Marriage Markets, Timing of First Marriage, and Child Outcomes

Mariusz Kolczykieiwcz
Bialystok, Poland

M.A., University of Virginia, 2008
B.A, University of Michigan, 2004

A Dissertation presented to the Graduate Faculty
of the University of Virginia in Candidacy for the Degree of
Doctor of Philosophy

Department of Economics

University of Virginia
August, 2013

Abstract

I develop an estimation framework for a model tracking population changes over time where observations are recorded by multiple sources with differing frequency and precision. I apply this framework to recover the number of unmarried men and women by age and race in each state for every year over the last three decades by combining measurements from the Decennial Census, the Current Population Survey, and the American Community Survey. Estimates of the model parameters are used to generate yearly measures of sex ratios and to estimate the impact of this marriage market feature on the marriage timing decision of women. Previous research has relied on the infrequent snapshots of marriage markets features. Using the yearly estimates of sex ratios I find a decrease in waiting time to first marriage for National Longitudinal Survey of Youth 1979 (NLY79) women when the relative number of potential spouses increases.

Yearly estimates of potential partner supply are also used to identify and estimate the effect of mother's marriage on children born out of wedlock. I use the Child and Young Adult Survey (CNLSY79), a survey that collects information on children of the NLSY79 women, to link children with their mothers and the marriage markets these women experienced. Ordinary least squares (OLS) regressions show that black children whose mothers marry are more likely to obtain a high school diploma relative to children whose mothers remain unwed. The effect of marital union is also positive when sex ratios are used as instruments for the mother's marriage decision. However, the instrumental variable results are not robust to the inclusion of other explanatory variables that can affect marriage market features.

JEL-Classification: I20, J11, J12, J13

Keywords: Demographics, Marriage, Census, Children, Education

Acknowledgments

I would like to express deep gratitude to my advisors Steven Stern, John Pepper, and Wayne-Roy Gayle. Their generous advice, guidance, and support made this work possible. I am especially thankful to Steven Stern for encouraging me to pursue the research ideas in my doctoral thesis and to John Pepper for helping me to express them clearly. I am also grateful to Melvin Wilson for his comments and suggestions.

This dissertation has benefited from discussions with Asli Senkal, Michael Schreck, Christopher Clapp, Catherine Alford, Leora Friedberg, Sarah Turner, Amalia Miller, Daniel Fried, and Michael LaForest.

I would also like to thank my family for their support and encouragement during this process. All errors are my own.

This research was conducted with restricted access to Bureau of Labor Statistics (BLS) data. The views expressed here do not necessarily reflect the views of the BLS.

Contents

Introduction

Empirical research on the effect of marriage market characteristics on the marriage decision often contends with a shortage of data available to describe markets. A decline in the relative availability of partners has been found to decrease the probability of marriage for females, however reliable data measuring potential partner supply is often infrequent. I address the deficit in reliable and frequent demographic data by developing and estimating a model that combines multiple sources of data to recover frequent partner availability measures. The model incorporates information from multiple sources of data, where each source has a different frequency of observation and propensity for accurate measurement of populations by demographic category. I find that lower supply of potential partners increases waiting time until first marriage for women, and that the model generated measures of partner supply have larger predictive power than measures constructed solely form the observed data. Besides helping to predict marriage, better data on partner availability can also be used to study the link between marriage and other outcomes. Whenever partner availability is uncorrelated with the outcomes of interest, the former can be used as an instrumental variable for marriage. The frequent partner availability measures generated by the model are used to investigate the effect of mother's marriage on child outcomes for a group of children born to unmarried women. I find some evidence that a mother's marriage increases the probability that her child obtains a high school diploma.

In the first chapter of this dissertation, I develop and estimate a model of observed population counts by demographic category. I combine information on population counts from the Decennial Census, the American Community Survey (ACS), and the Current

Population Survey (CPS). The model of the data generating process accounts for missing observations and allows for data-source-specific measurement error. Estimates of model parameters permit the construction of sex ratios, a measure of the number of unmarried men for each unmarried woman, for each marriage market and time period. The model generated ratios are then used to predict the effect of partner availability on waiting time to first marriage for the women of the National Longitudinal Survey of Youth 1979 (NLSY79). In accordance with previous research, I find that higher relative supply of potential partners decreases waiting time to first marriage. Additionally, compared to coefficients for ratios based purely on the data, the coefficients on the ratios from the model have higher predictive power, suggesting the possibility of attenuation bias due to measurement error in the ratios based solely on the data. The ratios generated from the model appear to succeed in reducing measurement errors and seem to reflect the true sex ratios faced by the NLSY79 respondents more accurately.

In the second chapter of this dissertation, I estimate the impact of mother's marriage on outcomes of children born to unwed mothers. Ordinary least squares (OLS) regressions show that children of the NLSY79 mothers who marry are more likely to graduate from high school or obtain a GED relative to their peers whose mothers remain unmarried. Because mother's marriage is likely to be correlated with unobservables in the high school graduation equation, I also use measures of sex ratios estimated by the model described in the first chapter as instrumental variables for the marriage decision. The results of Chapter 1 indicate that sex ratios have an impact on the waiting time until first marriage for the NLSY79 women. I use yearly measures of partner availability in the mother's marriage market in a General Method of Moments (GMM) estimation framework and find that mother's marriage has a large and positive effect on the high school graduation status of their children. However, the coefficient estimates decline and become insignificant when state-specific prison rates are included in the outcome equation, suggesting unobservable factors related to incarceration might affect the sex ratios in the mother's marriage market and high school graduation status of their children. These results suggest caution when considering the GMM coefficient estimates.

Chapter 1

Estimating Marriage Market Features Using Multiple Sources of Data Measured with Error

1.1 Introduction

Marriage is correlated with a range of outcomes for the individuals entering marital unions and with outcomes of children they might choose to have. Motivated by these correlations researchers have long been interested in learning about the connection between marital decisions and marriage market characteristics. Longitudinal datasets, which often contain detailed information about the timing into first marriage, are particularly suited for this task. To fully use the information on waiting times and transitions one also needs a good set of measures characterizing the marriage market over the observed period. The further the measure of the relevant market feature from the truth, the more likely one is to underestimate its impact, if any, on marital decisions.[1] Unfortunately for many of the periods spanned by older longitudinal datasets like the National Longitudinal Survey of Youth 1979 (NLSY79), there is no accurate yearly source of data on population counts disaggregated by many demographic categories. Population counts by category allow flexibility in

[1] See Brien (1997).

construction of marriage market features along many demographic dimensions of interest.

I address the lack of appropriate data by combining three different sources of data measuring the U.S. population. I focus on the sex ratio, a measure of the number of unmarried men for each unmarried woman, as the market feature that captures relative supply of available partners. To recover measures of sex ratios for each market and each time period, I use a system of equations framework to develop and estimate a model of the mean number of individuals in each demographic category of interest. The model accounts for missing observations, allows for and identifies data-source-specific measurement error, and is flexible enough to allow for sex ratio definition along any measured demographic characteristic. With the parameter estimates in hand, I construct market- and time-specific estimates of the sex ratios. Using the NLSY79 data I estimate the impact of sex ratios on waiting time until first marriage, and compare the coefficients when ratios are based solely on the observed data to coefficients on ratios generated by the model. Results suggest coefficient estimates on the sex ratios from the observed data are biased toward zero relative to estimates based on model generated ratios.

The most accurate source of demographic data comes from the Decennial Census. Every ten years the Census Bureau collects information about all the individuals living in the United States in that year. The Public Use Microdata Samples (PUMS), a subsample of the Census, includes individual level records which are weighted to be representative of the entire population. Using the Census PUMS allows one to generate total numbers of individuals by any measured demographic characteristic of interest. These in turn can be used to construct a wide variety of measures of marriage market features. However, the precision and flexibility afforded by the Census is offset by the infrequency of data collection.

In light of the data limitations, many researchers rely on infrequent but detailed observations of the U.S. population to connect aggregate marriage rates to marriage market characteristics like the sex ratios. For instance, South (1992) and Frieden (1974) use a single year of observations while Wood (1995) and Angrist (2002) rely on multiple Decennial Census surveys to connect aggregate marriage rates with marriage market characteristics. Some researchers use longitudinal datasets to model the effect of sex ratios on marriage

at the individual level, but find the lack of appropriate marriage market measures to be a hindrance. Brien (1997), who uses the National Longitudinal Study of 1972 (NLS72) and the 1980 Decennial Census, often relies on a measure at a single point in time to describe the marriage markets faced by the sample over nearly a 15 year period. Lichter et al. (1992) also use the 1980 Census data to construct sex ratios, but limit the waves of the NLSY79 survey they use to a 7 year period between 1979 to 1986 to assure more accuracy between their measures and the ones experienced by the women in the NLSY79. Lack of frequent demographic data has an impact on research outside the marriage market literature as well. Often linear interpolation between Census points is used to generate estimated population counts for the intercensal years.[2]

Instead of artificially censoring longitudinal data or relying solely on linear interpolation of demographic data, I mitigate the infrequency of the Census data by supplementing it with yearly information from the Current Population Survey (CPS) and the American Community Survey (ACS). While data frequency is an advantage over the Decennial Census, smaller sample size of the ACS and especially the CPS make them prone to larger deviations from the actual population counts. Sample deviations imply that each population count observed by any of the data sources is likely to be measured with error. Because of smaller sample size, the ACS and the CPS are also more likely to have missing observations. In these surveys, individuals with a certain set of characteristics might not be sampled. For instance, the CPS might not sample the number of 20 year old black females in Indiana in 1990 but will sample the number of 20 year old black males. I develop and estimate a model that accounts for both of these features in the observed data, the presence of data-source-specific measurement error and missing observations.

The remainder of my paper is organized as follows: First, I highlight issues with infrequent demographic data encountered in previous research. Next, I define marriage markets and sex ratios, discuss how limitations of available data affects the sex ratios, and describe the data used. Later, I present the model and the estimation strategy. Last, I discuss the results.

[2]See Wolfers (2006), Caceres-Delpiano (2012), Johnson (2009).

1.2 Literature Review

1.2.1 Demographic Data Limitations in Research

Research on the impact of marriage market features on marital decisions often depends on infrequent demographic data. Observing market characteristics at only a few points in time influences, and sometimes limits, the ability to identify the effects of interest. Some studies, like Freiden (1974) and South and Lloyd (1992), use data from a single point in time, and rely on cross sectional covariation in market characteristics, like sex ratios and marriage rates.[3] Others, like Wood (1993) and Angrist (2002), use repeated cross section data measured once every decade, and add covariation in time to estimate the relationship between marriage and market features.[4] Research that utilizes detailed information on waiting times to marriage employs infrequent data to characterize market features as well. Lichter et al. (1992) use the NLSY79 data waves between 1979 to 1986 to connect sex ratios generated from the 1980 Census to the timing of first marriage for each female respondent. Brien (1997) uses the NLS72 data waves between 1972 and 1986 and often just one point in time, the 1980 Census, to characterize marriage markets.[5] Studies that use longitudinal data sets are a particularly good example of how lack of frequent demographic data can be a hindrance. Survey respondents are observed over multiple periods of time, but infrequent data on marriage markets lacks information on how markets change over time. Lichter et al. (1992) limit the waves of the NLSY79 they use so that the marriage markets experienced by the respondents are close to the measures from the 1980 Census, while Brien (1997) often uses measures at only one point in time to describe markets over a 14 year period.

A more frequent demographic data source for the time frame of the older longitudinal data sets like the NLSY79 or NLS72 is available. The Current Population Survey (CPS) collects data every year and can be used to construct yearly marriage market characteristics.

[3]Freiden (1974) uses the 1960 U.S. Census data, while South and Lloyd (1992) use the 1980 U.S. Census to estimate the impact of marriage market conditions on marriage rates.

[4]Wood (1993) uses the 1970 and 1980 U.S. Census data, while Angrist (2002) uses the PUMS data files from 1910, 1920 and 1940

[5]Some of Brien's measures rely on two points in time, but are also constructed from the 1980 Decennial Census. The 1980 Census asked a subsample of respondents questions about their residential status five years prior to the Census, allowing him to construct the 1975 ratios for some definitions.

However, the measures derived from the CPS are likely to be inaccurate.[6] For example, the work of Seitz (2009) relies on measuring sex ratios every year during a 15 year period for a cohort of NLSY79 respondents.[7] To construct the sex ratios for each sample year Seitz (2009) reweights the NLSY79 sampling weights so that the number of married and unmarried men and women in each marriage market match the CPS counts for that category and year.[8] After reweighting, new ratios are constructed from the counts of unmarried individuals for each market and year. The resulting ratios, however, appear to be measured with error. Seitz divides the U.S. into four regional marriage markets and in three of the four regions reports an increase in sex ratios for whites as the NLSY79 respondents grow older. This result runs counter to other empirical evidence, where sex ratios tend to decline with age.[9] Besides the direction of change, the magnitude and volatility of changes in the ratios cast additional doubt on the suitability of the sex ratios used. For instance, ratios for whites in the Western region of the U.S. increased from approximately 1 to 1.4 in a single year; a 40 percent increase in the relative availability of unmarried men in a region encompassing 14 states of the union.

Inaccurate ratios can be detrimental to estimation, and the propensity of the CPS to measure population counts with error is a likely contributor to the poor features of the ratios used by Seitz (2009). Brien (1997) presents evidence that measurement errors in population counts might cause attenuation bias in the predicted effect of sex ratios on marriage. He compares the predictive power of sex ratios defined at three geographic levels, the state, SMSA, and the county group. In theory, smaller geographic areas should approximate the actual marriage market faced by individuals more accurately. However, Brien (1997) finds that coefficient estimates decline as the geographic areas decrease. Using a larger subsample

[6]The small sample size of the CPS is one of the reasons for the large propensity for measurement error, suggesting one possible reason why it was not chosen by Lichter et al. (1992) and Brien (1997). See Section 1.8.

[7]Analysis is from 1979 to 1994. She derives a dynamic equilibrium model, where men and women make marriage and employment choices every period, and are able to predict future sex ratios. Unfavorable ratios in the future imply higher search frictions for an adequate partner and can impact the decisions made in the current period.

[8]Seitz (2009) explains that attrition in the NLSY79 might cause mismeasurement of individuals in various categories, necessitating the reweighting.

[9]See Lichter et al. (1992) and Table 1.11. Higher mortality and incarceration rates for men cause sex ratios to decline as cohorts age.

of the Decennial Census he constructs a different measure of marriageable men and uses it as an instrumental variable for his preferred ratios.[10] The instrumental variable results support the measurement error hypothesis, smaller geographic areas depend on smaller number of observations and are likely to be more volatile and poorly measured.

Brien (1997) and Lichter et al. (1992) serve as a jumping off point for some of the work in this paper. Both studies would have benefited from access to data on market characteristics for all period actually experienced by the longitudinal survey respondents. They use many observations on respondents over time, but only one or two observations on marriage market characteristics experienced by the survey participants. However, frequent data is not enough. As shown by Brien (1997), the coefficient predicting entry into marriage will be biased toward zero if the data used has large measurement error. The model I develop and estimate allows for construction of marriage market features that are available yearly and are likely to reduce the amount of measurement error.

The results of this paper are also applicable to a wide range of other research topics. Besides research on sex ratios and marriage markets, lack of frequent and accurate demographic data has an impact on a variety of applied work. Many times, to fill the gaps in the explanatory variables, researchers use Census data and linearly interpolate between two, or more, data points. Wolfers (2006) investigates the impact of divorce laws on divorce. One of his explanatory variables uses the proportion of married individuals by state, which is constructed using the Decennial Census data and interpolated for the non-Census years. While studying the impact of divorce laws on crime Caceres-Delpiano and Giolito (2012) also linearly interpolate part of a series on population counts for each state (s), race (b) and year (t) combination. Drewianka (2008) uses linear interpolation to generate yearly measures of population characteristics when researching the impact of divorce laws on family formations. He uses the Census micro data for 1950- 2000 to generate frequent measures on characteristics like the distribution of the population in various age or race categories.

[10]Brien (1997) uses the 5 percent PUMS file to create his original ratios. He uses the State Summary Tape File (STF) from the 1980 Census to construct the instruments. The STF data includes 19 percent of the U.S. population, unfortunately one is forced to use tabulations for categories already constructed by the Census limiting how marriage markets are defined.

The lack of timely demographic data also impacts fields outside of economics. Johnson and Raphael (2009) investigate the relationship between incarceration rates and AIDS infection rates. To construct the rate of AIDS infections, they use linear interpolation to connect the Census micro data for each state (s), race (b), age (a), sex (x) category. Linear interpolation imposes what might be a strong assumption on the evolution of populations between Census years. Without more frequent data, researchers give up any non-linear variation between time periods and between groups. I incorporate yearly information on population characteristics to supplement the infrequent data, and recover population counts in the intercensal years.

1.2.2 Marriage and Marriage Market Characteristics

Despite data limitations researchers find that marriage market conditions have an impact on the marital decision. In a wide range of time periods, data sources, and empirical specifications considered, higher availability of potential partners increases the probability that women marry. Freiden (1974) derives and tests the predictions of a static model relating marriage market features to martial behavior.[11] His model predicts that an increase in the number of males, will generate an increase in the number of marriages. Freiden (1974) uses the 1960 U.S. Census data to regress female marriage ratio on sex ratios, earnings of men relative to women, and cost of divorce indicators and finds that higher ratios have a statistically significant and positive impact on the proportion of females married. South and Lloyd (1992) use the National Center for Health Statistics (NCHS) from 1980 and 1981, and the 1980 U.S. Census data to investigate differences in the predictive power of a variety of partner availability measures on rates of new marriage formations. They find that higher ratios of unmarried males to females increases the rate of black marriages. When marriage markets are further disaggregated by education level the effect of higher male availability is positive and statistically significant for both whites and blacks. Wood (1993)

[11] Individuals decide to marry when utility flow from marriage exceeds the utility of remaining single. Marriages form as long as both members of the couple are weakly better off when married. Differences in utility of remaining single across individuals implies some individuals with high values of being single remain unmarried.

uses the 1970 and 1980 waves of the U.S. Census to study the effects of relative availability of marriageable men on the fraction of women ever married at the SMSA level. Data from multiple time periods allows him to control for unobservable SMSA-level fixed effects that generate different marriage rates and might be correlated with partner availability measures.[12] He finds that an increase in marriageable men increases the share of females that marry.[13]

More recent work utilizes natural experiments as a cause for variation in the supply of potential partners. Angrist (2002) uses changes in the US immigration policy in the early 20th century as exogenous variation that changed sex ratios for various ethnic groups in the U.S.. Marriages of second generation immigrants, who to a very large degree still married within their ethnic group, were effected by changes in the supply of partners. His results show that a decrease in sex ratios decreased the probability of marriage for women.[14] Abramitzky et al. (2011) use the differential mortality rates in World War I for French men to investigate the impact of sex ratios on a variety of marital market outcomes. Male mortality during World War I was high and varied considerably between different regions of France. French soldiers usually served in regiments with others from their region, and regiments suffered different casualty rates based on the battles in which they participated. The authors find that relative scarcity of men allowed French men to marry women of higher social class, that in regions with higher mortality rates women were less likely to marry, and that women married at older ages.

Research on the effect of sex ratios on timing of first marriage also finds that a higher relative supply of partners induce women to enter marriage earlier. Lichter et al. (1992) combine the 1980 Census data with the NLSY79 data between 1979 to 1986 to investigate the timing of the transition to first marriage for female respondents. Sex ratios, defined as

[12] Only factors that are constant over a decade are controlled for because of the spacing of the U.S. Census.

[13] Because his definitions depend on employment and income characteristics, which can be correlated with marriage, Wood instruments changes in his measures of relative availability of marriageable men using changes in industrial structure of SMSA and the change in relative number of men in the military. The two stage least squares coefficients decrease in magnitude relative to OLS estimates and are insignificant.

[14] The immigrant counts in the Census are subject to measurement error, due to return migration for instance. Angrist instruments these with measures of immigrants admitted into the United States. The 2SLS estimates are significant and also predict that lower sex ratios decrease probability of marriage.

number of men over women, are age specific and change as respondents get older.[15] The authors find that higher ratios reduce the waiting time to first marriage for NLSY79 women. However, they use only one point in time, the 1980 Census, to construct the sex ratios. Brien (1997) builds on the work of Lichter et al. (1992). Using National Longitudinal Study of 1972 (NLS72) data he compares the predictive power of various definitions of the relevant marriage markets in explaining waiting times to first marriage. Definitions vary based on the geographic size of the market, or the economic characteristics considered, but are all constructed using the 1980 Census. All of the definitions considered imply that higher male to female ratios reduce waiting time to first marriage for women.

1.3 Definition of the Relevant Markets and Sex Ratios

I use sex ratios to capture the relative availability of potential partners from the perspective of a woman of race (b) and age (a) living in state (s) at time (t). In order to measure the relative supply of potential partners, I must first define the relevant market. While the relevant marriage market varies from one individual to another, there are some common features and constraints which I attempt to address. First, I assume that women prefer unmarried men of the same race (b).[16] Next, I assume that most search occurs within a limited geographic region. Ideally one would like to define the area to be small enough to accurately reflect the true pool of available partners. However, due to data limitations and possible heterogeneity in the geographic limitations across individuals, I choose to use the state (s) as the appropriate unit.[17] Finally, the relevant marriage markets vary with the age of the woman. I follow Lichter et al. (1992) and assume that a woman of age (a) considers unmarried men between the ages (a) and $(a + 10)$ as potential partners and unmarried women between the ages $(a - 2)$ and $(a + 8)$ as potential competitors.

[15] The ratios include only individuals close to the age of the respondent.

[16] In 1980 for instance, 98.15 percent of white women were married to white men, 98.5 percent of black women were married to black men, and 80 percent of Hispanic women were married to Hispanic men. See Finlay and Neumark (2010) Table 1.

[17] The smaller the area, the more potential for individuals endogenously selecting to move. This is less likely to happen across states. A broad geographic definition also allows for less measurement error in the sex ratio even if statewide sex ratios are imprecise measures of the relevant marriage market.

Let $\hat{U}^M_{s,b,a,t,d}$ and $\hat{U}^F_{s,b,a,t,d}$ represent the observed number of unmarried males (M) and females (F) from data source (d), at time (t), for age (a), state (s), and race (b). Using data source (d) the sex ratio at time (t) is constructed as

$$\hat{r}^d_{s,b,a,t} = \frac{\Sigma^{10}_{k=0}\hat{U}^M_{s,b,a+k,t,d}}{\Sigma^{8}_{k=-2}\hat{U}^F_{s,b,a+k,t,d}}. \tag{1.1}$$

The observed data cells, $\hat{U}^M_{s,b,a,t,d}$ and $\hat{U}^F_{s,b,a,t,d}$, are measures of the true underlying number of unmarried males and females, $U^M_{s,b,a,t}$ and $U^F_{s,b,a,t}$, measured with error. The size of the errors is likely to vary depending on the data source. The CPS, the only source available every year during the NLSY79 time frame, is likely to be the least precise. Unlike the Decennial Census and the ACS, the CPS has a much smaller sample size and was not designed to be primarily a census of the U.S. population.

Each observed sex ratio, $\hat{r}^d_{s,b,a,t}$, is made up of twenty different data cells, and ratios that share the same (s, b, t) and are close enough in age will have many observed data cells in common. For example, $\hat{r}^d_{s,b,a,t}$ and $\hat{r}^d_{s,b,a+1,t}$ will share eighteen of the same data cells in their constructions.[18] The number of data cells in common decreases as sex ratios move farther apart in age and only when the difference in age exceeds ten are the sex ratios not reliant on common observations. Therefore, a measurement error in any of the observed data cells will propagate across many of the observed sex ratios. Further, even if measurement errors are data cell-specific and are independent across cells, because sex ratios share common data cells in their construction, the measurement error in $\hat{r}^d_{s,b,a,t}$ will be correlated with measurement errors in sex ratios for other age categories.

In addition to errors from mismeasurement of cells, missing observations on data cells $\hat{U}^M_{s,b,a,t,d}$ and $\hat{U}^F_{s,b,a,t,d}$ will induce more measurement errors in $\hat{r}^d_{s,b,a,t}$. In any given year some of the relevant data cells might not be sampled by data source (d). For instance, the CPS might not measure the number of 20 year old black females in Indiana in 1990 but will measure the number of 20 year old black males. Just like data cell-specific measurement

[18] The overlapping observations for males are $\hat{U}^M_{s,b,a+1,t,d}, \hat{U}^M_{s,b,a+2,t,d}, \ldots, \hat{U}^M_{s,b,a+10,t,d}$, and for females, $\hat{U}^F_{s,b,a-1,t,d}, \hat{U}^F_{s,b,a,t,d}, \ldots, \hat{U}^F_{s,b,a+7,t,d}$.

errors, errors in $\hat{r}^d_{s,b,a,t}$ due to missing observations will also propagate across ages and induce correlation in errors for sex ratios across ages within a given (s, b, t, d) category.

Large measurement error in $\hat{r}^d_{s,b,a,t}$ will invariably be detrimental when estimating the effect of the sex ratios on marriage.[19] The problem is more likely to affect sex ratios constructed using the CPS, the primary source of data for many of the years of interest in the older longitudinal data sets. To diminish the impact of measurement errors but still obtain a measure of sex ratios for every year, I develop and estimate a model that generates the observed $\hat{U}^M_{s,b,a,t,d}, \hat{U}^F_{s,b,a,t,d}$. I then recover the conditional expectation of sex ratios for each (s, b, a, t)

$$E[\hat{r}^d_{s,b,a,t} | \Sigma^8_{k=-2} \hat{U}^F_{s,b,a+k,t,d} > 0], \tag{1.2}$$

where the conditioning statement ensures that the observed sex ratios exist.[20] The expectation integrates out measurement errors providing a more accurate measure of sex ratios every year in each marriage market.

1.4 Data

1.4.1 Data on Populations by Demographic Category

I combine three different sources of measurement for data cells of interest. The U.S. Census, the American Community Survey (ACS), and the Current Population Survey (CPS). At any given time (t), at most only two measures for the same observation will be available, and, for many of the years, only one source of data is available. To ensure comparability of the sample being measured by each data source, I exclude the members of the military and those living in institutionalized group quarters like correction facilities and mental institutions.[21] While prisoners and inmates in other types of institutions are unlikely to participate in marriage

[19] See Brien (1997).

[20] For notational convenience, in the remainder of this paper I suppress the conditioning statement in the expectation. Whenever an expectation of a ratio is discussed the conditional statement guaranteeing that the denominator is not zero should be assumed. For example, the above conditional expectation will be expressed as $E[\hat{r}^d_{s,b,a,t}]$.

[21] The CPS does not sample the same groups as the Census and ACS. To ensure comparability, individuals living in the following quarter types are excluded from the Census and the ACS: institution, correction institution, mental institution, institutions for the elderly, institutions for the handicapped, institutions for the poor, and the military.

markets at time (t), the exclusion of members of the military might be problematic. For instance, if the military accounts for a large fraction of the state's population for a given race (b), then exclusion of the military might produce biased estimates of the true sex ratios in that state. This problem is likely to affect only small states with very large military bases.

