity on attitudes regarding sex. We find that both dimensions of religiosity have a significant association with negative sexual attitudes and the dimension of religiosity related to one's beliefs has a significant association with positive attitudes concerning sex. In addition, both dimensions of attitudes regarding sex remain significant predictors of the hazard of first intercourse. Therefore, as one would anticipate from our theoretical model, religiosity has an indirect effect on the hazard of first intercourse through it's effect on attitudes related to sex.

– Table 7 about here –

In our final model, we reestimate our third model using all of the available cases. This increases our sample size by roughly 4,000 cases, almost entirely due

to the missing data for the sexual attitudes items. This approach also allows us to examine cases initiated sexual activity as early as age 12. Although the precise parameter estimates differ, we obtain substantively the same pattern of results as in our complete case analysis.

– Table 8 about here –

5.3 Summary

We chose this example to demonstrate some of the benefits of estimating discrete-time survival models as structural equation models. In this case, we see that accounting for measurement error in one of the key concepts is consequential for the substantive interpretation of the results. Without addressing measurement error in the religiosity scale, one might incorrectly conclude that religiosity has a direct effect on the age of first intercourse.

6 Conclusion

In this paper we have demonstrated how researchers can estimate discrete-time survival models in a structural equation modeling framework. In our first set of analyses, we documented that the parameter estimates from SEM discrete-time survival models match those obtained from the standard approach to estimating discrete-time survival models. In our second set of analyses, we illustrated some of the potential benefits estimating discrete-time survival models in a SEM framework. In particular, we provided an example in which accounting

for measurement error in the covariates had an appreciable effect on the results. Furthermore, we discussed another benefit of the SEM approach, the ability to explore mediational relationships and indirect effects. Given the potential presence of measurement error and indirect relationships in many areas of sociological and demographic research as well as the utility of survival analysis, the ability to estimate these models as SEMs may prove widely beneficial.

7 References

Allison, Paul D. 1984. *Event History Analysis: Regression for Longitudinal Event Data*. Newbury Park, CA: Sage Publications Inc.

Allison, Paul D. 1995. *Survival Analysis Using the SAS System: A Practical Guide*. Cary, NC: SAS Institute Inc.

Allison, Paul D. 2005. *Fixed Effects Regression Methods for Longitudinal Data Using SAS*. Cary, NC: SAS Institute Inc.

Andrews, Kenneth T. and Michael Biggs. 2006. "The Dynamics of Protest Diffusion: Movement Organizations, Social Networks, and News Media in the 1960 Sit-Ins." *American Sociological Review* 71: 752-777.

Asparouhov, Tihomir, Katherine Masyn, and Bengt Muthén. 2006. "Continuous Time Survival in Latent Variable Models." Proceedings of the Joint Statistical Meeting in Seattle, August 2006.

Bearman, Peter and Hannah Brückner. 2001. "Promising the Future: Virginity Pledges and First Intercourse." American Journal of Sociology 106: 859-

912.

Behrens, Angela, Christopher Uggen, and Jeff Manza. 2003. "Ballot Manipulation and the "Menace of Negro Domination": Racial Threat and Felon Disenfranchisement in the United States, 1850-2002." *American Journal of Sociology* 109: 559-605.

Bollen, Kenneth A. 1989. *Structural Equations with Latent Variables*. New York: Wiley.

Day, Randall. 1992. "The Transition to First Intercourse among Racially and Culturally Diverse Youth." *Journal of Marriage and the Family* 54: 749-762.

Guo, Guang. 1993. "Event-History Analysis for Left-Truncated Data." *Sociological Methodology* 23: 217-243.

Harris, Kathy M., F. Florey, J. Tabor, Peter Bearman, J. Jones, and J. R. Udry. 2003. "The National Longitudinal Study of Adolescent Health: Research Design." Available online at http://www.cpc.unc.edu/projects/addhealth/design.

Little, Roderick J. A. and Donald B. Rubin. 2002. *Statistical Analysis with Missing Data*. (2nd ed.). New York: Wiley.

Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage.

Masyn, Katherine. 2003. *Discrete-time Survival Mixture Analysis for Single and Recurrent Events Using Latent Variables*. Unpublished doctoral dissertation, University of California, Los Angeles.

Meier, Ann M. 2003. "Adolescents' Transition to First Intercourse, Religiosity, and Attitudes about Sex." *Social Forces* 81: 1031-1052.