U.S. Census

I use the 5% Public Use Microdata Samples (PUMS) for the years 1980, 1990, and 2000. These represent a random sample covering 5% of the housing units in the United States at the time of the survey. The PUMS are drawn from the full Census sample, which is the fraction of the population that receives the long form of the Census (approximately 1 in 6 or 16.7% of the entire population). To tabulate an estimated number of indiviudals in a given data cell ($\hat{U}^x_{s,b,a,t}$), I sum the weights for the unmarried individuals that share the same (s, b, a, x, t) characteristics. Only geographic units that contain at least 100,000 people can be identified in the data.

The American Community Survey (ACS)

I use yearly waves of the American Community Survey (ACS) starting with the year 2001. Each year between 2001 and 2004, the ACS sampled around 1 in 240 individuals in the United States. No geographic unit smaller than the state could be identified. Starting in 2005, the sample size improved to include 1 in every 100 individuals in the US with users able to identify geographic units that are subsections of the state. Estimates for each data cell are obtained by summing over the individuals that share the same demographic characteristics of interest.

Current Population Survey (CPS)

I use the CPS waves starting with the year 1979 sample. The survey size increased from roughly 150,000 individuals to about 200,000 individuals in the United States since 1979. The national sample consists of independent samples for all states and the District of Colombia. Each CPS household is interviewed for 4 consecutive months, given an eight month

break from interviews, and then interviewed again for another 4 months. The CPS interviews individuals who are not institutionalized and not in the Armed Forces. Individuals who live in prisons and other types of institutions are also excluded. Each person sampled in the survey is assigned a CPS derived weight. An estimated number of people in any given data cell is obtained by summing the weights for unmarried males and females with the same (s, b, a, t) characteristics. For a discussion of CPS weights and the possible impact of using weights in estimation see Appendix 1.A.

Data Cleaning

Each of the three above mentioned data sets is obtained from the Integrated Public Use Microdata Series (iPUMS). iPUMS preforms a considerable amount of work to standardize definitions and categories across data sets and across years within each data source. The number of unmarried individuals by state of residence, race, and sex is constructed by summing population weights for all individuals that report those characteristics. Individuals are considered unmarried if they reported never being married, if they are divorced, or widowed at the time of the interview. I estimate the model for three race categories: whites, blacks, and Hispanics. The Hispanic category includes any individual that reported being Hispanic regardless of the race association. The whites category includes individuals who reported white but not Hispanic, and the black category includes individuals who reported black but not Hispanic. All other race-ethnicity categories, American Indian or Chinese for example, were excluded due to the relatively small size of those groups. In addition, because the number of observations for blacks in the the state of Montana was small, I exclude that category from estimation. Only individuals between the ages of 14 and 60 at the time of the interview were included.

There is considerable variation in the number of data cells observed by each data source across geographic locations. A data cell is defined as the number of individuals in state (s), race (b), sex (x), and age (a) at the time period when a survey was fielded. Figure 1.1 shows the fraction of data cells observed by data source and state for the 20 most populous states. The fraction for each state and data source is obtained by calculating the number

of race, sex, and age categories for the years the survey was actually fielded. For instance, because the Census was fielded only 3 times then for each state the denominator includes 3 race categories times 46 age categories times 2 sex categories over 3 years.[22] The numerator includes the actual number of cells observed. Figure 1.1 shows that the CPS has the highest propensity to not observe data cells.

1.4.2 Marriage Timing Data

To ascertain the impact of sex ratios on marital transitions I use the National Longitudinal Survey of Youth 1979 (NLSY 79). The NLSY79 has followed a nationally representative sample of young men and women since 1979. The respondents were between the ages of 14 to 22 years old at the time of the first survey in 1979. The surveys were conducted on a yearly basis between 1979 and 1994, and every two years thereafter. NLSY79 contains a rich set of variables about the respondents, like demographic and family background characteristics, household composition, educational status and attainment, intelligence scores, labor market activity, fertility, health, and marital history. At each interview date respondents provide information about any changes in marital status since the last interview and the dates of those changes. In addition, on a restricted basis, respondents' geographic area of residence is also made available. Location of residence allows me to connect respondents to the marriage markets they experienced.

Over the course of the survey administration, the poor whites and the military subsamples were dropped from the NLSY79. I also exclude these subsamples from the analysis. In addition, I exclude respondents that were incarcerated during the study period, or served in the military, as these groups are more likely to deviate from the general population in ways in which they participate in marriage markets. Further, I also exclude observations that were marked by the interviewers as deaf, blind, mentally or physically handicapped as these observations are also likely to deviate from the general population in their participation in marriage markets. I exclude individuals that were reported as deceased. Finally, I also drop respondents who reported a marital transition but the date of first marriage was

[22] In total 828 categories.

unavailable.

Geographic Location

NLSY79 collects data on the respondents' county, SMSA, and state of residence at the time of each interview. Information on state and county location at the time of birth and at age 14 is also collected. The geographic data at the time of interview is available to researchers on a restricted basis and allows one to connect marriage market features to respondents. However, unlike marital history, which contains exact dates of transition between marital states, information on transitions between geographic locations is not as robust. Geographic location is available for all respondents at the time of the first interview in 1979. Subsequently, at each interview time, geographic location is available only when the respondent was actually interviewed. If a respondent was not interviewed in a particular round of data collection, geographic location during the absence might not be available.

Given the information available, I develop a set of rules to derive geographic information when respondents were not interviewed. For years 1980, 1983, and 2000 through 2010 respondents were asked if they moved since the last interview date, and if so did they change city, county, or state, and the date of their moves between interviews.[23] For a given survey year (t) that includes the moving question, if a responded was interviewed, if the state of residence for that year can be determined, and if they have a non-interview spell prior to the survey year (t), I establish the state at the date of the last interview prior to the spell. Therefore, I have the state of residence prior to the non-interview spell and after the non-interview spell. Whenever a respondent did not move, or moved but did not report a change in city, county, or state, the state of residence at survey year (t) is used for all the missing years between (t) and last interview date. However, if they did move at some point during the non-interview spell and did report a change in city, county, or state, I proceed to establish the date of the first move and the date of the last move. This allows me to extend the geographic information at the beginning and end of the non interview spell. For the years missing prior to the date of first move the geographic information for the state as

[23]Unfortunately, information on which of the three, the city, county, or state changed was not asked.

of the date of the last interview is used. For years missing after the date of the last move, I use information from the survey time (t).

In years the question about moves was not asked, I do not know if respondents moved between interview dates. While this is less likely to be a problem up to 1994 when interviews were conducted annually, two year periods between interview dates after 1994 might be more problematic. For 1995 and 1997, I use the following rules: (i) if respondent was observed in the same state between two interview dates, use that state; (ii) if respondent was observed in two different states between interview dates, I assume respondent moved at the end of the first interview year (i.e., moved at the end of 1994 if lived in a different state in 1996). Because more detailed moving questions were consistently asked in 2000 and later, I use the following rule for odd years starting with 1999: (i) if a respondent lived in the same state in two interview years and did not report a move, use that state; (ii) if a respondent reported two different states at the interview dates, establish a date of first move between states and assume that after the first move the respondent lived in the new state.[24]

The final sample includes women who were 18 years old or less at the time of the first interview, and for whom state of residence can be determined at all times while single and participating in the survey. Table 1.1 summarizes the number and shares of women by race. The sample includes observations on 2,336 women with whites accounting for 54 percent of observations, blacks for 30 percent, and Hispanics for 16 percent.

1.5 The Model Generating the Data

Measurement errors and missing data cells will likely cause the sex ratios observed in the CPS $\hat{r}^d_{s,b,a,t}$, the only data source for many of the relevant years, to deviate significantly from the actual ratios,

$$r_{s,b,a,t} = \frac{\Sigma_{k=0}^{10} U^M_{s,b,a+k,t,d}}{\Sigma_{k=-2}^{8} U^F_{s,b,a+k,t,d}}. \tag{1.3}$$

[24]For example, if the respondent moved after June 1999 I use year 2000 information for 1999, and if the respondent moved prior to June 1999 I use 1998 information for 1999.

Therefore I propose using an estimate of the expectation of the observed sex ratio, $E[\hat{r}^d_{s,b,a,t}]$ as the more appropriate measure of market tightness.[25] The most convenient approach to modeling $E[\hat{r}^d_{s,b,a,t}]$ would be to assume that measurement error enters $\hat{r}^d_{s,b,a,t}$ in an additively separable way. Such an assumption is justifiable if the measurement errors for each data cell do not display heteroskedasticity with respect to age.[26] Unfortunately this assumption is rejected by the data.

Instead of working with sex ratios directly, I proceed by modeling the data generating process for each data cell used in the definition of $\hat{r}^d_{s,b,a,t}$ in equation (1). To do so, I must recover information about the number of unmarried males and females ($U^M_{s,b,a,t}$, $U^F_{s,b,a,t}$) in the presence of missing data and measurement error. I model the natural log of observed measures as

$$ln(\hat{U}^x_{s,b,a,t,d}) = \mu^x_{s,b,a,t} + e^x_{s,b,a,t,d} \tag{1.4}$$

where $x = \{M, F\}$. $\mu^x_{s,b,a,t}$ is the mean I aim to estimate, and $e^x_{s,b,a,t,d}$ is the deviation from the mean. I allow the deviation from the mean to have the form,

$$e^x_{s,b,a,t,d} = v^x_{s,b,a,t} + \varepsilon^x_{s,b,a,t,d}. \tag{1.5}$$

The term $\varepsilon^x_{s,b,a,t,d}$ represents the data-source-specific measurement error which is independent across all observations. $v^x_{i,a,t}$ is the portion of the deviation from the mean common across sources of data.

I model sex ratios for three racial groups: blacks, whites, and Hispanics. Given data source (d) for each state (s), age (a), and time (t), I collect the observations for both sexes and all three races into a 6×1 vector of observed data $ln(\vec{\hat{U}}_{s,a,t,d})$. Then I define

$$ln(\vec{\hat{U}}_{s,a,t,d}) = \vec{\mu}_{s,a,t} + \vec{e}_{s,a,t,d} \tag{1.6}$$

[25] Depending on modeling choices, it might also be possible to estimate $E[r_{s,b,a,t}]$, the expectation of the actual population sex ratios.

[26] See Appendix 1.B.

where $\vec{e}_{s,a,t,d}$ represents a 6×1 vector of deviations from the means $\vec{\mu}_{s,a,t}$. I model the joint distribution of deviations as multivariate normal

$$\begin{bmatrix} \vec{v}_{s,a,t} \\ \vec{\varepsilon}_{s,a,t,d} \end{bmatrix} \sim N(0, \Omega^{v,\varepsilon}_{s,a,d}), \tag{1.7}$$

where $\vec{v}_{s,a,t}$ is a 6×1 vector of non-measurement error deviation from the mean, and $\vec{\varepsilon}_{s,a,t,d}$ is a 6×1 vector of the data-source-specific measurement errors. Given the definition of $e^{x}_{s,b,a,t,d}$ in equation (1.5) and the distributional assumptions on the errors in 1.7 the deviations from the mean have the following distribution,

$$\vec{e}_{s,a,t,d} \sim N(0, \Omega^{e}_{s,a,d}) \tag{1.8}$$

where $\Omega^{e}_{s,a,d}$ is a 6×6 covariance matrix.[27] For each state (s), age (a), and data source (d), the deviations $\vec{e}_{s,a,t,d}$ are independent across time periods. I assume that the measurement errors are independent across all observations, and with the additive structure on $e^{x}_{s,b,a,t,d}$, I can express the covariance matrix as

$$\Omega^{e}_{s,a,d} = \Omega^{v}_{s,a} + \Omega^{\varepsilon}_{s,a,d}. \tag{1.9}$$

$\Omega^{\varepsilon}_{s,a,d}$ is a diagonal matrix of variances for measurement errors in state (s), age (a) and data source (d). $\Omega^{v}_{s,a}$ is common to all data sources and has a flexible structure that can allow for correlation in deviations at time (t) in state (s) and age (a) between sex and race categories. The flexibility of $\Omega^{v}_{s,a}$ allows for period-specific events, an idiosyncratic shock to marriage rates for instance, which impacts both the number of unmarried men and women across race categories.

For any given time period (t), at most two different data sources measuring the same category of unmarried individuals will be observed. The Census is observed every decade, the CPS and ACS are observed every year with the latter starting in 2001. For the majority of years, the CPS is the only source of data. Modeling choices on the deviations from the

[27] For derivation of the distribution of $e^{x}_{s,b,a,t,d}$ see Section 1.6.

mean $\vec{e}_{s,a,t,d}$ imply a structure on the covariances between data sources. In periods when the Census (d=1) and CPS (d=3) both measure the same observations, the errors have the following joint distribution

$$\begin{bmatrix} \vec{e}_{s,a,t,d=1} \\ \vec{e}_{s,a,t,d=3} \end{bmatrix} \sim N(0, \Omega_{s,a,d=1,3}) \tag{1.10}$$

where the covariance matrix is

$$\Omega_{s,a,d=1,3} = \begin{bmatrix} \Omega^v_{s,a} + \Omega^\varepsilon_{s,a,d=1} & \Omega^v_{s,a} \\ \Omega^v_{s,a} & \Omega^v_{s,a} + \Omega^\varepsilon_{s,a,d=3} \end{bmatrix}. \tag{1.11}$$

In periods when the ACS (d=2) and CPS (d=3) are jointly observed, we have

$$\begin{bmatrix} \vec{e}_{s,a,t,d=2} \\ \vec{e}_{s,a,t,d=3} \end{bmatrix} \sim N(0, \Omega_{s,a,d=1,2}) \tag{1.12}$$

with

$$\Omega_{s,a,d=2,3} = \begin{bmatrix} \Omega^v_{s,a} + \Omega^\varepsilon_{s,a,d=2} & \Omega^v_{s,a} \\ \Omega^v_{s,a} & \Omega^v_{s,a} + \Omega^\varepsilon_{s,a,d=3} \end{bmatrix}. \tag{1.13}$$

Finally, in periods when only the CPS data is observed we have

$$\vec{e}_{s,a,t,d=3} \sim N(0, \Omega_{s,a,d=3}) \tag{1.14}$$

and

$$\Omega_{s,a,d=3} = \left[\Omega^v_{s,a} + \Omega^\varepsilon_{s,a,d=3} \right]. \tag{1.15}$$

Within each state (s) and age (a) category, the number of observed measures and the structure of their covariance will depend on the time period (t). Missing observations add further variability to the covariance structure across time periods (t). The vector of observed

measures $ln(\vec{\hat{U}}_{s,a,t,d})$ will oftentimes have less than 6 observed elements. For instance, the CPS might not have a record for 20 year old black females in Indiana in 1991 but will have measures for whites, Hispanics, and black men. When the CPS does not measure the 20 year old black females in Indiana in 1991, the observed data has a 5×5 covariance matrix that is a subset of $\Omega_{s,a,d=3}$.

1.6 Likelihood Equation and Estimation

Given the distributional assumptions on the deviations from the mean $e^x_{s,b,a,t,d}$, I estimate the model using the maximum likelihood framework. To construct the appropriate likelihood I must consider the events that render any data cell of interest observed or unobserved. First, a data cell (s,b,a,t) can be observed by data source (d) when there is at least one individual in the sample with those attributes, $U^x_{s,b,a,t} \geq 1$. This implies,

$$ln(\hat{U}^x_{s,b,a,t,d}) \geq 0 \text{ only if } ln(U^x_{s,b,a,t}) \geq 0. \tag{1.16}$$

The probability of a vector $ln(\vec{\hat{U}}_{s,a,t,d})$ with all 6 elements observed is expressed as

$$\begin{aligned} P(ln(\vec{\hat{U}}_{s,a,t,d}) - \vec{\mu}_{s,a,t} = \vec{e}_{s,a,t,d},\ \vec{\mu}_{s,a,t} + \vec{v}_{s,a,t} > 0) = \\ P(ln(\vec{\hat{U}}_{s,a,t,d}) - \vec{\mu}_{s,a,t} = \vec{e}_{s,a,t,d} \mid \vec{v}_{s,a,t} > -\vec{\mu}_{s,a,t}) P(\vec{v}_{s,a,t} > -\vec{\mu}_{s,a,t}). \end{aligned} \tag{1.17}$$

The conditioning statement on the observed data vector excludes the possibility that a data cell is sampled when in fact no individuals with attributes (s,b,a,t) exist.[28]

Second, not all data cells will be observed. One way to explicitly model the missing

[28] In other words, for a given $\mu^x_{s,b,a,t}$, the restriction guarantees that the pair $(v^x_{s,b,a,t}, \varepsilon^x_{s,b,a,t,d})$ does not fall in a region where $\mu^x_{s,b,a,t} + v^x_{s,b,a,t} \leq 0$, but due to large measurement error $\mu^x_{s,b,a,t} + v^x_{s,b,a,t} + \varepsilon^x_{s,b,a,t,d} > 0$.

data mechanism is to think of missing cells as censored. An observations is missing if,

$$ln(\hat{U}^x_{s,b,a,t,d}) < 0 \tag{1.18}$$
$$\mu^x_{s,b,a,t} + e^x_{s,b,a,t,d} < 0.$$

Let (i) be an index of all the age and year combinations observed in the data.[29] The number of elements in the vector $ln(\vec{\hat{U}}_i)$, will depend on the year. If CPS is the only survey fielded, $ln(\vec{\hat{U}}_i)$ is a 6×1 vector, otherwise $ln(\vec{\hat{U}}_i)$ has 12 elements. Let $\vec{D}_i$ represent a $G \times 1$ vector[30] of indicators for missing observations where each element (g) is defined

$$D_{i,g} = \begin{cases} 1 \text{ when } \mu^x_{s,b,a,t,d} + e^x_{s,b,a,t,d} \geq 0 \\ 0 \text{ when } \mu^x_{s,b,a,t,d} + e^x_{s,b,a,t,d} < 0 \end{cases}. \tag{1.19}$$

The pattern of missing data also dictates the restrictions on the observed data. Let $O(\vec{D}_i)$ represent a set of restrictions on observed data given a missing data pattern, $\vec{D}_i$. For each race and sex (b,x) combination, if $ln(\hat{U}^x_{s,b,a,t,d})$ is observed by any data source the set $O(\vec{D}_i)$ will include the restriction $\mu^x_{s,b,a,t} + v^x_{s,b,a,t} \geq 0$. For example, if the Census and CPS are fielded in the same year and number of 20 year old black females in Indiana is recorded by at least one data source, we know that the restriction on sampling only observed cells is binding for that demographic group. The number of restrictions in $O(\vec{D}_i)$ can vary from six, where both males and females of all three races are observed by at least one data source, to just one, where only one sex and race combination is observed.[31] Given the index of observed data restrictions, $o = \{1, 2, \ldots, 32\}$, the elements of the set $O(\vec{D}_i)$ can be expressed as a vector inequality

$$\vec{v}_{i,o} \geq -\vec{\mu}_{i,o} \tag{1.20}$$

[29] (i) represents an index over all the elements in a set that is a cross product of age and time $i \in I = \{A \times T\}$ for state (s) where n is the number of those elements. I use data from 1979 to 2010 for ages 16 to 50, so $i = \{1, \ldots, 1120\}$.

[30] G is either 6 or 12.

[31] Depending on the sampling units in a given survey and the geographic definition of the data cells, theoretically it might be possible that $O(\vec{D}_i)$ contains no elements because no data cell was sampled. I use the state as the geographic unit and each data source samples the population in each state.

where the vector dimensions depend on the number of restriction in the set $O(\vec{D}_i)$.

For each observation (i), the observed and missing data cells imply the following joint probability

$$P(ln(\vec{\hat{U}}_i), \vec{D}_i, O(\vec{D}_i)). \tag{1.21}$$

Consider the case where (i) corresponds to a year where both CPS and Census are observed, and all six race and sex categories are observed by each data source. The joint probability for (i) is

$$P(ln(\vec{\hat{U}}_i), \vec{D}_i, O(\vec{D}_i)) = P(ln(\vec{\hat{U}}_i) - \vec{\mu}_i = \vec{e}_i, \vec{v}_{i,o} \geq -\vec{\mu}_{i,o}), \tag{1.22}$$

and depends on the joint distribution of the data cell errors $\vec{e}_i$ and the non-measurement error deviations from the mean, $\vec{v}_{i,o}$. The probability will involve a statement on 18 elements, 12 $e^x_{s,b,a,t,d}$ elements, and 6 $v^x_{s,b,a,t}$ elements. To model the probability I derive the joint distribution for $(\vec{v}_{i,o}, \vec{e}_i)$ and I rely on the assumption that

$$\begin{bmatrix} \vec{v}_{i,o} \\ \vec{\varepsilon}_i \end{bmatrix} \sim N(0, \Omega^{v,\varepsilon}). \tag{1.23}$$

In periods where two sources of data are observed, Census (d=1) and CPS (d=3) for instance,

$$\Omega^{v,\varepsilon}_i = \begin{bmatrix} \Omega^v_{s,a} & 0 & 0 \\ 0 & \Omega^\varepsilon_{s,a,d=1} & 0 \\ 0 & 0 & \Omega^\varepsilon_{s,a,d=3} \end{bmatrix}. \tag{1.24}$$

In periods where only one sources of data is observed,

$$\Omega^{v,\varepsilon}_i = \begin{bmatrix} \Omega^v_{s,a} & 0 \\ 0 & \Omega^\varepsilon_{s,a,d} \end{bmatrix}. \tag{1.25}$$

The off-diagonal matrices have all elements equal to zero because of the independence between data source specific measurement errors, $\varepsilon^x_{s,b,a,t,d}$, and non-measurement error de-

viations from the mean, $v^x_{s,b,a,t}$. By definition each deviation from the mean is

$$e^x_{s,b,a,t,d} = v^x_{s,b,a,t} + \varepsilon^x_{s,b,a,t,d}, \tag{1.26}$$

therefore, the mapping from $(\vec{v}_{i,o}, \vec{\varepsilon}_i)$ to $(\vec{v}_{i,o}, \vec{e}_i)$ is just a linear transformation of a multivariate normal, and can be expressed as

$$\begin{bmatrix} \vec{v}_{i,o} \\ \vec{e}_i \end{bmatrix} = B \begin{bmatrix} \vec{v}_{i,o} \\ \vec{\varepsilon}_i \end{bmatrix}. \tag{1.27}$$

A linear transformation of a multivariate normal vector has a multivariate normal distribution, with

$$\begin{bmatrix} \vec{v}_{i,o} \\ \vec{e}_i \end{bmatrix} \sim N(0, \Omega^{v,e}) \tag{1.28}$$

where $\Omega^{v,e} = B\Omega^{v,\varepsilon}B^T$.[32] To derive the structure of the covariance matrix $\Omega^{v,e}$, let I be a 6×6 identity matrix, then the linear transformation can be expressed as

$$B = \begin{bmatrix} I & 0 & 0 \\ I & I & 0 \\ I & 0 & I \end{bmatrix}. \tag{1.29}$$

Given the independence assumption on the non-measurement error deviations from the mean, I obtain

$$\Omega^{v,e} = B\Omega^{v,\varepsilon}B^T = \begin{bmatrix} I & 0 & 0 \\ I & I & 0 \\ I & 0 & I \end{bmatrix} \begin{bmatrix} \Omega^v_{s,a} & 0 & 0 \\ 0 & \Omega^\varepsilon_{s,a,d=1} & 0 \\ 0 & 0 & \Omega^\varepsilon_{s,a,d=3} \end{bmatrix} \begin{bmatrix} I & I & I \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix}, \tag{1.30}$$

[32] Greene(2003).

and block multiplication yields

$$\Omega^{v,e} = B\Omega^{v,\varepsilon}B^T = \begin{bmatrix} \Omega^v & \Omega^v & \Omega^v \\ \Omega^v & \Omega^v + \Omega^{\varepsilon}_{d=1} & \Omega^v \\ \Omega^v & \Omega^v & \Omega^v + \Omega^{\varepsilon}_{d=3} \end{bmatrix}. \tag{1.31}$$

Any observed and missing data combination will have the distribution in (1.28), or a marginal distribution derived from (1.28). Therefore, for observation (i) where Census and CPS are observed and no data cells are missing,

$$\begin{aligned} P(ln(\vec{\hat{U}}_i), D_i, O(D_i)) &= P(ln(\vec{\hat{U}}_i) - \vec{\mu}_i = \vec{e}_i, \vec{v}_{i,o} \geq -\vec{\mu}_{i,o}) \\ &= \int_{-\vec{\mu}_{i,o}}^{\infty} f(\vec{v}_o, \vec{e}_i) d\vec{v}_{i,o} = \int_{-\vec{\mu}_{i,o}}^{\infty} f(\vec{v}_o|\vec{e}_i) f(\vec{e}_i) d\vec{v}_{i,o} \end{aligned} \tag{1.32}$$

where the integration sign is a shorthand for multiple integrals.

Next, I derive the probability for (i) given that at least one observation is missing. For any observation (i) and pattern of missing data $\vec{D}_i$ it is helpful to reorder the elements $ln(\vec{\hat{U}}_i)$ and put the missing observations first. Define the mapping of the vectors $ln(\vec{\hat{U}}_i)$, $\vec{D}_i$

$$C : (ln(\vec{\hat{U}}_i), \vec{D}_i) \mapsto ln(\vec{\hat{U}}_{i,c}) \tag{1.33}$$

where $ln(\vec{\hat{U}}_{i,c})$ has been reordered such that the first (l) data cells are missing, and (c) is an index of all possible missing data patterns.[33] If (i) represents the vector of 20 year olds in Indiana in 1991 and the CPS does not sample 20 year old black females in Indiana in 1991, then $ln(\vec{\hat{U}}_{i,c})$ has the first element missing and the remaining five elements include the CPS observations on 20 year old whites, Hispanics and black men in Indiana in 1991. Given missing data pattern (c), I can write

$$ln(\vec{\hat{U}}_{i,c}) = \vec{\mu}_{i,c} + \vec{e}_{i,c} \tag{1.34}$$

[33]Possible combinations of missing data cells will depend on the year of observation. In years where only the CPS is observed there are $2^6 = 64$ combinations of missing data cells. In years where two data sources are observed, CPS and Census or CPS and ACS, there are $2^{12} = 4096$ possible combinations of missing data cells.

where

$$ln(\vec{\hat{U}}_{i,c}) = \begin{bmatrix} ln(\vec{\hat{U}}_{i,c,1}) \\ ln(\vec{\hat{U}}_{i,c,2}) \end{bmatrix}, \vec{\mu}_{i,c} = \begin{bmatrix} \vec{\mu}_{i,c,1} \\ \vec{\mu}_{i,c,2} \end{bmatrix} \text{ , and } \vec{e}_{i,c} = \begin{bmatrix} \vec{e}_{i,c,1} \\ \vec{e}_{i,c,2} \end{bmatrix}. \tag{1.35}$$

The elements $ln(\vec{\hat{U}}_{i,c,1})$, $\vec{\mu}_{i,c,1}$ and $\vec{e}_{i,c,1}$ correspond to missing data cells.

The probability of data vector (i) with missing data pattern (c) and observed cell restrictions (o) involves the p.d.f. for the vector $(\vec{v}_{i,o}, \vec{e}_{i,c,1}, \vec{e}_{i,c,2})^T$, which has a multivariate normal distribution with a covariance matrix that is a subset of $\Omega^{v,e}$. Therefore, for observation (i)

$$\begin{aligned} P(ln(\vec{\hat{U}}_i), \vec{D}_i, O(\vec{D}_i)) &= P(ln(\vec{\hat{U}}_{i,c,2}) - \vec{\mu}_{i,c,2} = \vec{e}_{i,c,2},\ \vec{e}_{i,c,1} < -\vec{\mu}_{i,c,1},\ \vec{v}_{i,o} \geq -\vec{\mu}_{i,o}) \\ &= \int_{-\vec{\mu}_{i,o}}^{\infty} \int_{-\infty}^{-\vec{\mu}_{i,c,1}} f(\vec{v}_{i,o}, \vec{e}_{i,c,1}, \vec{e}_{i,c,2}) d\vec{e}_{i,c,1} d\vec{v}_{i,o} \\ &= \int_{-\vec{\mu}_{i,o}}^{\infty} \int_{-\infty}^{-\vec{\mu}_{i,c,1}} f(\vec{v}_{i,o}, \vec{e}_{i,c,1} | \vec{e}_{i,c,2}) f(\vec{e}_{i,c,2}) d\vec{e}_{i,c,1} d\vec{v}_{i,o} \end{aligned} \tag{1.36}$$

where the integration signs are a shorthand for multiple integrals, depending on the dimensions $\vec{\mu}_{i,o}$ and $\vec{\mu}_{i,c,1}$.

Both $f(\vec{v}_{i,o}|\vec{e}_i)$ and $f(\vec{v}_{i,o}, \vec{e}_{i,c,1}|\vec{e}_{i,c,2})$ represent the density conditional on observed values, and will have a multivariate normal p.d.f. with a mean that depends on the observed data. Given $\vec{D}_i$ and $O(\vec{D}_i)$ where at least one data cell is missing, the covariance matrix for the vector $(\vec{v}_{i,o}, \vec{e}_{i,c,1}, \vec{e}_{i,c,2})^T$ can be partitioned

$$\Omega^{v,e}_{c,o} = \begin{bmatrix} \Omega^{v,e}_{c,o,11} & \Omega^{v,e}_{c,o,12} \\ \Omega^{v,e}_{c,o,21} & \Omega^{v,e}_{c,o,22} \end{bmatrix} \tag{1.37}$$

where $\Omega^{v,e}_{c,o,11}$ is the covariance matrix for $(\vec{v}_{i,o}, \vec{e}_{i,c,1})^T$, and $\Omega^{v,e}_{c,o,22}$ is the covariance matrix for the observed data cells. The distribution conditional on the observed data is

$$\begin{bmatrix} \vec{v}_{i,o} \\ \vec{e}_{i,c,1} \end{bmatrix} \Big| \quad \vec{e}_{i,c,2} \sim N(\vec{\mu}_{i,o,c,1.2}, \Omega^{v,e}_{i,o,c,1.2}) \tag{1.38}$$

where

$$\vec{\mu}_{i,o,c,1.2} = \Omega^{v,e}_{c,o,12}(\Omega^{v,e}_{c,o,22})^{-1}\vec{e}_{i,c,2}, \tag{1.39}$$

$$\Omega^{v,e}_{i,o,c,1.2} = \Omega^{v,e}_{c,o,11} - \Omega^{v,e}_{c,o,12}(\Omega^{v,e}_{c,o,22})^{-1}\Omega^{v,e}_{c,o,21}, \tag{1.40}$$

and

$$\vec{e}_{i,c,2} = ln(\vec{\hat{U}}_{i,c,2}) - \vec{\mu}_{i,c,2}. \tag{1.41}$$

In the case where none of the data cells is missing, the likelihood contribution is,

$$\mathcal{L}_{i,c,o} = f(\vec{e}_i)\int_{-\vec{\mu}_o}^{\infty} f(\vec{v}_{i,o}|\vec{e}_i)d\vec{v}_{i,o} \tag{1.42}$$

with a log likelihood

$$L_{i,c,o} = ln(\mathcal{L}_{i,c,o}) = ln(f(\vec{e}_i)) + ln(\int_{-\vec{\mu}_o}^{\infty} f(\vec{v}_{i,o}|\vec{e}_i)d\vec{v}_{i,o}). \tag{1.43}$$

The likelihood contribution for an observation (i) with a missing data pattern (c) and observed restriction (o), can be denoted as

$$\mathcal{L}_{i,c,o} = f(\vec{e}_{i,c,2})\int_{-\vec{\mu}_o}^{\infty}\int_{-\infty}^{-\vec{\mu}_{i,c,1}} f(\vec{e}_{i,c,1},\vec{v}_{i,o}|\vec{e}_{i,c,2})d\vec{e}_{i,c,1}d\vec{v}_{i,o}. \tag{1.44}$$

with a log likelihood

$$L_{i,c,o} = ln(\mathcal{L}_{i,c,o}) = ln(f(\vec{e}_{i,c,2})) + ln\left(\int_{-\vec{\mu}_o}^{\infty}\int_{-\infty}^{-\vec{\mu}_{i,c,1}} f(\vec{e}_{i,c,1},\vec{v}_{i,o}|\vec{e}_{i,c,2})d\vec{e}_{i,c,1}d\vec{v}_{i,o}\right). \tag{1.45}$$

The log likelihood function for the observed data values and the missing and observed data patterns is expressed as,

$$L = \Sigma_i^n L_{i,c,o} = \Sigma_i^n ln(f(\vec{e}_{i,c,2})) + \Sigma_i^n ln\left(\int_{-\vec{\mu}_o}^{\infty}\int_{-\infty}^{-\vec{\mu}_{i,c,1}} f(\vec{e}_{i,c,1},\vec{v}_{i,o}|\vec{e}_{i,c,2})d\vec{e}_{i,c,1}d\vec{v}_{i,o}\right). \tag{1.46}$$

1.6.1 Simplified Likelihood

Estimation of equation (1.46) requires integration over multiple dimensions which can be computationally burdensome. A substantial simplification of the likelihood equation is possible whenever the underlying data generating process satisfies two assumptions about the observed and missing data cells. First, one can take advantage of working with state level data and assume that non-measurement error deviations from the mean $v^x_{s,b,a,t}$ are very small relative to the observed means $\mu^x_{s,b,a,t}$. When all data cells are observed the probability of the observed data values and the observability restriction can be expressed

$$P(ln(\vec{\hat{U}}^x_{s,b,a,t,d}) - \vec{\mu}_{s,a,t} = \vec{e}_{s,a,t,d},\ \vec{v}_{s,a,t} > -\vec{\mu}_{s,a,t}) = \tag{1.47}$$
$$P(ln(\vec{\hat{U}}^x_{s,b,a,t,d}) - \vec{\mu}_{s,a,t} = \vec{e}_{s,a,t,d} | \vec{v}_{s,a,t} > -\vec{\mu}_{s,a,t}) P(\vec{v}_{s,a,t} > -\vec{\mu}_{s,a,t}).$$

Because

$$P(\vec{v}_{s,a,t} > -\vec{\mu}_{s,a,t}) \approx 1 \tag{1.48}$$

and

$$P(ln(\vec{\hat{U}}^x_{s,b,a,t,d}) - \vec{\mu}_{s,a,t} = \vec{e}_{s,a,t,d} | \vec{v}_{s,a,t} > -\vec{\mu}_{s,a,t}) \approx P(ln(\vec{\hat{U}}^x_{s,b,a,t,d}) - \vec{\mu}_{s,a,t} = \vec{e}_{s,a,t,d}), \tag{1.49}$$

when the elements of $\vec{\mu}_{s,a,t}$ are large relative to $\vec{v}_{s,a,t}$, the censoring of the joint distribution of $\vec{v}_{s,a,t}$ is negligible and one can use the following approximation to probability of the observed data,

$$P(ln(\vec{\hat{U}}_{s,a,t,d}) - \vec{\mu}_{s,a,t} = \vec{e}_{s,a,t,d},\ \vec{\mu}_{s,a,t} + \vec{v}_{s,a,t} > 0) \approx \tag{1.50}$$
$$P(ln(\vec{\hat{U}}_{s,a,t,d}) - \vec{\mu}_{s,a,t} = \vec{e}_{s,a,t,d}).$$

The same approximation applies when some data cells in vector (i) are missing. The approximation will induce bias in estimation, although when the measured population is large and stable over (t), the size of the bias should be small.

Second, modeling the missing data mechanism can omitted. Little and Rubin (2002)

show that, if the data is Missing at Random (MAR), researchers can forgo explicitly modeling the process generating the missing data and still obtain consistent estimates of their model. MAR requires that, conditional on the observed data, the mechanism generating the missing observations does not depend on the values of those missing observations.[34] In cases where all the cells in $ln(\vec{\hat{U}}_{s,a,t,d})$ are unobserved because the Census or the ACS were not fielded that year, MAR is easily satisfied. However, when observations are missing because they were not sampled, it is possible that MAR fails if the unobservables are large enough to influence the probability of being sampled.

Whenever MAR holds and the approximation to the probability of the observed data is appropriate, the log likelihood can be represented as

$$L_2 = \Sigma_i^n ln(f(\vec{e}_{i,c,2})) = Constant - \frac{1}{2}\Sigma_i^n ln\,|\Omega_i| - \frac{1}{2}\Sigma_i^n(\vec{e_i})'\Omega_i^{-1}(\vec{e_i}) \tag{1.51}$$

where $\vec{e_i}$ is a vector of errors for males and females of all three race categories and data sources available at time (t) and Ω_i is the covariance of these elements. The size of $\vec{e_i}$ can vary from twelve elements at most to just one or two elements. Every element of the vector is defined as $e^x_{s,b,a,t,d} = ln(\hat{U}^x_{s,b,a,t,d}) - \mu^x_{s,b,a,t}$. Each Ω_i is some subset of $\Omega_{s,a,d}$ as defined in equations (1.11), (1.13), and (1.15). The size and structure of Ω_i depends on the data sources available at time (t) and race and sex combination these data sources actually measured in their sample. The approximate likelihood does not require integration and is computationally much faster to evaluate.

1.6.2 Parametrization

The estimation framework relies on the independence of e_i between (i), in other words deviations from the mean are allowed to be correlated between sex, race, and data source categories but are independent across age and time within a given state. However, it seems natural for any given $U^x_{s,b,a,t}$ to be correlated with $U^x_{s,b,a+1,t+1}$. For instance, many of the 19 year old unmarried white women in Indiana in 1990 become the 20 year old unmarried

[34] For a further discussion see Appendix 1.C.

white women in Indiana in 1991. As any cohort ages its numbers diminish in time through marriage and mortality, and shrink or grow depending on the size of net migration between states. It is also very likely that $U^x_{s,b,a,t}$ will be correlated with $U^x_{s,b,a,t+1}$. Population growth and slow changes in marriage patterns also imply that the number of 19 year old unmarried women in 1990 will be correlated with 19 year old unmarried women in 1991. I adopt a flexible functional form for $\mu^x_{s,b,a,t}$ to capture correlations within groups across time and age and allow for the unobservables to be uncorrelated across age and time. Let,

$$\begin{aligned}\mu^x_{s,b,a,t} =& \beta_0^{x,s,b} + \beta_1^{x,s,b} * age + \beta_2^{x,s,b} * age^2 + \beta_3^{x,s,b} * age^3 \\ &+ \beta_4^{x,s,b} * year + \beta_5^{x,s,b} * year^2 + \beta_6^{x,s,b} * (year * age).\end{aligned} \tag{1.52}$$

I estimate 42 mean parameters per state. Let β_s represent the set of mean parameters for state (s).

For each data source (d), $\Omega^{\varepsilon}_{s,a,d}$ is a diagonal 6×6 matrix. I allow for heteroskedasticity in age by modeling two age groups with the age of 25 as the dividing line. In total, for each (s) there are 36 variance parameters for the measurement errors $\Omega^{\varepsilon}_{s,a,d}$. $\Omega^{v}_{s,a}$ is a 6×6 matrix and given that I allow for two age groups, will at most involve 42 parameters for each state (s). The model allows for a flexible structure on $\Omega^{v}_{s,a}$. A diagonal $\Omega^{v}_{s,a}$ assumes independence for the deviations while a fully parametrized $\Omega^{v}_{s,a}$ allows for covariance between all the deviations across sex and race categories. Let θ_s represent the set of covariance parameters to be estimated for state (s).

1.6.3 Identification

Independence of measurement errors across time (t), states (s), race (b), age (a), and sex (x) allows me to separately identify the parameters of $\Omega^{v}_{s,a}$ and $\Omega^{\varepsilon}_{s,a,d}$. Whenever the same data cell is measured by two data sources covariance between the residuals will identify the variance of the deviation $v_{s,b,a}$. With diagonal elements of $\Omega^{v}_{s,a}$ identified, the remainder of elements will be identified by the covariance of residuals across categories within and across datasets. Once $\Omega^{v}_{s,a}$ is identified, variances of the measurement errors can be separately

identified by using the variance of the residuals.

1.6.4 Estimation Details

For each state (s) parameters that maximize the likelihood function are found using a quasi-newton algorithm.[35] The probability of missing and observed data cells, conditional on observed data cell values, requires multidimensional integration and is approximated using the GHK algorithm. For each (i)

$$ln(P(\vec{D}_i, O(\vec{D}_i)|ln(\hat{\vec{U}}_i)) = ln\left(\int_{-\vec{\mu}_o}^{\infty}\int_{-\infty}^{-\vec{\mu}_{i,c,1}} f(\vec{e}_{i,c,1}, \vec{v}_{i,o}|\vec{e}_{i,c,2})d\vec{e}_{i,c,1}d\vec{v}_{i,o}\right) \tag{1.53}$$
$$\approx ln\left(\frac{1}{Q}\Sigma_{q=1}^{Q}P_{i,q}\right)$$

where q represents the simulated draws of random variables and $P_{i,q}$ is the GHK approximation to the integral of a multivariate normal distribution. Antithetic acceleration is used to reduce the additional variance in estimated parameters due to simulation.[36]

Evaluation of the likelihood equation involves calculating the natural log of the determinant of Ω_i for each (i). This requires that for all (i), Ω_i be positive definite at each parameter step (w).[37] To ensure a positive definite Ω_i, I use Cholesky decomposition and find the parameters of the triangular matrix.[38] I start the iterative algorithm with parameters $\beta^{w=0}$, $\theta^{w=0}$ such that all Ω_i are positive definite.

[35]D.C. is also estimated separately. I use the fmincon function with the sqp algorithm option in Matlab to find the minimum of the negative likelihood equation. The Hessian of the likelihood at parameter values is approximated using the BFGS method.

[36]See Stern (1997). I set Q=5, which results in 10 random variable draws given the use of antithetic acceleration.

[37]Statistical theory also requires that $\Omega^{v}_{s,a}$ for both age groups, all six $\Omega^{\varepsilon}_{s,a,d}$ and all their submatrices be positive definite.

[38]The diagonal elements have lower bounds to ensure that all of the variance parameters are bounded away from zero, and upper bounds to ensure a smooth stepping of the maximization algorithm.

1.7 Expectation of the Sex Ratios

Combining the definitions in equations (1) and (4), the observed sex ratios can be written as

$$\hat{r}^d_{s,b,a,t} = \frac{\Sigma^{10}_{k=0}\hat{U}^M_{s,b,a+k,t,d}}{\Sigma^{8}_{k=-2}\hat{U}^F_{s,b,a+k,t,d}} = \frac{\Sigma^{10}_{k=0}e^{\mu^M_{s,b,a+k,t}+e^M_{s,b,a+k,t,d}}}{\Sigma^{8}_{k=-2}e^{\mu^F_{s,b,a+k,t}+e^F_{s,b,a+k,t,d}}}. \tag{1.54}$$

Generating an expectation of the sex ratio involves integrating out 20 different error terms $e^x_{s,b,a,t,d}$ for each (s,b,a,t). Instead I simulate the expectation using the parameter estimates for $\hat{\mu}^x_{s,b,a,t}$ and $\hat{\Omega}_{s,a,d}$. I draw a vector $\vec{e}_{s,a,t}$ from a multivariate normal distribution $N(0, \hat{\Omega}_{s,a,d})$ (q) times. The simulated expectation of each sex ratio for data source (d) is

$$\tilde{E}[\hat{r}^d_{s,b,a,t}] = \frac{1}{Q}\Sigma^{Q}_{q=1}\frac{\Sigma^{10}_{k=0}exp(\hat{\mu^M}_{s,b,a+k,t} + e^M_{s,b,a+k,t,d,q})}{\Sigma^{8}_{k=-2}exp(\hat{\mu^F}_{s,b,a+k,t} + e^F_{s,b,a+k,t,d,q})}, \tag{1.55}$$

where I set $Q = 100$.

1.8 Results

I estimate two versions of the likelihood equation, the full model in equation (1.46) which explicitly deals with missing and observed data cells, and the simplified model in equation (1.51) which assumes data is missing at random (MAR) and uses approximations for the probability of the observed data. To illustrate some general results, I discuss a subset of parameter estimates for two states, New York and Virginia, and start with the full model results. Tables 1.2 and 1.3 present the parameter estimates for the state of New York for two parameterizations of the model, and Tables 1.4 and 1.5 present estimates Virginia. Tables 1.2 and 1.4 present results from a parametrization that allows for covariance in non-measurement errors between men and women within the same category. Tables 1.3 and 1.5 present results from a parametrization that does not allow for any covariances. Because I allow for heteroskedasticity in age, each table has two sets of variance parameters: for populations less than 25 years old and older than 25. Measurement errors appear to

account for much of the deviation from the mean. For the majority of demographic groups in Virginia and New York, the variance for non-measurement deviations from the mean, σ_v^2, is considerably smaller than variances associated with the measurement errors, σ_ε^2. The variance parameters differ between the states, race categories, and age groups suggesting the importance of allowing for heteroskedasticity along those dimensions. The model parameters are also informative about some features of the observed data. for instance, there is more volatility in population counts when cohorts age and exit the unmarried state. Variance estimates for younger populations are smaller than older ones, therefore sex ratios based on the data are likely to deviate from the true ratios more for older individuals. Moreover, as expected, the variance of the CPS measurement errors is frequently larger than Census and ACS variances, suggesting the small sample size of the CPS makes it prone to larger deviations from actual population counts. Sex ratios based on the CPS data alone are likely to contain more measurement error relative to the other two data sources.

Relative to the estimates of measurement error variance parameters for the Census and ACS data, the CPS estimates display large differences between demographic groups, and a propensity for large parameter values. The variance parameters for the CPS in the state of Virginia in Table 1.4 are the most dramatic example of this. The smallest variance parameter is 0.11 while the largest one is nearly 200. The increase in the size of measurement errors corresponds to a decrease in the proportion of the data cells observed by that data source. The left panel of Table 1.6 shows the number of observations for each data source by demographic category in the state of Virginia while the right one shows the fraction of cells observed relative to the number that would be observed if that cell was sampled every year the survey was fielded. The measurement error variance estimates for Virginia are largest for the CPS blacks and Hispanics, two categories that have the lowest fraction of observed cells. A similar pattern is observed for the state of New York in Table 1.7, where CPS observations for black and Hispanic men are low. For Virginia and New York, the propensity for unobserved data cells is mostly attributed to the measurement error part of the deviation from the mean. Tables 1.8 and 1.9 present variance estimates for the simplified model where

I do not account for the mechanisms that render data cells unobserved.[39] As illustrated in those tables, the CPS measurement error variance estimates also exceed Census and ACS variances, and heteroskedasticity in age and race persists. However, because the simplified model does not explicitly address the occurrence of unobserved data cells, often the CPS variance parameters are far smaller compared to the full model. Covariance parameters from the two models are closer to each other for New York than for Virginia. The results from the simplified model suggest that in states where the fraction of observed data cells is high, the simplified likelihood might be a good approximation to the full likelihood. Figure 1.1 illustrates the proportion of cells observed by data source for each of the 20 most populous states. Virginia is far more likely to have missing CPS observations than New York, therefore states with missing data patterns similar to New York are better candidates for the use of the simplified model.

1.8.1 Waiting Time for First Marriage

Next, I discuss the results when sex ratios are constructed from parameter estimates and used as explanatory variables for waiting time until first marriage. I use the National Longitudinal Survey of Youth 1979 (NLSY79) to estimate the impact of sex ratios on the transition to first marriage. In particular, I compare the magnitude and significance of coefficient estimates when model generated expectations, $\tilde{E}[\hat{r}^d_{s,b,a,t}]$, are used as explanatory variables, to results when ratios constructed from the data alone, $\hat{r}^d_{s,b,a,t}$, are used. The majority of the NLSY79 respondents entered their first marriage in the 1980's and 1990's, a time frame for which the CPS is the only source of yearly data on marriage market features. There are considerable differences in marital behavior and the distribution of sex ratios between races. As illustrated in Table 1.10, black women are the least likely to marry, followed by Hispanics and whites. Only 6 percent of the white women in the NLSY79 sample did not marry, while nearly 27 percent of black women did not report a transition to first marriage. Age at marriage also differs by race. Nearly 67 percent of whites and Hispanics entered first marriage before the age of 25, while only 43 percent of black women

[39] I assume MAR holds, and use approximations to the probability of the observed data.

entered marriage by that age. The distribution of sex ratios varies considerably by race as well. I use the definition of sex ratio in equation (1.1) to capture the relative availability of potential partners for each woman in the sample. Figure 1.2 illustrates the estimated density for sex ratios constructed using the Census data. Blacks have the lowest sex ratios, followed by Hispanics and whites. Hispanics have the most variance in the sex ratios, while sex ratios for whites appear to have the least dispersion. As illustrated in Table 1.11, ratios also tend to decline when cohorts grow older. Different incarceration rates and mortality rates between men and women contribute to this imbalance. Moreover, as cohorts age and enter marriage the overall number of unmarried men and women declines, exacerbating any initial differences between the sexes.[40]

I estimate a parametric survival model with log-normal errors to determine the impact of sex ratios on waiting time to first marriage. The model allows for sex ratios that vary over time and for the presence of person-specific unobserved heterogeneity in preferences for marriage. Each woman enters the analysis at the age 18. I compare results for six different sources of sex ratios. The first two are constructed using observed data only. $rCNACS$ combines the Census and the ACS observations. Whenever available I use the ACS or Census numbers, and in years where neither survey collected data I use the Census data that is closest in absolute distance to the year considered.[41] $rCPS$ relies solely on the CPS data which is available for every year. The next four measures of partner availability are based on simulated expectations from models estimated under different assumptions. $Er1$ and $Er2$ are estimates from the simplified likelihood function, where I assume MAR holds for unobserved data cells and I can use the approximation to the probability of the observed data cells. $Er3$ and $Er4$ are estimated from the full likelihood function, where I model the process generating missing and observed data cells explicitly. For each likelihood specification I also estimate models with different assumptions on the covariances between unobservables $\Omega^v_{s,a}$. $Er1$ and $Er3$ assume that all shocks are independent.[42] $Er2$ and $Er4$

[40] Consider a market where there are 900 unmarried men and 1000 unmarried women. The initial sex ratio is $\frac{900}{1000} = .9$. Over time let 600 marriages take place. The new number of unmarried men is 300 and unmarried women is 400 giving a sex ratio of $\frac{300}{400} = .75$.

[41] For instance, for 1994 I use the 1990 Census sex ratios and for 1996 I use the 2000 Census sex ratios.

[42] In other words, $\Omega^v_{s,a}$ has nonzero terms only on the diagonal.

allow for within race covariance in unobservables between men and women.

Table 1.12 presents the estimated coefficients for different sources of sex ratios on entry into first marriage. Each column represents a separate survival analysis where explanatory variables include race indicators, log of AFQT score, and the sex ratios indicated in each row. The coefficients for all the definitions have the predicted sign. Higher sex ratios, meaning more unmarried men per unmarried woman, result in shorter waiting times for transition into marriage. The coefficient on Census ratios, $rCNACS$, is significantly higher than the coefficient on the CPS ratios, $rCPS$. Brien (1997) finds evidence of bias towards zero for the coefficient on sex ratios when the ratios are measured with error. The small size of the estimated coefficient when CPS ratios are used suggests the presence of considerable measurement error in the CPS sex ratios. The Census measures, even with the arbitrary assumption on the evolution of sex ratios through time, fare much better in predicting entry into first marriage for NLSY79 women. However, using simulated expectations generates larger and more significant coefficients relative to coefficients when ratios are based solely on the data. The expectations appear to succeed in reducing the measurement error associated with observed ratios, and appear to better predictors of the true ratios for the intercensal years. The coefficients on ratios from the full model, $Er3$ and $Er4$, are higher than the coefficients on ratios from the simplified model, $Er1$ and $Er2$, suggesting the importance of using models that are closer to approximating the true data generating process for the observed population counts.

Estimation results in Table 1.13 replicate the analysis in Table 1.12 but also include state fixed effects. The state fixed effects are meant to capture the effect of time invariant state-specific factors influencing marriage which might be correlated to sex ratios. Similarly to results in Table 1.12, simulated expectations of sex ratios have higher coefficients than the sex ratios from the observed data sources, $rCNACS$ and $rCPS$, and the ratios from the full likelihood, $Er3$ and $Er4$, generate higher coefficients than the simplified model ratios, $Er1$ and $Er2$. The simulated sex ratios from parameterizations that allow for greater flexibility in correlations between shocks across groups, $Er2$ and $Er4$, also produce higher coefficients, suggesting more flexible specification might approximate the true sex ratios

better. State fixed effects might not account for all sources of unobservables that are possibly correlated with sex ratios. There might exist time varying covariates that impact the marriage decision and sex ratios. In an attempt to account for these, I also include prison rates and unemployment rates by state and year. Unemployment rates are meant to control for differences in economic climate that might impact sex ratios, while prison rates are meant to capture the effect on marital behavior, if any, of different criminal justice systems.[43] The results are summarized in Table 1.14. The simulated sex ratios have higher coefficients relative to the coefficients on sex ratios from observed data alone. In this specification the coefficients on ratios from the full likelihood are also larger relative to the simplified likelihood, and coefficients on ratios from parameterizations with more flexible covariance structures are also higher. In all three survival analysis specifications, the coefficients predicting entry into first marriage are higher when expected ratios are used as explanatory variables relative to coefficients when ratios are constructed just from the observed data.

1.9 Conclusion

In many research situations ideal data might not be available. I develop a model that combines multiple data sources, each with individual strengths and weaknesses, to recover the features of interest for the data generating process. The model is applied to mitigate the infrequency of the Decennial Census data by supplementing it with yearly information from the CPS and the ACS. I account for data-source-specific measurement error, and allow for missing observations. Using the model estimates, I construct sex ratios, a measure of market tightness, and use them to predict entry into first marriage for NLSY79 women. I find evidence that model generated ratios diminish the presence of measurement error in sex ratios relative to the ratios constructed from yearly CPS data.