Muthén, Bengt and Katherine Masyn. 2005. "Discrete-Time Survival Mixture Analysis." *Journal of Educational and Behavioral Statistics* 30: 27-58.

Muthén, Bengt and Linda Muthén. 2007. *Mplus* (Version 5.1) [computer software]. Los Angeles: Muthén and Muthén.

Singer, Judith D. and John B. Willett. 2003. *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence.* New York: Oxford University Press.

Sørensen, Jesper B. 2004. "The Organizational Demography of Racial Employment Segregation." *American Journal of Sociology* 110: 626-671.

Sweeney, Megan M. 2002. "Two Decades of Family Change: The Shifting Economic Foundations of Marriage." *American Sociological Review* 67: 132-147.

Upchurch, Dawn M., Lee A. Lillard, Carol S. Aneshensel, and Nicole Fang Lee. 2002. "Inconsistencies in Reporting the Occurrence and Timing of First Intercourse among Adolescents." *The Journal of Sex Research* 39: 197-206.

8 Tables and Figures

Table 1: Descriptive Statistics

	N	Mean	SD		N	Mean	SD
Ever Promoted	301	0.72	0.45				
Promoted y1	301	0.00	0.06	Tot Articles y1	301	4.03	3.72
Promoted y2	299	0.00	0.06	Tot Articles y2	299	5.05	4.41
Promoted y3	292	0.06	0.23	Tot Articles y3	292	6.25	5.20
Promoted y4	263	0.16	0.37	Tot Articles y4	263	7.37	5.57
Promoted y5	211	0.25	0.43	Tot Articles y5	211	8.42	6.16
Promoted y6	149	0.31	0.46	Tot Articles y6	149	9.36	6.74
Promoted y7	96	0.32	0.47	Tot Articles y7	96	9.71	7.42
Promoted y8	59	0.25	0.44	Tot Articles y8	59	10.25	8.09
Promoted y9	42	0.17	0.38	Tot Articles y9	42	10.12	9.70
Promoted y10	29	0.14	0.35	Tot Articles y10	29	9.41	8.06
Undergrad Select.	301	5.03	1.34				
PhD Med School	301	0.63	0.48				
PhD Prestige	301	3.20	0.98				

Table 2: Comparison of Parameter Estimates

	Baseline Model				Model with Covariates			
	Mplus		Stata		Mplus		Stata	
	Est	SE	Est	SE	Est	SE	Est	SE
τ_1 / int	5.70	1.00	-5.70	1.00	6.66	1.08	-6.66	1.08
τ_2 / y2	5.70	1.00	0.01	1.42	6.75	1.08	-0.08	1.42
τ_3 / y3	2.78	0.25	2.92	1.03	3.93	0.49	2.74	1.03
τ_4 / y4	1.66	0.17	4.04	1.02	2.86	0.45	3.81	1.02
τ_5 / y5	1.09	0.16	4.61	1.01	2.35	0.45	4.31	1.02
τ_6 / y6	0.81	0.18	4.90	1.03	2.11	0.45	4.55	1.02
τ_7 / y7	0.74	0.22	4.96	1.03	2.06	0.47	4.60	1.03
τ_8 / y8	1.08	0.30	4.63	1.05	2.41	0.51	4.25	1.05
τ_9 / y9	1.61	0.41	4.09	1.08	2.98	0.59	3.68	1.10
τ_{10} / y10	1.83	0.54	3.87	1.14	3.11	0.68	3.55	1.15
Und Select					0.17	0.06	0.17	0.06
PhD Med					-0.23	0.17	-0.23	0.17
PhD Prest					-0.03	0.09	-0.03	0.09
Cumul. Art.					0.07	0.01	0.07	0.01
N	301		301		301		301	
LL	-529.16		-529.16		-505.23		-505.23	