The modeling framework is used to recover information between distant time periods, however other notions of distance can be accommodated. For instance, when smaller geo-

[43]Data on prison rates and unemployment rates by state was obtained from John J. Donohue.

graphic units are of interest and are measured with error, but accurate data on large geographic units is available. In addition, the modeling framework can accommodate measures other than just population levels. Total population counts might be available infrequently and can be supplemented with imprecise measures of changes in the population.

1.A Appendix: Survey Weights

For the ACS and the CPS, I obtain the number of unmarried individuals $\hat{U}^x_{s,b,a,t,d}$ in state (s) of race (b), age (a), time (t), sex (x) from data source (d) by summing the weights for all unmarried individuals in data source (d) at survey time (t) that are in the category (s,b,a). Using the CPS as the main example I briefly describe the process of creating the weights in the CPS and discuss whether using survey weights violates any of the assumptions I have made on $ln(\vec{\hat{U}}_{s,a,t,d})$. I find that some of the assumptions made on the measurement errors might be violated. It is possible that the expectation of these errors is not zero, and it's unlikely that the measurement errors are independent. The modeling framework allows me to relax the independence assumption and to model possible correlations between the errors.

The CPS is a probability sample[44] where the CPS knows the probability of selecting each unit in the sample.[45] These probabilities are used in the construction of estimators. According to the CPS:

"An unbiased estimator of the population total for any characteristic investigated in the survey can be obtained by multiplying the value of that characteristic for each sample unit (person or household) by the reciprocal of the probability with which that unit was selected and summing the products over all units in the sample (Hansen, Hurwitz, and Madow, 1953)."[46]

The inverse probability of selection for each unit is used as the base for the weights. The CPS then makes a set of adjustments to the weights, like corrections for nonresponse. To reduce the variance of their estimates the CPS also adjust the weights in order to match the sample distribution for a certain set of characteristics with a "known population distribution" for these characteristics.[47] At the state level the weights are adjusted so that sex (x), age (a) and race (b) groups in the sample match independent estimates of the state

[44] "A probability sample is defined as a sample that has a known nonzero probability of selection for each sample unit." Zbikowski (2006) at p10-1.

[45] "...the probability of selecting each unit in the CPS is known, and every attempt is made to keep departures from true probability sampling to a minimum." Zbikowski (2006) at p10-2.

[46] Zbikowski (2006) at p10-2.

[47] The ACS also adjusts its weights to match "population characteristics".

population.[48] The independent estimates of the population distribution are projections based on the Census numbers supplemented with information from others sources on births, deaths and migration.[49]

The weight creation process might have several implications for the assumptions made on $ln(\vec{\hat{U}}_{s,a,t,d})$. I assume that $E(\vec{e}_{s,a,t,d}) = 0$. If the projection of the population distribution, on which sample weights are based, for the cell (s,b,a,x) at year (t) is unbiased, then the number of people who are unmarried in (s,b,a,x) should also be unbiased because that is just a fraction of the total population in (s,b,a,x). However, if the projection of the population distribution is biased, then it is likely that the estimate for the number of people who are unmarried is also biased and $E(\vec{e}_{s,a,t,d}) \neq 0$.

I also make assumptions on the structure of $e^x_{s,b,a,t,d}$, which divides into a measurement error component $\varepsilon^x_{s,b,a,t,d}$ and a deviation from the mean that is not part of measurement error $v^x_{s,b,a,t}$, and its correlation between categories. I assume that measurement errors are independent across data cells. This might no longer be true, especially since the CPS relies on the Census information to make projections. It is likely that any measurement error in the Census will be correlated with measurement error in the CPS. Measurement errors in the CPS might also be correlated through time as the population projections move further away from the year of the Census. Moreover, the adjustment steps made on the weights to match various characteristics of the population distribution might induce correlations between measurement errors across other categories. For instance, the adjustment of the sample weights to match the "population" estimates at the national level might induce

[48] "In the first-stage ratio adjustment, weights are adjusted so that the distribution of the single-race Black population and the population that is not single-race Black (based on the Census) in the sample PSUs in a state corresponds to the same population groups Census distribution in all PSUs in the state. In the national-coverage ratio adjustment, weights are adjusted so that the distribution of age-sex-race-ethnicity groups match independent estimates of the national population. In the state-coverage ratio adjustment, weights are adjusted so that the distribution of age-sex-race groups match independent estimates of the state population. In the second-stage ratio adjustment, weights are adjusted so that aggregated CPS sample estimates match independent estimates of population in various age/sex/race and age/sex/ethnicity cells at the national level. Adjustments are also made so that the estimated state populations from the CPS match independent state population estimates by age and sex." Zbikowski (2006) at p10-3.

[49] "The independent population controls used in the second-stage ratio adjustment and in the coverage adjustment steps are prepared by projecting forward the population figures derived from Census 2000 using information from a variety of other sources that account for births, deaths, and net migration. ... Prepared in this manner, the controls are themselves estimates. However, they are derived independently of the CPS and provide useful information for adjusting sample estimates." Zbikowski (2006) at p10-8.

correlations in measurement errors across states.[50] I can relax some of the constraints on the correlations between measurement errors, which will imply a new covariance structure on $e^x_{s,b,a,t,d}$. Because the CPS does not match the sample distribution to population projections for the unmarried category the measurement error can be decomposed into two components, the part of error attributable to the CPS population projections $\phi^x_{s,b,a,t,d}$ and the independent sampling error $\psi^x_{s,b,a,t,d}$.

$$\varepsilon^x_{s,b,a,t,d} = \phi^x_{s,b,a,t,d} + \psi^x_{s,b,a,t,d} \tag{1.56}$$

$\phi^x_{s,b,a,t,d}$ will be correlated across categories depending on the weight adjustment process. However, it is unlikely that I would be able to separately identify variance-covariance parameters for $v^x_{s,b,a,t}$ and $\phi^x_{s,b,a,t,d}$.

[50] If the projection for the number of Hispanics for the nations has an error, then it is likely that there is correlation in measurement error across states for the Hispanics.

1.B Appendix: Additively Separable Measurement Error in Sex Ratios

A very convenient modeling approach when estimating $E[\hat{r}^d_{s,b,a,t}]$ is to assume that measurement error enters the observed measures $\hat{r}^d_{s,b,a,t}$ in an additively separable way. For instance, using natural log of the sex ratios as the dependent variable, let

$$ln(\hat{r}^d_{s,b,a,t}) = ln(r_{s,b,a,t}) + \xi_{s,b,t,d} \tag{1.57}$$

where $\xi_{s,b,t,d}$ is the measurement error. If I assume that measurement error is the same for all age groups for a given (s,b,t,d) and has the following form

$$\hat{U}^M_{s,b,a,t,d} = U^M_{s,b,a,t} * \epsilon^M_{s,b,t,d} \tag{1.58}$$

then

$$\Sigma^{10}_{k=0}\hat{U}^M_{s,b,a+k,t,d} = \epsilon^M_{s,b,t,d} * (\Sigma^{10}_{k=0}U^M_{s,b,a+k,t}). \tag{1.59}$$

If I assume the same for women, the log of the sex ratio is equal to

$$ln(\hat{r}^d_{s,b,a,t}) = ln\left(\frac{\Sigma^{10}_{k=0}U^M_{s,b,a+k,t}}{\Sigma^{8}_{k=-2}U^W_{s,b,a+k,t}}\right) + ln\left(\frac{\epsilon^M_{s,b,t,d}}{\epsilon^W_{s,b,t,d}}\right) = ln(r_{s,b,a,t}) + \xi_{s,b,t,d}. \tag{1.60}$$

With a normalizing assumption $E[\xi_{s,b,t,d}] = 0$, the additive error has the additional desirable feature of $E[ln(\hat{r}^d_{s,b,a,t})] = E[ln(r_{s,b,a,t})]$.[51] Assuming linearity on $E[ln(r_{s,b,a,t})]$ and the covariance structure for errors $\xi_{s,b,t,d}$ leads to a Generalized Least Squares approach to estimation. An additively separable measurement error can be justified if the measurement errors for each data cell do not display heteroskedasticity with respect to age (a); an assumption that is, unfortunately, rejected by the data.

[51] This feature should persist even if some of the data cells are missing, since the measurement error is proportional to the true unobserved $U^x_{s,b,a,k,t}$ for all age groups within a (s, b, t, d) category.

1.C Appendix: The Missing Data Problem and Estimation

Supplementing the infrequent Census with more frequent ACS and CPS introduces a missing data problem. Smaller sample size implies that not all data cells will be observed by some data sources. Little and Rubin (2002) develops conditions under which researchers can ignore explicitly modeling the process generating the missing data and still obtain consistent estimates of their model. The most important condition, Missing at Random (MAR), requires that conditional on observed data the mechanism generating the missing data does not depend on the values of the missing observations. Following Little and Rubin, let the joint distribution of data (Y) and missing data indicators (D) be written as $f(Y, D|\theta, \psi) = f(Y|\theta)f(D|Y, \psi)$, where θ, ψ are the parameters describing the data generating process. The data are divided into the observed Y_{obs} and missing components Y_{mis} and so

$$f(Y, D|\theta, \psi) = f(Y_{obs}, Y_{mis}|\theta)f(D|Y_{obs}, Y_{mis}|\psi). \tag{1.61}$$

The MAR assumption requires that $f(D|Y_{obs}, Y_{mis}, \psi) = f(D|Y_{obs}, \psi)$ allowing the joint density of (Y) and (D) to be expressed

$$f(Y, D|\theta, \psi) = f(Y_{obs}, Y_{mis}|\theta) * f(D|Y_{obs}, \psi). \tag{1.62}$$

The density of the actual observed data $f(Y_{obs}, D|\theta, \psi)$ is obtained by integrating out the missing data.

$$f(Y_{obs}, D|\theta, \psi) = \int f(Y, D|\theta, \psi)dY_{mis} = \tag{1.63}$$
$$f(D|Y_{obs}, \psi) \int f(Y_{obs}, Y_{mis}|\theta)dY_{mis} = f(D|Y_{obs}, \psi)f(Y_{obs}|\theta)$$

If θ are the parameters of interest, likelihood estimation ignoring the missing data mechanism will rely on $f(Y_{obs}|\theta)$ and will be proportional to maximizing the full likelihood that explicitly models the missing data.

I model the number of unmarried individuals in each data cell, for notational simplicity denoted by (i), as $\hat{U}_i = e^{\mu_i(Z_i)+v_i+\varepsilon_i}$ where Z_i is a set of indicators for sex (x), state (s), race (b), age (a), and year of observation (t). v_i is a deviation of observation (i) from the mean $\mu_i(Z_i)$, and ε_i is the measurement error. Let D_i be an indicator equal to 1 if the observation is missing and zero otherwise. Z_i is always observed. The probability that observation (i) is missing can be denoted by $Pr(D_i = 1|Z_i, ln(\hat{U}_i))$ which is equivalent to $Pr(D_i = 1|Z_i, v_i, \varepsilon_i)$ given the assumptions made on $\hat{U}_i$. Missing data mechanism is MAR if D_i is independent of deviations from the mean, v_i, and measurement errors, ε_i. For data cells (i) missing because the surveys were not fielded in those years MAR is easily satisfied, probability of missing does not depend on the unobserved number of individuals measured with error. However if a data cell is more or less likely to be missing given the direction and the size of the deviation from the mean v_i, then MAR fails. A large positive shock for instance implies a large number of individuals in that data cell, increasing the probability that some of these individuals are sampled by a survey.

References

Abramitzky, Ran, Adeline Delavande, and Luis Vasconcelos. "Marrying Up: The Role of Sex Ratio in Assortative Matching." *American Economic Journal: Applied Economics 3.3* (2011): 124-157.

Angrist, Josh. "How do sex ratios affect marriage and labor markets? Evidence from America's second generation." *The Quarterly Journal of Economics* 117.3 (2002): 997-1038.

Brien, Michael J. "Racial differences in marriage and the role of marriage markets." *Journal of Human Resources* (1997): 741-778.

Bureau of Labor Statistics, U.S. Department of Labor. *National Longitudinal Survey of Youth 1979 cohort, 1979-2008 (rounds 1-23)* [computer file]. Produced and distributed by the Center for Human Resource Research, The Ohio State University. Columbus, OH: 2002.

Bureau of Labor Statistics, U.S. Department of Labor, and National Institute for Child Health and Human Development. *Children of the NLSY79, 1979-2008* [computer file]. Produced and distributed by the Center for Human Resource Research, The Ohio State University. Columbus, OH: 2002.

Caceres-Delpiano, Julio, and Eugenio Giolito. "The impact of unilateral divorce on crime." *Journal of Labor Economics* 30. 1 (2012): 215-248.

Donohue, John J, Aneja, Abhay, and Alexandria Zhang. "The Impact of Right-to-Carry Laws and the NRC Report: Lessons for the Empirical Evaluation of Law and Policy" *American Law and Economics Review* 13.2 (2011): 565-631.

Drewianka, Scott. "Divorce law and family formation." *Journal of Population Economics* 21.2 (2008): 485-503.

Finlay, Keith, and David Neumark. "Is Marriage Always Good for Children? Evidence from Families Affected by Incarceration." *Journal of Human Resources* 45.4 (2010): 1046-1088.

Freiden, Alan. "The United States Marriage Market." *Journal of Political Economy* 82.2

(1974): S34-S53.

Gaines, Leonard M., "A Compass for Understanding and Using American Community Survey Data" U.S. Census Bureau, Washington DC (2009).

Greene, William H. Econometric Analysis. Vol. 5. Upper Saddle River, NJ: Prentice Hall, 2003.

Jennrich, Robert I., and Mark D. Schluchter. "Unbalanced repeated-measures models with structured covariance matrices." *Biometrics* (1986): 805-820.

Johnson, Rucker C., and Steven Raphael. "The effects of male incarceration dynamics on acquired immune deficiency syndrome infection rates among African American women and men." *Journal of Law and Economics* 52.2 (2009): 251-293.

Lichter, Daniel T., Diane K. McLaughlin, George Kephart, and David J. Landry. "Race and the retreat from marriage: A shortage of marriageable men?." *American Sociological Review* (1992): 781-799.

Little, Roderick JA, and Donald B. Rubin. *Statistical analysis with missing data.* Vol. 333. Wiley-Interscience, 2002.

Paxson, Christina, and Jane Waldfogel. "Work, Welfare, and Child Maltreatment." *Journal of Labor Economics* 20.3 (2002): 435-474.

Ruggles, Steven, Alexander, Trent J., Genadek, Katie, Goeken, Ronald, Schroeder, Matthew B., and Matthew Sobek.*Integrated Public Use Microdata Series: Version 5.0* [Machine-readable database]. Minneapolis: University of Minnesota, 2010.

Ruggles, Steven, Alexander, Trent J., Genadek, Katie, Schroeder, Matthew B., King, Miriam, Flood, Sarah, Trampe, Brandon, and and Rebecca Vick. *Integrated Public Use Microdata Series, Current Population Survey: Version 3.0.* [Machine-readable database]. Minneapolis: University of Minnesota, 2010.

Seitz, Shannon. "Accounting for racial differences in marriage and employment." *Journal of Labor Economics 27.3* (2009): 385-437.

South, Scott J., and Kim M. Lloyd. "Marriage opportunities and family formation: Further implications of imbalanced sex ratios." *Journal of Marriage and the Family* (1992): 440-451.

Wolfers, Justin. "Did Unilateral Divorce Laws Raise Divorce Rates? A Reconciliation and New Results." *American Economic Review* 96.5 (2006): 1802-1820.

Wood, Robert G. "Marriage rates and marriageable men: A test of the Wilson hypothesis." *Journal of Human Resources* (1995): 163-193.

Stern, Steven. "Simulation-based estimation." *Journal of Economic Literature* 35.4 (1997): 2006-2039.

Zbikowski, Andrew, and Antoinette Lubich. "Current Population Survey: Design and Methodology". *Technical Paper 66*, Bureau of Labor Statistics and U.S. Census Bureau, Washingotn DC (2006).

Tables

Table 1.1: NLSY Sample by Race

	White	Black	Hipanic	Total
number	1,266	702	368	2,336
% of total	54%	30%	16%	

NLSY79 sample of women whose marital decision was investigated. Excludes observations with gaps in residence while at risk of getting married.

Table 1.2: New York

	White		Black		Hispanic	
	Male	Female	Male	Female	Male	Female
Age <25						
σ_v^2	0.016	0.034	0.010	0.012	0.014	0.019
σ_ε^2 CN	0.010	0.010	0.010	0.010	0.010	0.010
σ_ε^2 ACS	0.019	0.019	0.038	0.031	0.029	0.026
σ_ε^2 CPS	0.036	0.035	0.147	0.099	0.083	0.068
Age ≥ 25						
σ_v^2	0.033	0.033	0.030	0.021	0.012	0.015
σ_ε^2 CN	0.010	0.010	0.010	0.010	0.010	0.010
σ_ε^2 ACS	0.046	0.035	0.346	0.132	0.290	0.110
σ_ε^2 CPS	0.213	0.119	11.892	1.745	8.391	1.198

Full model with covariances in errors within race

Table 1.3: New York

	White		Black		Hispanic	
	Male	Female	Male	Female	Male	Female
Age <25						
σ_v^2	0.010	0.029	0.010	0.010	0.010	0.010
σ_ε^2 CN	0.010	0.010	0.010	0.010	0.010	0.010
σ_ε^2 ACS	0.021	0.021	0.039	0.032	0.030	0.028
σ_ε^2 CPS	0.043	0.045	0.148	0.101	0.089	0.077
Age ≥ 25						
σ_v^2	0.022	0.010	0.010	0.010	0.010	0.010
σ_ε^2 CN	0.030	0.024	0.037	0.011	0.015	0.010
σ_ε^2 ACS	0.047	0.038	0.347	0.133	0.291	0.110
σ_ε^2 CPS	0.224	0.142	11.951	1.771	8.403	1.206

Full model with no covariances

Table 1.4: Virginia

	White		Black		Hispanic	
	Male	Female	Male	Female	Male	Female
Age <25						
σ_v^2	0.010	0.018	0.012	0.015	0.031	0.037
σ_ε^2 CN	0.010	0.010	0.010	0.010	0.023	0.010
σ_ε^2 ACS	0.034	0.062	0.215	0.184	0.524	0.595
σ_ε^2 CPS	0.112	0.385	4.584	3.372	27.329	35.229
Age ≥ 25						
σ_v^2	0.029	0.033	0.063	0.050	0.660	0.220
σ_ε^2 CN	0.010	0.010	0.010	0.010	1.582	0.010
σ_ε^2 ACS	0.290	0.240	0.695	0.461	3.229	1.633
σ_ε^2 CPS	8.363	5.737	48.051	21.172	150.904	187.453

Full model with covariances in errors within race

Table 1.5: Virginia

	White		Black		Hispanic	
	Male	Female	Male	Female	Male	Female
Age <25						
σ_v^2	0.010	0.010	0.010	0.010	0.010	0.010
σ_ε^2 CN	0.022	0.023	0.017	0.010	0.056	0.051
σ_ε^2 ACS	0.034	0.063	0.215	0.184	0.523	0.594
σ_ε^2 CPS	0.112	0.390	4.584	3.380	27.202	35.131
Age ≥ 25						
σ_v^2	0.010	0.010	0.010	0.010	0.010	0.010
σ_ε^2 CN	0.035	0.031	0.080	0.062	2.964	0.163
σ_ε^2 ACS	0.291	0.241	0.695	0.461	3.891	1.895
σ_ε^2 CPS	8.409	5.765	48.108	21.140	140.831	192.876

Full model with no covariances

Table 1.6: Virginia

Count of Observed Data Cells							Fraction of Observed Data Cells*						
	White		Black		Hispanic			White		Black		Hispanic	
	Male	Female	Male	Female	Male	Female		Male	Female	Male	Female	Male	Female
Age <25							*Age <25*						
Census	33	33	33	33	33	33	Census	100%	100%	100%	100%	100%	100%
ACS	99	99	99	99	99	99	ACS	100%	100%	100%	100%	100%	100%
CPS	352	351	333	338	246	221	CPS	100%	100%	95%	96%	70%	63%
Age ≥ 25							*Age ≥ 25*						
Census	108	108	108	108	101	108	Census	100%	100%	100%	100%	94%	100%
ACS	324	324	324	324	279	308	ACS	100%	100%	100%	100%	86%	95%
CPS	1,026	1,068	681	874	308	347	CPS	89%	93%	59%	76%	27%	30%

VA, *Numerator is the number of observed data cells, denominator is the number of cells that would be observed if the cell was observed every period data was fielded.

Table 1.7: New York

Count of Observed Data Cells							Fraction of Observed Data Cells*						
	White		Black		Hispanic			White		Black		Hispanic	
	Male	Female	Male	Female	Male	Female		Male	Female	Male	Female	Male	Female
Age <25							*Age <25*						
Census	33	33	33	33	33	33	Census	100%	100%	100%	100%	100%	100%
ACS	99	99	99	99	99	99	ACS	100%	100%	100%	100%	100%	100%
CPS	352	352	352	352	352	352	CPS	100%	100%	100%	100%	100%	100%
Age ≥ 25							*Age ≥ 25*						
Census	108	108	108	108	108	108	Census	100%	100%	100%	100%	100%	100%
ACS	324	324	324	324	324	324	ACS	100%	100%	100%	100%	100%	100%
CPS	1,151	1,152	981	1,128	1,015	1,136	CPS	100%	100%	85%	98%	88%	99%

NY, *Numerator is the number of observed data cells, denominator is the number of cells that would be observed if the cell was observed every period data was fielded.

Table 1.8: New York

	White		Black		Hispanic	
	Male	Female	Male	Female	Male	Female
Age <25						
σ_v^2	0.010	0.017	0.026	0.029	0.024	0.028
σ_ε^2 CN	0.012	0.102	0.024	0.017	0.138	0.035
σ_ε^2 ACS	0.034	0.023	0.028	0.028	0.056	0.057
σ_ε^2 CPS	0.045	0.055	0.119	0.079	0.069	0.054
Age ≥ 25						
σ_v^2	0.031	0.037	0.035	0.028	0.027	0.042
σ_ε^2 CN	0.076	0.074	0.054	0.068	0.142	0.163
σ_ε^2 ACS	0.084	0.077	0.074	0.068	0.199	0.213
σ_ε^2 CPS	0.132	0.139	0.265	0.262	0.262	0.188

Simplified model with covariances in errors within race

Table 1.9: Virginia

	White		Black		Hispanic	
	Male	Female	Male	Female	Male	Female
Age <25						
σ_v^2	0.014	0.012	0.012	0.019	0.010	0.012
σ_ε^2 CN	0.055	0.180	0.056	0.073	0.183	0.207
σ_ε^2 ACS	0.039	0.076	0.044	0.061	0.117	0.105
σ_ε^2 CPS	0.125	0.149	0.272	0.322	0.410	0.338
Age ≥ 25						
σ_v^2	0.060	0.033	0.014	0.020	0.103	0.023
σ_ε^2 CN	0.120	0.181	0.131	0.283	0.190	0.292
σ_ε^2 ACS	0.190	0.170	0.175	0.193	0.191	0.269
σ_ε^2 CPS	0.260	0.346	0.501	0.469	1.247	1.381

Simplified model with covariances in errors within race

Table 1.10: Marriage by Race and Age for NLSY79 Women

	Number				% by race			
	White	Black	Hispanic	Total	White	Black	Hispanic	Total
never married	79	186	47	312	6.39	26.72	12.98	13.60
up to 20	350	89	132	571	28.32	12.79	36.46	24.89
20 to 24	481	211	112	804	38.92	30.32	30.94	35.05
25 to 29	209	107	47	363	16.91	15.37	12.98	15.82
30 to 34	77	52	13	142	6.23	7.47	3.59	6.19
35 or older	40	51	11	102	3.24	7.33	3.04	4.45
Total	1236	696	362	2294	100.00	100.00	100.00	100.00

Table 1.11: Census Sex Ratios by Age

age	mean	median
18	1.098	1.000
19	1.090	0.997
20	1.097	1.005
21	1.103	1.008
22	1.091	0.997
23	1.097	1.004
24	1.090	1.018
25	1.090	1.024
26	1.091	1.034
27	1.079	1.036
28	1.080	1.036
29	1.069	1.039
30	1.069	1.048
31	1.059	1.045
32	1.055	1.037
33	1.034	1.020
34	1.026	1.008
35	1.018	0.997
36	0.994	0.963
37	0.975	0.958
38	0.958	0.934
39	0.963	0.926
40	0.941	0.897
41	0.918	0.876
42	0.903	0.860
43	0.885	0.844
44	0.870	0.822
45	0.853	0.806
46	0.826	0.775
47	0.808	0.746
48	0.788	0.729
49	0.758	0.703
50	0.743	0.682

Table 1.12: Waiting Time for First Marriage: Coefficients on Sex Ratios

	1	2	3	4	5	6
rCNACS	-1.045**					
(se)	(0.330)					
p-value	0.002					
rCPS		-0.140				
(se)		(0.096)				
p-value		0.143				
Er1			-2.157***			
(se)			(0.420)			
p-value			0.000			
Er2				-2.397***		
(se)				(0.422)		
p-value				0.000		
Er3					-2.341***	
(se)					(0.339)	
p-value					0.000	
Er4						-2.793***
(se)						(0.404)
p-value						0.000
N	18911	18911	18911	18911	18911	18911

The errors have a log-normal distribution. Time constant person-specific unobserved heterogeneity is included. Only observations for whom state of residence is available for all periods while at risk are used. Race indicators and log of AFQT score are included in all specifications.

Table 1.13: Waiting Time for First Marriage: Coefficients on Sex Ratios

	1	2	3	4	5	6
rCNACS	-1.230**					
(se)	(0.445)					
p-value	0.006					
rCPS		-0.0962				
(se)		(0.096)				
p-value		0.317				
Er1			-2.001***			
(se)			(0.506)			
p-value			0.000			
Er2				-2.374***		
(se)				(0.519)		
p-value				0.000		
Er3					-2.954***	
(se)					(0.401)	
p-value					0.000	
Er4						-3.807***
(se)						(0.503)
p-value						0.000
State FE	YES	YES	YES	YES	YES	YES
N	18911	18911	18911	18911	18911	18911

The errors have a log-normal distribution. Time constant person-specific unobserved heterogeneity is included. Only observations for whom state of residence is available for all periods while at risk are used. Race indicators and log of AFQT score are included in all specifications.

Table 1.14: Waiting Time for First Marriage: Coefficients on Sex Ratios

	1	2	3	4	5	6
rCNACS	-1.144**					
(se)	(0.371)					
p-value	0.002					
rCPS		-0.101				
(se)		(0.086)				
p-value		0.240				
Er1			-1.582***			
(se)			(0.414)			
p-value			0.001			
Er2				-1.863***		
(se)				(0.422)		
p-value				0.000		
Er3					-2.298***	
(se)					(0.349)	
p-value					0.000	
Er4						-2.845***
(se)						(0.437)
p-value						0.000
State FE	YES	YES	YES	YES	YES	YES
Unemp Rate	YES	YES	YES	YES	YES	YES
Prison Rate	YES	YES	YES	YES	YES	YES
N	18911	18911	18911	18911	18911	18911

The errors have a log-normal distribution. Time constant person-specific unobserved heterogeneity is included. Only observations for whom state of residence is available for all periods while at risk are used. Race indicators and log of AFQT score are included in all specifications.

Figures

Figure 1.1: Fraction of Observed Data Cells

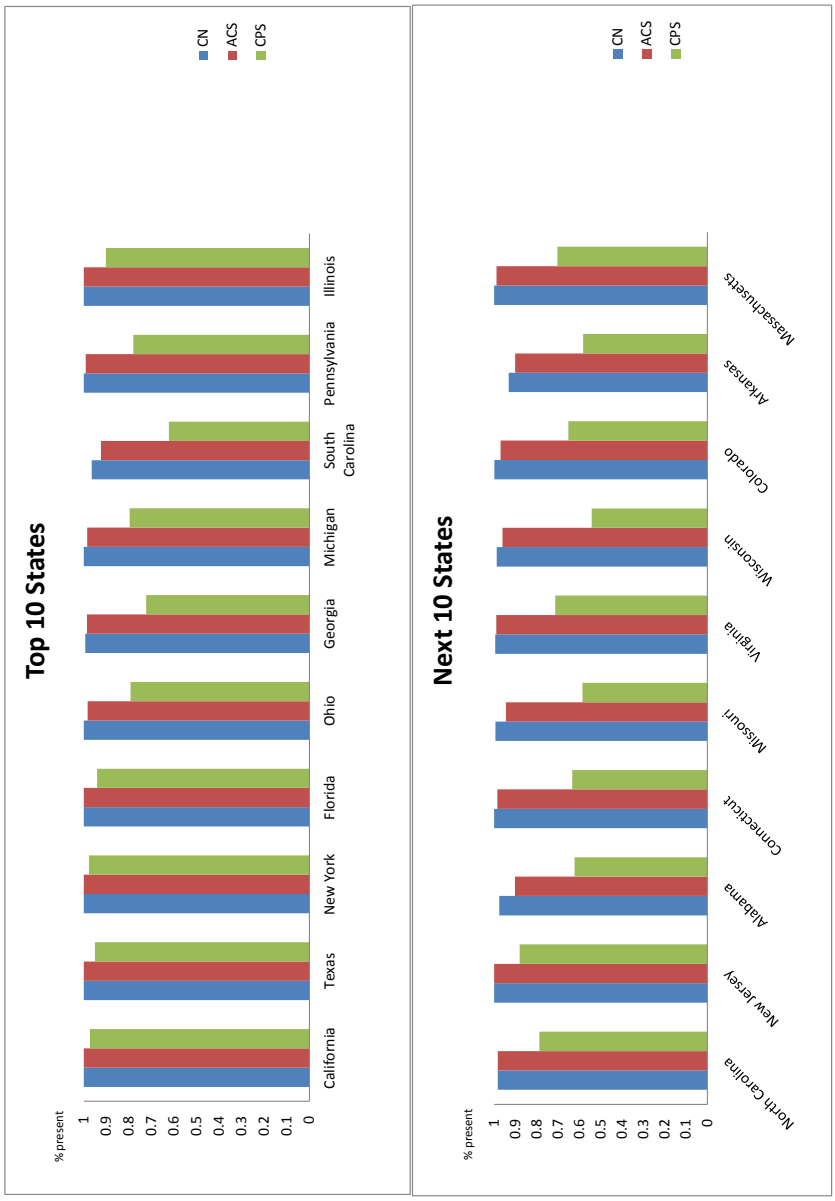
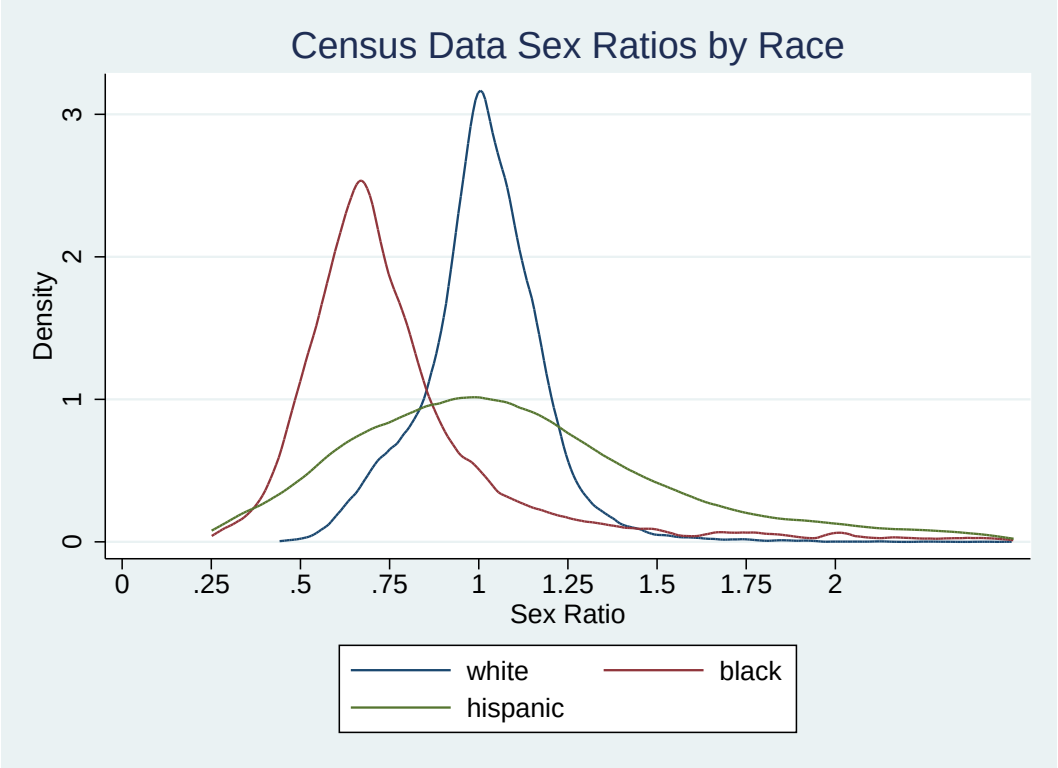


Figure 1.2: Census Data Sex Ratios by Race



Chapter 2

The Impact of Mothers' Decision to Marry on Child Outcomes

2.1 Introduction

In the United States, children raised in families with two biological parents have better outcomes than children in other types of family structures.[1] However, non-traditional family structures have been on the rise. In particular, the percentage of births to unmarried women doubled since 1980, and by 2007 represented 40 percent of all births.[2] Children in single mother households are more likely to live below the poverty line, prompting some to suggest marriage promotion as a policy to reduce childhood poverty.[3] While it might reduce poverty, the impact of a mother's marriage on children born out of wedlock is difficult to determine. Empirical research must address the possibility that a host of other factors, potentially unobserved, is also correlated with child outcomes and mother's marriage. I use the National Longitudinal Survey of Youth 1979 (NLSY79) and the Child and Young Adult supplement (CNLSY79) to study the effects of mother's marriage on children born out of wedlock. To address the endogeneity of the mother's marital status, I use an instrumental

[1] See McLanahan and Sandefur (1994), and Ginther and Pollak (2004).

[2] According to the Center of Disease Control (CDC), 18 percent of all births were to unmarried women in 1980.

[3] See Sigle-Rushton and McLanahan (2002).

variable approach. Previous research has found that relative availability of partners has an effect on the marital decisions of women, and the results of Chapter 1 allow for a construction of frequent measures for marriage market features experienced by the NLSY79 women. I use these measures as instrumental variables for mother's marital status.

Marriage is likely to affect the child in a variety of ways. For one, it is likely to change the level of resources in the household which in turn might change the amount of resources devoted to the child. In a household with two potential earners instead of one, the mother might choose to work less and devote more of her leisure time to the child. Marriages are also likely to be more stable than other types of relationships, increasing the stability in household resources available and potentially decreasing the number of future partners for the mother.[4] Mother's marriage can also have negative consequences. For instance, marriage can generate additional conflict in the household as the new partner begins to participate in raising the child. While mother's marriage can have some positive and some negative consequences, ordinary least squares (OLS) regressions show that mother's marriage is positively correlated with high school graduation for a sample of black children born to the unwed mothers of the NLSY79.[5] Relative to children whose mothers remain unmarried, children whose mothers married were more likely to obtain a high school diploma.[6] The size of the marriage coefficient in the high school diploma equation, however, changes considerably depending on the set of other covariates included in the regression. The importance of other variables underscores how family structure and child outcomes depend on a wide range of other factors, not all of which will be observed. The potential of unobservables correlated with both variables of interest necessitates a different estimation approach.

In order to address the endogeneity of the mother's transition to first marriage, I use

[4]In the Fragile Families and Child Wellbeing Study (FFCWS) data only 35 percent of couples who were unmarried at the birth of the child were living together five years after the birth. Many mothers formed new relationships and had children with the new partner. The work of Craigie et al. (2012) suggests that transitions might be detrimental to children.

[5]The majority of black women in the NLSY79 chose to have a first birth prior to first marriage, if any. For Hispanic and white respondents, however, only a small minority chose to have a first birth prior to first marriage.

[6]I also use math test scores and an indicator if the child became an unwed parent as dependent variables. OLS results show no significant difference between the two groups in test scores or children's propensity to become unwed parents.

a set of instrumental variables. Previous research has found that sex ratios, the relative availability of potential spouses, impacts women's transitions into marriage.[7] Fewer potential partners relative to potential competitors translates to a longer waiting time until first marriage. Variation in sex ratios across marriage markets is used to predict the mother's entry into first marriage. However, for the NLSY79 and many other older data sets, timely and accurate demographic data used to construct sex ratios is not available. In Chapter 1 of this dissertation, I develop and estimate a model that combines multiple sources of data measuring the demographic characteristics of the population during the NLSY79 survey time frame. The model estimates allow me to construct yearly measures of expected sex ratios by marriage market. I obtain access to restricted geographic location of the NLSY79 women which allows me to connect estimated sex ratios to the women for each year they are in the sample. The sex ratios observed every year in the mother's marriage market are then used as a set of instruments for the marriage status indicator in the child outcome equation.

Using all the instrumental variables in a General Method of Moments (GMM) estimation framework, I find that mother's marriage has a large and positive effect on the high school graduation status of their children. Concerns over unobservable factors that are correlated with child outcomes and mother's marriage markets prompt further investigation of whether sex ratios make suitable instruments. The initial results are robust to the inclusion of state fixed effects or unemployment rates in the outcome equation. State fixed effects are meant to capture any time-constant variables that might drive greater imbalance in sex ratios and educational outcomes, while unemployment rates are meant to capture time varying differences in economic climate across states. However, the marriage coefficient estimate declines when state-specific prison rates are included in the analysis. Underlying factors generating criminal behavior could impact high school graduation, and higher male incarceration rates decrease the supply of unmarried men. These results suggest caution when considering the GMM coefficient estimates. The exogeneity of the instrumental variables is undermined by the sensitivity of coefficient estimates when other marriage-market-specific variables are

[7]See Freiden (1974), South and Lloyd (1992), Wood (1995), Lichter et al. (1992), and Brien (1997).

included in the outcome equations.

I proceed by describing previous research results on the effects of family structure on child outcomes. Next, I describe the data used, define the family structure variable, and present the equations to be estimated. Last, I describe the OLS results and the results of GMM estimation using sex ratios as instruments for the endogenous variables.

2.2 Literature Review

The correlation between family structure and child outcomes has been extensively explored and documented. McLanahan and Sandefur (1994) compare outcomes between children who experience family disruptions and children who do not. Relative to their peers who grow up with both parents, children who experience a divorce or are born to unwed mothers have worse outcomes. They are more likely to drop out of high school and, if they do graduate from high school, are less likely to enroll and graduate from college. They are also more likely to be out of school and out of work, and young women are more likely to become teen mothers. Ginther and Pollak (2004) add nuance to these descriptive results using the NLSY79 and PSID data sources. They show that a child-centric classification system of family structure might mask important differences between family structures. For instance, a family with two children, one from a previous marriage and the other a biological offspring of both spouses, will be classified differently depending on the child. One child will live in a household with two biological parents, while the other will live in a household with a stepparent. Ginther and Pollak show that, in these blended families, half-siblings (biological children of both parents) have similar educational outcomes, like years of schooling or high school graduation status, as their step-children siblings. Both types of children, however, have substantially worse outcomes than kids reared in traditional nuclear families where all children share the same biological parents.

McLanahan and Sandefur (1994), and Ginther and Pollak (2004) document the correlation between child outcomes and family structure measured when the child is older. For the most part however, their work fails to consider the effect of transitions between family

structures.[8] It is likely that the number and types of transitions between family structures will have an impact on the child. Liu and Heiland (2006) investigate the relationship between marital transitions and child outcomes for children born to unmarried parents. They use the Fragile Families and Child Wellbeing Study (FFCWS) which collects information about both biological parents regardless of marital or cohabitation status at the birth of the child.[9] Liu and Heiland (2006) find that child health outcomes are not observed to improve when cohabiting parents marry within the first year of the child's birth. However, children whose parents continue to stay together in that first year, either in marriage, cohabitation, or a visiting relationship, fare better than children whose mothers end the relationship with the biological father. Craigie et al. (2012) go further and attempt to attribute differences in outcomes to initial family structure separately from the transitions in family structure until the child is five years old. They use the FFCWS and find small amount of correlation between outcomes and initial family structure. Relative to children born to married parents, children born to single mothers or cohabiting parents fare worse. However, they find a strong correlation between the relationship status of the mother when the child is five and child outcomes. Children living with married parents at age five had higher test scores compared to children whose parents were cohabiting, the mother was single, or the mother was with a new partner (either cohabiting or married). The results of Craigie et al. (2012) suggest that marital transitions away from a nuclear family might be detrimental to child outcomes.

The work of McLanahan and Sandefur (1994), Ginther and Pollak (2004), Liu and Heiland (2006), and Craigie et al. (2012) is informative but cannot tell us the effect of family structure or transitions on child outcomes. Unobservable characteristics, like preferences for the amount of time spent with children and desire for a spouse, might be responsible for the observed correlation. Manski et al. (1992) attempt to go beyond descriptive correlations and lay out a framework to identify the effects of family structure on child outcomes. They use

[8] McLanahan and Sandefur do show that children of divorced parents have worse outcomes than children of parents that stay together, suggesting transitions matter and outcomes might not depend on whether a child starts in a two parent household.

[9] Many other data sets collect the information about the father only if he is a member of the household because of marriage or cohabitation.

the NLSY79 data and estimate the effects of family structure under a variety of assumptions. First, they assume that unobservables affecting the selection into family structure when the child is 14 years old and unobservables for high school graduation status are multivariate normal.[10] Given the parametric assumption on the distribution of the unobservables, they estimate two models. In one model they assume family status is exogenous to high school graduation, and in the other unobservables for family status and graduation are freely correlated.[11] Estimates from both models suggest living in a non-intact family at age 14 substantially lowers the probability that the child graduates from high school. Second, because the normality assumption on the probability distribution of the unobservables is rather arbitrary, the authors assume no information on the joint probability distribution but, to achieve point identification, they assume family structure is exogenous to outcomes. Under exogeneity, the nonparametric estimates support the estimates from the parametric models. Living in an intact family improves the probability of high school graduation for kids but the positive effects seem to decline as the educational level of the parents increases.

The work of Manski et al. (1992) highlights the difficulty in identifying the effects of family structure on child outcomes. Both assumptions used for point identification are unsatisfactory. The normality assumption on the distribution of unobservables seems arbitrary, while the exogeneity of family structure to high school graduation seems implausible. It is easy to suggest unobservable factors that affect family structure and the educational outcome of the child. For instance, employment status of both parents might influence family structure and resources available to the child. Manski et al. (1992) recognize the difficulty in achieving identification and suggest progress can be made if researchers use additional information on the relationship between family structure and child outcomes. Much of subsequent work follows that suggestion. Researchers have used a variety of other strategies, besides making distributional assumptions or relying on exogeneity, to model the relationship and attempt to address the presence of unobservables that can generate corre-

[10] The unobservables are assumed to have zero mean and variances equal to one.

[11] In their context exogeneity implies that there is zero covariance between family status and unobservables on outcomes. In the model that allows for correlation in unobservables, they also make cross equation exclusion restrictions on observed covariates. For example, family structure does not depend on sex of the child or education status of the parents.

lation between family structure and child outcomes. One approach relies on access to a rich set of data which measures a wide range of the relevant factors that influence both variables of interest. Liu and Heiland (2012) use the FFCWS to estimate the impact of marriage on children born to unmarried parents. They compare child outcomes at age three for the group whose biological parents married each other after the child's birth, to the group that did not report a marriage. Unlike most data sets, the FFCWS contains information on both biological parents, regardless of the presence of the father in the household.[12] Liu and Heiland (2010) use a propensity score approach which assumes that assignment into the married state is random after conditioning on a rich set of observable covariates. They find that children's cognitive ability increased when the biological parents married. While the FFCWS data allow researchers to control for characteristics of both parents, the propensity score approach is invalid if some factors that influence the transition into marriage and child outcomes remain unobserved. It is unlikely that any data set will ever collect all the possible variables that influence child outcomes and family structure.[13]

Aughinbaugh et. al. (2005) do not depend on observing all relevant variables. They use a fixed effects model to analyze the impact of mother's marriage on children born to unwed mothers, and the impact of divorce on children born in two-parent households. The fixed effects approach allows researchers to control for the presence of unobserved person-specific variables that remain constant and affect outcomes. They use the PIAT Math and Reading test scores and the Behavioral Problems Index (BPI) scores from the CNLSY79 sample as the outcome variables. In their model, test scores at time (t) depend on a set of indicators for when the marital transition took place relative to the test time.[14] They find that children born to unwed mothers experience a decline in reading test scores after the marriage, decline in math test scores before the marriage, and some evidence of worse behavior prior to the

[12]Many data sets include biological parent characteristics only if they both live in the same household, regardless if they are married.

[13]For example, amount of effort a mother exerts in her relationships, with the child and partners, would be very difficult to quantify and observe. In addition, the mere knowledge that behavior is observed might induce individuals to change how they interact with others.

[14]The OLS specification allows test scores to also depend on an indicator if the mother transitioned to marriage and three additional explanatory variables: child's gender, the child's race, and the mother's AFQT score. In the fixed effects specification the time-constant marital indicator and the explanatory variables drop out due to differencing.

mother's marriage. The work of Aughinbaugh et. al. (2005) suggests marriage of unwed mothers might be a disruptive event in the life of the child born out of wedlock. Gennetian (2005) also uses panel data on PIAT test scores for the NLSY79 children to control for time-constant unobservables possibly correlated with family structure. Family structure is defined from the child perspective, because it is possible that two children residing in the same household can report two different family structures.[15] Gennetian (2005) models two different fixed effects specifications, a specification with a fixed effect for each child, and a specification with a fixed effect for each family. The outcome variable is defined as the average of math and reading recognition scores and she finds only one precisely estimated effect; children born to unwed mothers have lower test scores than children living in nuclear families.[16] In the child-specific fixed effect model, the single mother coefficient is identified when an unwed mother transitions to a different marital state sometime between the first and the last period her child takes the tests. Gennetian (2005) allows for 5 family structures, but only a small fraction of children in her sample actually experience a transition between these states. Overall only 7 percent of the children in her sample experience any family structure change between the first and last test assessment.

Wagmiller et al. (2010) use a different data source, the Early Childhood Longitudinal Study (ECLS-K), to investigate the impact of a mother's marriage on children who were living with an unwed mother at the time of the first interview. Their identification strategy assumes a quadratic trend for math and reading scores for the children during the early parts of their lives, and marriage is allowed to cause a deviation from this baseline trend. Because test scores and family structure are observed at six points in time, the authors are able to control for child-specific time-constant unobservables that impact the trend for each child. They find that, when marriage occurs sometime between kindergarten and fifth grade, the initial impact on scores is relatively small and positive but declines with the passage of time. The authors also find that response to marriage is heterogeneous where the benefits are much higher for whites than for minorities, are more positive for children

[15]For example, if one sibling is from a previous marriage, she will have a step-parent and a biological parent family structure, while the younger sibling will report two biological parents.

[16]Nuclear families have two biological parents and siblings that have the same two parents.

with higher educated mothers, and are higher among children who do not exhibit behavioral problems. Their assumption of a quadratic trend in the growth of test scores however seems relatively arbitrary, and the identification of marriage effects relies on a change in status during the window that the children are assessed. Potentially only a small fraction of the total sample might experience such a change.

The work of Gennetian (2005), Aughinbaugh et. al. (2005), and Wagmiller et. al. (2010) suffers from shortcomings associated with the fixed effects models. Often the subsample of children where transitions between marital states occur while the children are tested is small. Additionally, because fixed effects models depend on observing the same outcome over multiple periods, only the impact on educational outcomes that change over time can be observed. The impact of family structure on whether a child is a high school dropout, for example, cannot be estimated as the dropout status is a time-constant binary variable. Finally, and perhaps most importantly, fixed effects models assume that marital status can be correlated with time-constant unobservables affecting outcomes but is uncorrelated with unobservables that change over time. This assumption is strong. For instance, a change in geographic location and marital status can occur simultaneously, and both can impact a child's outcomes. In light of these limitations, other researchers have pursued an instrumental variable strategy. With access to variables that are correlated with family structure but independent of any unobservables that jointly influence child outcomes and family structures, the fixed effects limitations can be avoided. Unfortunately, plausible instrumental variables are difficult to find.

Bjorklund et al. (2007) use a change in pension benefits for spouses as an instrument for the marriage decision. They use Swedish data to estimate the impact of a parent's transition to marriage on the child's school GPA. An administrative change in the pension system in Sweden abolished a generous lifetime pension for widows and replaced it with a limited pension. Older cohorts of women, however, were still eligible to qualify for the lifetime pension if they married prior to the end of 1989. The reform thus made marriage financially beneficial to some unmarried couples. The authors observe a spike in marriages in December of 1989, but only some of the women who got married in that month were

actually eligible for the financial benefits from marriage.[17] Their sample is restricted to children born between 1977 and 1987, and who lived with both biological parents, whether cohabiting or married. The children whose parents were cohabiting and eligible for the financial benefit from marriage did not experience any changes in their GPA due to the parents' transition to marriage.[18] Bjorklund et al. (2007) results suggest that marriage for couples that already chose to cohabit and raise children together is unlikely to have any effect on the children. The validity of the instrument might be undermined if other factors that are correlated with child outcomes influenced the spike in marriages at the end of 1989. After all, couples not directly affected by the pension reform also decided to get married at much higher rates than usual.

Finlay and Neumark (2010) use Decennial Census data to test the relationship between never married motherhood and the dropout decision.[19] Their identification strategy depends on variation in incarceration rates across states, time, age, and race as an instrument for the never married status of the mother. Higher incarceration rates decrease the relative supply of partners which, in turn, impacts the marriage decision. Because the Bureau of Justice Statistics, which tracks the prison population in the United States, does not publish data disaggregated by state and race, they use Census data to construct institutionalization rates. Institutionalized individuals are meant to proxy for individuals that are incarcerated. The sample is restricted to children living with their mothers at the time of the interview. Using the institutionalized rate from the previous decade to instrument the mother's marital status, the authors find that never married motherhood decreases the probability of dropping out for Hispanics, but has no effect for blacks. The results of Finlay and Neumark (2010) suggest that any transitions into marriage for Hispanic women who give birth prior to marriage is disruptive to the children born out of wedlock.

Finlay and Neumark (2010) use the Decennial Census data and construct institutional-

[17] Many younger women also married, even though they would not qualify for the widow's pension in the event of their husband's death.

[18] Interestingly, the effect on the GPA of boys from parental marriage in the sample that was not eligible for financial benefits was positive. However, these effects are not robust to the inclusion of birth order and presence of half-siblings in the regressions.

[19] Dropouts are defined as children who are not enrolled in school and have not completed grade 12.

ized rates once every 10 years. The lack of more frequent data sources limits their ability to describe marriage markets in intercensal years and thus limits their ability to use information on marriage timing and children's outcomes available in older longitudinal data sets. In this chapter, I extend the identification idea used in Finlay and Neumark (2010). I use variation in the relative availability of spouses to instrument for the marriage decision, and I overcome the limitation of having infrequent data on marriage markets.[20] In Chapter 1, I combine multiple sources of demographic data and generate instrumental variables for all years, not just for every 10 year interval. With more frequent instruments, I use the NLSY79 and the CNLSY79 samples to estimate the effect of mother's marriage on child outcomes.

2.3 Data

I use the National Longitudinal Survey of Youth 1979 (NLSY79) and the Child and Young Adult supplement (CNLSY79). The NLSY79 started as a nationally representative sample of men and women. Respondents were between the ages of 14 to 22 when they were first interviewed in 1979. The surveys were conducted on a yearly basis between 1979 and 1994, and once every two years thereafter. Starting in 1986, the NLSY79 Child and Young Adult supplement began collecting data on the children born to the NLSY79 mothers. Mothers were asked about their young children, while children older than 10 years old were interviewed separately once every two years. In 1994, children 15 years and older were asked to participate in interviews which resembled the interviews of the main NLSY79 respondents. Over time as the children mature, similar information about the education, marital, and employment history that is available about the mothers, will become available for their children. The NLSY79 contains identifiers allowing researchers to connect mothers with their children.

[20] Incarceration rates rely on variation in criminal behavior and punishment across demographic categories. Other measures of partner availability, sex ratios of unmarried individuals for example, rely on additional forces generating variation in partner availability, like differences in mortality rates.

2.3.1 Sample Selection

I start with women in the NLSY79 that reported having children.[21] The initial sample consists of 4,043 mothers. I exclude observations that were incomplete or unsuitable for the analysis. Table 2.1 illustrates the sample restrictions and number of observations dropped. The first set of sample selection rules in Table 2.1 relates to mothers. I exclude respondents that were reported as deceased. Respondents that were incarcerated during the study period or served in the military are excluded. Both groups are more likely to deviate from the general population in the ways that they participate in marriage markets. Likewise, individuals marked by the interviewers as deaf, blind, mentally handicapped, or physically handicapped are excluded. Respondents who reported a marriage but for whom the date of first marriage was unavailable are dropped. NLSY79 respondents were administered an Armed Forces Qualification Test (AFQT). The test is often used as a measure of intelligence. Observations for whom AFQT score could not be determined are dropped. Out of the 3,586 women remaining, only 1,122 chose to become mothers prior to any marriage. Timing of first childbirth and marriage varies considerably between races. As illustrated in Table 2.2, motherhood before marriage is an exception for whites and Hispanics. Only 12 percent of white mothers, and 27 percent of Hispanic mothers have a child before first marriage. However, birth prior to first marriage is by far the most common choice of family formation for black women. Two-thirds of black mothers in the NLSY79 data choose to have a child prior to marriage. Because of such large differences in the distribution of unmarried motherhood between races, I restrict the sample to black mothers. I also exclude mothers who reported a first marriage within 12 months of the first birth. These mothers were likely to have stable relationships at the time of the child's birth and were unlikely to participate in marriage markets.