Table 3: Comparison of Models Relaxing Proportional Hazard

	Mplus		Stata	
	Est	SE	Est	SE
τ_1 / int	9.38	5.59	-9.38	5.59
τ_2 / y2	5.63	3.81	3.75	6.76
τ_3 / y3	4.95	1.21	4.43	5.71
τ_4 / y4	1.28	0.69	8.10	5.62
τ_5 / y5	2.74	0.77	6.64	5.63
τ_6 / y6	1.94	0.77	7.44	5.63
τ_7 / y7	2.36	0.95	7.02	5.66
τ_8 / y8	3.17	1.20	6.21	5.71
τ_9 / y9	5.44	1.86	3.94	5.88
τ_{10} / y10	4.68	1.93	4.70	5.90
PhD Med	-0.21	0.17	-0.21	0.17
PhD Prest	-0.03	0.09	-0.03	0.09
Cumul. Art.	0.08	0.01	0.08	0.01
Und. Sel. y1 / Und. Sel.	0.64	0.92	0.64	0.92
Und. Sel. y2 / Und. Sel. x y2	-0.06	0.74	-0.70	1.18
Und. Sel. y3 / Und. Sel. x y3	0.35	0.21	-0.29	0.94
Und. Sel. y4 / Und. Sel. x y4	-0.15	0.13	-0.79	0.93
Und. Sel. y5 / Und. Sel. x y5	0.23	0.13	-0.41	0.93
Und. Sel. y6 / Und. Sel. x y6	0.12	0.14	-0.52	0.93
Und. Sel. y7 / Und. Sel. x y7	0.22	0.17	-0.42	0.93
Und. Sel. y8 / Und. Sel. x y8	0.31	0.17	-0.33	0.94
Und. Sel. y9 / Und. Sel. x y9	0.64	0.33	0.00	0.97
Und. Sel. y10 / Und. Sel. x y10	0.47	0.35	-0.17	0.98
N	301		301	
LL	-499.63		-499.62	

Table 4: Likelihood Ratio Tests

	Log Likelihood
Model relaxing proportional hazard constraint	-499.62
Model with proportional hazard constraint	-505.23
-2*(Model 2 LL - Model 1 LL)	11.22
df	9
Chi-square test	0.26

Table 5a: Descriptive Statistics

	N	Mean	SD
Ever had sex	9914	0.82	0.38
Age 12 first sex	365	0.02	0.15
Age 13 first sex	1886	0.05	0.22
Age 14 first sex	3560	0.10	0.29
Age 15 first sex	5251	0.17	0.37
Age 16 first sex	6277	0.24	0.43
Age 17 first sex	6193	0.27	0.44
Age 18 first sex	5342	0.33	0.47
Age 19 first sex	3616	0.22	0.41
Age 20 first sex	2544	0.19	0.39
Age 21 first sex	1780	0.21	0.41
Age 22 first sex	1088	0.12	0.33
Age 23 first sex	675	0.12	0.33
Age 24 first sex	307	0.10	0.31
Age 25 first sex	77	0.09	0.29

Table 5b: Descriptive Statistics

	N	Mean	SD
Age at wave 1	9914	15.17	1.66
Female	9914	0.54	0.50
Black	9914	0.20	0.40
Hispanic	9914	0.08	0.27
Asian	9914	0.08	0.27
Other race	9914	0.10	0.29
Mother's education	9914	5.52	2.41
How often attend services	9905	2.87	1.19
How often attend yth grp	9908	2.21	1.26
How often pray	9904	3.72	1.51
How important religion	9911	3.13	1.02
Religiosity scale	9892	2.98	1.04
If you had sex:			
Partner would lose respect	6078	3.30	1.10
You would feel guilty	6096	2.64	1.23
Upset your mother	6095	1.85	1.06
Friends respect you more	6119	2.31	1.04
Make you feel more attractive	6055	2.35	0.99
Feel less lonely	6062	2.45	1.03
Neg Sex Attitudes scale	6043	2.60	0.88
Pos Sex Attitudes scale	6034	2.37	0.81

Table 6: Model Fit Statistics for Measurement Models

	N	Chi-Sq	df	sig	CFI	TLI	RMSEA	BIC
Religiosity:								
1 latent var	5980	552.20	2	0.00	0.95	0.86	0.21	534.80
2 latent vars	5980	1.71	1	0.19	1.00	1.00	0.01	-6.99
Attitudes:								
1 latent var	5980	2752.53	9	0.00	0.60	0.34	0.23	2674.26
2 latent vars	5980	150.70	8	0.00	0.98	0.96	0.06	81.13
2 lat vars, cor err	5980	131.78	7	0.00	0.98	0.96	0.06	70.91

Table 7: Parameter Estimates Using Complete Case Data (N=5980)

	Model 1		Model 2		Model 3					
					Neg Sex		Pos Sex		Age First Sex	
	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE
Rel Scale	-0.10	0.02								
Lat Attend			-0.06	0.07	-0.11	0.03	-0.02	0.02	-0.05	0.05
Lat Belief			-0.11	0.08	-0.10	0.03	-0.06	0.03	-0.07	0.06
Neg Sex Scale	0.12	0.02								
Lat Neg Sex			0.21	0.07					0.16	0.06
Pos Sex Scale	0.27	0.02								
Lat Pos Sex			0.95	0.00					0.70	0.07

Notes

SEs are robust standard errors.