After the restrictions on mothers, the sample consists of 657 unwed mothers and 1,736 children. I exclude children that were designated by the interviewers as deaf, blind, physically, or mentally handicapped. Children born after the date of first marriage are excluded.

[21] The poor white and military subsamples were discontinued from the NLSY79, and I exclude those as well.

Out of the 1,307 children remaining, 978 report living with their mother between the ages of 14 or 15.[22] It is unlikely that mother's marital behavior has an impact on a child who does not live with her. I also exclude observations because of inconsistencies in the data on child outcome variables or geographic location.[23] The instrumental variable approach relies on access to geographic location for the first 18 years of the child's life. Out of 906 children remaining, only 602 have geographic location every year for the first 18 years.

2.3.2 Sample Description

Summary statistics for the mothers in the sample are presented in Table 2.3. Unwed mothers entered motherhood at a relatively young age; the average age at first birth is nearly 20. They had on average 2.8 children in their lifetime, 73 percent reported having graduated from high school, and 60 percent eventually married. Descriptive child statistics are presented in Table 2.4. Around 50 percent of the children are the first-born child, implying many of the mothers continued to have children out of wedlock after their first child. Half of the children experienced their mother marrying at least once prior to their 18th birthday. The majority of the children, nearly 78 percent of the sample, reported having a high school diploma or a GED by the age of 20. Only 68 percent grew up with a mother that was a high school graduate suggesting educational attainment has increased over the generations. The NLSY79 Child and Young Adult Survey also asks respondents if they have ever been convicted of any crime. While this self-reported variable is likely to suffer from measurement error, nearly 21 percent of the sample reported a conviction, suggesting crime might be commonplace in their surroundings and is likely to have an impact on child outcomes as well.

Omitted variables are often a concern when using OLS to estimate the effect of family structure on child outcomes. Other variables excluded from the analysis might be correlated with child outcomes and family structure. For instance, more educated mothers might be

[22] Young children who were less than 14 years old at the time of their last interview are also excluded.

[23] I exclude observations if the data contains conflicting reports of educational attainment. I also exclude observations for which the data on geographic location were problematic. For instance, respondents that lived outside of the US are excluded.

more likely to have more educated children and be more likely to marry, or preferences for the number of children might be correlated with preferences for the quality of children and preferences for marriage. Further analysis of the mothers in the final sample reveals variables that are correlated with family structure. Mothers' marital status appears to be related to high school graduation status and lifetime fertility. Mothers who graduated from high school are more likely to marry. Table 2.5 shows that only 35 percent of women with a high school diploma remained unmarried, while 54 percent of women who did not graduate from high school never married. High school status is also associated with a lower lifetime fertility. Figure 2.1 illustrates that 70 percent of women with no high school diploma reported three or more children, while only 48 percent of mothers with a high school diploma reported three or more children. However, marriage appears to be correlated with additional fertility. Table 2.6 includes summary statistics for the married subsample. On average, the married subsample had two children before marriage and one child after the marriage. Figure 2.2 presents further evidence that marriage is positively correlated with lifetime fertility. Only 45 percent of women who never marry had three or more children, while 60 percent of women that report a marital transition had three or more children.

Timing of the first birth also appears to be correlated with marriage, lifetime fertility, and mother's education status. Figure 2.3 illustrates the distribution of the mothers' age at first birth by eventual marital status. Relative to the married women, the women who remained unmarried were more likely to give birth at age 20 or later. Women who entered motherhood at older ages are also likely to have fewer children. As illustrated in Table 2.7, the share of women that become mothers before age 20 increases with the number of children. Only 40 percent of women with 1 child gave birth prior to age 20, while nearly 75 percent of women with 4 children gave birth prior to age 20. Additionally, as illustrated in Figure 2.4, lower educated women were more likely to have their first child at a younger age.

2.4 Definition of the Family Structure Variable

I define the family structure variable as an indicator equal to one if the mother transitions to first marriage. The excluded category is therefore a never married mother. The division of children into the mother never married and mother married categories might mask some heterogeneity within the two groups. For example, the never married category includes children that live with mothers who have no stable relationship with the children that live in households with two parents cohabiting. Transition to first marriage might be followed by a long and stable relationship of the couple, or it could be the first of many volatile changes in family structure.[24] Both of these groups will be classified under the same married category. While imperfect, using the measure of first transition to marriage is both informative and convenient. The effect of entry into first marriage should be of interest to policy makers because promotion of marriage has been suggested as a policy tool for poverty reduction.[25] Households headed by single mothers are especially prone to fall into poverty. According to the Census Bureau, in 2011, 31 percent of households headed by a female without a spouse present were living below the poverty level, relative to 6 percent for married couples. Additionally, nearly 50 percent of children in households headed by a single mother were below the poverty level. Marriage promotion policies will impact these children and their effects should be considered prior to policy implementation.

Using transition to first marriage as the relevant family structure is also convenient. Unlike information on cohabitation or the level of all the various resources invested in children, marital status is more readily available in most data sources. NLSY79 contains a detailed marital history throughout the sample period, while only some cohabitation information is available. Prior to 2002 women were asked about cohabitation only if they chose to report a partner on the household enumeration sheet. If they did report a partner, information on the date cohabitation began with the current partner was asked starting in 1990. Thus, while some cohabitation information is available, it is incomplete and a full

[24] For instance, 16 percent of the NLSY79 mothers who reported a first marriage in my sample also reported a second marriage.

[25] See Sigle-Rushton and McLanahan (2002).

cohabitation history cannot be reconstructed.

2.5 Model

The relationship between an outcome Y_i for each child (i), an indicator for the mother's marital status M_i, and a set of other explanatory variables X_i can be expressed using the following linear specification

$$Y_i = X_i'\beta + \alpha M_i + u_i \tag{2.1}$$

where u_i represents the unobservables that impact the outcomes. Given the specification in equation (2.1), the goal is to obtain an estimate for the coefficient α. Ordinary Least Squares (OLS) estimation will not generate a consistent estimate of the parameters because the mother's marriage decision M_i is likely to be correlated with the unobservables in the outcome equation u_i. However, with access to instrumental variables Z_i such that the following conditions hold,

$$\begin{aligned} E(X_i'u_i) &= 0 \\ E(Z_i'u_i) &= 0 \\ Cov(Z_i'M_i) &\neq 0, \end{aligned} \tag{2.2}$$

α and β can be identified and estimated. A scalar Z_i is sufficient for identification and estimation because M_i is also a scalar. Whenever the number of moment conditions exceeds the number of parameters to be estimated, the model is overidentified. For instance, if Z_i is a vector of two variables, $Z_i = Z_{i,1},\ Z_{i,2}$, then either $E(Z_{i,1}u_i) = 0$ or $E(Z_{i,2}u_i) = 0$, in addition to $E(X_i'u_i) = 0$, would suffice to generate a consistent estimator for α and β. However, a GMM estimator that combines information from all the available instruments can derived. A consistent estimator is obtained as the result of the following minimization problem

$$min_c\ Q(c) = min_c\ (O'u(c))'W(O'u(c)), \tag{2.3}$$

where $c = \{a, b\}$, $u(c) = Y - X'b - aM$, O is a matrix of all the exogenous variables (X, Z), and W is any positive definite matrix.[26]

2.6 OLS Results

The NLSY79 and CNLSY79 surveys record a variety of child outcomes. I start by using OLS to investigate the relationship between mother's marriage and three outcome variables: the child's PIAT math scores, an indicator for whether the child obtained a high school diploma or a GED, and an indicator of whether the child became an out of wedlock parent. The coefficient on mother's marital status is significant only in the child's high school graduation status equation. The coefficient on mother's marriage in the child's PIAT Math test score equation, or whether the child became an unwed parent equation, was not significant. The results of the high school graduation outcome are discussed below.

The outcome variable is an indicator equal to one if the child graduated from high school or obtained a GED by age 20. The marriage indicator is equal to one if the child's mother reported a transition to first marriage before the child was 19 years old. The OLS regression results for equation

$$Y_i = X_i'\beta + \alpha M_i + u_i$$

are summarized in Table 2.8. Each column corresponds to a separate regression with a different set of explanatory variables X_i. The coefficient estimate in the first row and column shows mother's marriage at or before the child's 18th birthday is positively correlated with the child's educational attainment. Relative to children whose mothers remain unmarried, children whose mother marries are on average 14 percentage points more likely to have a high school diploma or a GED by the age of 20. However, the size of the mother's marriage coefficient declines to 7 percentage points, as illustrated by the results in the last column, when other explanatory variables are added to the regression. The 50 percent drop in the magnitude of the coefficient reveals the importance of other factors that are both correlated

[26] An efficient GMM estimator is obtained if W is equal to the inverse of the asymptotic variance covariance matrix of the moment equations. Greene (2003).

with mother's marriage and children's educational attainment. For example, the sample analysis in Section 2.3.2 revealed that mother's high school graduation status and lifetime fertility are both correlated with her marital status. The inclusion of these variables in the child's outcome equation accounts for more than half, nearly 4 percentage points, of the drop in the size of the marriage coefficient. It is likely that variables not included in X_i in Table 2.8 are also correlated with child outcomes and mother's marriage.[27] Although the covariates included in Table 2.8 might be just a fraction of the relevant variables, it is still instructive to explore how they are correlated with child outcomes and mother's marriage.

Educational attainment appears to be strongly correlated through the generations. Children whose mother graduated from high school are on average 14 percentage points more likely to have a high school diploma or a GED by the time they are 20 years old. As discussed in Section 2.3.2, women with a high school diploma are also more likely to marry and have fewer children. Likewise, children born to older mothers, who are less likely to marry and have fewer children, are more likely to graduate from high school as well. First-born children are on average 7 percent more likely to graduate from high school. The magnitude of the first-born coefficient is comparable to the magnitude of the mother's marriage coefficient. First-born children might enjoy the benefits of being the only child for at least a fraction of their lives, whereas any subsequent siblings do not get their mother's undivided attention. The first-born indicator might also pick up the differences between women who have only one child out of wedlock and the ones with multiple children.[28] However, this seems unlikely as the size of the first-born coefficient is still positive and quite large even after controlling for the number of children the mother had before any marital transition. Higher fertility of the mother before the date of first marriage is negatively correlated with educational attainment, while the number of children after the date of first marriage is positively correlated with child outcomes. The latter might suggest that marriages which result in children are more likely to be stable and have a positive effect on the child born

[27]For example, if a mother spends a lot of her time with the child she is less likely to participate in marriage markets and more likely to have a high preforming child.

[28]These women might on average have a higher preference for "high quality" children and so they invest more in each child and have fewer children.

out of wedlock. An indicator if the child was ever convicted of a crime has a strong negative correlation with high school graduation. The conviction status is self-reported by the CNLSY79 respondents and is likely to be measured with error. The magnitude of the coefficient is very large and comparable in size with the mother's high school graduation status. Including the conviction status in the regressions decreases the negative correlation between boys and graduation, suggesting that part of lower educational attainment for men is due to higher propensity for criminal behavior. These correlations are robust to the inclusion of state fixed effects.[29]

The child's sex, mom's high school graduation status, the first-born indicator, and child's conviction status are strongly correlated with the child's educational attainment. I interact each one with the marriage variable and repeat the OLS regressions to explore the possibility of heterogeneous impact of mother's marriage. As illustrated in Table 2.9, only the interaction with first-born appears to be statistically significant; the magnitude of the marriage coefficient nearly doubles to 13.4 percent. Subsequent children born out of wedlock whose mothers marry are on average more likely to graduate from high school relative to children whose mothers remain unmarried. The positive association with marriage dissipates for first-born children. They are no more likely to graduate from high school if their mother marries relative to first-born children whose mothers remain unmarried.[30] I investigate the heterogeneous effects further and I repeat the OLS regression with outcome variable defined as high school graduation by different age of the child for ages 19 to 22. Table 2.10 illustrates that correlations between first-born, marriage, and educational attainment are less pronounced for other ages although the general pattern still holds, mother's marriage is far less advantageous for first-born children relative to their subsequent siblings.

Outcomes might also correlate with the age of the child at mother's marriage. In addition, age of the child at mother's marriage could explain the possible heterogeneity in marriage effects by birth order. First-born children are more likely to be older than their subsequent siblings, and if any marriage occurs it might be more disruptive to older chil-

[29]State of residence when the child was 18, the standard age for high school graduation, is used.

[30]The sum of the marriage effect is slightly negative, $0.134 - 0.141 = -0.007$. First born children are still observed to have a 15 percentage points higher chance of high school graduation.

dren outweighing any positive effects. I define two marital event indicators, one where the mother married before the child turns 10, and another if she married when the child was older than 10. Results in Table 2.11 indicate that marriage of the mother at younger child ages has a positive association with the eventual high school graduation status. Children with mothers that marry later appear to be no more likely to graduate than children of mothers who do not marry at all. The positive association between first-born and educational attainment persists. In Table 2.12, I present OLS regression results when the two marital indicators are interacted with the first-born indicator. The pattern of differences in child outcomes by marriage and birth order remain, mother's marriage when the child is young has a positive association with educational attainment but the effect is much lower for first-born children. Marriage when the child is older does not appear to have much effect on the child, regardless of birth order.

The OLS results are informative. There are considerable differences in high school graduation status between children whose mothers marry and the ones whose mothers remain unmarried. These differences diminish for children who are first-born in their family and who are older when the marriage occurs. Large changes in the magnitude of the marriage coefficient when more variables are included in X also suggests a strong possibility for correlation between mother's marriage and unobservables affecting the child's educational attainment. To obtain a consistent estimate for the α coefficient in equation (2.1), I rely on an instrumental variable approach discussed below. First, the instrumental variable is defined and explained. Second, I discuss the appropriate specification for estimation given the OLS results. Last, I present the estimation results and robustness checks.

2.7 Instrumental Variables: Definition and Estimation Results

2.7.1 Sex Ratios and Marriage

I use sex ratios, the ratio of unmarried men over unmarried women, to instrument for the unwed mother's marriage decision. When men are relatively scarce in a given marriage

market, women have a harder time finding an adequate husband. Many researchers present empirical evidence for the presence of these search frictions. Freiden (1974), South and Lloyd (1992), and Wood (1993) find that higher relative availability of partners in a given geographic location in the U.S. increases the share of married women.[31] Angrist (2002) finds that changes in the U.S. immigration policy decreased the supply of immigrant spouses for some second generation immigrant groups and decreased the probability of marriage. Abramitzky et al. (2011) find that women in regions of France which experienced higher male mortality in World War I battles, and thus a lower relative supply of partners, were less likely to marry. Lichter et al. (1992), and Brien (1997) find that an increase in the relative availability of partners reduces the waiting time until first marriage for women in the U.S.. Likewise, in Chapter 1 of this dissertation, I also find that an increase in the relative availability of partners reduces the waiting time to first marriage for NLSY79 women.

Edin and Kefalas (2005) explore the marriage decision of the unwed mothers specifically. Based on a set of interviews with 162 unwed mothers in the city of Philadelphia, the authors provide qualitative information on the aspirations and obstacles to marriage for this specific population. Most importantly, the vast majority of unwed mothers find marriage desirable. In their sample 70 percent of all women, and 77 percent of the black women, wanted to eventually marry. However, many had difficulty finding a suitable partner for marriage. Oftentimes the potential partners available had substance abuse issues and problems keeping steady employment. Based on the interviews, Edin and Kefalas (2005) conclude that faced with a poor supply of marriageable men and unable to discern which men are of higher quality, unwed mothers often decide to wait and let the trials of the relationship act as a screening mechanism for the quality of the men. Because marriage is desirable but finding adequate partners can be difficult, unwed mothers are also likely to respond to changes in the relative supply of unmarried men.[32]

Sex ratios impact the marriage decision and it is instructive to consider why sex ratios vary. Besides immigration policies and the tragedy of war, differences in sex ratios can

[31] State and Standard Metropolitan Statistical Area (SMSA) are used.

[32] I also find evidence that the marriage decision for the women in my sample is correlated with partner availability measures. See Section 2.7.4.

emerge due to a host of factors. Different mortality and incarceration rates for men and women across demographic categories can generate different sex ratios. Sex ratios can differ because of differential migration between the sexes across geographic units. Men might migrate to a certain state in much higher proportion due to the presence of industries that disproportionately hire men. The impact of any event generating an inequality between unmarried men and women will impact the sex ratios differentially depending on the size of the group considered. For example, consider a group of 100 men and 100 women. If 10 men move to a different geographic location, the sex ratio is 0.9. However, if there are a 1000 men and 1000 women and if 10 men move the sex ratio is 0.99. Random variation in incarceration, mortality, and migration decisions could generate substantial variation in sex ratios, especially for small groups. In addition, initial imbalances between the sexes are exacerbated as absolute numbers of unmarried individuals declines with age due to marriage. Consider a market where initially there are 900 unmarried men and 1000 unmarried women. The initial sex ratio is $\frac{900}{1000} = .9$. Subsequently, over time 500 men and women marry, leaving 400 unmarried men and 500 unmarried women for a sex ratio of $\frac{400}{500} = .8$. As cohorts age and many of its members exit, the marriage markets sex ratios change.

2.7.2 Instrumental Variable Definition

Sex ratios are defined to capture the variation in relative availability of potential partners in marriage markets. For each mother observed in the NLSY79 at time (t) the marriage market is defined by her state of residence (s), race (b), and age (a). I follow Lichter et al. (1992) and assume that a woman of age (a), race (b), living in state (s) at time (t), considers unmarried men between the ages (a) and $(a + 10)$ as potential partners and unmarried women between the ages $(a-2)$ and $(a+8)$ as potential competitors. Let $\hat{U}^{M}_{s,b,a,t,d}$ and $\hat{U}^{F}_{s,b,a,t,d}$ represent the observed number of unmarried males (M) and females (F) from data source (d), at time (t), for age (a), state (s), and race (b). Using data source (d) the

sex ratio at time (t) is defined as

$$\hat{r}^{d}_{s,b,a,t} = \frac{\Sigma_{k=0}^{10} \hat{U}^{M}_{s,b,a+k,t,d}}{\Sigma_{k=-2}^{8} \hat{U}^{F}_{s,b,a+k,t,d}}. \tag{2.4}$$

Much of the marriage market activity for NLSY79 women occurred in the 1980's and the 1990's. The Current Population Survey (CPS) and the Decennial Census are the only two data sources available from which to construct sex ratios by marriage market for those two decades. The Census provides the most accurate estimates of population counts by demographic category but is only available once every 10 years. On the other hand, the CPS is available every year but, in part due to its small sample size, its population estimates are likely to deviate significantly from the true underlying population counts. In addition, the CPS is also most likely not to sample certain categories of individuals. For instance, 20 year old black females living in Indiana might not be sampled in 1992. Because of measurement error in the observed population counts and the propensity not to sample certain groups, the CPS observed sex ratios ($\hat{r}^{d}_{s,b,a,t}$) can deviate considerably from the true sex ratios. In turn, large measurement error in the observed sex ratios is likely to generate attenuation bias in the estimated effect of sex ratios on the women's marriage decision.[33] Given the shortcomings of each data source, researchers often decide to use the accurate but infrequent Decennial Census data.

I develop and estimate a model generating the observed population counts, $\hat{U}^{M}_{s,b,a,t,d}$ and $\hat{U}^{F}_{s,b,a,t,d}$, to address the scarcity in accurate and frequent demographic data during a large portion of the NLSY79 time frame. The model accounts for unobserved data cells and the presence of data source specific measurement error. I combine demographic data from the Decennial Census, the CPS, and the American Community Survey (ACS) to estimate the model parameters. Using the parameter estimates I construct an expectation of the sex ratio $E(\hat{r}^{d}_{s,b,a,t})$ for every state (s), race (b), age (a), and time period (t). I find that expectations of sex ratios based on the model generate a higher predicted effect on the waiting time to first marriage relative to coefficients when only the observed data ratios are

[33] For example see results of Chapter 1 or Brien (1997).

used. The larger coefficient estimates are most likely the result of smaller deviations from the true sex ratios relative to the CPS generated ratios.[34]

The expected sex ratios are a source of instrumental variables for the mother's marriage decision. I connect the mothers and their children to a set of marriage markets, and in turn to the expected sex ratios, using restricted access geographic information in the NLSY79 survey data. The model generates ratios for every year of the NLSY79 survey, so an instrumental variable is available for each year the child is in the sample. Because the marriage indicator is defined from the child's perspective, the instruments for each child are also defined based on the age perspective of the child. In other words, two children born at the same time (t) in the same state (s) will have different ratios if their mothers were of different age. Thus, for each child (i), of age (A), at time (t), I have $Z_{i,1}, Z_{i,2}, \ldots, Z_{i,A}$, where $Z_{i,A}$ is the expected sex ratio experienced by the mother who is (a) years old when the child is of age (A).[35]

2.7.3 Choosing the Appropriate Empirical Specification

The instrumental variable approach depends on the moment conditions

$$\begin{aligned} E(X_i' u_i) &= 0 \\ E(Z_i' u_i) &= 0 \\ Cov(Z_i' M_i) &\neq 0, \end{aligned} \tag{2.5}$$

where $u_i = Y_i - X_i'\beta - \alpha M_i$. X_i includes a constant, the child's gender, and the log of the mother's AFQT score. The moment conditions require that after "subtracting out" the impact on outcomes from covariates in X_i and mother's marriage M_i, other determinants of outcomes not controlled for explicitly are uncorrelated with sex ratios. The OLS results discussed in Section 6 identify variables observed in the CNLSY79 that have an impact on high school graduation but are unlikely to be exogenous so they cannot be used as

[34] For a further discussion of the model or the predicted effects on waiting time until first marriage see Chapter 1.

[35] For example if the child is 5 when the mother is 27, I use the ratios for 27 year old women when the child is 5, 28 year old women when the child is 6, and so on.

instruments and are left as unobservables. If these variables are correlated with Z_i, the moment conditions no longer hold and parameter identification fails.[36] Fortunately, because I generate 18 instruments for each child in my sample, the number of moment conditions exceeds the number of endogenous parameters. The surplus of instruments allows me to explicitly include the endogenous variables in the outcome equation and instrument them with the sex ratios. This approach in effect redefines the unobservable for each individual to $u_i = Y_i - X_i'\beta - \alpha M_i - V_i'\gamma$, where V_i are other observed endogenous variables correlated with sex ratios.

I include the first-born indicator, mother's age at the birth of the child, and the number of children a mother had out of wedlock as additional endogenous explanatory variables in the high school graduation equation. OLS regression results in Table 2.8 indicate that each one of these is correlated with the child's high school graduation status, and it is possible that each one is correlated with sex ratios. A relatively low supply of men might change the willingness of women to bear children out of wedlock, even after conditioning on the marriage decision. The first-born indicator and mother's age at the child's birth are correlated with the sex ratios because of more "mechanical" reasons. Sex ratios decrease as mothers age. Differences in mortality and imprisonment rates between the sexes contribute to this trend. In addition, any initial imbalances between the sexes are exacerbated as absolute numbers of unmarried individuals decline with age due to marriage.[37] First-born children are likely to be born earlier than subsequent children and as a result will have higher sex ratios. Table 2.14 illustrates that for each of the 18 years the sex ratios for first-born children tend to be higher relative to subsequent children. Likewise, children born to older women will have lower sex ratios compared to children born to younger women. Table 2.15 illustrates the result of regressing mother's age at child's birth on sex ratios. Results in the first column, where mother's age at birth is regressed on sex ratio at child's age one, show that younger mothers experience higher sex ratios. The regression in the second column

[36] Parameter estimates will not be consistent if any unobservables are correlated with sex ratios, not just the variables that are observed in the CNLSY79 but not explicitly included in the outcome equation.

[37] Consider one market where initially there are 900 unmarried men and 1000 unmarried women. The initial sex ratio is $\frac{900}{1000} = .9$. Subsequently over time 500 men and women marry, leaving 400 unmarried men and 500 unmarried women for a sex ratio of $\frac{400}{500} = .8$.

includes sex ratios for all 18 years. An F test rejects the null hypothesis that sex ratios are uncorrelated with mother's age at birth.

2.7.4 "First-stage" Results

In research that utilizes instrumental variables, besides exogeneity of the instruments, there is a concern about weak instruments. Stock et al. (2002) point out that the asymptotic distribution for the estimates is a poor approximation to the sampling distribution, if the instruments and the endogenous variable are weakly correlated. They develop tests for weak instruments and compute thresholds for the size of the F statistic in the first-stage regression when the model is linear with homoskedastic error.[38] However, according to Stock et al. (2002), no such test is available when the errors are heteroskedastic, as is the case when the dependent variable is an indicator for high school graduation status. Nevertheless, I report the results of regressing the mother's marriage indicator on sex ratios and other explanatory variables in Table 2.13. While not a formal test for weak instruments, the regressions allow for testing of correlation between the instruments and mother's marriage indicator conditional on other variables. Column 1 contains the result when only the sex ratios and a constant are included in the regression, Column 2 contains results when all the independent variables are included, and Column 3 contains the results when the three additional endogenous variables discussed above are also included in the regression. While an F test on the null hypothesis that the sex ratio coefficients are jointly equal to zero is rejected in all three regressions, the F statistic in all three regression is below the rule-of-thumb value of 10. These results suggest that sex ratios are correlated with the mother's marriage decision and make plausible instruments, but the 18 ratios could be a set of weak instruments.

[38]The threshold size of the F statistic depends on the number of instruments and the test used. The general rule-of-thumb is for the F statistic to be larger than 10. Stock et al. also provide a test statistic for weak instruments when the errors are i.i.d. and multiple endogenous variables are present.

2.7.5 Instrumental Variable Results

The outcome equation is modified to

$$Y_i = X_i'\beta + \alpha M_i + V_i'\gamma + u_i \tag{2.6}$$

where X_i includes a constant, the child's gender, and the log of mother's AFQT score. M_i is an indicator if the mother married prior to the child's 18th birthday, and V_i are the additional three endogenous explanatory variables, the first-born indicator, age of the mother at birth of the child, and number of kids the mother had out of wedlock. Y_i is an indicator variable equal to one if the child graduated from high school or obtained a GED by a certain age. Sex ratios for all 18 years and the variables in X_i are used as the instrumental variables. Because the number of instruments exceeds the number of endogenous parameters I use GMM estimation described in Section 5.[39] The results are summarized in Table 2.16. Each column in the table corresponds to an estimation on the child's educational attainment measured at a different age. For example, the estimate of the coefficient on mother's marriage in column 2 indicates that children whose mothers remain unmarried are nearly 24 percentage points less likely to obtain a high school diploma by age 20. Overall the results in Table 2.16 indicate that the impact of mother's marriage on high school graduation is positive, very large, and mostly statistically significant.[40]

Robustness Checks

The GMM estimate of the effect of mother's marriage on the child's high school graduation probability by age 20 is three times larger than the OLS coefficient. The magnitude of the difference in the estimates can generate concerns that unobservables might still bias the estimates. Unobserved factors that generate differences in sex ratios, like differences in

[39] I use a two-step estimation procedure in STATA. In the first step, consistent estimates of α, β, γ, are obtained using inverse of the product of instrument matrix as the weights, $W = (O'O)^{-1}$. In the second step α, β, γ are reestimated for a weight matrix equal to the inverse of the estimated moment covariance matrix. The moment covariance matrix estimate is obtained by using the first step estimates of α, β, γ.