All models control for gender, race/ethnicity, and mother's education.

Table 8: Parameter Estimates Using Full Sample (N=9914)

	Model 3					
	Neg. Sex Att.		Pos. Sex Att.		Age First Sex	
	Est	SE	Est	SE	Est	SE
Latent attendance	-0.12	0.02	-0.04	0.02	-0.04	0.05
Latent beliefs	-0.09	0.03	-0.02	0.02	-0.07	0.06
Latent negative sex att.					0.19	0.06
Latent positive sex att.					0.80	0.07

Notes:

SEs are robust standard errors.

All models control for gender, race/ethnicity, and mother's education.

Figure 1: Discrete-Time Survival Model Represented as a Structural Equation Model

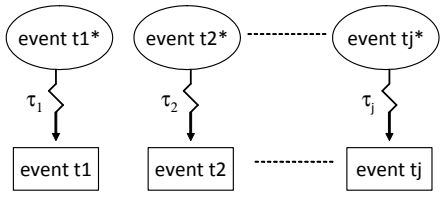

Figure 2: Discrete-Time Survival Model Data Formats

Case-Period Format

ID	time	event	x1	x2
1	1	0	1	4
1	2	0	1	2
1	3	1	1	3
2	1	0	0	5
2	2	1	0	9
3	1	0	1	6
3	2	0	1	8

Case Format for SEM

ID	e_1	e_2	e_3	x1	x2_1	x2_2	x2_3
1	0	0	1	1	4	2	3
2	0	1	.	0	5	9	.
3	0	0	.	1	6	8	.

Figure 3: Discrete-Time Survival Model with Covariates

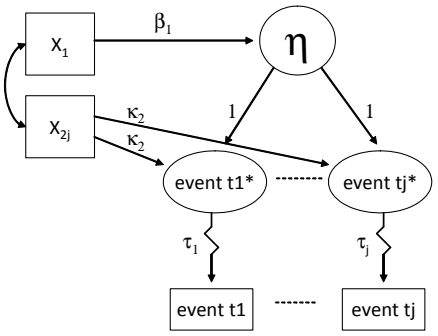

Figure 4: Models for Careers of Biochemists

Panel A: Baseline Model

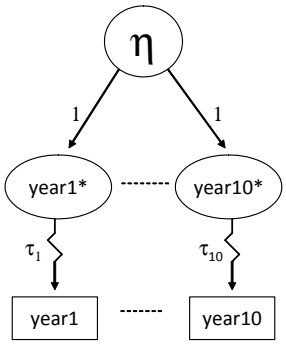

Figure 4: Models for Careers of Biochemists

Panel B: Model with Covariates

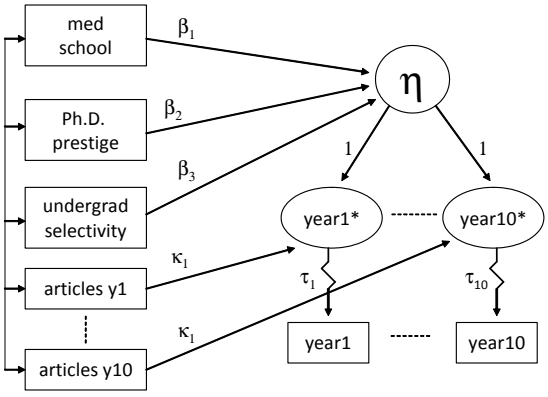

Figure 4: Models for Careers of Biochemists

Panel C: Model Relaxing Proportional Hazard Odds Constraint for Undergraduate Selectivity