[40] I also use the GMM approach to estimate the impact of mother's marriage on PIAT math scores and on an indicator if the child became an unwed parent. Similarly to OLS results, the coefficient estimates for mother's marriage were not significant.

incarceration rates, mortality rates for men, or economic opportunities, might also impact the educational outcomes of children. Table 2.17 summarizes the results of estimating a specification similar to equation (2.6) with the inclusion of state fixed effects when the child was 18 in X_i. The inclusion of state fixed effects controls for any unobservables specific to a marriage market that affect both the sex ratios and high school graduation. The estimated coefficient on marriage, especially for high school graduation by age 20 and older, remains large. The standard errors, however, increase substantially.

State fixed effects are meant to capture unobservables that might affect children's outcomes and sex ratios. Differences in economic opportunities across states could be one such factor. High unemployment rates might cause poor child outcomes and differential migration between state lines in search of better labor markets. I estimate a specification similar to equation (2.6) and include unemployment rates for the child's state of residence for all 18 years in X_i. The results are summarized in Table 2.18. The marriage coefficients remain large and significant.

Differences in incarceration rates between states are also likely to contribute to differences in sex ratios.[41] Higher incarceration rates could indicate that in some marriage markets crime is more prevalent which might impact high school graduation. I reestimate equation (2.6) but also include prison rates for the child's state of residence for all 18 years in X_i.[42] The results, summarized in Table 2.19, show a large decrease in the estimated marriage coefficient. For instance, in the second column the coefficient on high school graduation by age 20 declines from 24 percent to 9 percent. The coefficient estimates for marriage are also no longer significant. It appears that forces causing differences in prison rates also impact mother's sex ratios and child's high school graduation status.

The results of this section suggest caution when considering the GMM coefficient estimates. The exogeneity of the instrumental variables is undermined by the sensitivity of coefficient estimates when marriage-market-specific "environmental variables" are included in the outcome equations.[43] Forces that generate differences in sex ratios for the mother

[41]The sex ratio is based on the mother's marriage markets, so if the mother is 38 when the child is 18, based on the definition of sex ratios the incarceration of men that are between 38 and 48 matters.

[42]Data on prison rates and unemployment rates by state was obtained from John J. Donohue.

[43]The coefficient estimates are also inconsistent with OLS results when marriage is interacted with birth

might also influence children's educational attainment. In addition, sex ratios could potentially be a weak set of instruments. As discussed in Section 2.7.4, the ratios are correlated with the mother's decision, but the correlation could be weak, leading to unreliable point estimates.[44]

2.8 Conclusion

I use the National Longitudinal Survey of Youth 1979 (NLSY79) and the Child and Young Adult supplement (CNLSY79) to study the effects of mother's marriage on black children born out of wedlock. Ordinary least squares (OLS) regressions reveal that mother's marriage is positively correlated with high school graduation status of the children. However, the possibility that unobserved factors can be correlated with child outcomes and mother's marriage remains. To address the likely endogeneity of mother's marriage decision I use sex ratios in the mother's marriage markets as instrumental variables. Previous research has found that sex ratios, the relative availability of potential spouses, impacts women's transitions into marriage. Higher relative availability of spouses leads to shorter waiting times until first marriage. Using estimation results of Chapter 1, I construct yearly measures of sex ratios for the period the women in the NLSY79 are observed. The General Method of Moments (GMM) estimation framework allows me to use information from multiple measures of sex ratios to estimate the effect of mother's marriage. I find that mother's marriage has a large and positive effect on the high school graduation status of their children, although the results seem sensitive to common concerns associated with instrumental variables. The exogeneity of the instrumental variables is undermined by the sensitivity of coefficient estimates when other marriage-market-specific variables are included in the outcome equations. When state-specific prison rates are included, the size of the marriage coefficients decrease and becomes insignificant. In addition, while the sex ratios are correlated with the mother's marriage decision, the correlation might be insufficient to overcome the weak instruments problem.

order of the child or the age of the child. For more details see Appendix 2.A.

[44] See Stock et al. (2002).

2.A Appendix: Instrumental Variables and Heterogeneity in the Marriage Effect

OLS regression results of high school graduation on a variety of explanatory variables suggest that outcomes for children differ when mothers marry, and depend on the birth order of the child and age at which marriage occurs.[45] First-born children are no more likely to graduate from high school when their mother marries, however subsequent children have higher probability of graduating high school if their mothers marry. Heterogeneity in birth order is not observed when first-born is interacted with mother's marriage in the instrumental variables approach. The GMM coefficient estimates for the marriage indicator, and the first-born and marriage interaction term are insignificant. OLS results in Table 2.11 also indicate that younger children have a higher probability of high school graduation when their mothers marry relative to children whose mothers remain unmarried. I reestimate equation (2.6) but allow for two marriage indicators, if the mother married before the child was 10 or if the child was between 10 and 18 years old, and instrument with sex ratios. The GMM estimation results in Table 2.20 are inconsistent with the OLS results, mother's marriage when the child is older has a positive, large, and statistically significant effect on child's high school graduation status. Coefficients on marriage at younger ages are positive but insignificant. When state fixed effects are added to the independent variables, the size of the marriage coefficient at older ages increases, in many cases above one, and becomes insignificant.[46] The coefficients remain large and significant when unemployment rates are added to the regressions, but decline in significance and magnitude when prison rates are included.[47] Inclusion of state fixed effects and prison rates reduce the significance of the marriage coefficient at older ages, suggesting that unobservable factors in mother's marriage markets might be correlated with sex ratios and child outcomes. GMM coefficient estimates in Table 2.24, for a specification that includes both state fixed effects and prison rates, support the possibility that sex ratios might be correlated with other factors that are

[45] See Tables 2.10 and 2.11.
[46] See Table 2.21.
[47] See Tables 2.22 and 2.23.

correlated with child outcomes. The coefficient estimates are still large, and in one case above one, but they are also insignificant. The instability of estimates when prison rates and state fixed effects are added suggests caution when considering the relevance of the results in Section 2.7.5.

References

Abramitzky, Ran, Adeline Delavande, and Luis Vasconcelos. "Marrying Up: The Role of Sex Ratio in Assortative Matching." *American Economic Journal: Applied Economics 3.3* (2011): 124-157.

Angrist, Josh. "How do sex ratios affect marriage and labor markets? Evidence from America's second generation." *The Quarterly Journal of Economics* 117.3 (2002): 997-1038.

Aughinbaugh, Alison, Charles R. Pierret, and Donna S. Rothstein. "The impact of family structure transitions on youth achievement: Evidence from the children of the NLSY79." *Demography* 42.3 (2005): 447-468.

Bjorklund, Anders, Donna K. Ginther, and Marianne Sundstrm. *Does marriage matter for children? Assessing the causal impact of legal marriage.* No. 3189. IZA Discussion Papers, 2007.

Brien, Michael J. "Racial differences in marriage and the role of marriage markets." *Journal of Human Resources* (1997): 741-778.

Bureau of Labor Statistics, U.S. Department of Labor. *National Longitudinal Survey of Youth 1979 cohort, 1979-2008 (rounds 1-23)* [computer file]. Produced and distributed by the Center for Human Resource Research, The Ohio State University. Columbus, OH: 2002.

Bureau of Labor Statistics, U.S. Department of Labor, and National Institute for Child Health and Human Development. *Children of the NLSY79, 1979-2008* [computer file]. Produced and distributed by the Center for Human Resource Research, The Ohio State University. Columbus, OH: 2002.

Craigie, Terry-Ann L., Jeanne Brooks-Gunn, and Jane Waldfogel. "Family structure, family stability and outcomes of five-year-old children." *Families, Relationships and Societies* 1.1 (2012): 43-61.

Donohue, John J, Aneja, Abhay, and Alexandria Zhang. "The Impact of Right-to-Carry Laws and the NRC Report: Lessons for the Empirical Evaluation of Law and Policy"

American Law and Economics Review 13.2 (2011): 565-631.

Edin, Kathryn, and Maria Kefalas. *Promises I can keep: Why poor women put motherhood before marriage.* University of California Press, 2011.

Finlay, Keith, and David Neumark. "Is Marriage Always Good for Children? Evidence from Families Affected by Incarceration." *Journal of Human Resources* 45.4 (2010): 1046-1088.

Freiden, Alan. "The United States Marriage Market." *Journal of Political Economy* 82.2 (1974): S34-S53.

Gennetian, Lisa A. "One or two parents? Half or step siblings? The effect of family structure on young children's achievement." *Journal of Population Economics* 18.3 (2005): 415-436.

Ginther, Donna K., and Robert A. Pollak. "Family structure and childrens educational outcomes: Blended families, stylized facts, and descriptive regressions." *Demography* 41.4 (2004): 671-696.

Greene, William H. Econometric Analysis. Vol. 5. Upper Saddle River, NJ: Prentice Hall, 2003.

Lichter, Daniel T., Diane K. McLaughlin, George Kephart, and David J. Landry. "Race and the retreat from marriage: A shortage of marriageable men?." *American Sociological Review* (1992): 781-799.

Heiland, Frank, and Shirley H. Liu. "Family structure and wellbeing of out-of-wedlock children: The significance of the biological parents' relationship." *Demographic Research* 15.4 (2006): 61-104.

Liu, Shirley H., and Frank Heiland. "SHOULD WE GET MARRIED? THE EFFECT OF PARENTS'MARRIAGE ON OUT-OF-WEDLOCK CHILDREN." *Economic inquiry* 50.1 (2012): 17-38.

Manski, Charles F., Gary D. Sandefur, Sara McLanahan, and Daniel Powers. "Alternative estimates of the effect of family structure during adolescence on high school graduation." *Journal of the American Statistical Association* 87.417 (1992): 25-37.

McLanahan, Sara, and Gary D. Sandefur. *Growing up with a single parent: What hurts,*

what helps. Harvard University Press, 1994.

Sigle-Rushton, Wendy, and Sara McLanahan. "For richer or poorer? Marriage as an anti-poverty strategy in the United States." *Population (english edition)* 57.3 (2002): 509-526.

South, Scott J., and Kim M. Lloyd. "Marriage opportunities and family formation: Further implications of imbalanced sex ratios." *Journal of Marriage and the Family* (1992): 440-451.

Stock, James H., Jonathan H. Wright, and Motohiro Yogo. "A survey of weak instruments and weak identification in generalized method of moments." *Journal of Business & Economic Statistics* 20.4 (2002).

Wagmiller, Robert L., Elizabeth Gershoff, Philip Veliz, and Margaret Clements. "Does Childrens Academic Achievement Improve when Single Mothers Marry?." *Sociology of education* 83.3 (2010): 201-226.

Wood, Robert G. "Marriage rates and marriageable men: A test of the Wilson hypothesis." *Journal of Human Resources* (1995): 163-193.

Tables

Table 2.1: Sample Restrictions

	Mothers		Children	
	dropped	remaining	dropped	remaining
Initial Sample		4043		
Deceased	104	3939		
Incarcerated, or in military	6	3933		
Mother marked as deaf, blind, mentally, or physically handicapped	64	3869		
Missing marriage date	95	3774		
Missing AFQT score	188	3586		
1st marriage before 1st birth	2464	1122		
Exclude whites and Hispanics	409	713		
Married within a year of 1st birth	56	657		
Unwed mothers subsample		657		1736
Child marked as deaf, blind, mentally, or physically handicapped	2	655	33	1703
Child born after 1st marriage	2	653	396	1307
Do not live with mother at 14 or 15	95	558	329	978
Difficulties with respondent's geographic location	27	531	44	934
Difficult to determine child's HS status	9	522	28	906
Geographic location of the child not observed for all 18 years	138	384	304	602

Table 2.2: Birth Before Marriage by Race

	Number			% of Total		
	White	Black	Hispanic	White	Black	Hispanic
1st child before any marriage	218	713	191	12%	66%	27%
1st child after 1st marriage	1577	360	527	88%	34%	73%
Total	1795	1073	718	100%	100%	100%

Table 2.3: Unwed Mother Sample Descriptions

	mean
mom's age at 1st birth	19.31
# of children ever born	2.78
mom HS graduate	0.73
mom reported a marriage	0.60
N = 384	

Table 2.4: Summary Statistics for Children

	mean
male	0.48
log of mom's AFQT	9.32
HS diploma or GED by age 20	0.78
mom married	0.50
mom HS graduate	0.68
first-born	0.49
child ever convicted	0.21
N = 602	

Table 2.5: Marital Status and Age at 1st Marriage by Mother's HS Status

	% by HS Category		
Age at 1st marriage:	NO HS	HS	Total
never married	53.85	34.64	39.84
up to 20	2.88	1.79	2.08
20 to 24	12.50	17.86	16.41
25 to 29	10.58	18.93	16.67
30 to 34	6.73	12.14	10.68
35 or older	13.46	14.64	14.32
Total	100.00	100.00	100.00

Table 2.6: Unwed Mothers that Eventually Married

	mean
age at 1st marriage	29.14
# of children ever	2.88
# of kids out of wedlock	2.07
# of kids after 1st marriage	0.81
N = 231	

Table 2.7: Mother's Age at First Birth by Number of Children Born

	% by Number of Children				
Age at 1st birth:	1	2	3	4 or more	Total
14 or less	0.00	0.00	5.08	4.55	2.60
15 to 19	39.29	46.72	63.56	70.45	56.25
20 to 24	44.64	41.80	28.81	22.73	33.85
25 to 29	12.50	9.84	2.54	2.27	6.25
30 to 34	3.57	1.64	0.00	0.00	1.04
Total	100.00	100.00	100.00	100.00	100.00

Table 2.8: OLS, High School Diploma or GED by Age 20

	1	2	3	4	5	6	7	8	9	10	11
Explanatory Variables:											
mom married	0.140***	0.134***	0.117***	0.105***	0.115***	0.094***	0.101***	0.084**	0.072*	0.080**	0.071*
	(0.034)	(0.034)	(0.033)	(0.033)	(0.033)	(0.033)	(0.033)	(0.034)	(0.037)	(0.038)	(0.038)
male		-0.125***	-0.131***	-0.128***	-0.123***	-0.129***	-0.127***	-0.127***	-0.126***	-0.128***	-0.096***
		(0.034)	(0.033)	(0.033)	(0.033)	(0.032)	(0.032)	(0.032)	(0.032)	(0.033)	(0.033)
log of mom's AFQT			0.073***	0.065***	0.058***	0.032**	0.033**	0.025	0.025	0.024	0.022
			(0.015)	(0.015)	(0.015)	(0.016)	(0.016)	(0.016)	(0.016)	(0.017)	(0.017)
first-born				0.107***	0.080**	0.068**	0.120**	0.102**	0.099**	0.083	0.076
				(0.033)	(0.034)	(0.034)	(0.049)	(0.050)	(0.050)	(0.050)	(0.051)
mom's age at 1st birth					0.014***	0.011**	0.002	-0.002	-0.001	0.004	0.003
					(0.005)	(0.005)	(0.008)	(0.008)	(0.008)	(0.009)	(0.009)
mom HS graduate						0.164***	0.164***	0.146***	0.146***	0.141***	0.140***
						(0.045)	(0.045)	(0.046)	(0.046)	(0.048)	(0.048)
mom's age at birth							0.011	0.014*	0.014*	0.011	0.009
							(0.007)	(0.007)	(0.007)	(0.008)	(0.008)
# of kids out of wedlock								-0.032**	-0.030**	-0.026*	-0.021
								(0.014)	(0.015)	(0.016)	(0.016)
# of kids after 1st marriage									0.021	0.038**	0.039**
									(0.019)	(0.019)	(0.018)
child ever convicted											-0.136***
											(0.050)
State FE[1]										YES	YES
N	592	592	592	592	592	592	592	592	592	585	585
R^2	0.028	0.051	0.093	0.109	0.118	0.143	0.147	0.157	0.158	0.206	0.221

Each column represents a seperate regression. Outcome is high school graduate or GED by age 20. Every regression includes a constant.
Standard errors are in parentheses. Signifcant at: *** 1%, ** 5%, * 10%.
[1] State fixed effects at the child's age 18.

Table 2.9: OLS, High School Diploma or GED by Age 20

	1	2	3	4	5
Explanatory Variables:					
mom married	0.070*	0.061	0.068	0.134**	0.051
	(0.038)	(0.046)	(0.077)	(0.052)	(0.040)
male×mom married		0.020			
		(0.067)			
mom HS graduate×mom married			0.003		
			(0.082)		
first-born×mom married				-0.141**	
				(0.066)	
child ever convicted×mom married					0.092
					(0.093)
male	-0.096***	-0.106**	-0.096***	-0.090***	-0.094***
	(0.033)	(0.051)	(0.033)	(0.034)	(0.033)
log of mom's AFQT	0.022	0.022	0.022	0.022	0.021
	(0.017)	(0.018)	(0.017)	(0.017)	(0.017)
first-born	0.076	0.076	0.076	0.152**	0.079
	(0.051)	(0.051)	(0.051)	(0.063)	(0.051)
mom's age at 1st birth	0.003	0.003	0.003	0.001	0.003
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
mom HS graduate	0.140***	0.140***	0.138**	0.133***	0.141***
	(0.048)	(0.048)	(0.060)	(0.048)	(0.048)
mom's age at birth	0.009	0.009	0.009	0.011	0.009
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
# of kids out of wedlock	-0.021	-0.021	-0.021	-0.022	-0.020
	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)
# of kids after 1st marriage	0.039**	0.039**	0.039**	0.049***	0.040**
	(0.018)	(0.018)	(0.018)	(0.018)	(0.019)
child ever convicted	-0.136***	-0.135***	-0.136***	-0.139***	-0.177***
	(0.050)	(0.050)	(0.050)	(0.049)	(0.065)
State FE[1]	YES	YES	YES	YES	YES
N	585	585	585	585	585
R^2	0.221	0.221	0.221	0.227	0.223

Each column represents a seperate regression. Outcome is high school graduate or GED by age 20. Every regression includes a constant. Standard errors are in parentheses. Signifcant at: *** 1%, ** 5%, * 10%.

[1] State fixed effects at the child's age 18.

Table 2.10: OLS, High School Diploma or GED by Age

	Graduated from HS by age:			
	19	20	21	22
Explanatory Variables:				
mom married	0.111**	0.134***	0.085*	0.103**
	(0.053)	(0.052)	(0.050)	(0.049)
first-born	0.088	0.152**	0.111*	0.113*
	(0.067)	(0.063)	(0.060)	(0.060)
first-born×mom married	-0.093	-0.141**	-0.077	-0.078
	(0.071)	(0.066)	(0.064)	(0.063)
male	-0.102***	-0.090***	-0.084***	-0.075**
	(0.037)	(0.034)	(0.032)	(0.032)
log of mom's AFQT	0.027	0.022	0.020	0.027*
	(0.018)	(0.017)	(0.017)	(0.016)
mom's age at 1st birth	0.007	0.001	0.002	0.000
	(0.009)	(0.009)	(0.008)	(0.008)
mom HS graduate	0.139***	0.133***	0.163***	0.100**
	(0.051)	(0.048)	(0.047)	(0.046)
mom's age at birth	0.008	0.011	0.012	0.015**
	(0.008)	(0.008)	(0.007)	(0.007)
# of kids out of wedlock	-0.021	-0.022	-0.020	-0.016
	(0.016)	(0.016)	(0.015)	(0.015)
# of kids after 1st marriage	0.044**	0.049***	0.037**	0.035**
	(0.020)	(0.018)	(0.018)	(0.017)
child ever convicted	-0.148***	-0.139***	-0.123***	-0.121***
	(0.053)	(0.049)	(0.048)	(0.047)
State FE[1]	YES	YES	YES	YES
N	588	585	579	573
R^2	0.223	0.227	0.223	0.190

Each column represents a seperate regression. Outcome is high school graduate or GED by the age in the column heading. Every regression includes a constant. Standard errors are in parentheses. Signifcant at: *** 1%, ** 5%, * 10%.

[1] State fixed effects at the child's age 18.

Table 2.11: OLS, High School Diploma or GED by Age

	Graduated from HS by age:			
	19	20	21	22
Explanatory Variables:				
mom married up to age 10[1]	0.112**	0.142***	0.094**	0.129***
	(0.045)	(0.041)	(0.040)	(0.038)
mom married after age 10[2]	0.010	-0.029	-0.011	-0.017
	(0.055)	(0.052)	(0.048)	(0.050)
male	-0.100***	-0.088***	-0.082**	-0.072**
	(0.037)	(0.033)	(0.032)	(0.031)
log of mom's AFQT	0.026	0.020	0.019	0.026
	(0.019)	(0.018)	(0.017)	(0.017)
first-born	0.052	0.099**	0.084*	0.091*
	(0.054)	(0.050)	(0.048)	(0.047)
mom's age at 1st birth	0.007	0.001	0.002	-0.000
	(0.009)	(0.008)	(0.008)	(0.008)
mom HS graduate	0.139***	0.134***	0.162***	0.099**
	(0.051)	(0.047)	(0.046)	(0.046)
mom's age at birth	0.008	0.010	0.011	0.015**
	(0.008)	(0.008)	(0.007)	(0.007)
# of kids out of wedlock	-0.020	-0.020	-0.019	-0.015
	(0.016)	(0.016)	(0.015)	(0.015)
# of kids after 1st marriage	0.022	0.014	0.016	0.008
	(0.021)	(0.018)	(0.018)	(0.017)
child ever convicted	-0.149***	-0.140***	-0.124***	-0.124***
	(0.053)	(0.049)	(0.048)	(0.047)
State FE[3]	YES	YES	YES	YES
N	588	585	579	573
R^2	0.224	0.235	0.227	0.200

Each column represents a seperate regression. Outcome is high school graduate or GED by the age in the column heading. Every regression includes a constant. Standard errors are in parentheses. Signifcant at: *** 1%, ** 5%, * 10%.

[1] Indicator if entered first marriage before the child's was 10.

[2] Indicator if entered first marriage when the child was between 10 and 18.

[3] State fixed effects at the child's age 18.

Table 2.12: OLS, High School Diploma or GED by Age

	Graduated from HS by age:			
	19	20	21	22
Explanatory Variables:				
mom married up to age 10[1]	0.148**	0.202***	0.128**	0.164***
	(0.059)	(0.056)	(0.054)	(0.051)
mom married up to age 10×first-born	-0.089	-0.148**	-0.083	-0.08
	(0.075)	(0.067)	(0.066)	(0.063)
mom married after age 10[2]	0.034	-0.017	-0.011	-0.028
	(0.085)	(0.085)	(0.079)	(0.080)
mom married after age 10×first-born	-0.053	-0.040	-0.011	0.008
	(0.111)	(0.106)	(0.098)	(0.100)
male	-0.098***	-0.086**	-0.081**	-0.071**
	(0.037)	(0.034)	(0.032)	(0.031)
log of mom's AFQT	0.026	0.020	0.019	0.026
	(0.019)	(0.018)	(0.017)	(0.017)
first-born	0.094	0.162**	0.117*	0.122**
	(0.068)	(0.063)	(0.060)	(0.060)
mom's age at 1st birth	0.006	-0.000	0.001	-0.001
	(0.009)	(0.008)	(0.008)	(0.008)
mom HS graduate	0.135***	0.127***	0.159***	0.095**
	(0.051)	(0.047)	(0.046)	(0.046)
mom's age at birth	0.008	0.012	0.012*	0.016**
	(0.008)	(0.007)	(0.007)	(0.007)
# of kids out of wedlock	-0.021	-0.021	-0.020	-0.015
	(0.016)	(0.016)	(0.015)	(0.015)
# of kids after 1st marriage	0.031	0.028	0.024	0.016
	(0.021)	(0.018)	(0.017)	(0.017)
child ever convicted	-0.150***	-0.141***	-0.125***	-0.124***
	(0.053)	(0.049)	(0.047)	(0.047)
State FE[3]	YES	YES	YES	YES
N	588	585	579	573
R^2	0.226	0.240	0.229	0.202

Each column represents a seperate regression. Outcome is high school graduate or GED by the age in the column heading. Every regression includes a constant. Standard errors are in parentheses. Signifcant at: *** 1%, ** 5%, * 10%.

[1] Indicator if entered first marriage before the child's was 10.

[2] Indicator if entered first marriage when the child was between 10 and 18.

[3] State fixed effects at the child's age 18.

Table 2.13: OLS, Mom Married by Child Age 18

	1	2	3
Regression Results:			
Includes sex ratios for child ages 1 to 18?	YES	YES	YES
male		-0.086**	-0.076*
		-0.041	(0.039)
log of mom's AFQT		0.045***	0.019
		-0.016	(0.018)
first-born			-0.057
			(0.048)
mom's age at birth			-0.020**
			(0.008)
# of kids out of wedlock			-0.071***
			(0.013)
constant	-0.715**	-1.163***	0.354
	(0.309)	-0.321	(0.506)
N	596	596	596
R^2	0.072	0.090	0.141
F statistic	4.757	5.348	7.405
Test of Joint Significance for the Sex Ratios:			
F statistic	4.76	5.24	3.99
p-value	0.00	0.00	0.00

Standard errors are in parentheses. Signifcant at: *** 1%, ** 5%, * 10%.

Table 2.14: Sex Ratios by Child's Birth Order

	mean		median	
sex ratio at age:	1st born	Subsequent	1st born	Subsequent
1	0.733	0.693	0.737	0.687
2	0.717	0.682	0.717	0.673
3	0.704	0.672	0.706	0.664
4	0.692	0.662	0.693	0.653
5	0.680	0.653	0.679	0.644
6	0.669	0.646	0.669	0.637
7	0.659	0.637	0.658	0.627
8	0.650	0.630	0.647	0.622
9	0.641	0.623	0.639	0.615
10	0.634	0.617	0.630	0.610
11	0.626	0.613	0.625	0.606
12	0.620	0.609	0.618	0.603
13	0.615	0.606	0.613	0.600
14	0.610	0.602	0.608	0.598
15	0.605	0.599	0.604	0.597
16	0.602	0.597	0.602	0.594
17	0.599	0.595	0.599	0.593
18	0.596	0.595	0.597	0.590

Table 2.15: Mother's Age at Birth of Child Regressed on Sex Ratios

	1	2
Sex ratio age 1	-44.781***	-68.925***
	(4.340)	(15.160)
Includes sex ratios for ages 2 to 18?	NO	YES
constant	53.570***	51.661***
	(3.094)	(2.341)
N	596	596
R^2	0.554	0.633
F statistic	104.070	41.319

Standard errors are in parentheses. Signifcant at: *** 1%, ** 5%, * 10%.

Table 2.16: Instrumental Vairable Results, HS Graduation by Age

	Graduated from HS by age:			
	19	20	21	22
Explanatory Variables:				
mom married	0.320**	0.237*	0.257**	0.155
	(0.148)	(0.128)	(0.120)	(0.112)
male	-0.113***	-0.114***	-0.101***	-0.112***
	(0.039)	(0.033)	(0.032)	(0.031)
log of mom's AFQT	0.066	0.074**	0.052	0.066*
	(0.042)	(0.037)	(0.036)	(0.034)
first-born	0.417	0.103	0.123	-0.029
	(0.257)	(0.230)	(0.224)	(0.213)
mom's age at birth	0.029*	0.013	0.020	0.010
	(0.016)	(0.014)	(0.013)	(0.012)
# of kids out of wedlock	0.063	0.024	-0.009	-0.005
	(0.086)	(0.079)	(0.076)	(0.077)
constant	-0.989	-0.362	-0.215	0.000
	(0.647)	(0.600)	(0.577)	(0.585)
N	595	592	586	580

Each column represents GMM estimation results on indicator if child was a high school graduate or obtained a GED by the age in the column heading. Standard errors are in parentheses. Signifcant at: *** 1%, ** 5%, * 10%.

Instruments:

Esr4 - expected sex ratios from the model with covariance in shocks between men and women of the same race for child ages 1 to 18, male - child's gender, log of mom's AFQT score.

Endogenous variables:

mom married - indicator if mother married by when the child was 18 or younger, first-born - indicator if child was the first-born, mom's age at birth - mother's age at the birth of the child, # kids out of wedlock - number of children the mother had before date of any marital transition.

Table 2.17: Instrumental Vairable Results, HS Graduation by Age

	Graduated from HS by age:			
	19	20	21	22
Explanatory Variables:				
mom married	0.186	0.233	0.282*	0.198
	(0.184)	(0.166)	(0.167)	(0.155)
male	-0.123***	-0.118***	-0.105***	-0.110***
	(0.040)	(0.035)	(0.035)	(0.033)
log of mom's AFQT	0.027	0.038	0.035	0.037
	(0.037)	(0.033)	(0.033)	(0.030)
first-born	0.381	0.229	0.219	0.167
	(0.259)	(0.217)	(0.219)	(0.208)
mom's age at birth	0.033**	0.022	0.026*	0.025**
	(0.016)	(0.013)	(0.013)	(0.013)
# of kids out of wedlock	-0.040	-0.033	-0.038	-0.036
	(0.078)	(0.065)	(0.067)	(0.064)
constant	-0.457	-0.147	-0.214	-0.136
	(0.704)	(0.592)	(0.606)	(0.583)
State FE	YES	YES	YES	YES
N	531	528	522	516

Each column represents GMM estimation results on indicator if child was a high school graduate or obtained a GED by the age in the column heading. Standard errors are in parentheses. Signifcant at: *** 1%, ** 5%, * 10%.
Instruments:

Esr4 - expected sex ratios from the model with covariance in shocks between men and women of the same race for child ages 1 to 18, male - child's gender, log of mom's AFQT score, State FE - state fixed effects when the children were 18 years old.
Endogenous variables:

mom married - indicator if mother married by when the child was 18 or younger, first-born - indicator if child was the first-born, mom's age at birth - mother's age at the birth of the child, # kids out of wedlock - number of children the mother had before date of any marital transition.

Table 2.18: Instrumental Vairable Results, HS graduation by Age

	Graduated from HS by age:			
	19	20	21	22
Explanatory Variables:				
mom married	0.300**	0.199	0.265**	0.201
	(0.150)	(0.136)	(0.134)	(0.127)
male	-0.107***	-0.116***	-0.098***	-0.107***
	(0.041)	(0.036)	(0.036)	(0.034)
log of mom's AFQT	0.030	0.055	0.035	0.044
	(0.048)	(0.043)	(0.042)	(0.042)
first-born	0.232	-0.041	0.032	-0.026
	(0.240)	(0.215)	(0.222)	(0.205)
mom's age at birth	0.032**	0.012	0.022*	0.017
	(0.014)	(0.012)	(0.012)	(0.011)
# of kids out of wedlock	-0.051	-0.041	-0.066	-0.047
	(0.090)	(0.083)	(0.084)	(0.085)
constant	-0.637	-0.048	-0.141	0.032
	(0.703)	(0.643)	(0.633)	(0.622)
Unemp Rates	YES	YES	YES	YES
N	566	563	562	556

Each column represents GMM estimation results on indicator if child was a high school graduate or obtained a GED by the age in the column heading. Standard errors are in parentheses. Signifcant at: *** 1%, ** 5%, * 10%.

Instruments:

Esr4 - expected sex ratios from the model with covariance in shocks between men and women of the same race for child ages 1 to 18, male - child's gender, log of mom's AFQT score, Unemp Rates - unemployment rates for all 18 years.

Endogenous variables:

mom married - indicator if mother married by when the child was 18 or younger, first-born - indicator if child was the first-born, mom's age at birth - mother's age at the birth of the child, # kids out of wedlock - number of children the mother had before date of any marital transition.

Table 2.19: Instrumental Vairable Results, HS Graduation by Age

	Graduated from HS by age:			
	19	20	21	22
Explanatory Variables:				
mom married	0.152	0.094	0.226	0.094
	(0.174)	(0.163)	(0.159)	(0.151)
male	-0.138***	-0.133***	-0.111***	-0.123***
	(0.040)	(0.035)	(0.034)	(0.033)
log of mom's AFQT	0.067	0.076**	0.069*	0.075**
	(0.041)	(0.038)	(0.038)	(0.036)
first-born	0.463*	0.077	0.118	-0.009
	(0.273)	(0.248)	(0.243)	(0.227)
mom's age at birth	0.031	0.008	0.015	0.006
	(0.019)	(0.017)	(0.016)	(0.015)
# of kids out of wedlock	0.076	0.024	0.033	0.017
	(0.084)	(0.077)	(0.076)	(0.073)
constant	-1.138	-0.326	-0.502	-0.140
	(0.747)	(0.673)	(0.660)	(0.643)
Prison Rates	YES	YES	YES	YES
N	566	563	562	556

Each column represents GMM estimation results on indicator if child was a high school graduate or obtained a GED by the age in the column heading. Standard errors are in parentheses. Signifcant at: *** 1%, ** 5%, * 10%.

Instruments:

Esr4 - expected sex ratios from the model with covariance in shocks between men and women of the same race for child ages 1 to 18, male - child's gender, log of mom's AFQT score, Prison Rates - prison rates for all 18 years.

Endogenous variables:

mom married - indicator if mother married by when the child was 18 or younger, first-born - indicator if child was the first-born, mom's age at birth - mother's age at the birth of the child, # kids out of wedlock - number of children the mother had before date of any marital transition.

Table 2.20: Instrumental Vairable Results, HS Graduation by Age

	Graduated from HS by age:			
	19	20	21	22
Explanatory Variables:				
mom married up to age 10	0.162	0.100	0.155	0.062
	(0.162)	(0.148)	(0.132)	(0.128)
mom married after age 10	0.506***	0.431**	0.377**	0.258*
	(0.176)	(0.183)	(0.165)	(0.156)
male	-0.127***	-0.132***	-0.110***	-0.119****
	(0.040)	(0.036)	(0.033)	(0.033)
log of mom's AFQT	0.054	0.070*	0.051	0.062*
	(0.042)	(0.039)	(0.037)	(0.036)
first-born	0.418	0.040	0.076	-0.065
	(0.268)	(0.246)	(0.224)	(0.221)
mom's age at birth	0.030*	0.009	0.017	0.008
	(0.017)	(0.015)	(0.013)	(0.013)
# of kids out of wedlock	0.016	-0.010	-0.032	-0.033
	(0.090)	(0.086)	(0.081)	(0.082)
constant	-0.752	-0.104	-0.043	0.196
	(0.676)	(0.642)	(0.599)	(0.614)
N	595	592	586	580

Each column represents GMM estimation results on indicator if child was a high school graduate or obtained a GED by the age in the column heading. Standard errors are in parentheses. Signifcant at: *** 1%, ** 5%, * 10%.

Instruments:

Esr4 - expected sex ratios from the model with covariance in shocks between men and women of the same race for child ages 1 to 18, male - child's gender, log of mom's AFQT score.

Endogenous variables:

mom married up to age 10 - indicator if entered first marriage before the child's was 10, mom married after age 10 - indicator if entered first marriage when the child was between 10 and 18, first-born - indicator if child was the first-born, mom's age at birth - mother's age at the birth of the child, # kids out of wedlock - number of children the mother had before date of any marital transition.

Table 2.21: Instrumental Vairable Results, HS Graduation by Age

	Graduated from HS by age:			
	19	20	21	22
Explanatory Variables:				
mom married up to age 10	-0.092	0.037	0.035	0.060
	(0.309)	(0.274)	(0.276)	(0.227)
mom married after age 10	1.340	1.106	1.127	0.770
	(0.905)	(0.852)	(0.828)	(0.656)
male	-0.192***	-0.174***	-0.152**	-0.140***
	(0.074)	(0.064)	(0.063)	(0.051)
log of mom's AFQT	0.135	0.106	0.111	0.087
	(0.096)	(0.078)	(0.080)	(0.062)
first-born	-0.086	-0.158	-0.141	-0.064
	(0.440)	(0.388)	(0.376)	(0.319)
mom's age at birth	0.005	0.001	0.005	0.014
	(0.027)	(0.022)	(0.022)	(0.018)
# of kids out of wedlock	0.023	-0.007	0.001	-0.009
	(0.136)	(0.108)	(0.110)	(0.089)
constant	-0.900	-0.285	-0.496	-0.377
	(1.139)	(0.929)	(0.955)	(0.784)
State FE	YES	YES	YES	YES
N	531	528	522	516

Each column represents GMM estimation results on indicator if child was a high school graduate or obtained a GED by the age in the column heading. Standard errors are in parentheses. Signifcant at: *** 1%, ** 5%, * 10%.

Instruments:

Esr4 - expected sex ratios from the model with covariance in shocks between men and women of the same race for child ages 1 to 18, male - child's gender, log of mom's AFQT score, State fixed effects at child's age 18.

Endogenous variables:

mom married up to age 10 - indicator if entered first marriage before the child's was 10, mom married after age 10 - indicator if entered first marriage when the child was between 10 and 18, first-born - indicator if child was the first-born, mom's age at birth - mother's age at the birth of the child, # kids out of wedlock - number of children the mother had before date of any marital transition.

Table 2.22: Instrumental Vairable Results, HS Graduation by Age

	Graduated from HS by age:			
	19	20	21	22
Explanatory Variables:				
mom married up to age 10	0.036	-0.043	0.054	0.058
	(0.202)	(0.202)	(0.177)	(0.158)
mom married after age 10	0.662**	0.615**	0.578**	0.405*
	(0.274)	(0.310)	(0.275)	(0.234)
male	-0.146***	-0.150***	-0.126***	-0.123***
	(0.050)	(0.049)	(0.044)	(0.039)
log of mom's AFQT	0.068	0.085	0.066	0.061
	(0.058)	(0.059)	(0.053)	(0.047)
first-born	-0.011	-0.273	-0.168	-0.153
	(0.298)	(0.295)	(0.270)	(0.232)
mom's age at birth	0.006	-0.009	0.003	0.004
	(0.020)	(0.020)	(0.018)	(0.015)
# of kids out of wedlock	-0.037	-0.041	-0.053	-0.050
	(0.102)	(0.103)	(0.097)	(0.093)
constant	-0.114	0.467	0.252	0.376
	(0.812)	(0.811)	(0.734)	(0.697)
Unemp Rates	YES	YES	YES	YES
N	566	563	562	556

Each column represents GMM estimation results on indicator if child was a high school graduate or obtained a GED by the age in the column heading. Standard errors are in parentheses. Signifcant at: *** 1%, ** 5%, * 10%.

Instruments:

Esr4 - expected sex ratios from the model with covariance in shocks between men and women of the same race for child ages 1 to 18, male - child's gender, log of mom's AFQT score, unemployment rates for all 18 years.

Endogenous variables:

mom married up to age 10 - indicator if entered first marriage before the child's was 10, mom married after age 10 - indicator if entered first marriage when the child was between 10 and 18, first-born - indicator if child was the first-born, mom's age at birth - mother's age at the birth of the child, # kids out of wedlock - number of children the mother had before date of any marital transition.

Table 2.23: Instrumental Vairable Results, HS Graduation by Age

	Graduated from HS by age:			
	19	20	21	22
Explanatory Variables:				
mom married up to age 10	-0.019	-0.081	0.064	-0.035
	(0.192)	(0.198)	(0.181)	(0.174)
mom married after age 10	0.458	0.474	0.490*	0.338
	(0.282)	(0.308)	(0.272)	(0.263)
male	-0.155***	-0.155***	-0.126***	-0.135***
	(0.041)	(0.041)	(0.037)	(0.036)
log of mom's AFQT	0.070*	0.072*	0.070*	0.071*
	(0.042)	(0.043)	(0.040)	(0.039)
first-born	0.175	-0.207	-0.100	-0.193
	(0.312)	(0.291)	(0.276)	(0.256)
mom's age at birth	0.010	-0.014	-0.002	-0.009
	(0.022)	(0.020)	(0.019)	(0.018)
# of kids out of wedlock	0.001	-0.057	-0.029	-0.048
	(0.091)	(0.094)	(0.088)	(0.087)
constant	-0.394	0.490	0.125	0.454
	(0.827)	(0.804)	(0.755)	(0.747)
Prison Rates	YES	YES	YES	YES
N	566	563	562	556

Each column represents GMM estimation results on indicator if child was a high school graduate or obtained a GED by the age in the column heading. Standard errors are in parentheses. Signifcant at: *** 1%, ** 5%, * 10%.

Instruments:

Esr4 - expected sex ratios from the model with covariance in shocks between men and women of the same race for child ages 1 to 18, male - child's gender, log of mom's AFQT score, Prison Rates - priosn rates for all 18 years.

Endogenous variables:

mom married up to age 10 - indicator if entered first marriage before the child's was 10, mom married after age 10 - indicator if entered first marriage when the child was between 10 and 18, first-born - indicator if child was the first-born, mom's age at birth - mother's age at the birth of the child, # kids out of wedlock - number of children the mother had before date of any marital transition.

Table 2.24: Instrumental Vairable Results, HS Graduation by Age

	Graduated from HS by age:			
	19	20	21	22
Explanatory Variables:				
mom married up to age 10	-0.116	0.003	0.072	0.102
	(0.284)	(0.288)	(0.237)	(0.194)
mom married after age 10	0.974	1.034	0.762	0.311
	(0.713)	(0.818)	(0.677)	(0.531)
male	-0.189***	-0.185***	-0.149***	-0.135***
	(0.061)	(0.064)	(0.053)	(0.042)
log of mom's AFQT	0.104	0.108	0.092	0.067
	(0.090)	(0.093)	(0.076)	(0.060)
first-born	-0.029	-0.305	-0.203	-0.052
	(0.412)	(0.449)	(0.374)	(0.308)
mom's age at birth	0.008	-0.011	0.000	0.016
	(0.025)	(0.026)	(0.021)	(0.018)
# of kids out of wedlock	-0.015	-0.034	-0.038	-0.037
	(0.128)	(0.123)	(0.100)	(0.080)
constant	-0.680	-0.219	-0.135	0.171
	(1.207)	(1.167)	(0.944)	(0.726)
State FE	YES	YES	YES	YES
Prison Rates	YES	YES	YES	YES
N	508	505	504	498

Each column represents GMM estimation results on indicator if child was a high school graduate or obtained a GED by the age in the column heading. Standard errors are in parentheses. Signifcant at: *** 1%, ** 5%, * 10%.

Instruments:

Esr4 - expected sex ratios from the model with covariance in shocks between men and women of the same race for child ages 1 to 18, male - child's gender, log of mom's AFQT score, State fixed effects at child's age 18, Prison Rates - prison rates for all 18 years.

Endogenous variables:

mom married up to age 10 - indicator if entered first marriage before the child's was 10, mom married after age 10 - indicator if entered first marriage when the child was between 10 and 18, first-born - indicator if child was the first-born, mom's age at birth - mother's age at the birth of the child, # kids out of wedlock - number of children the mother had before date of any marital transition.

Figures

Figure 2.1: Number of Children by Mother's HS Status

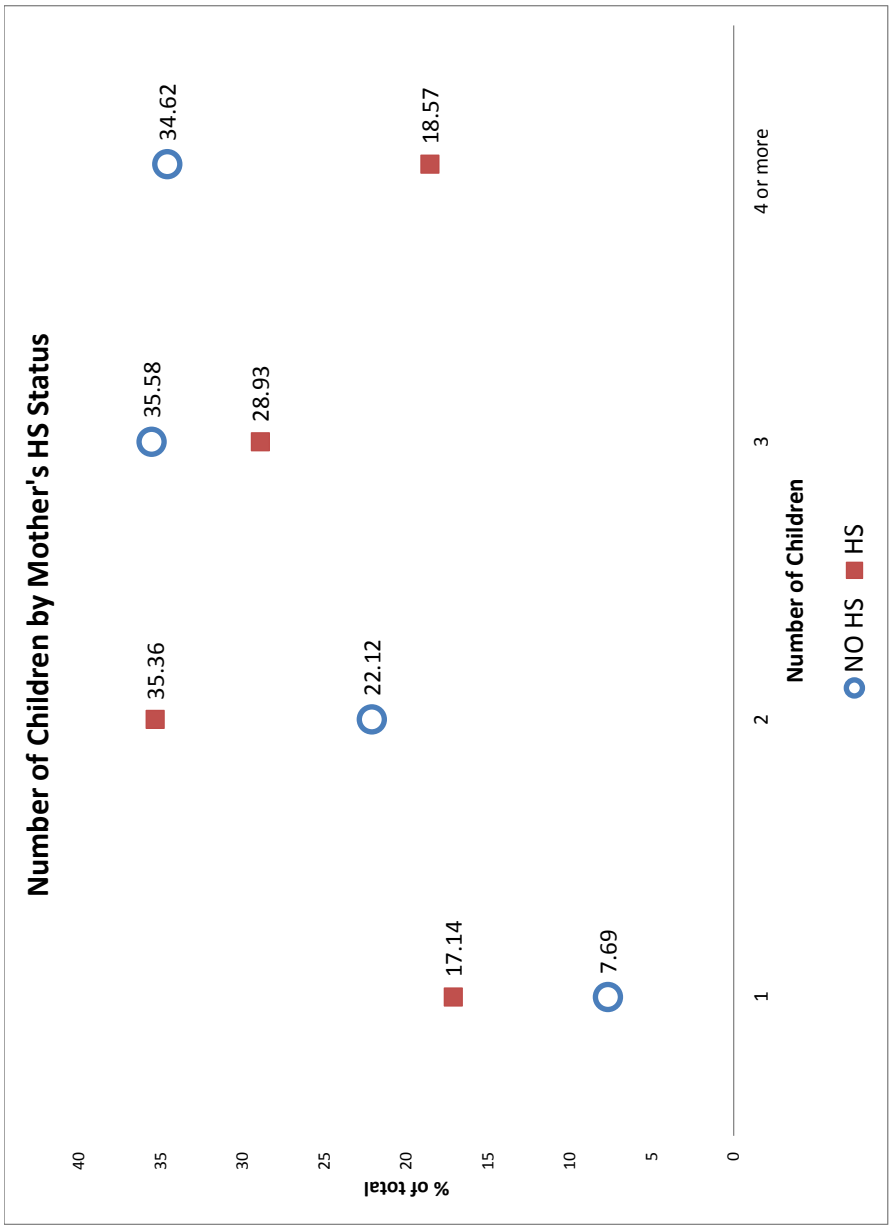


Figure 2.2: Number of Children by Mother's Marital Status

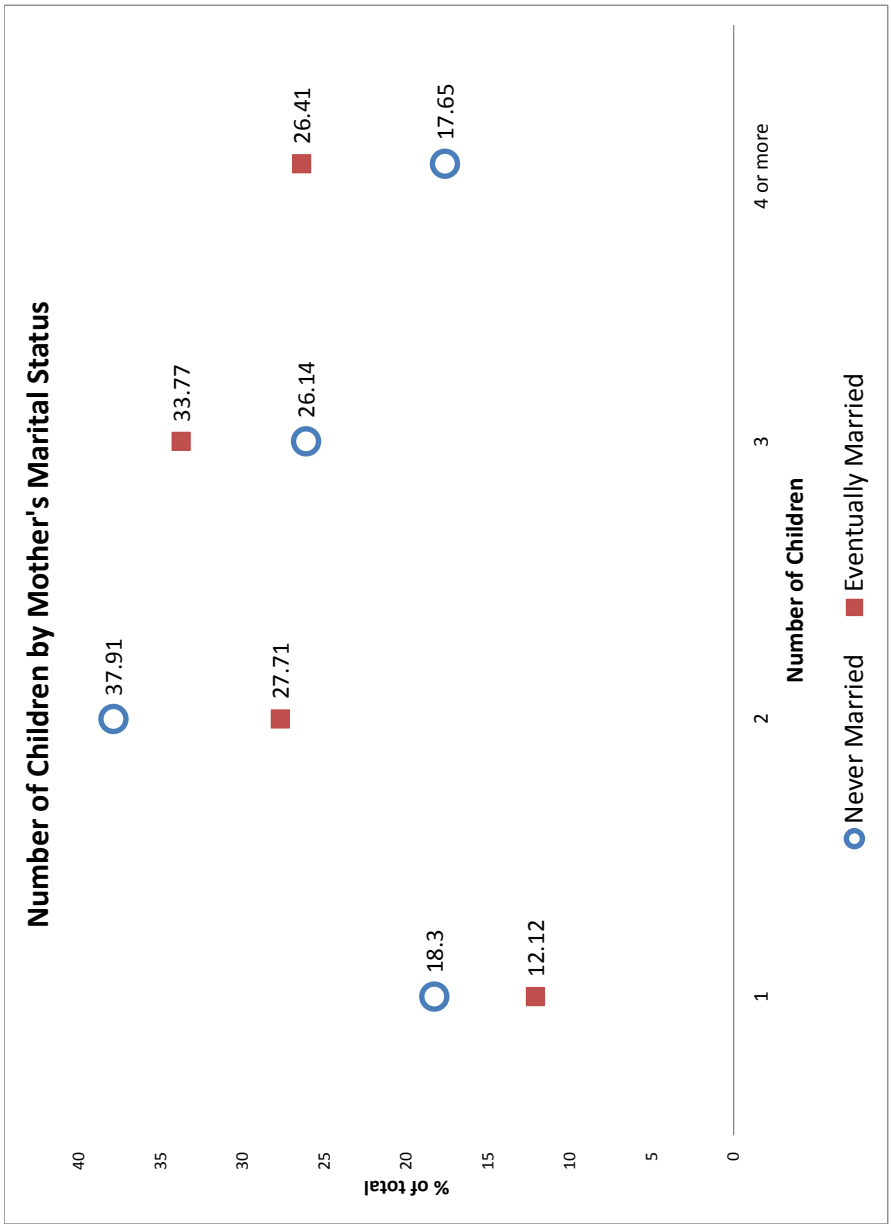


Figure 2.3: Mother's Age at 1st Birth by Marital Status

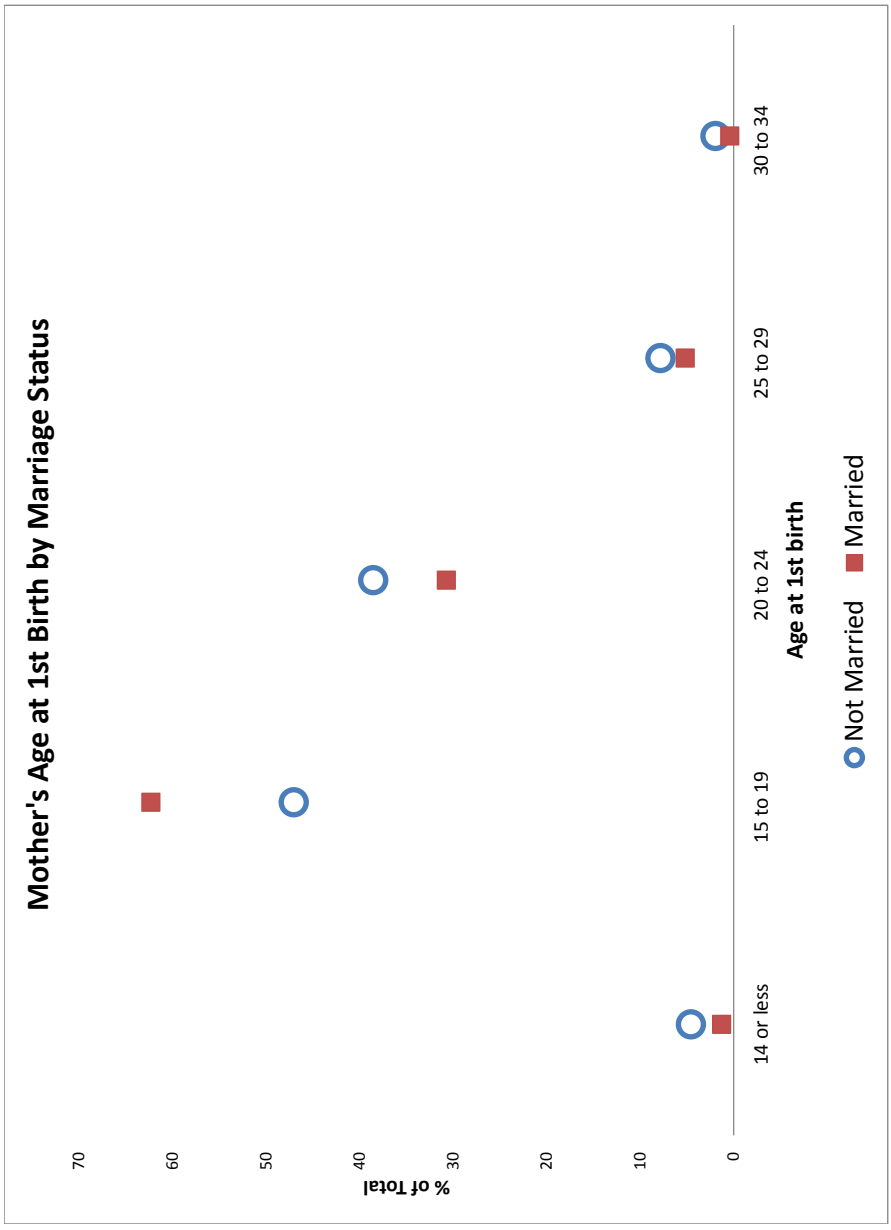


Figure 2.4: Mother's Age at 1st Birth by HS Status