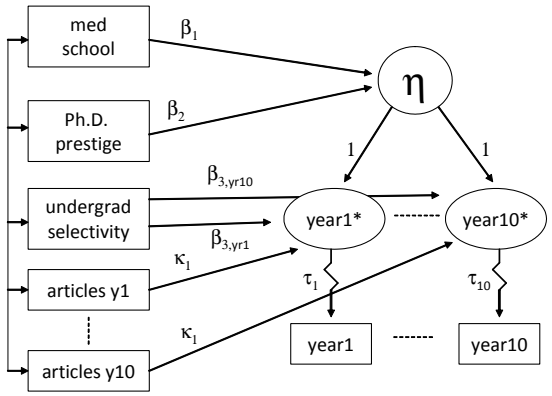

Figure 5: Religiosity Models

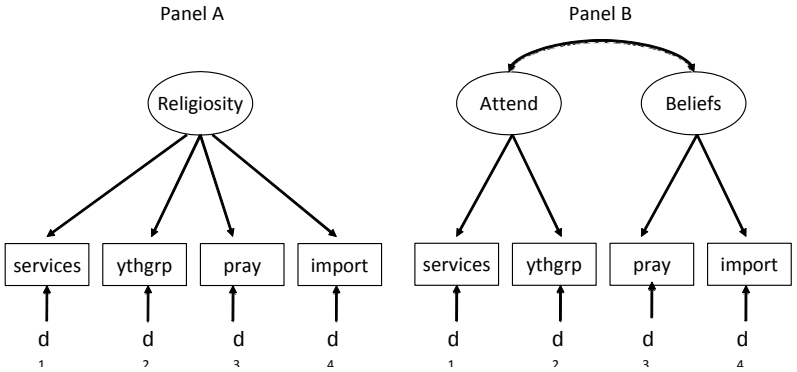

Figure 6: Sex Attitudes Models

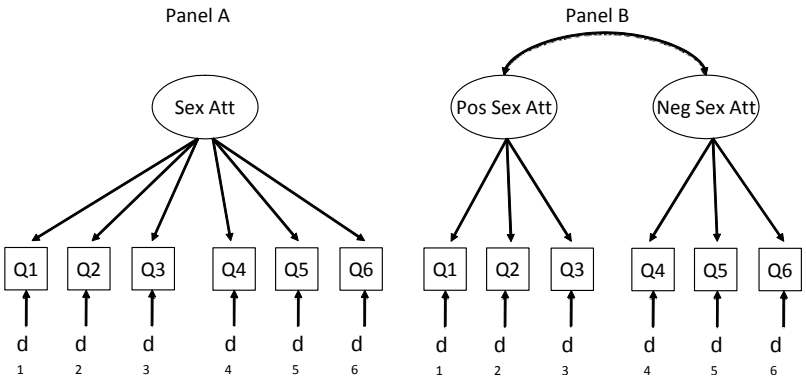

Figure 6: Sex Attitudes Models

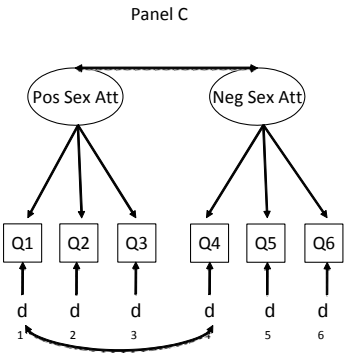

Figure 7: Models for Time to First Sex

Panel A: Model with Covariates (No Latent Variables)

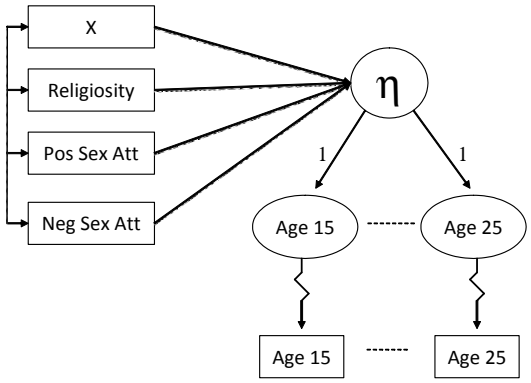

Figure 7: Models for Time to First Sex

Panel B: Model with Covariates (Latent Variables)

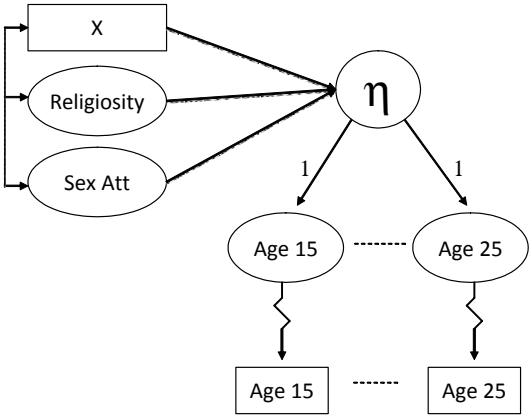

Figure 7: Models for Time to First Sex

Panel C: Model with Covariates (Latent Variables)